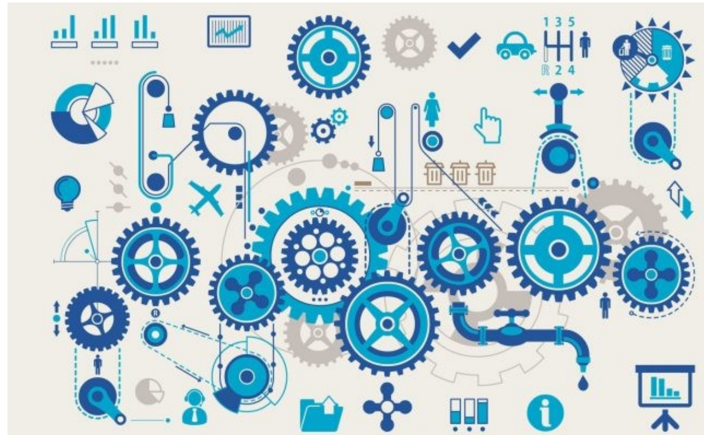
On the problems of interface

explainability, conceptual spaces, relevance

TELECOM ParisTech

DAUPHINE UNIVERSITÉ PARIS

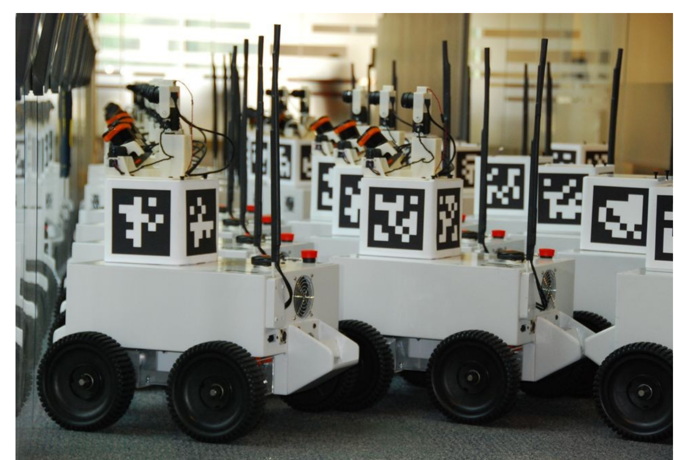
Giovanni Sileno
13 June 2018
gsileno@enst.fr


with the (supposedly) near advent of ***autonomous artificial entities***, or other forms of ***distributed automatic decision making***,

- humans less and less in the loop
- increasing concerns about ***unintended consequences***





Unintended consequences: bad or limited design

Unintended consequences: bad or limited design

- Wallet hacks, fraudulent actions and bugs in the in the ***blockchain*** sector during 2017:
  - CoinDash ICO Hack ($10 millions)
  - Parity Wallet Breach ($105 millions)
  - Enigma Project Scum
  - Parity Wallet Freeze ($275 millions)
  - Tether Token Hack ($30 millions)
  - Bitcoin Gold Scam ($3 millions)
  - NiceHash Market Breach ($80 millions)

Source: CoinDesk (2017), *Hacks, Scams and Attacks: Blockchain's 2017 Disasters*

Unintended consequences: the “artificial prejudice”

Unintended consequences: the "artificial prejudice"

- Several studies prove that associations extracted from linguistic corpora reproduce stereotypes.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases. Science,* 356(6334), 183–186.

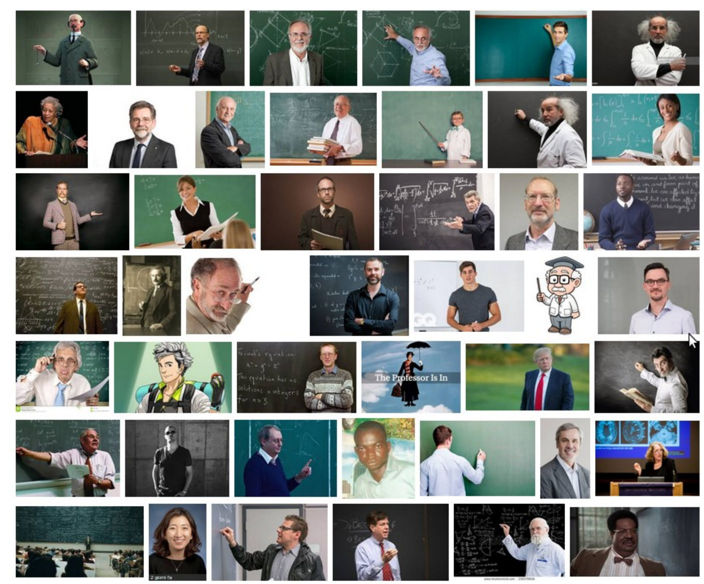
Unintended consequences: the "artificial prejudice"

- Several studies prove that associations extracted from linguistic corpora reproduce stereotypes.
- Ex.: a simple Google visual search a few days ago:

teacher

vs

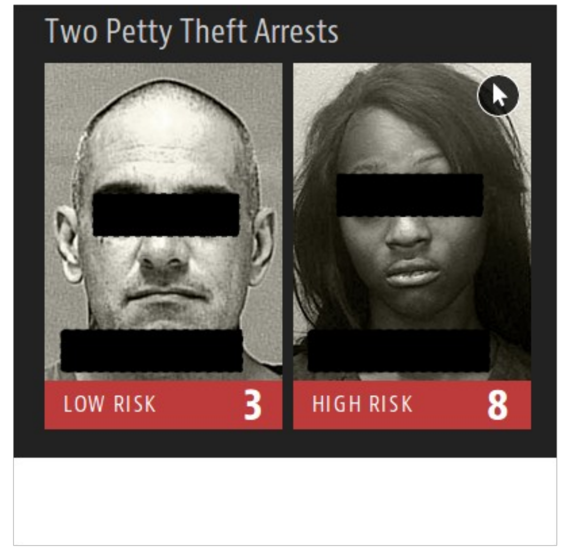
professor



Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases. Science,* 356(6334), 183–186.

Unintended consequences: the “artificial prejudice”

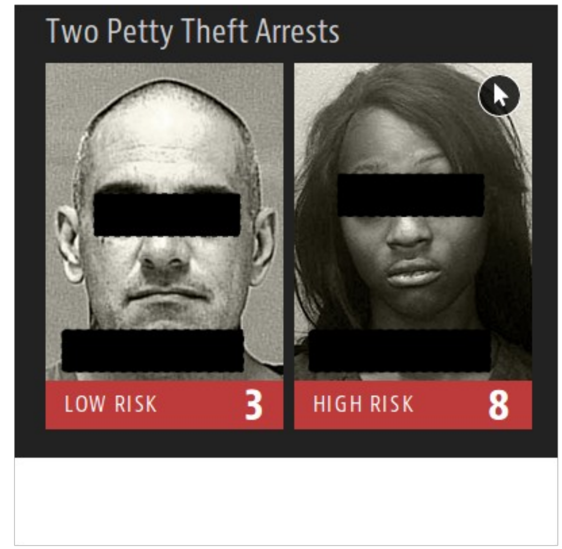
- Software used across the US predicting future crimes and criminals is biased against African Americans (2016).

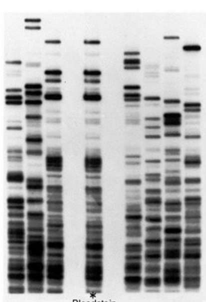


Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

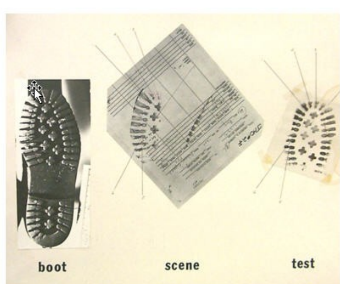
Unintended consequences: the “artificial prejudice”

- Software used across the US predicting future crimes and criminals is biased against African Americans (2016).
- Role of *circumstantial evidence*: how to integrate statistical inference in judgment?





DNA



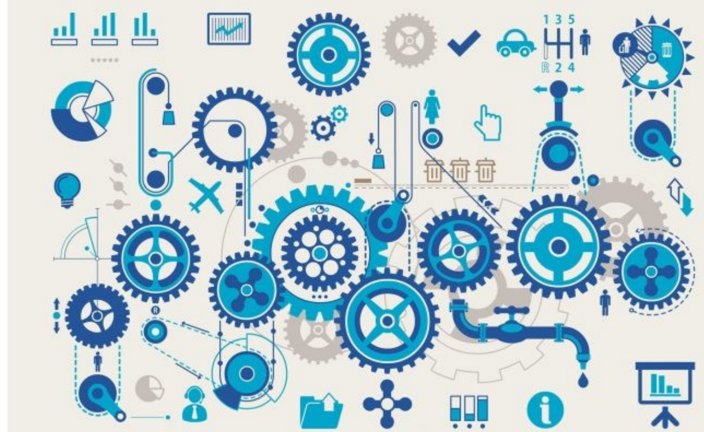
footwear

...

origin, gender,
ethnicity, wealth, ...

improper profiling?

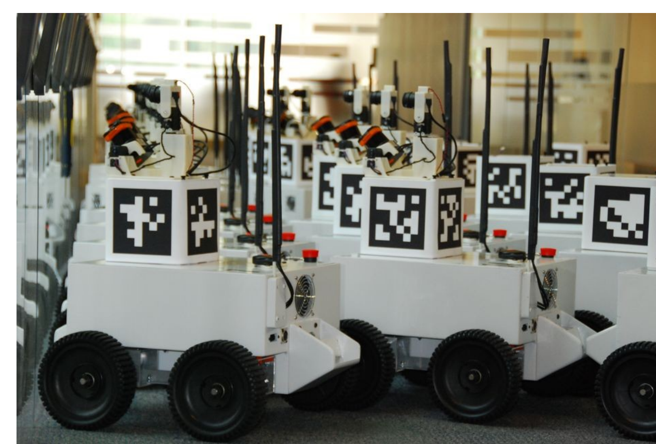
Angwin J. et al. ProPublica, May 23 (2016). *Machine Bias: risk assessments in criminal sentencing*

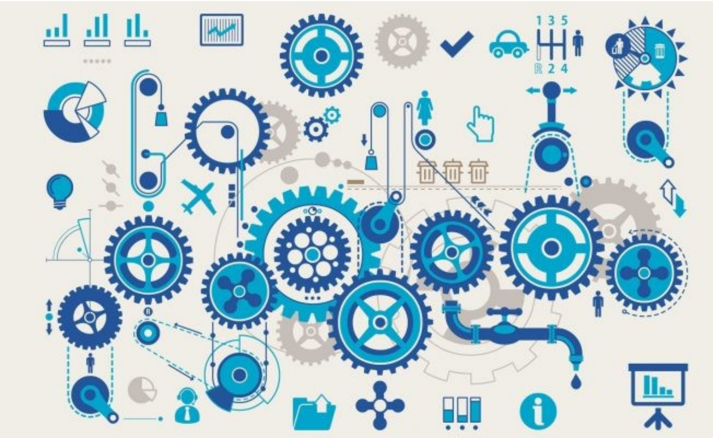
  

unintended consequences

with ubiquitous devices/services?

scaling → wider effects → increased risks

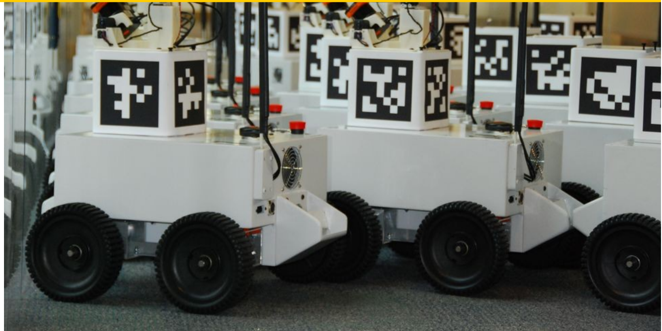




unintended consequences

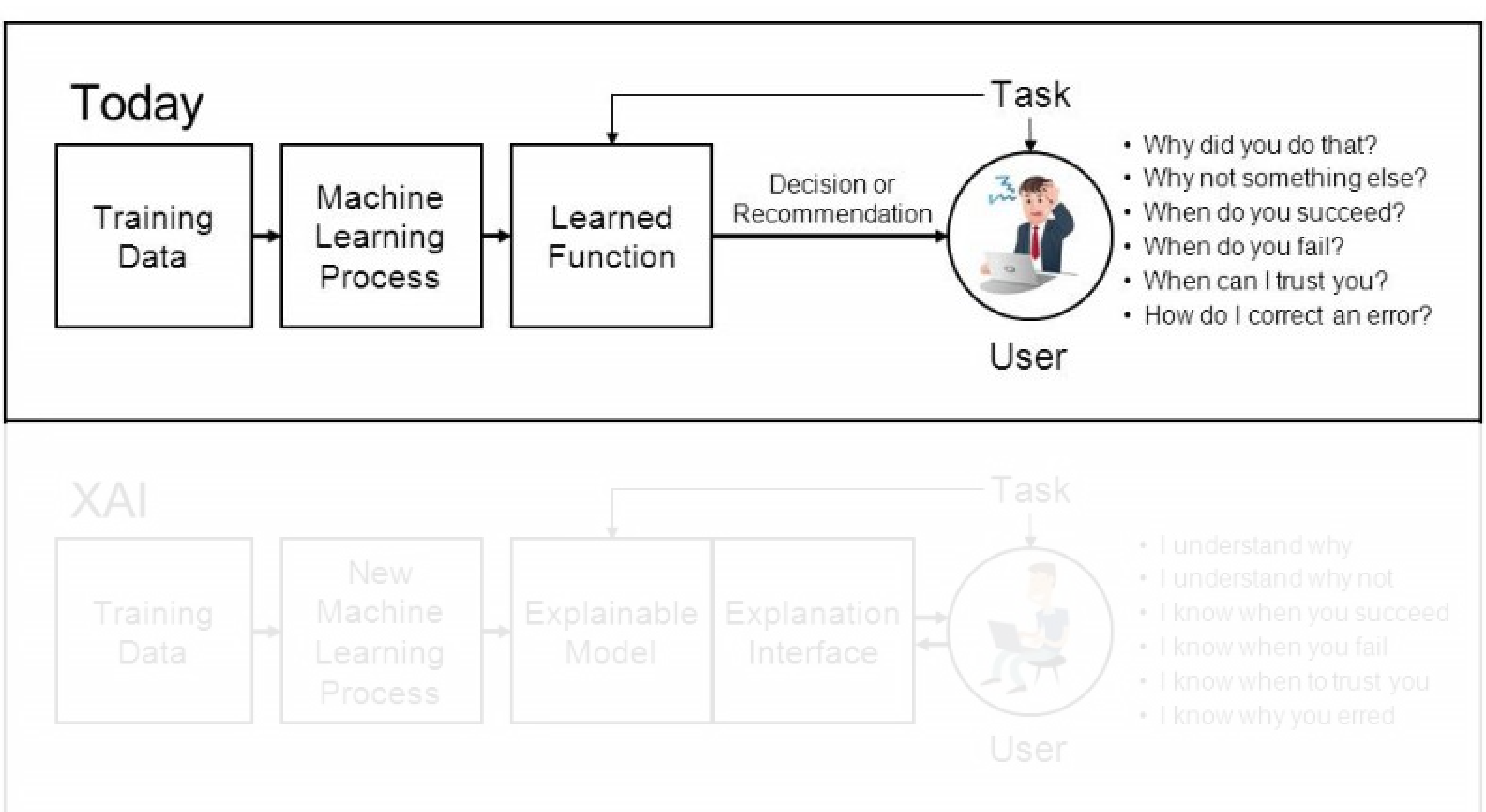
with ubiquitous devices/services?

scaling → wider effects → increased risks

necessity to review our conception methods!

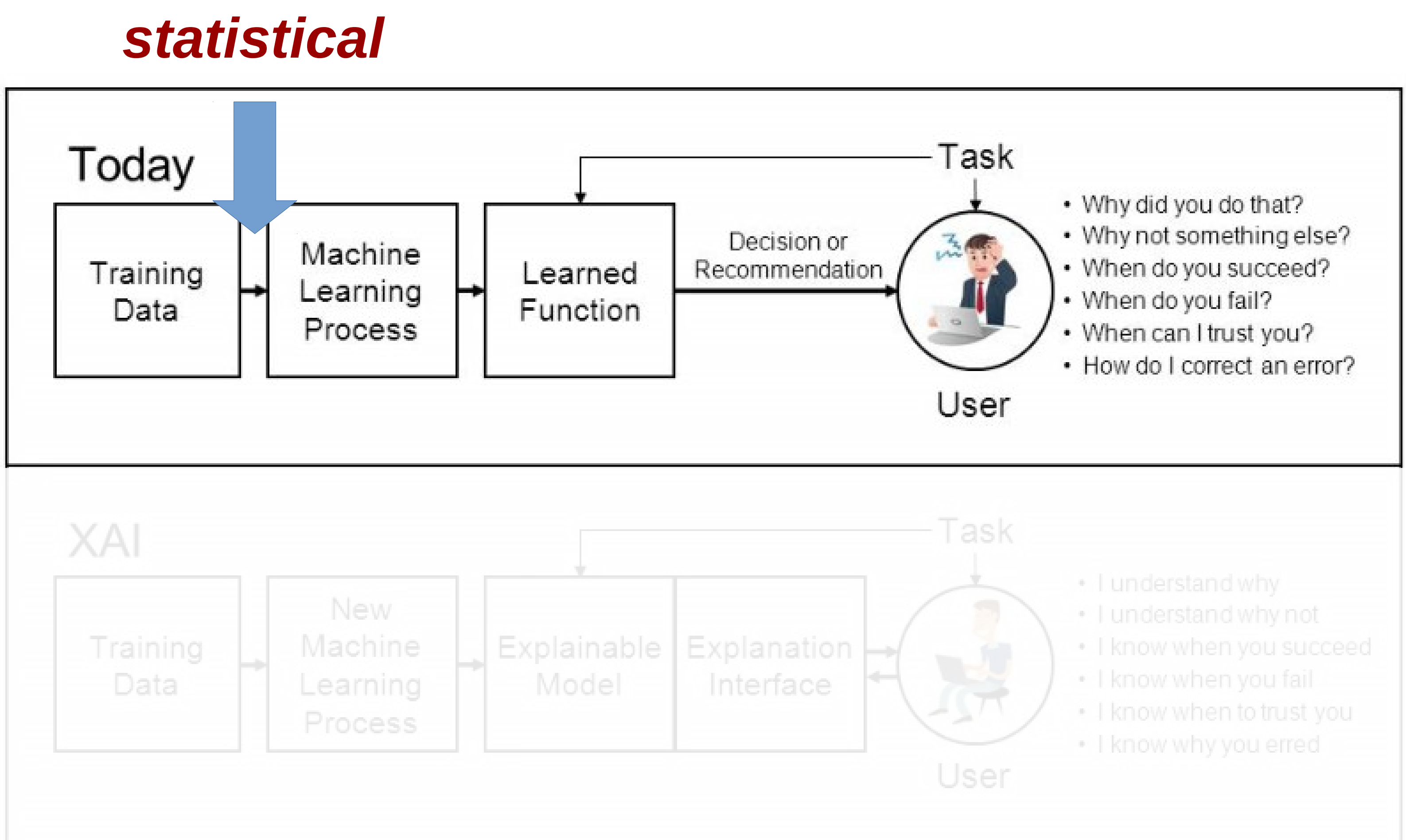



The call for *Explanaible AI* (XAI)



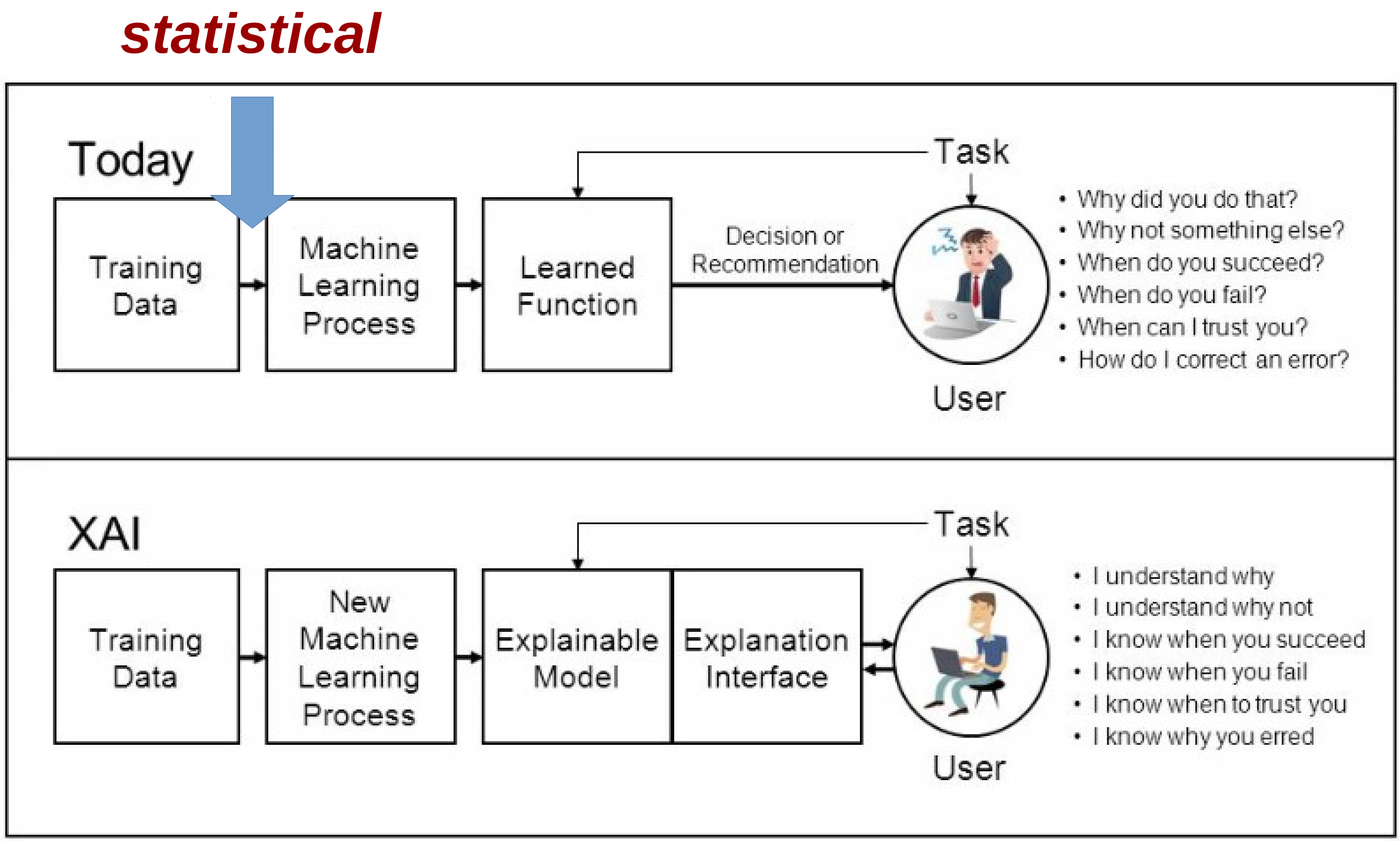
Source: DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence

The call for *Explanaible AI* (XAI)



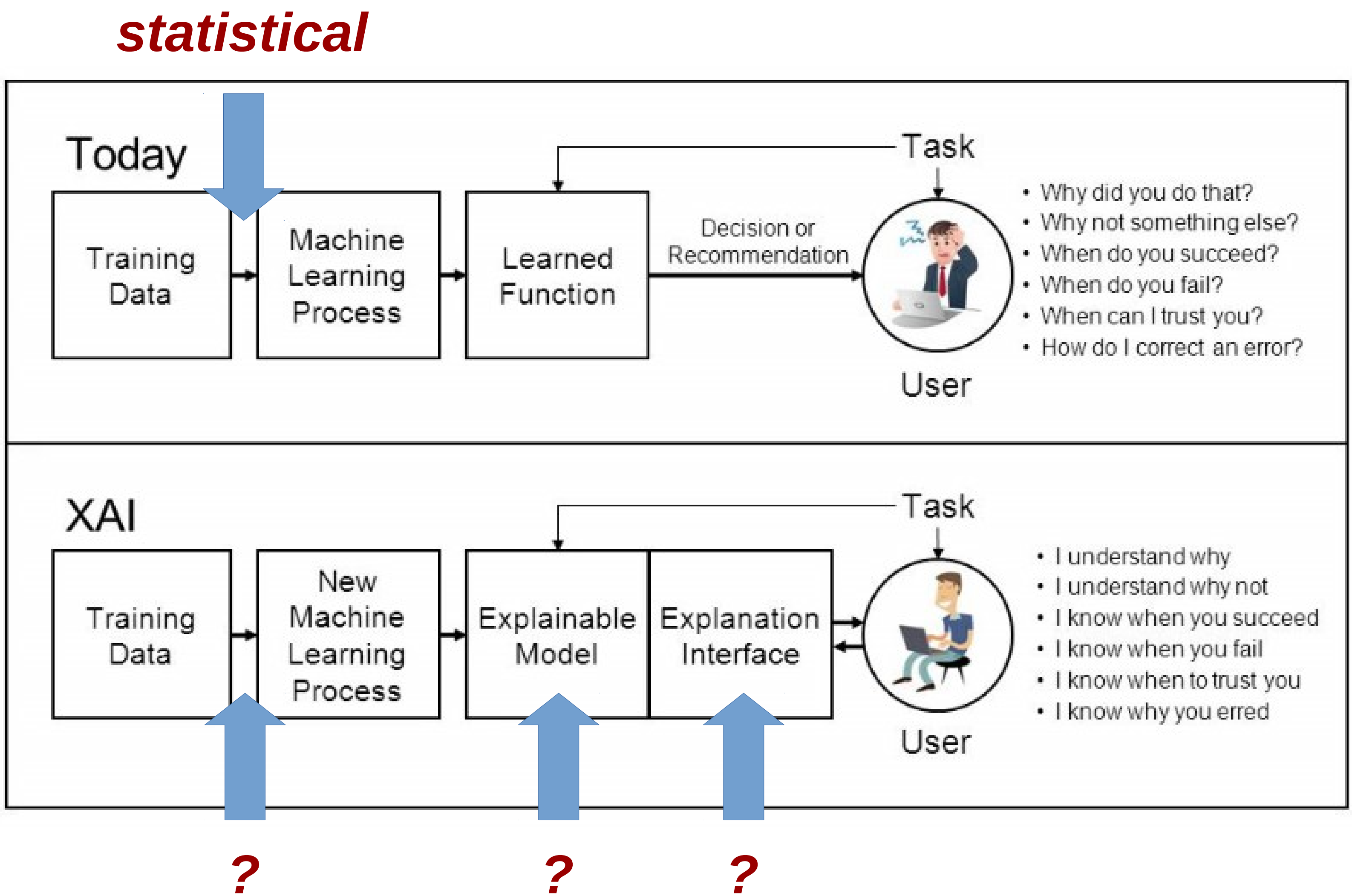
Source: DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence

The call for *Explanaible AI* (XAI)



Source: DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence

The call for *Explanaible AI* (XAI)



Source: DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence

Reasoning

- According to the "***argumentative theory***" of reasoning [Herbert & Spencer, 2011], *reasoning is not meant to take the **best decisions** or **true conclusions***, but to **justify** these choices **in front of the others**.

Reasoning

- According to the "***argumentative theory***" of reasoning [Herbert & Spencer, 2011], *reasoning is not meant to take the* ***best decisions*** *or* ***true conclusions***, but to **justify** these choices **in front of the others**.

- Two functions used in dual roles:
  - ***generate*** arguments that are accepted by the others
  - ***evaluate*** arguments given by others

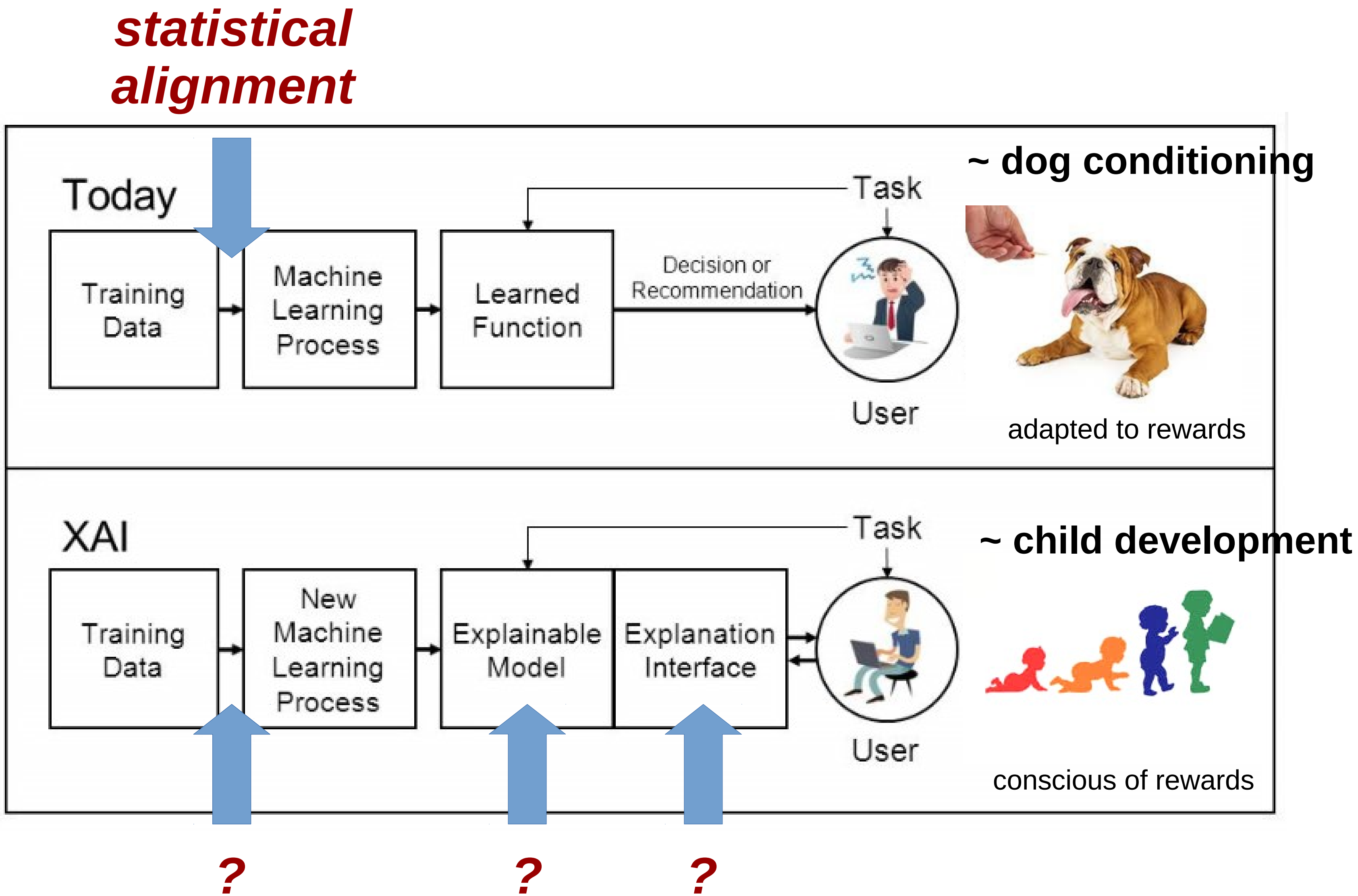
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. The Behavioral and Brain Sciences, 34(2), 57-74

Reasoning

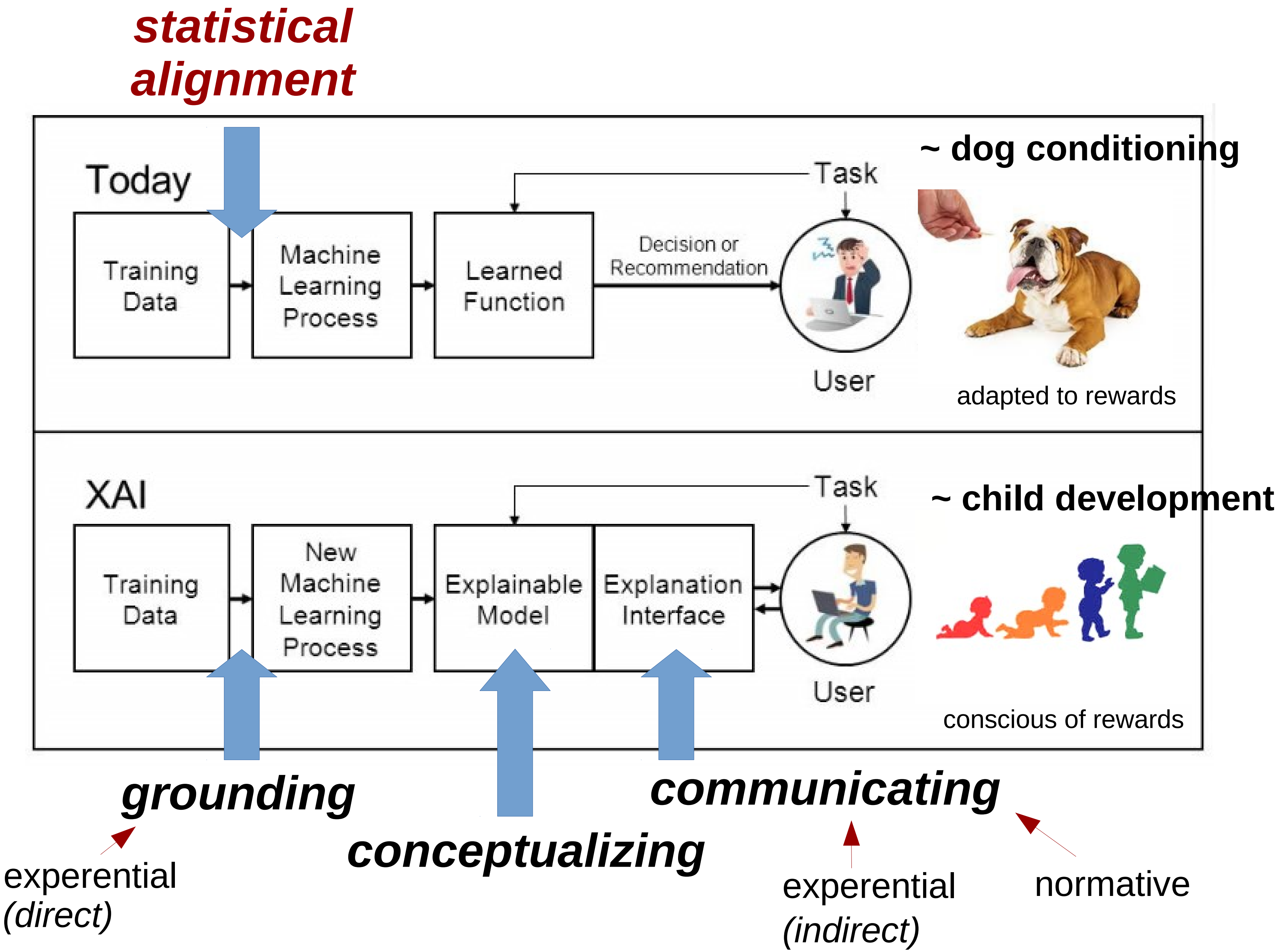
- Herbert & Spencer [2011] insist on the ***persuasion*** aspect:
  - **generation** ↔ convincing others
  - **evaluation** ↔ protecting against being persuaded to take positions resulting in negative outcomes

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. The Behavioral and Brain Sciences, 34(2), 57-74

The call for *Explanaible AI* (XAI)

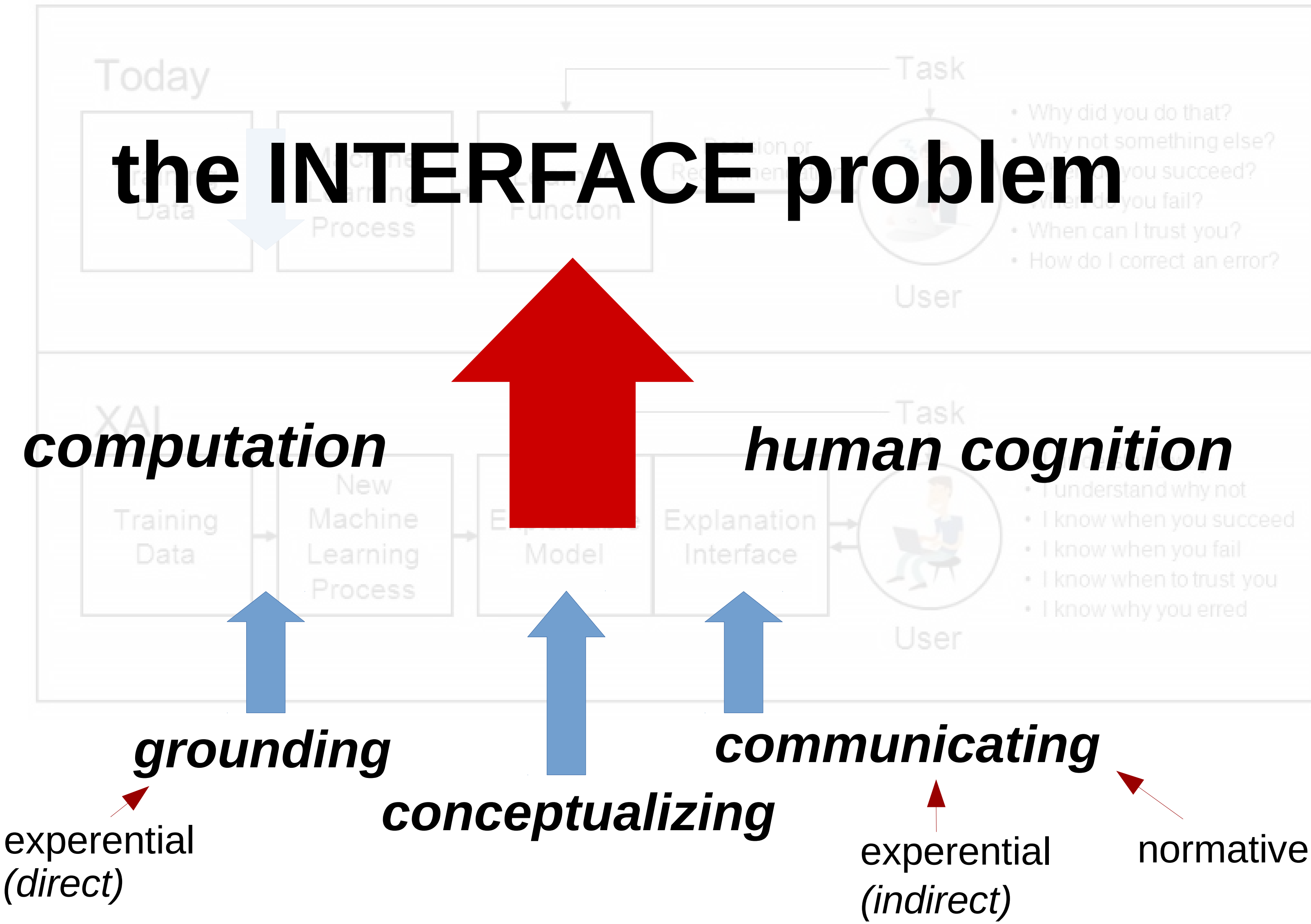


The call for *Explanaible AI* (XAI)



The call for *Explanaible AI* (XAI)

statistical alignment

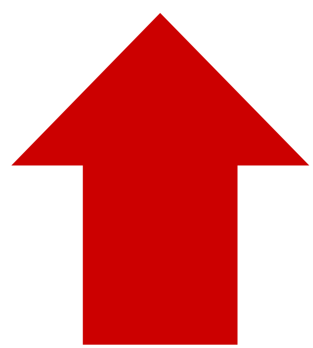


Possible research approaches

- *bottom-up*: use statistical ML to recreate functions ***mimicking*** to some extent human cognition
- *top-down*: conceive algorithms ***reproducing*** by design functions observable in human cognition

the INTERFACE problem

computation



human cognition

Possible research approaches

- *bottom-up*: use statistical ML to recreate functions ***mimicking*** to some extent human cognition
- *top-down*: conceive algorithms ***reproducing*** by design functions observable in human cognition

here we have control on what we want to reproduce

the INTERFACE problem

computation ***human cognition***

Outline of this presentation

- Problems and solutions about similarity [KI2017]
- Computing contrast [AIC2018]
- An introduction to *Simplicity theory* [ST]
  - Pertinence of causes [COG2018]
  - Moral responsibility [JURIX2017]

Sileno, G., Bloch, I., Atif, J., & Dessalles, J.-L. (2017). Similarity and Contrast on Conceptual Spaces for Pertinent Description Generation. Proceedings of the 2017 KI conference, 10505 LNAI

Sileno, G., Bloch, I., Atif, J., & Dessalles, J. (2018). Computing Contrast on Conceptual Spaces. In Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition (AIC2018)

https://simplicitytheory.telecom-paristech.fr/

Sileno, G., & Dessalles, J.-L. (2018). Qualifying Causes as Pertinent. Proceedings of the 40th Conference of the Cognitive Science Society (CogSci 2018)

Sileno, G., Saillenfest, A., & Dessalles, J.-L. (2017). A Computational Model of Moral and Legal Responsibility via Simplicity Theory. Proceedings of the 30th Int. Conf. on Legal Knowledge and Information Systems (JURIX 2017), FAIA 302, 171–176

Unveiling similarity

Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in *similar* context ➡ *similar* response

Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in ~~*similar*~~ context $\Rightarrow$ *similar* response

~ fixing the task

Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in ~~*similar*~~ context ➡ *similar* response

~ *fixing the task*

Practical uses: ***description generation***

proximate elements can be used as **reference** to identify a certain **target** (*object*, *situation*, etc.)

Similarity is crucial to cognition

General (often implicit) hypothesis:

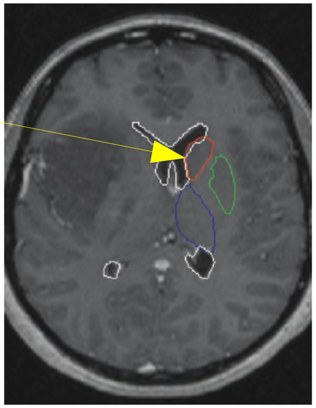
similar stimulus in ~~*similar*~~ context ➡ *similar* response

~ fixing the task

Practical uses: *description generation*

proximate elements can be used as **reference** to identify a certain **target** (*object*, *situation*, etc.)

the caudate nucleus is an internal brain structure which is very close to the lateral ventricles



Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in ~~*similar*~~ context $\Rightarrow$ *similar* response

~ fixing the task

but how two stimuli are defined *similar* ?

Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in ~~*similar*~~ context ➡ *similar* response

~ *fixing the task*

but how two stimuli are defined *similar* ?

psychology

- similarity is a function of a mental ***distance*** between conceptualizations [Shepard1962]

“psychological space” hypothesis

Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in ~~*similar*~~ context → *similar* response

~ fixing the task

but how two stimuli are defined *similar* ?

psychology

- similarity is a function of a mental ***distance*** between conceptualizations [Shepard1962]

“psychological space” hypothesis

machine learning

- relies on some ***metric*** to compare inputs
- offers ***pseudo-metric*** learning methods

Similarity is crucial to cognition

General (often implicit) hypothesis:

similar stimulus in ~~*similar*~~ context ➡ *similar* response

~ *fixing the task*

but how two stimuli are defined *similar* ?

psychology

- similarity is a function of a mental ***distance*** between conceptualizations [Shepard1962]

“psychological space” hypothesis

machine learning

- relies on some ***metric*** to compare inputs
- offers ***pseudo-metric*** learning methods

geometrical model of cognition

psychology **machine learning**

geometrical model of cognition

Problems:

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327–352.

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

- reasoning via artificial devices (still?) relies on ***symbolic*** processing

  e.g. through ontologies

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

- reasoning via artificial devices (still?) relies on ***symbolic*** processing

  e.g. through ontologies

  but.. **symbol grounding?**
  predicate selection?

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

- reasoning via artificial devices (still?) relies on ***symbolic*** processing

  e.g. through ontologies

  but.. **symbol grounding?**
  predicate selection?

Proposed solutions:

- enriching the metric model with additional elements (e.g. density [Krumhansl78])

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

- reasoning via artificial devices (still?) relies on ***symbolic*** processing

  e.g. through ontologies

  but.. **symbol grounding?**
  predicate selection?

Proposed solutions:

- enriching the metric model with additional elements (e.g. density [Krumhansl78])

  but.. **holistic distance?**

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

- reasoning via artificial devices (still?) relies on ***symbolic*** processing

  e.g. through ontologies

  but.. **symbol grounding?**
  predicate selection?

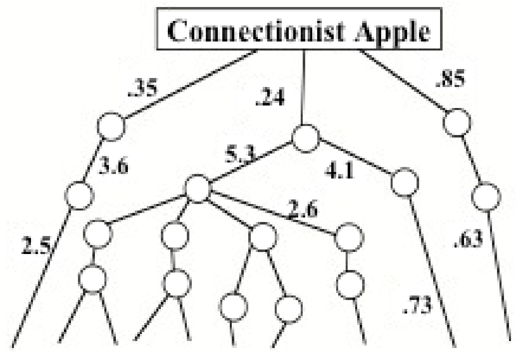
Proposed solutions:

- enriching the metric model with additional elements (e.g. density [Krumhansl78])

  but.. **holistic distance?**

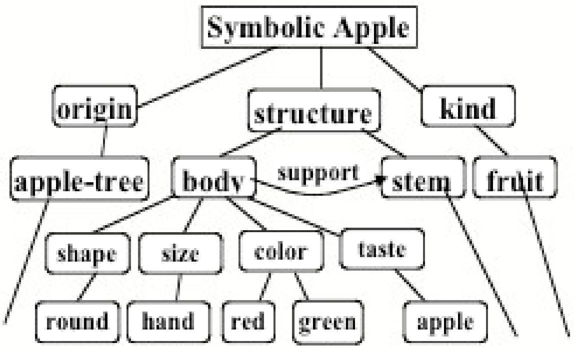
- approaching logical structures through geometric methods (e.g. [Distel2014])

Towards an alternative solution..



grounded
not intelligible

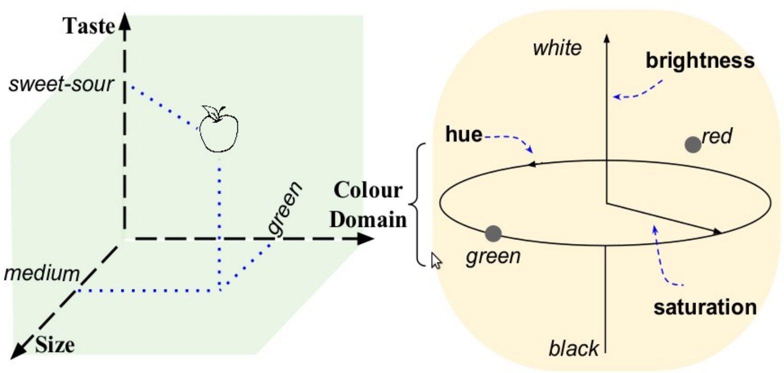
associationistic methods



symbolic methods

not grounded
intelligible

Towards an alternative solution..



grounded
not intelligible

associationistic methods

grounded
and intelligible

conceptual spaces

symbolic methods

not grounded
intelligible

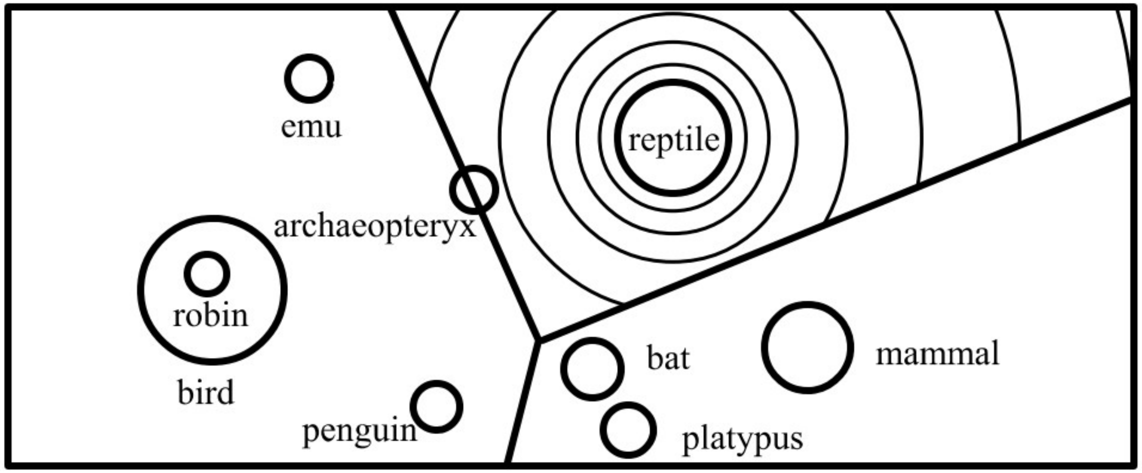
Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought. MIT Press.
Gärdenfors, P. (2014). The Geometry of Meaning: Semantics Based on Conceptual Spaces. MIT Press.

Overview on conceptual spaces

- Conceptual spaces stem from (continuous) **perceptive spaces**. → ***grounded***
- Natural **properties** emerge as **convex regions** over ***integral dimensions*** *(e.g. color).*
- **Concepts** are ***weighted combinations*** of properties
- **Prototypes** can be seen as **centroids** of convex regions (properties or concepts).

  Convex regions can be seen as resulting from the competition between prototypes (forming a *Voronoi Tessellation*).

conceptual spaces

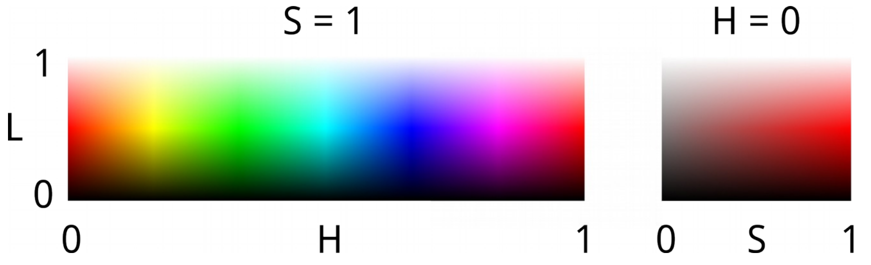


Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought. MIT Press.
Gärdenfors, P. (2014). The Geometry of Meaning: Semantics Based on Conceptual Spaces. MIT Press.

“small” problem

The standard theory of conceptual spaces insists to ***lexical meaning***: linguistic marks are associated to regions.
→ ***extensional*** as the standard symbolic approach.

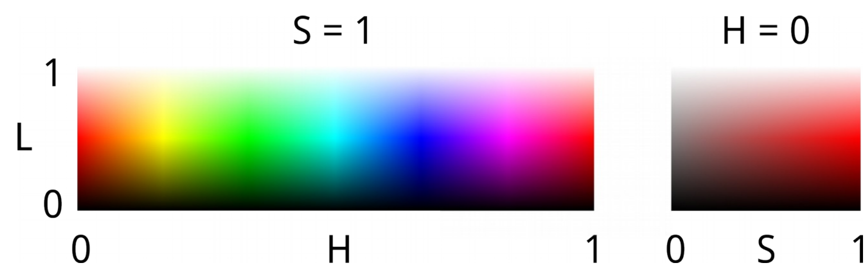
If ***red***, or ***green***, or ***brown*** correspond to regions in the color space...



“small” problem

The standard theory of conceptual spaces insists to ***lexical meaning***: linguistic marks are associated to regions.
→ ***extensional*** as the standard symbolic approach.

If ***red***, or ***green***, or ***brown*** correspond to regions in the color space...



why do we say “**red dogs**” even if they are actually brown?

images after Google

Predicates resulting from contrast

Alternative hypothesis [Dessalles2015]:

Predicates are generated *on the fly* after an operation of ***contrast***.

$$C = O - P$$

object ***prototype***

contrastor ***(target)*** ***(reference)***

Dessalles, J.-L. (2015). From Conceptual Spaces to Predicates. Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation, 17–31.

Predicates resulting from contrast

Alternative hypothesis [Dessalles2015]:

Predicates are generated *on the fly* after an operation of ***contrast***.

$$C = O - P \rightsquigarrow \text{“red”}$$

contrastor ***object (target)*** ***prototype (reference)***

These dogs are “red dogs”:

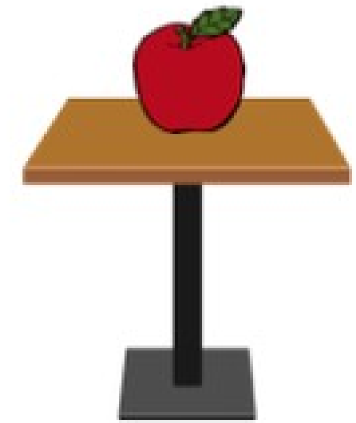


- not because their color is red (they are brown),
- because they are ***more red*** with respect to the dog prototype

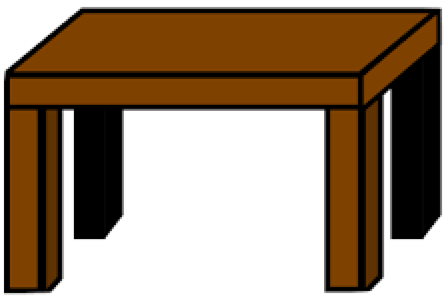
Predicates resulting from contrast

In logic, usually: *above(a, b)* $\leftrightarrow$ *below(b, a)*

However, people don't say



“the table is
below the apple.”

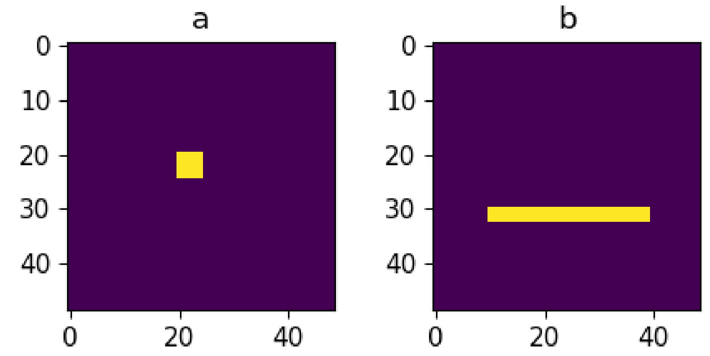


“the board is
above the leg.”

If the contrastive hypothesis is correct, $C = A - B \rightsquigarrow$ “above”

Directional relationships

We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).

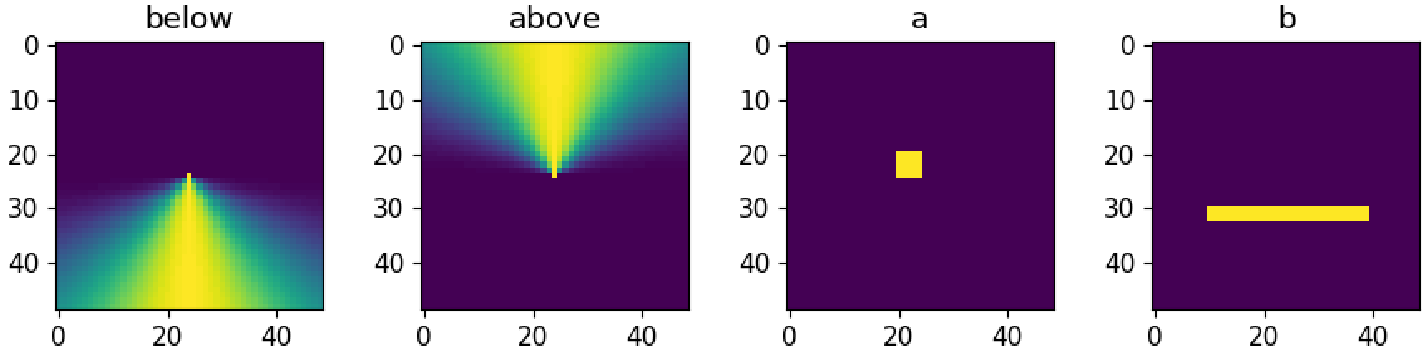


objects

Bloch, I. (2006). Spatial reasoning under imprecision using fuzzy set theory, formal logics and mathematical morphology. International Journal of Approximate Reasoning, 41(2), 77–95.

Directional relationships

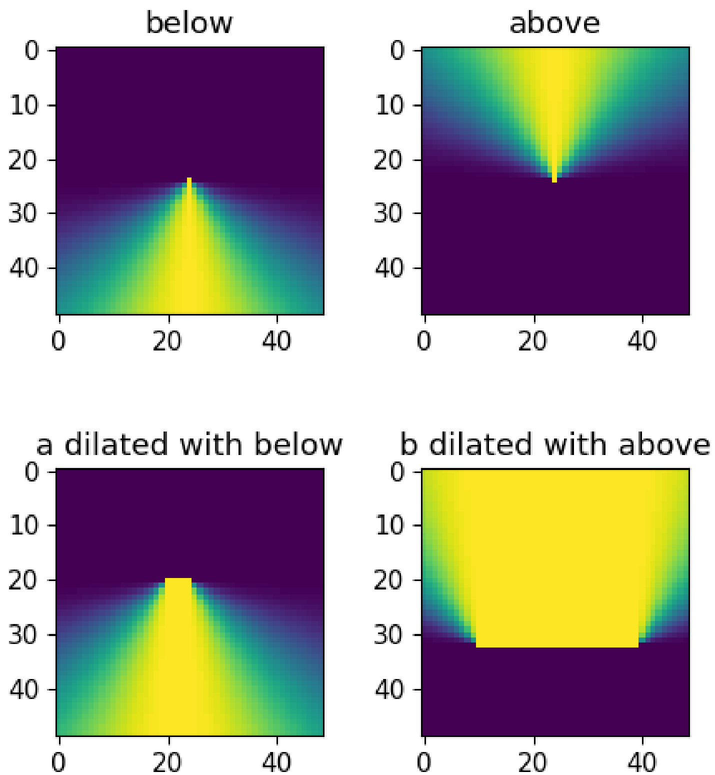
We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



models of relations
for a point centered
in the origin

Directional relationships

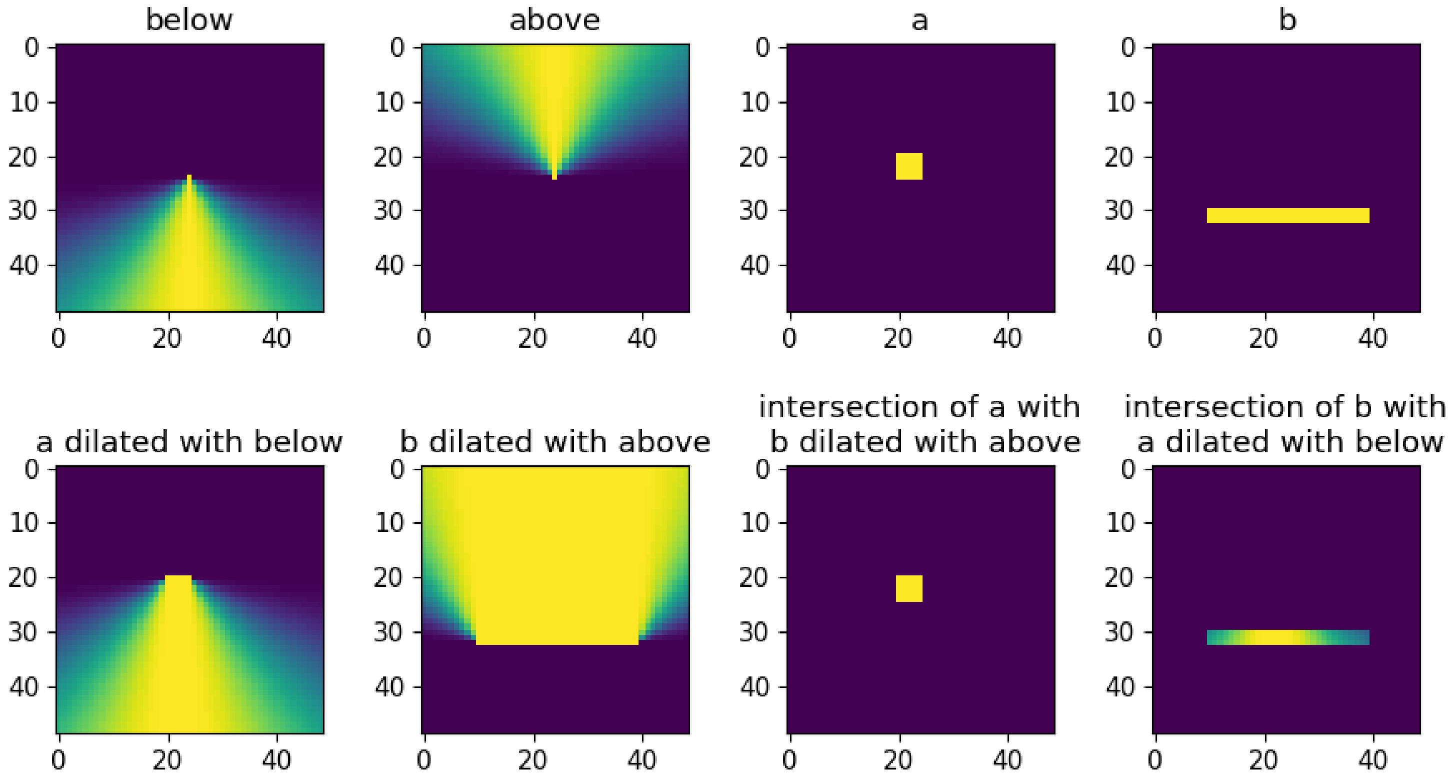
We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



“below a” “above b”

Directional relationships

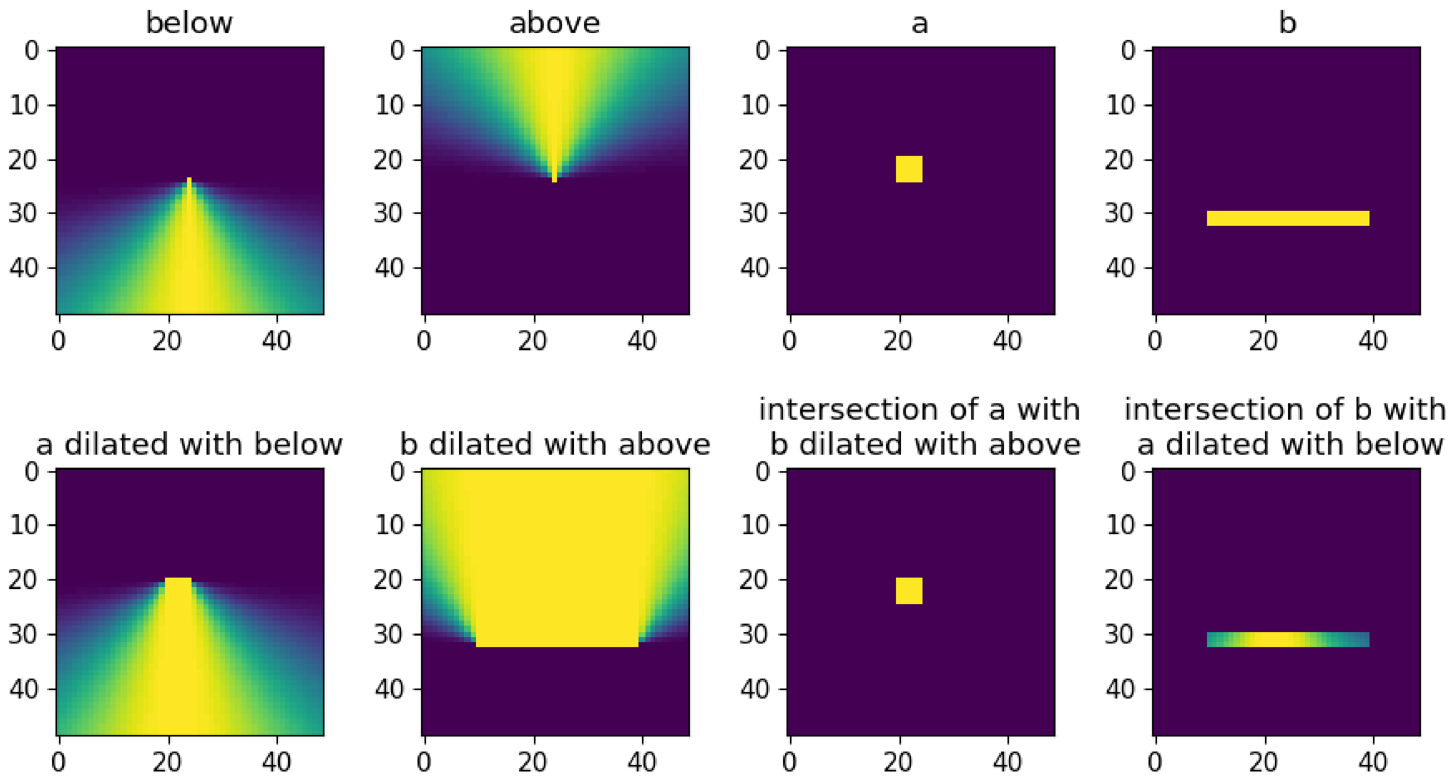
We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



"below a" "above b" how much a is (in) "above b" how much b is (in) "below a"

Directional relationships

We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).

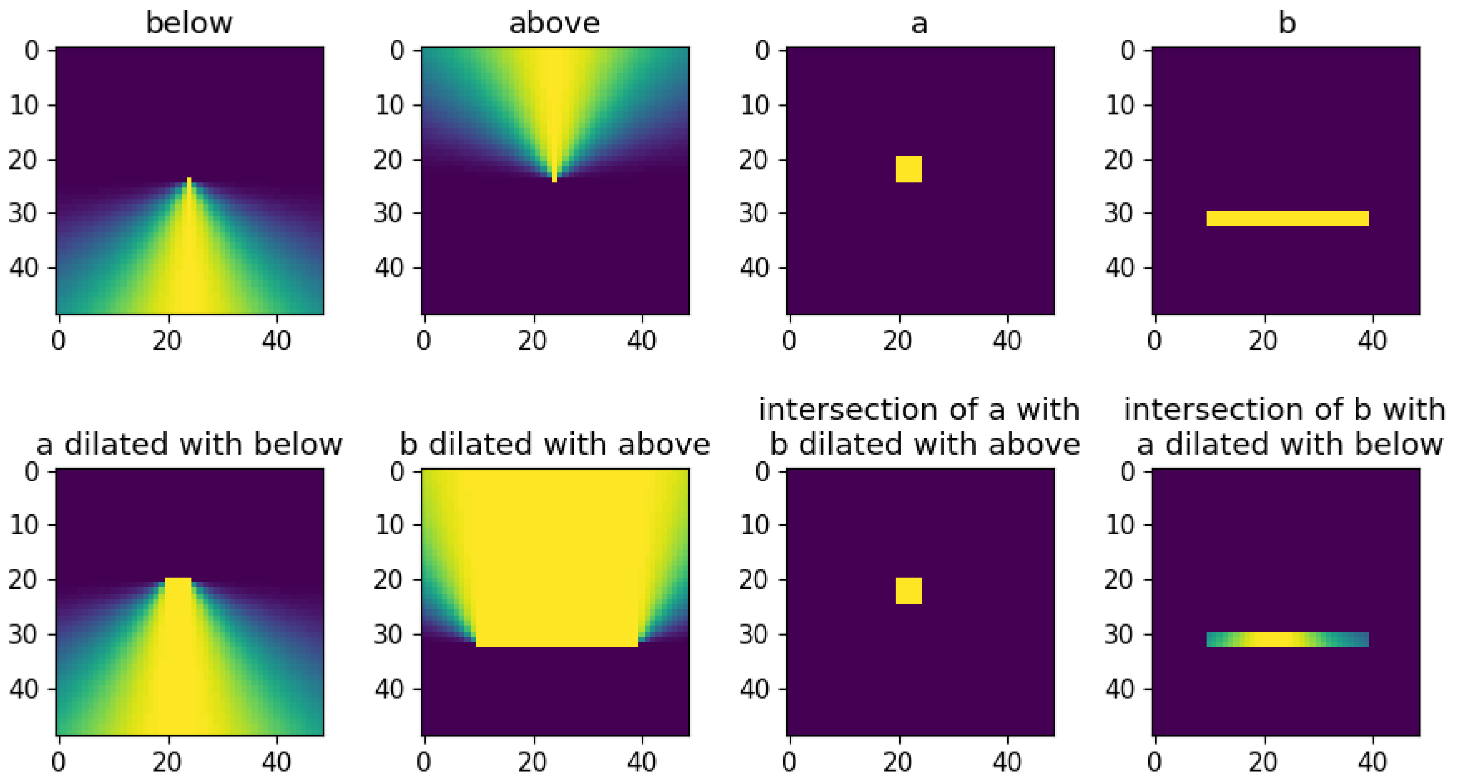


how much a is "above b"

operation scheme: $a \rightsquigarrow b + \text{“above”}$

Directional relationships

We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



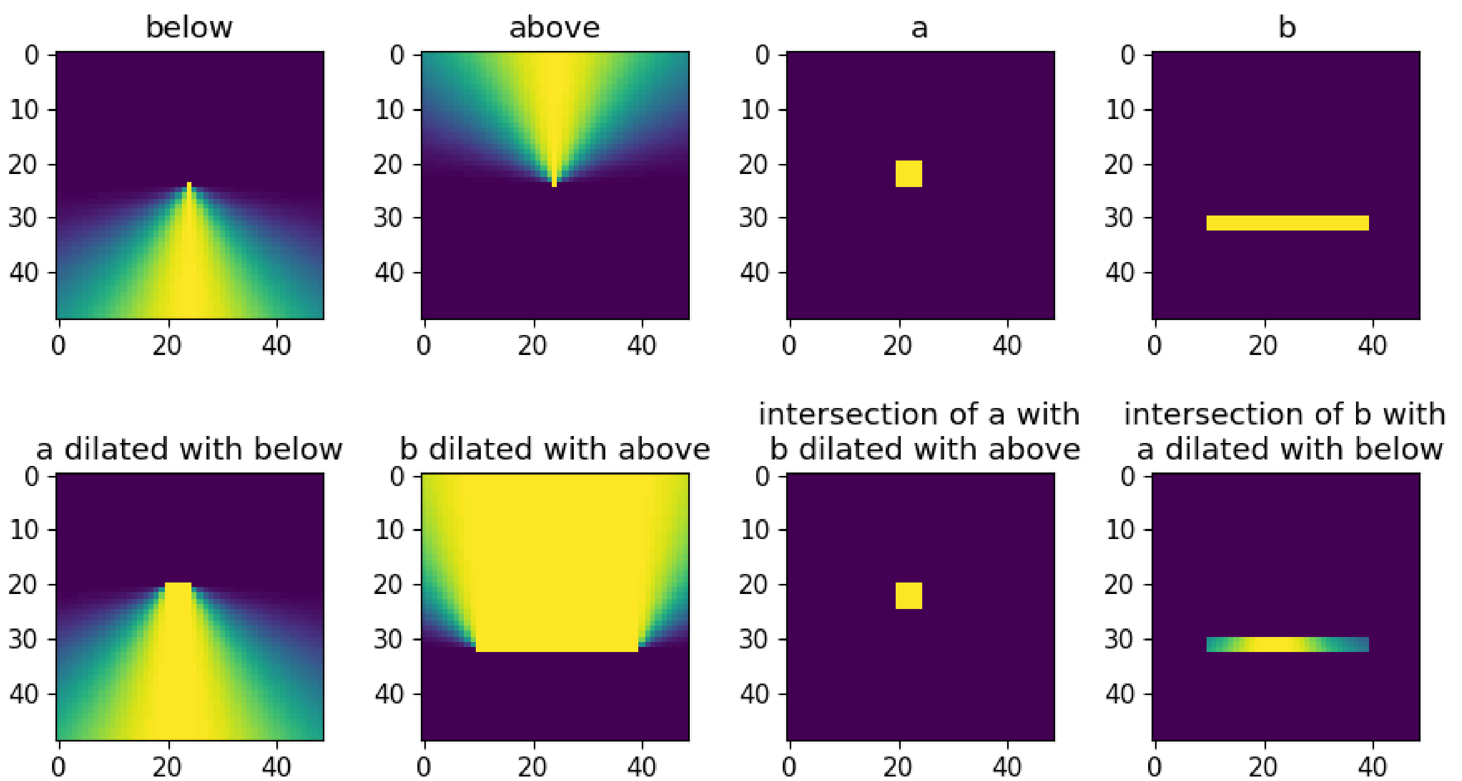
how much a is “above b”

inverse operation to contrast: ***merge***

operation scheme: $a \rightsquigarrow b + \text{“above”}$

Directional relationships

We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



how much a is "above b"

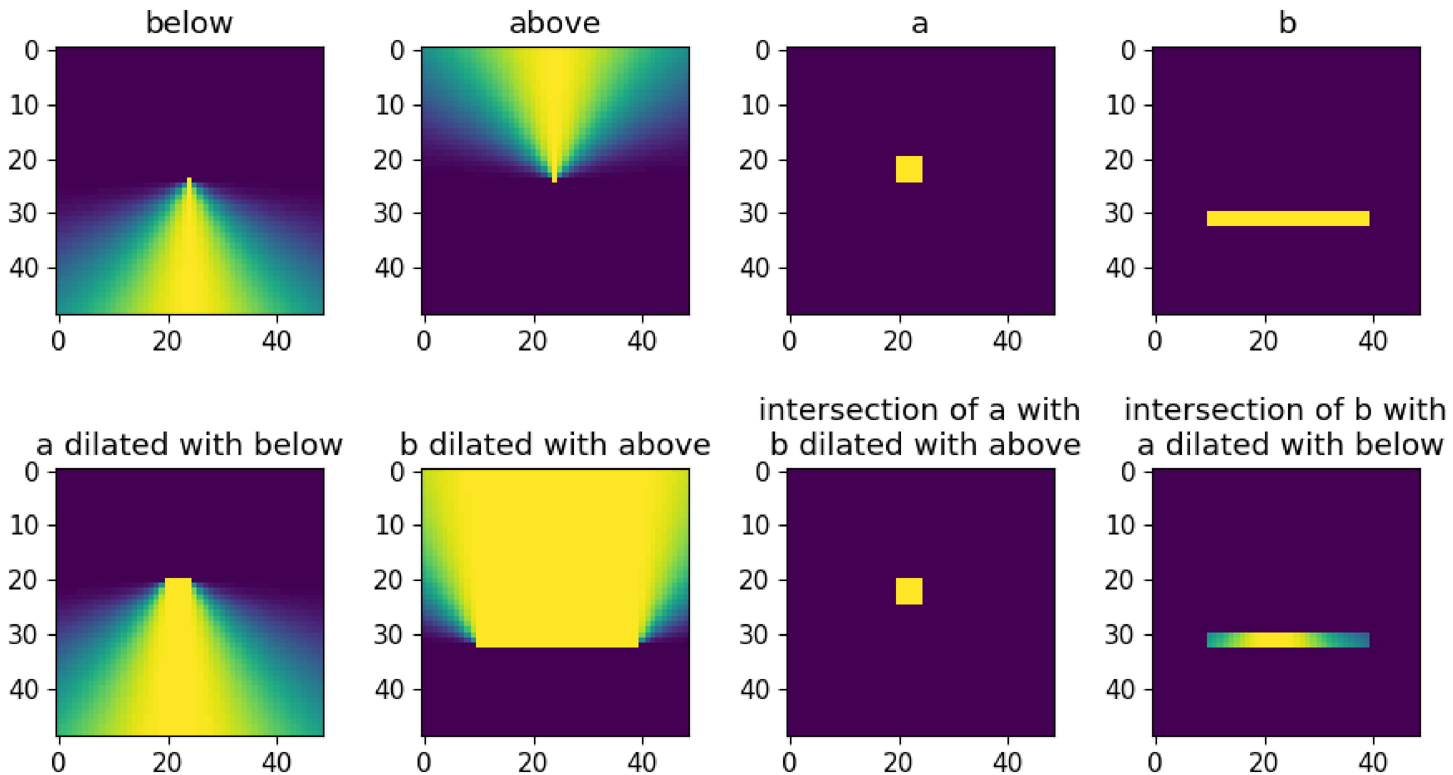
inverse operation to contrast: ***merge***

operation scheme: $a \rightsquigarrow b + \text{“above”}$

alignment as overlap

Directional relationships

We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



how much a is "above b"

inverse operation to contrast: ***merge***

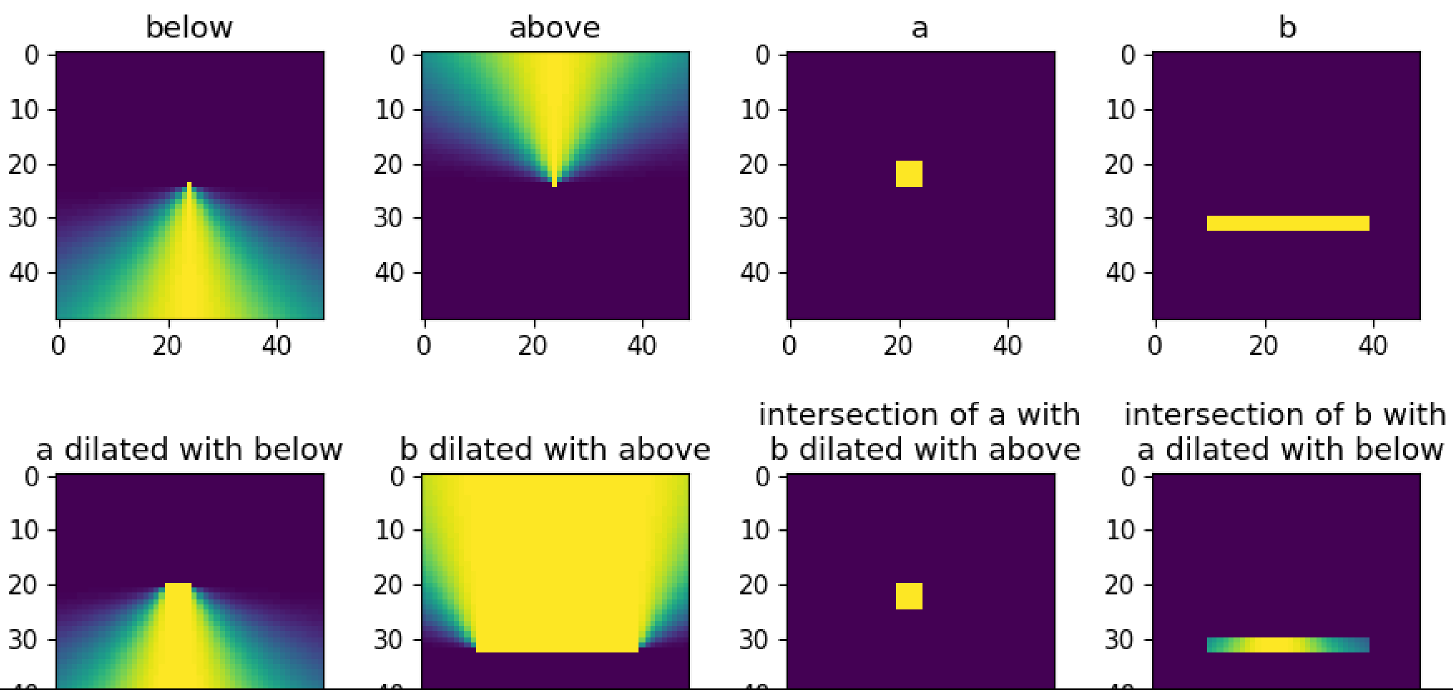
operation scheme: $a \leadsto b$ + *"above"*

cf. with o - $p \leadsto$ *"red"*

alignment as overlap

Directional relationships

We considered an existing method [Bloch2006] used in image processing to compute directional relative positions of visual entities (e.g. of biomedical images).



If we settle upon contrast, we can categorize its output for relations!

"above b"

operation scheme: $a \rightsquigarrow b$ + *"above"*

cf. with $o - p \rightsquigarrow$ *"red"*

alignment as overlap

From contrast to concept similarity

- Contrast has been computed by operations inherent to integral dimensions. These may be interpreted as related to local perceptual **dissimilarity**.
  - no need to define a ***holistic distance***
- *But what about concept (i.e. multi-dimensional) similarity?*

From contrast to concept similarity

“she is strong.”

this person − prototype person ↝ *“strong”*

From contrast to concept similarity

“she is strong.”

this person − prototype person ↝ *“strong”*

(metaphor as conceptual analogy)

“she is (like) a lion.”

From contrast to concept similarity

“she is strong.”

this person − prototype person ↝ *“strong”*

(metaphor as conceptual analogy)

“she is (like) a lion.”

double contrast

target

this person − prototype person ↝ *“strong”*, *etc.*

prototype lion − prototype animal ↝ *“strong”*, *etc.*

reference

comparison ground

From contrast to concept similarity

“she is strong.”

this person − prototype person ↝ *“strong”*

(metaphor as conceptual analogy)

“she is (like) a lion.”

double contrast

target

this person − prototype person ↝ *“strong”, etc.*

prototype lion − prototype animal ↝ *“strong”, etc.*

reference

comparison ground

*The reference activates certain **discriminating features.***

From contrast to concept similarity

“she is strong.”

this person − prototype person ↝ *“strong”*

(metaphor as conceptual analogy)

“she is (like) a lion.”

double contrast

target

this person − prototype person ↝ *“strong”, etc.*

prototype lion − prototype animal ↝ *“strong”, etc.*

reference

comparison ground

*The reference activates certain **discriminating features.***

Concept similarity is a sequential, multi-layered computation

psychology **machine learning**

geometrical model of cognition

Problems:

- similarity in human judgments does not satisfy **fundamental geometric axioms** [Tversky77]

  basis of feature-based models

  but.. **feature selection?**

- reasoning via artificial devices (still?) relies on ***symbolic*** processing

  e.g. through ontologies

  but.. **symbol grounding?**
  predicate selection?

Proposed solutions:

- enriching the metric model with additional elements (e.g. density [Krumhansl78])

  but.. **holistic distance?**

- approaching logical structures through geometric methods (e.g. [Distel2014])

1. Problems with **symmetry**

$$d(a, b) = d(b, a)$$

- *Distance between two points should be the same when inverting the terms of comparison.*

1. Problems with **symmetry**

$$d(a,b) = d(b,a)$$

- *Distance between two points should be the same when inverting the terms of comparison.*

However,

Tel Aviv is like New York

has a different meaning than:

New York is like Tel Aviv

1. Problems with **symmetry**

$$d(a, b) = d(b, a)$$

- *Distance between two points should be the same when inverting the terms of comparison.*

However,

Tel Aviv is like New York

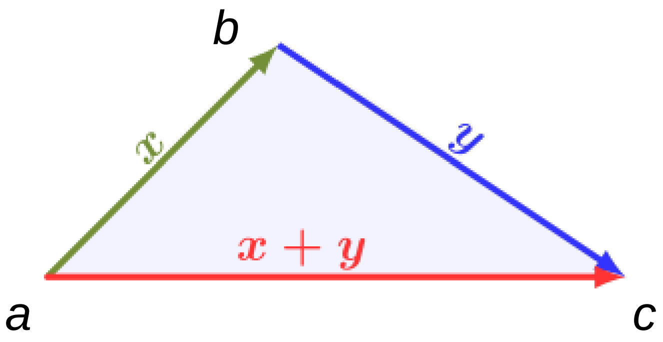
has a different meaning than:

New York is like Tel Aviv

Our explanation: **changing of reference activates different features**

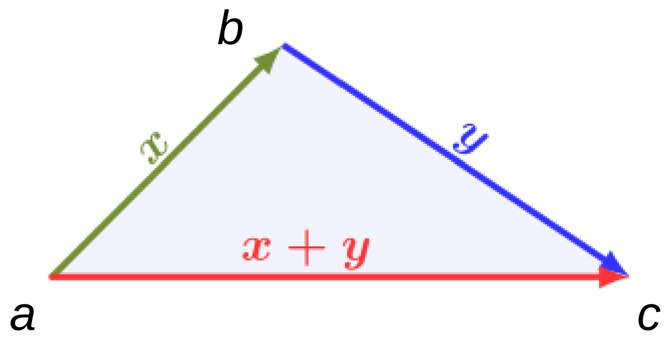
2. Problems with **triangle inequality**

$$d(a,b) + d(b,c) \geq d(a,c)$$



2. Problems with **triangle inequality**

$$d(a,b) + d(b,c) \geq d(a,c)$$

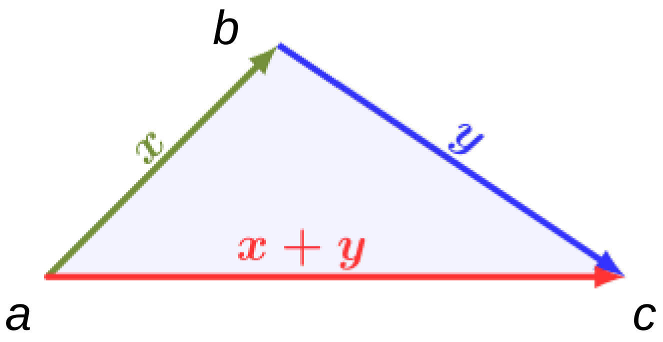


However,

Jamaica is similar to Cuba
Cuba is similar to Russia
Jamaica is ***not*** similar to Russia.

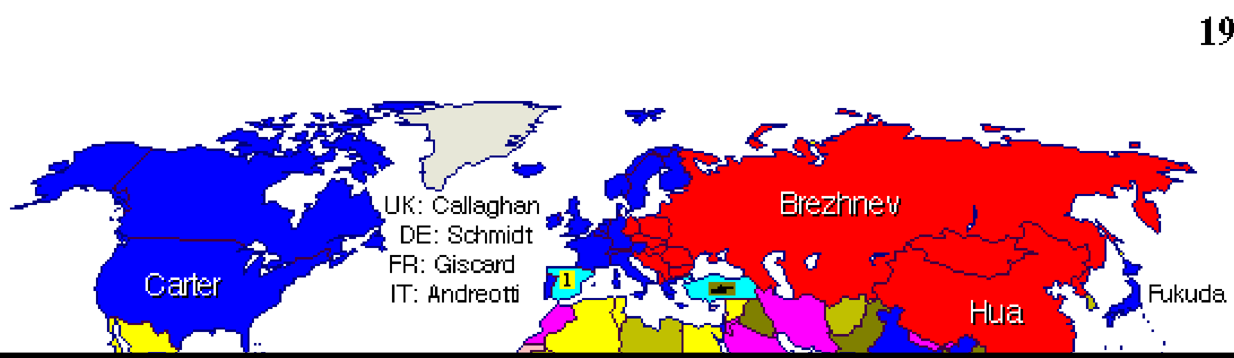
2. Problems with **triangle inequality**

$$d(a,b) + d(b,c) \geq d(a,c)$$



However,

Jamaica is similar to Cuba
Cuba is similar to Russia
Jamaica is ***not*** similar to Russia.



Our explanation: **different/no *comparison grounds* after contrast**

3. Problems with **minimality**

$$d(a, b) \geq d(a, a) = 0$$

- *Distance with a distinct point should be greater than with the point itself.*

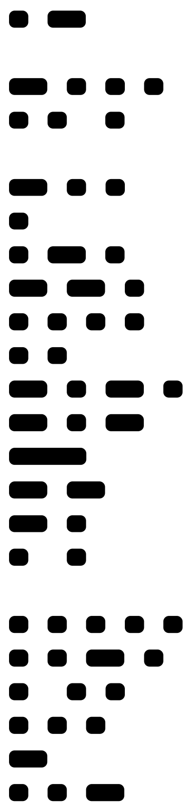
3. Problems with **minimality**

$$d(a, b) \geq d(a, a) = 0$$

- *Distance with a distinct point should be greater than with the point itself.*

However,

– when people were asked to find the most similar Morse code within a list, including the original one, they did not always return the object itself.



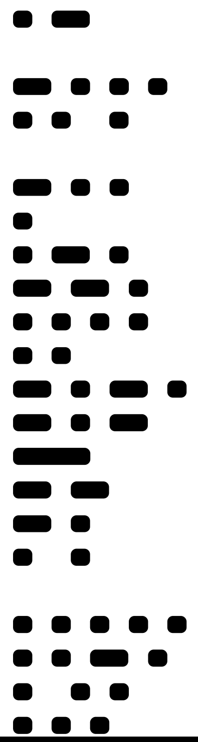
3. Problems with **minimality**

$$d(a,b) \geq d(a,a) = 0$$

- *Distance with a distinct point should be greater than with the point itself.*

However,

- when people were asked to find the most similar Morse code within a list, including the original one, they did not always return the object itself.



Our explanation: **sequential nature of similarity assessment.**

4. **Diagnosticity** effect

- *The distance between two points in a set should not change when changing the set.*

4. **Diagnosticity** effect

- *The distance between two points in a set should not change when changing the set.*

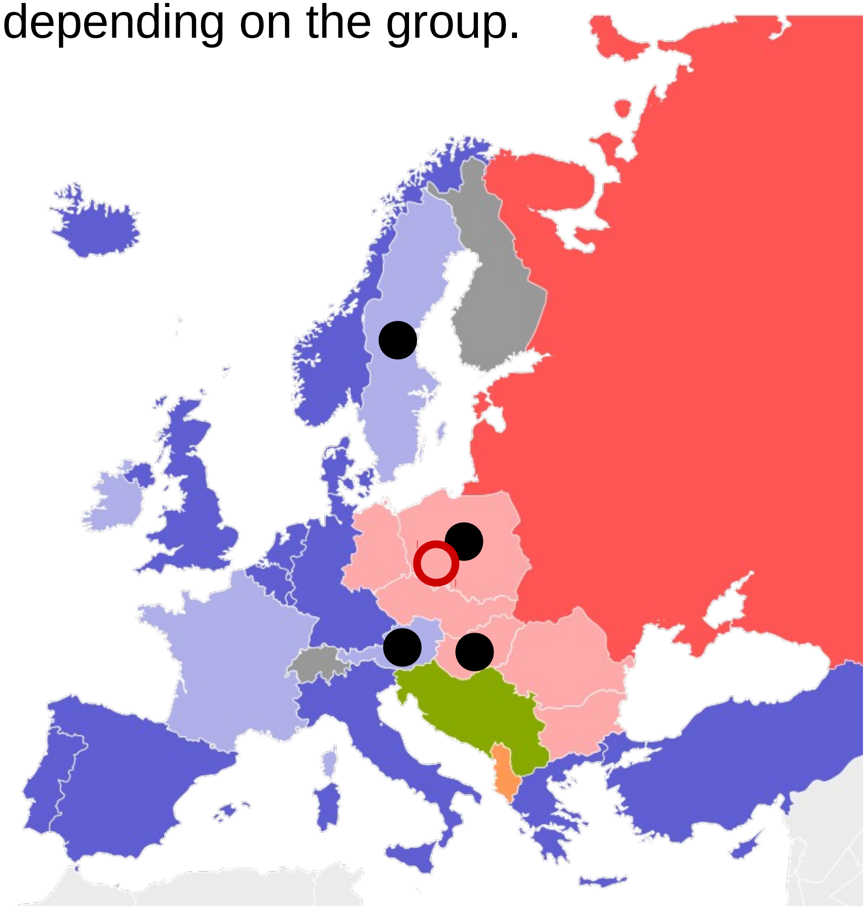
However,

– when people were asked for the country most similar to a reference amongst a given group of countries, they changed answers depending on the group.

Austria

most similar to

Hungary
Poland
Sweden



4. **Diagnosticity** effect

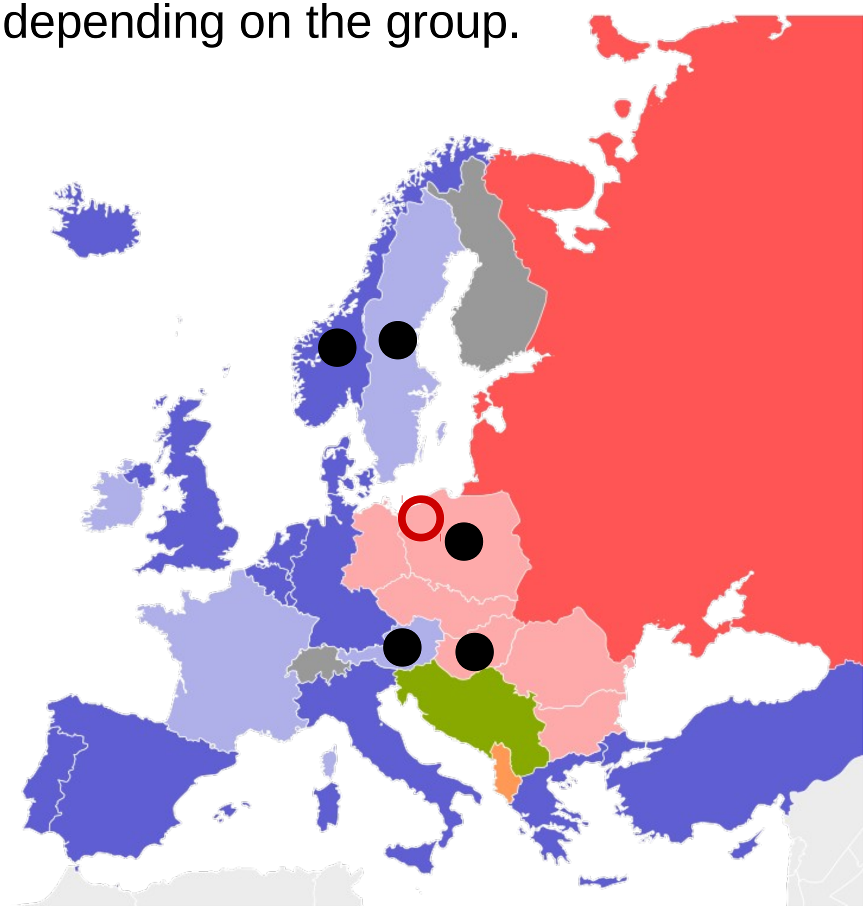
- *The distance between two points in a set should not change when changing the set.*

However,

– when people were asked for the country most similar to a reference amongst a given group of countries, they changed answers depending on the group.

Austria ← *most similar to*

Hungary
Poland
Sweden
Norway



4. **Diagnosticity** effect

- *The distance between two points in a set should not change when changing the set.*

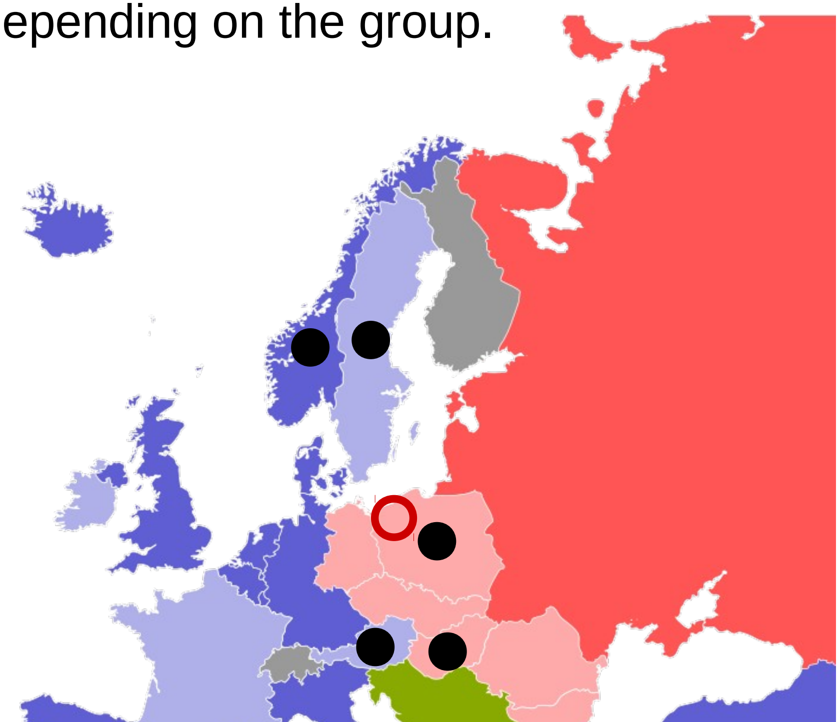
However,

– when people were asked for the country most similar to a reference amongst a given group of countries, they changed answers depending on the group.

Austria

most similar to

Hungary
Poland
Sweden
Norway



Our explanation: effect due to the **change of group prototype**

Two types of similarity

- There is a fundamental distinction between:
  - perceptual similarity
  - contrastively analogical similarity
- The two are commonly conflated:
  - by using MDS on people's similarity judgments to elicit dimensions of psychological (conceptual) spaces
  - in similar dimensional reduction techniques used in ML
- This hypothesis provides simple explanations to empirical experiences manifesting non-metrical properties, yet maintaining a geometric infrastructure.

Sileno, G., Bloch, I., Atif, J., & Dessalles, J.-L. (2017). Similarity and Contrast on Conceptual Spaces for Pertinent Description Generation. Proceedings of the 2017 KI conference, 10505 LNAI.

How does contrast work?

Computing contrast (1D)

- Consider coffees served in a bar. Intuitively, whether a coffee is qualified as being ***hot*** or ***cold*** depends mostly on what the speaker expects of coffees served at bars, rather than a specific absolute temperature.

$$c = o - p \rightsquigarrow \text{“hot”}$$

contrastor

object (target)

prototype (reference)

Sileno, G., Bloch, I., Atif, J., & Dessalles, J. (2018). Computing Contrast on Conceptual Spaces. In Proceedings of the 6th International Workshop on Artificial Intelligence and Cognition (AIC2018)

Computing contrast (1D)

- Consider coffees served in a bar. Intuitively, whether a coffee is qualified as being ***hot*** or ***cold*** depends mostly on what the speaker expects of coffees served at bars, rather than a specific absolute temperature.

$$c = o - p \rightsquigarrow \text{“hot”}$$

contrastor

object
(target)

prototype
(reference)

- For simplicity, we represent objects on 1D (temperature) with real coordinates.

Computing contrast (1D)

$$c = o - p \rightsquigarrow \text{“hot”}$$

object ***prototype***

contrastor ***(target)*** ***(reference)***

- Because prototypes are defined together with a concept region, let us consider some regional information, for instance represented as an *egg-yolk* structure.

Computing contrast (1D)

$$c = o - p \rightsquigarrow \text{“hot”}$$

contrastor ***object (target)*** ***prototype (reference)***

- Because prototypes are defined together with a concept region, let us consider some regional information, for instance represented as an *egg-yolk* structure.
  - internal boundary (*yolk*) $\boldsymbol{p \pm \sigma}$ for **typical** elements of that category of objects (e.g. coffee served at bar).

Computing contrast (1D)

$$c = o - p \rightsquigarrow \text{“hot”}$$

object ***prototype***

contrastor ***(target)*** ***(reference)***

- Because prototypes are defined together with a concept region, let us consider some regional information, for instance represented as an *egg-yolk* structure.
  - internal boundary (*yolk*) $\boldsymbol{p \pm \sigma}$ for **typical** elements of that category of objects (e.g. coffee served at bar).
  - external boundary (*egg*) $\boldsymbol{p \pm \rho}$ for **all** elements directly associated to that category of objects

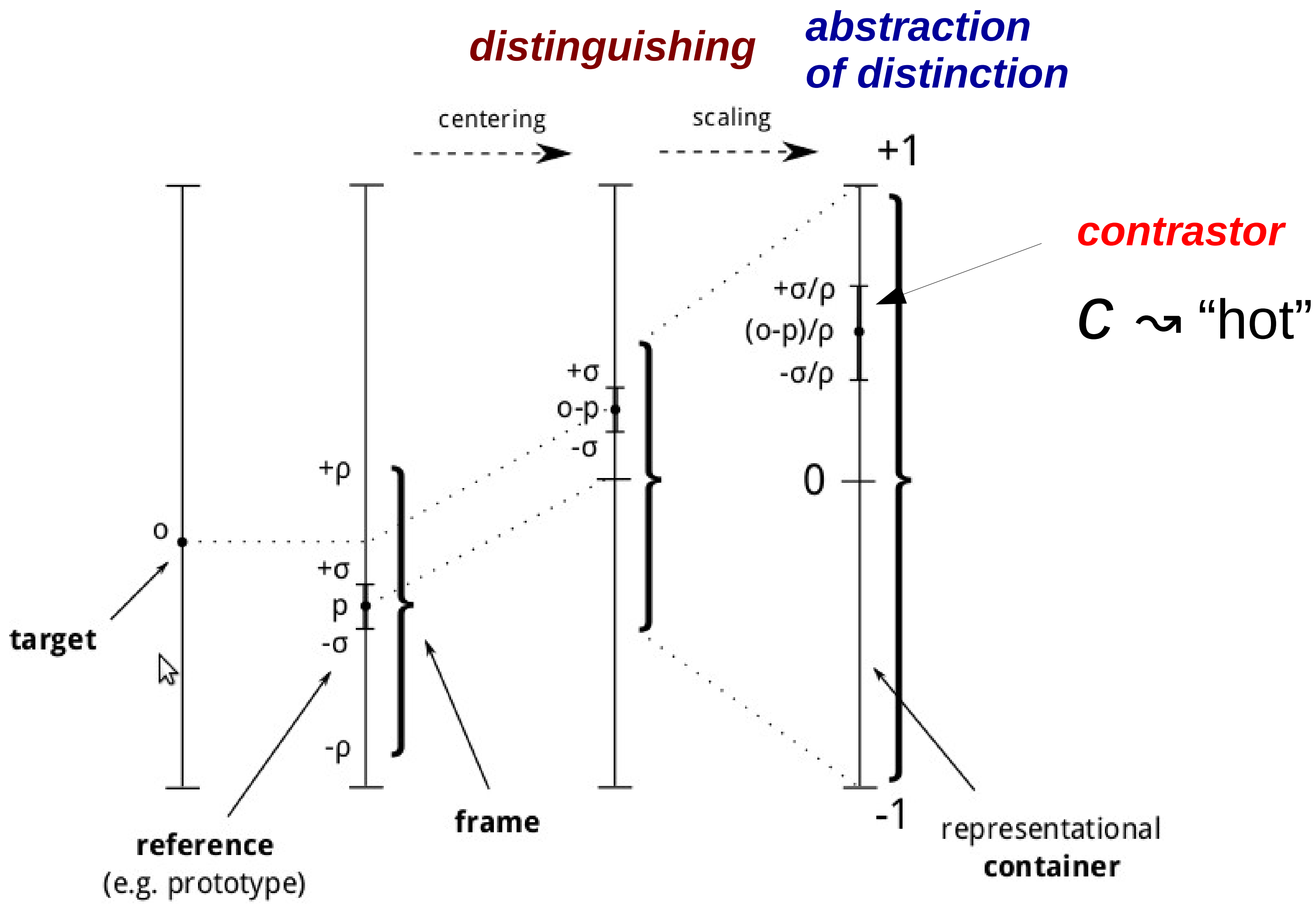
Computing contrast (1D)

$$c = o - p \rightsquigarrow \text{“hot”}$$

contrastor ***object (target)*** ***prototype (reference)***

- Two required functions:
  - ***centering*** of target with respect to typical region
  - ***scaling*** to neutralize effect of scale (e.g. “hot coffee”, “hot planet”)

Computing contrast (1D)



$$C = \mathsf{contrast}(o, \langle p, \sigma, \rho \rangle) = \odot^{\sigma}_{o-p} * \frac{1}{\rho} = \odot^{\sigma/\rho}_{(o-p)/\rho}$$

Computing contrast (1D)

- As *contrastors* are extended objects, they might be compared to model categories represented as regions by measuring their *degree of overlap:*

contrastor

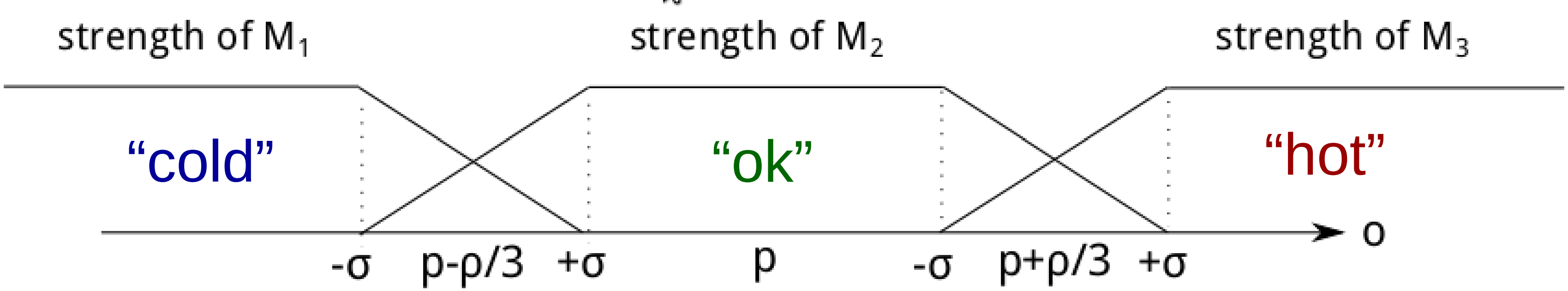
model region of property

$$\text{strength}(r) = \frac{|C \cap M^{(r)}|}{|C|}$$

property label

Computing contrast (1D)

- Applying the previous computation, we easily derive the ***membership functions*** of some general relations with respect to the objects of that category.
- For instance, by dividing the representational container in 3 equal parts, we have:



Computing contrast (1D)

- The previous formulation might be extended to consider contrast between two regions, by utilizing ***discretization*** ($\lfloor \cdot \rfloor$ denotes the approximation to the nearest integer):

$$C = \mathsf{contrast}_{\mathsf{R}}(\langle t, \tau \rangle, \langle r, \sigma \rangle, \langle f, \rho \rangle) \approx \mathsf{contrast}\left(\lfloor \frac{t}{2\tau} \rfloor, \left\langle \lfloor \frac{r}{2\tau} \rfloor, \lfloor \frac{\sigma}{\tau} \rfloor, \lfloor \frac{\rho}{\tau} \rfloor \right\rangle \right)$$

Computing contrast (>1D)

- If dimensions are ***perceptually independent***, we can apply contrast on each dimensions separately:

$$C = (C_1, \ldots, C_n) = (\mathsf{contrast}(o_1, \langle p_1, \sigma_1, \rho_1 \rangle), \ldots, \mathsf{contrast}(o_n, \langle p_n, \sigma_n, \rho_n \rangle))$$

- The result can be used to create a contrastive description of the object, i.e. its ***most distinguishing*** features.

- e.g. *apple* (as a fruit):
  red, spherical, quite sugared

Computing contrast (>1D)

- In the case of 2D visual objects, the two dimensions are ***not*** perceptually independent.
- Let us consider two objects A and B. We can apply contrast iteratively for each point of A with respect to B, and then ***aggregate*** the resulting contrastors.

Computing contrast (>1D)

- In the case of 2D visual objects, the two dimensions are ***not*** perceptually independent.
- Let us consider two objects A and B. We can apply contrast iteratively for each point of A with respect to B, and then ***aggregate*** the resulting contrastors.

accumulation set

$$\mathcal{H}(A, B)(z) = \{a \in A, b \in B | a - b = z\}$$

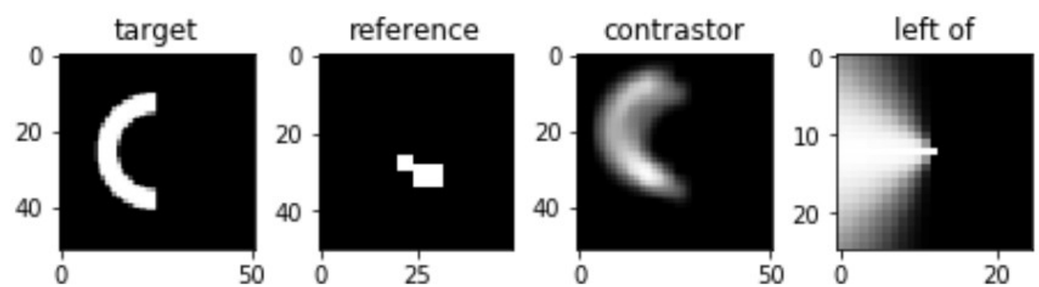
counting

$$\mu_{A \ominus B}(z) = \frac{|\mathcal{H}(A, B)(z)|}{\max_w |\mathcal{H}(A, B)(w)|}$$

normalization

Computing contrast (>1D)

- In the case of 2D visual objects, the two dimensions are ***not*** perceptually independent.
- Let us consider two objects A and B. We can apply contrast iteratively for each point of A with respect to B, and then ***aggregate*** the resulting contrastors.



accumulation set

$$\mathcal{H}(A, B)(z) = \{a \in A, b \in B | a - b = z\}$$

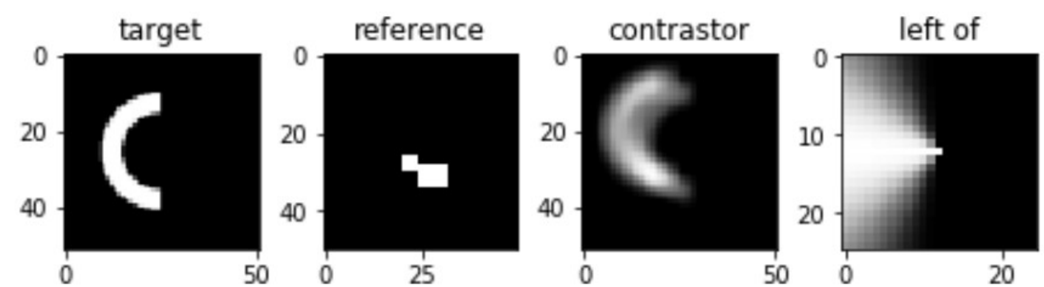
counting

$$\mu_{A \ominus B}(z) = \frac{|\mathcal{H}(A, B)(z)|}{\max_w |\mathcal{H}(A, B)(w)|}$$

normalization

Computing contrast (>1D)

- In the case of 2D visual objects, the two dimensions are ***not* perceptually independent**.
- Let us consider two objects A and B. We can apply contrast iteratively for each point of A with respect to B, and then ***aggregate*** the resulting contrastors.



accumulation set

$$\mathcal{H}(A,B)(z) = \{a \in A, b \in B | a - b = z\}$$

counting

$$\mu_{A \ominus B}(z) = \frac{|\mathcal{H}(A,B)(z)|}{}$$

Work in progress: use of ***erosion*** to compute contrastor!

Computing pertinence

Relevance

- Given a certain image,
  - what is relevant to be recognized?
  - what is relevant to be said?

Relevance

- Given a certain image,
  - what is <u>relevant</u> to be recognized?
  - what is <u>relevant</u> to be said?

- More in general, given a certain situation
  - what is <u>relevant</u> to be interpreted?
  - what is <u>relevant</u> to be done?

Relevance

- Given a certain image,
  - what is <u>relevant</u> to be recognized?
  - what is <u>relevant</u> to be said?

- More in general, given a certain situation
  - what is <u>relevant</u> to be interpreted?
  - what is <u>relevant</u> to be done?

- ***Simplicity Theory*** (ST) offers a computational cognitive model for computing relevance, based on ***unexpectedness*** and ***emotion***.

For a more detailed overview and further references see https://simplicitytheory.telecom-paristech.fr/

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness** $U(s) = C_W(s) - C_D(s)$

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

causal complexity

concerning how the world generates the situation

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

concerning how to identify the situation

causal complexity

concerning how the world generates the situation

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.
- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

concerning how to identify the situation

causal complexity

concerning how the world generates the situation

The two complexities are defined following Kolmogorov complexity.

Kolmogorov complexity

length in bits of the **shortest** program generating a string

description of an object

Kolmogorov complexity

length in bits of the **shortest** program generating a string
description of an object

string		equivalent programs
“2222222222222222222222222”	=	“2” + “2” + … + “2”
	=	“2” * 25
	=	“2” * 5^2

Kolmogorov complexity

length in bits of the **shortest** program generating a string description of an object

string		equivalent programs
"2222222222222222222222222"	=	"2" + "2" + … + "2"
	=	"2" * 25
	=	"2" * 5^2

depends on the available operators!!

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.
- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

about how to identify the situation

length of shortest program determining the situation

causal complexity

about how the world generates the situation

length of shortest program creating the situation

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

description complexity

about how to identify the situation

length of shortest program determining the situation

instructions = **mental operators**

causal complexity

about how the world generates the situation

length of shortest program creating the situation

instructions = **causal operators**

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness**

$$U(s) = C_W(s) - C_D(s)$$

SIMULATION

causal complexity

about how the world generates the situation

length of shortest program creating the situation

instructions = **causal operators**

description complexity

about how to identify the situation

length of shortest program determining the situation

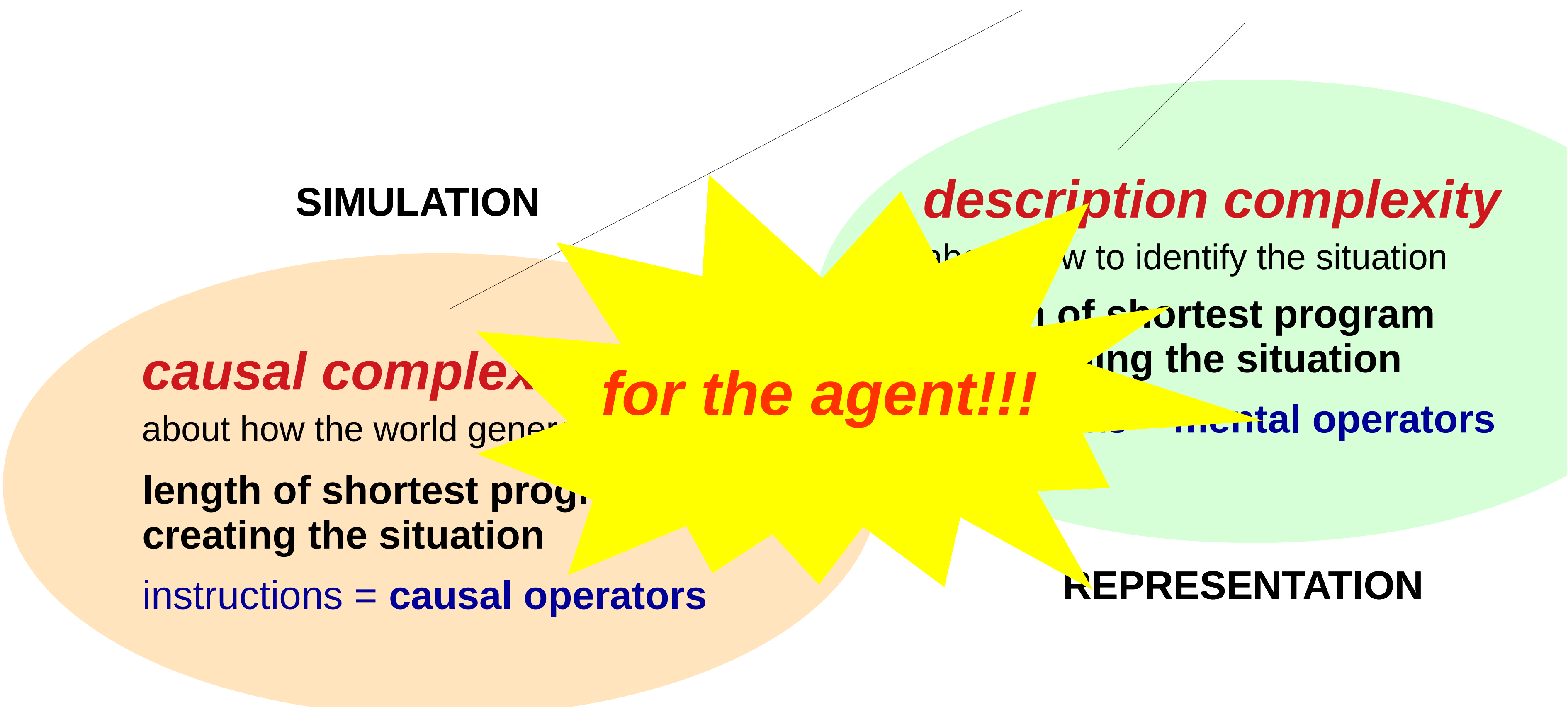
instructions = **mental operators**

REPRESENTATION

Simplicity theory: unexpectedness

- Human individuals are highly sensitive to **complexity drops**: i.e. to *situations that are simpler to describe than to explain*.

- Core notion: **Unexpectedness** $$U(s) = C_W(s) - C_D(s)$$



$$U(s) = C_W(s) - C_D(s)$$

(in a fair extraction)

22222222222222 is more unexpected than **21658367193445**

$$U(s) = C_W(s) - C_D(s)$$

(in a fair extraction)

22222222222222 is more unexpected than **21658367193445**

meeting Obama is more unexpected than **meeting Dupont**

(or any other famous person) (or any other unknown person)

meeting an old of friend of mine

(or any other known person)

Unexpectedness *captures* ***plausibility***

$$U(s) = C_W(s) - C_D(s)$$

informativity is maximized by maximizing unexpectedness

when C_W(s) is the same, we look for low C_D(s)

(in a fair extraction)

22222222222222 is more unexpected than **21658367193445**

meeting Obama is more unexpected than **meeting Dupont**

(or any other famous person) (or any other unknown person)

meeting an old of friend of mine

(or any other known person)

Unexpectedness *captures* ***plausibility***

Simplicity Theory: Emotion

- Focusing on intensity, we can capture anticipation as:

$$AE(s) = E(s) - U(s)$$

unexpectedness

emotion

what the situation induces to the agent

reward model

actualized emotion

Simplicity Theory: Emotion

- Focusing on intensity, we can capture anticipation as:

$$AE(s) = E(s) - U(s)$$

unexpectedness

emotion

what the situation induces to the agent

reward model

actualized emotion

- Attention is intuitively associated to situations that might occur depending on their emotional impact.

Simplicity Theory: Relevance

epithymically

- Fundamental principles:
  - situations with *high anticipated emotion* are ***relevant***
  - situations with *high unexpectedness* are ***relevant***

epistemically

Simplicity Theory: Relevance

- Fundamental principles:
  - situations with *high anticipated emotion* are ***relevant***
  - situations with *high unexpectedness* are ***relevant***

- Intuitively, contrast and similarity play a role with C_D as they function with the most accessible references, i.e.:

  target is determined as
  proximate
  to ***simple*** references
  with respect to ***simple*** relations

Simplicity Theory: Relevance

- Fundamental principles:
  - situations with *high anticipated emotion* are ***relevant***
  - situations with *high unexpectedness* are ***relevant***

- *Why it is relevant to speak of hot coffees, rather than normal coffees?*

Simplicity Theory: Relevance

- Fundamental principles:
  - situations with *high anticipated emotion* are ***relevant***
  - situations with *high unexpectedness* are ***relevant***

- *Why it is relevant to speak of hot coffees, rather than normal coffees?*

- Several factors play a role:
  - ***descriptively*** simple (qualitatively distinctive, accessible references)**,**
  - ***causally*** difficult (supposing a normal distribution of temperatures)**,**
  - ***emotionally*** intense (as we might get burned with it)**.**

Simplicity Theory: Relevance

- Fundamental principles:
  - situations with *high anticipated emotion* are ***relevant***
  - situations with *high unexpectedness* are ***relevant***

- *Why it is relevant to speak of hot coffees, rather than normal coffees?*

- Several factors play a role:
  - ***descriptively*** simple (qualitatively distinctive, accessible references)**,**
  - ***causally*** difficult (supposing a normal distribution of temperatures)**,**
  - ***emotionally*** intense (as we might get burned with it)**.**

- In the following I'll briefly present two additional tracks I've started studying, concerning $\boldsymbol{C_w(s)}$ and $\boldsymbol{E(s)}$

Identifying causes

An experiment

- Causes play a central role in the way we conceptualize the world.
- But there is no established model about how people qualify a cause as ***pertinent*** (literally, holding together) to a specific event.

Sileno, G., & Dessalles, J.-L. (2018). Qualifying Causes as Pertinent. Proceedings of the 40th Conference of the Cognitive Science Society (CogSci 2018)

An experiment

- Causes play a central role in the way we conceptualize the world.
- But there is no established model about how people qualify a cause as ***pertinent*** (literally, holding together) to a specific event.
- We performed an experiment to compare:
  - the computation of ***actual causation*** via
    - conterfactuals (*structural equations*)
    - Bayesian inference
    - Simplicity Theory
  - people's responses

Example of task

Johnny is 7 years old. In recent months his mother has been worried because he developed a craving for sweet things. She bought some pots of strawberry jam and put them into the larder (a small room near the kitchen). Then one afternoon she finds that Johnny has gone into the larder and has eaten half a pot of strawberry jam.

Q1. Why is ”half a pot of jam gone”?

a. because of Johnny’s gluttony
b. because Johnny ate it
c. because mother has put the pot in the larder

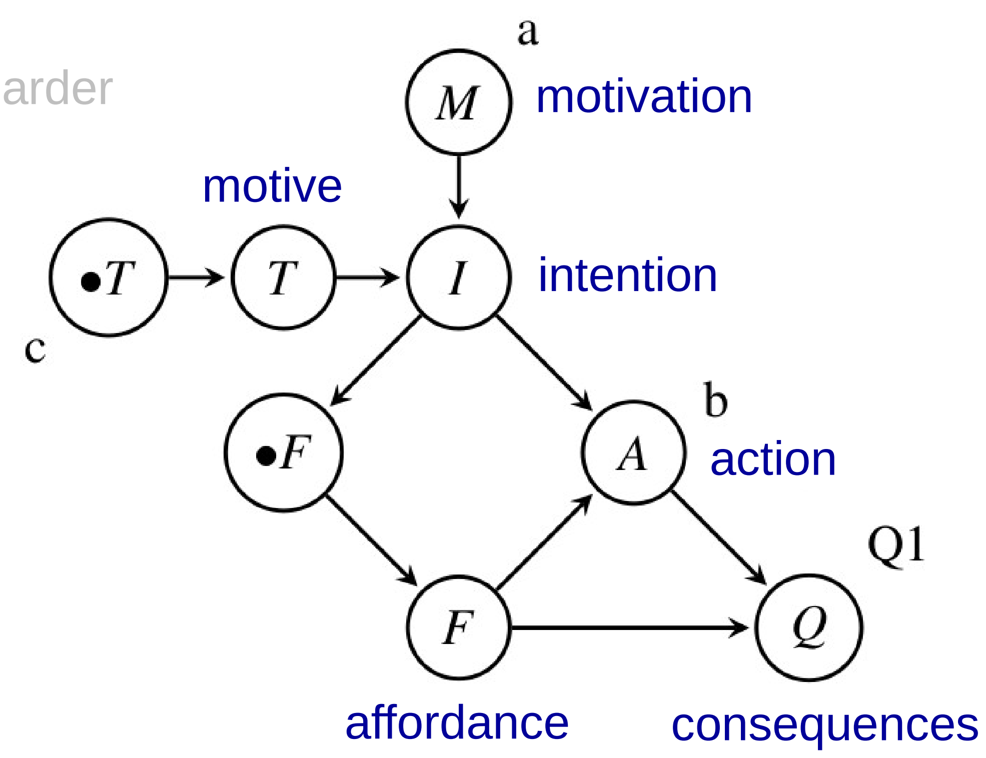
Example of task

Johnny is 7 years old. In recent months his mother has been worried because he developed a craving for sweet things. She bought some pots of strawberry jam and put them into the larder (a small room near the kitchen). Then one afternoon she finds that Johnny has gone into the larder and has eaten half a pot of strawberry jam.

Q1. Why is "half a pot of jam gone"?

a. because of Johnny's gluttony
b. because Johnny ate it
c. because mother has put the pot in the larder

- For each task, a model of the story is constructed, based on a general ***action-scheme***



Evaluation

- Measures based on probability:

$$p(E|C)$$

$$p(E|C) \cdot p(C)$$

$$\log \frac{p(E|C)}{p(E|\neg C)}$$

$$\frac{p(E|C) - p(E|\neg C)}{p(E|C) + p(E|\neg C)}$$

- Measure based on complexity:

$$C_W(E) - C_W(E||C)$$

Evaluation

- Measures based on probability:

given a certain model:

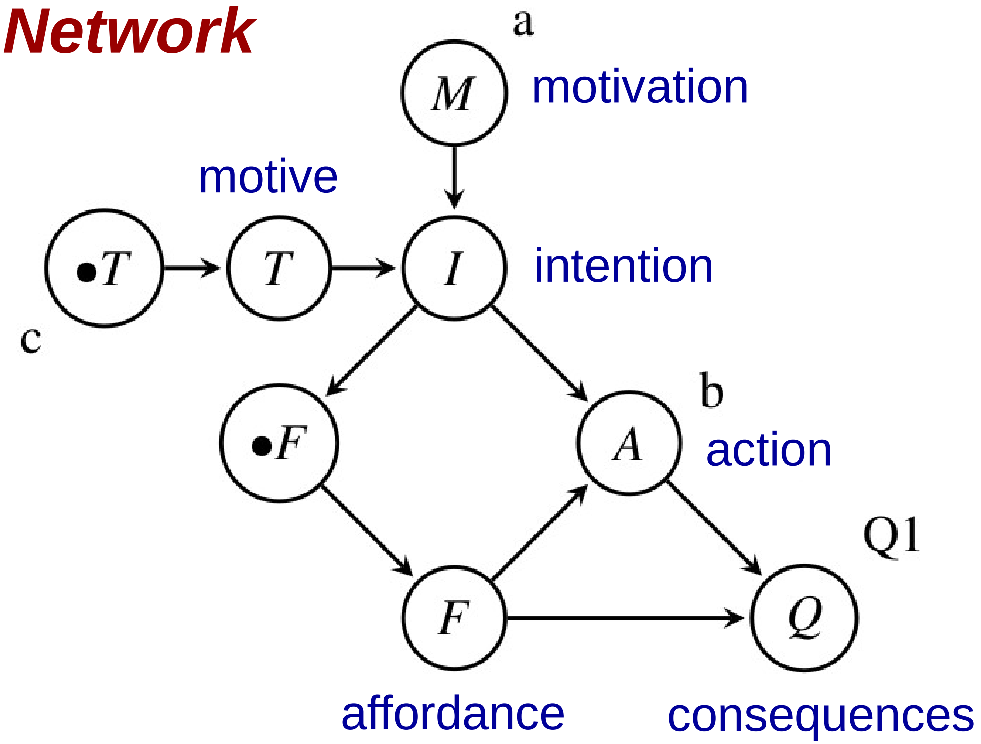
$$p(E|C)$$

$$p(E|C) \cdot p(C)$$

$$\log \frac{p(E|C)}{p(E|\neg C)}$$

$$\frac{p(E|C) - p(E|\neg C)}{p(E|C) + p(E|\neg C)}$$

computation using a ***Bayesian Network***



- Measure based on complexity:

$$C_W(E) - C_W(E||C)$$

computation of complexities using ***minimal path search***

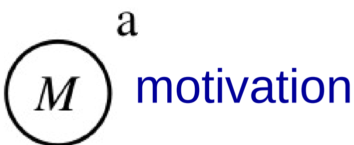
Evaluation

- Measures based on probability:

$p(E|C)$ → *computation using a* ***Bayesian Network***

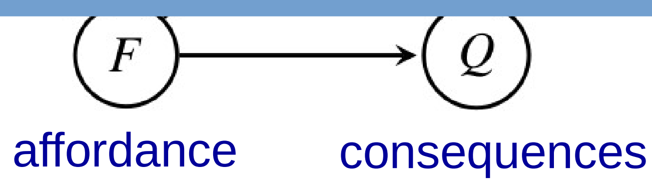
$p(E|C) \cdot p(C)$

given a certain model:



Results: No probabilistic measure is consistently aligned.

Causal contribution as defined by ST performs much better, and divergences can be explained by intervention of description complexity.



- Measure based on complexity:

$C_W(E) - C_W(E||C)$ → *computation of complexities using* ***minimal path search***

Attributing responsibility


Rashomon, 1950


12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.

Sileno, G., Saillenfest, A., & Dessalles, J.-L. (2017). A Computational Model of Moral and Legal Responsibility via Simplicity Theory. Proceedings of the 30th Int. Conf. on Legal Knowledge and Information Systems (JURIX 2017), FAIA 302, 171–176

Rashomon, 1950

12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.
- Non-related ancient legal systems bear much resemblance to modern law and seem perfectly sensible nowadays.

Rashomon, 1950

12 Angry Men, 1956

Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.
- Non-related ancient legal systems bear much resemblance to modern law and seem perfectly sensible nowadays.

→ *responsibility attribution* may be controlled by fundamental cognitive mechanisms.

Rashomon, 1950

12 Angry Men, 1956

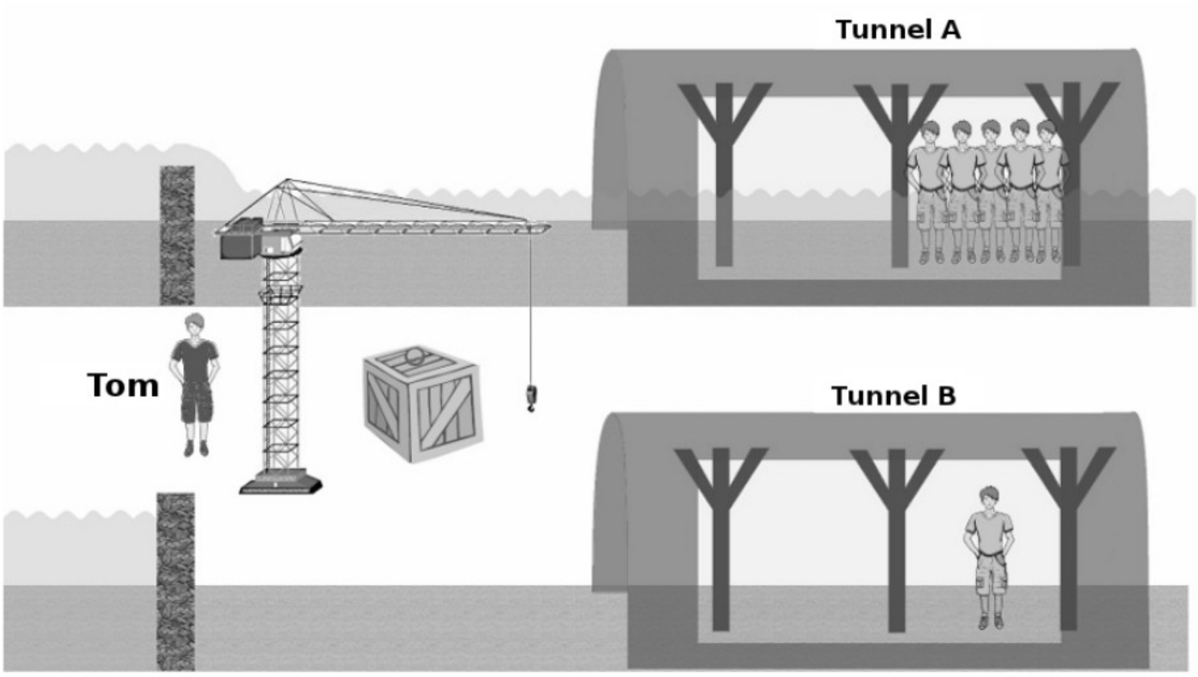
Responsibility attribution for humans

- In human societies, responsibility attribution is a ***spontaneous*** and ***seemingly universal*** behaviour.

- Non-related ancient legal systems bear much resemblance to modern law and seem perfectly sensible nowadays.

  → *responsibility attribution* may be controlled by fundamental cognitive mechanisms.

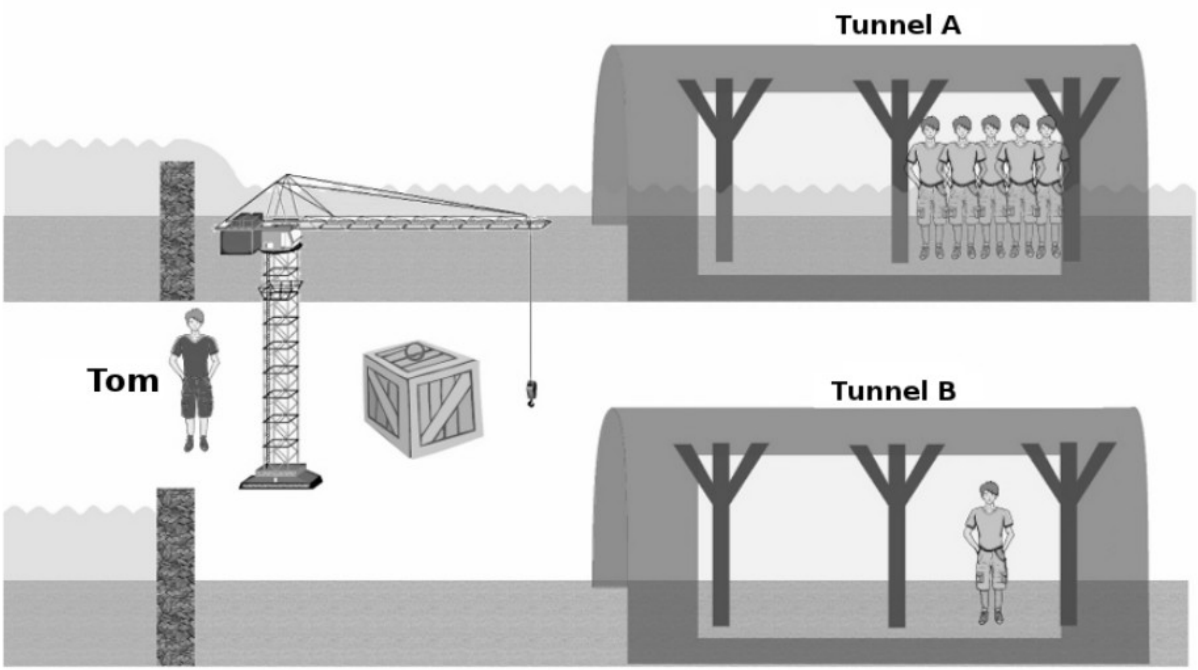
Working hypothesis: attributions of **moral** and **legal responsibility** share a similar cognitive architecture



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:

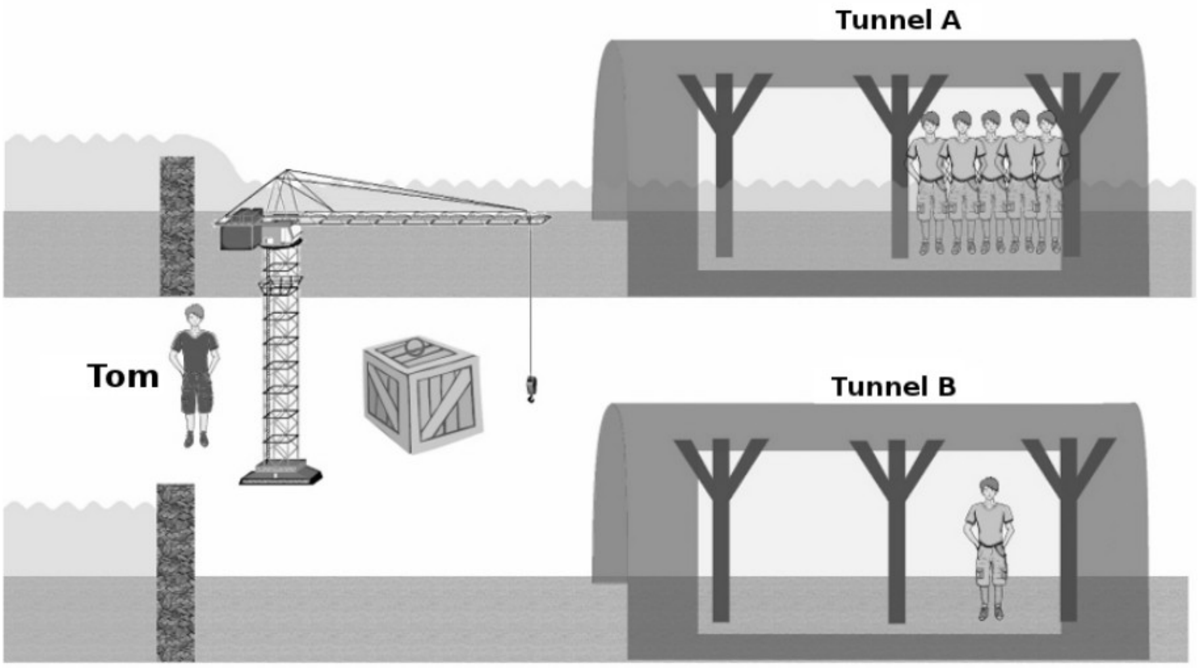
[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012, pages 947–952]



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:
  - the more the outcome is severe,
  - the more they are closer to the victims,
  - the more the outcome follows the action.

[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012, pages 947–952]



flooded mine dilemma *(trolley problem variation)*

- Experiments show that people are more prone to blame an agent for an action:
  - the more the outcome is severe,
  - the more they are closer to the victims,
  - the more the outcome follows the action.
- The cognitive model of ***Simplicity Theory*** predicts these results.

[A. Saillenfest and J.-L. Dessalles. Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. CogSCI 2012]

Simplicity Theory: Emotion

- Focusing on intensity, we can capture anticipation as:

$$AE(s) = E(s) - U(s)$$

unexpectedness

emotion

what the situation induces to the agent

reward model

- The anticipated emotion of doing *a* to reach *s*:

$$AE(a * s) = E(a * s) - U(a * s) = E(a * s) - U(s||a) - U(a)$$

$$I(a, s) = E(a * s) - U(s||a) - U(a)$$

intention as driven by anticipated emotional effects

Simplicity Theory: Moral responsibility

- Difference between intention* and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

* For simplicity, we assume here that the action a has only a relevant outcome s and it has no impact on emotion, i.e.   E(a*s) = E(s)

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

computed by an observer O

computed by a model of A

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

computed by an observer O

reward model

computed by a model of A

prescribed role, reasonable standard

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

Simplicity Theory: Moral responsibility

- Difference between intention and moral responsibility is one of **point of views**.

computed by A

$$I(a) = E^A(s) - U^A(s||a) - U^A(a)$$

computed by an observer O

reward model

computed by a model of A

prescribed role, reasonable standard

$$M(a) = E(s) - U^{\downarrow A}(s||a) - U^{\downarrow A}(a)$$

- Introducing causal responsibility $R^{\downarrow A}(a,s) = C_W(s) - C_W^{\downarrow A}(s||a)$

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion
for observer O
+

causal responsibility
attributed to A
+

conceptual remoteness
for observer O
–

inadvertence
attributed to A
–

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion
for observer O
+

causal responsibility
attributed to A
+

conceptual remoteness
for observer O
–

inadvertence
attributed to A
–

- From moral to legal responsibility:
  - **equity before the law**

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion	***causal responsibility***	***conceptual remoteness***	***inadvertence***
for observer O	*attributed to A*	*for observer O*	*attributed to A*
+	**+**	**–**	**–**

- From moral to legal responsibility:
  - **equity before the law**
  - law, as a reward system, defines emotion

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion *for observer O* **+**

causal responsibility *attributed to A* **+**

conceptual remoteness *for observer O* **–**

inadvertence *attributed to A* **–**

- From moral to legal responsibility:
  - **equity before the law**
  - law, as a reward system, defines emotion

This enables to consider ***extrinsic commitments*!**

Simplicity Theory: Moral responsibility

$$M(a) \approx E_h(s) + R^{\downarrow A}(a,s) - C_D(s) - U^{\downarrow A}(a)$$

actualized emotion
for observer O
+

causal responsibility
attributed to A
+

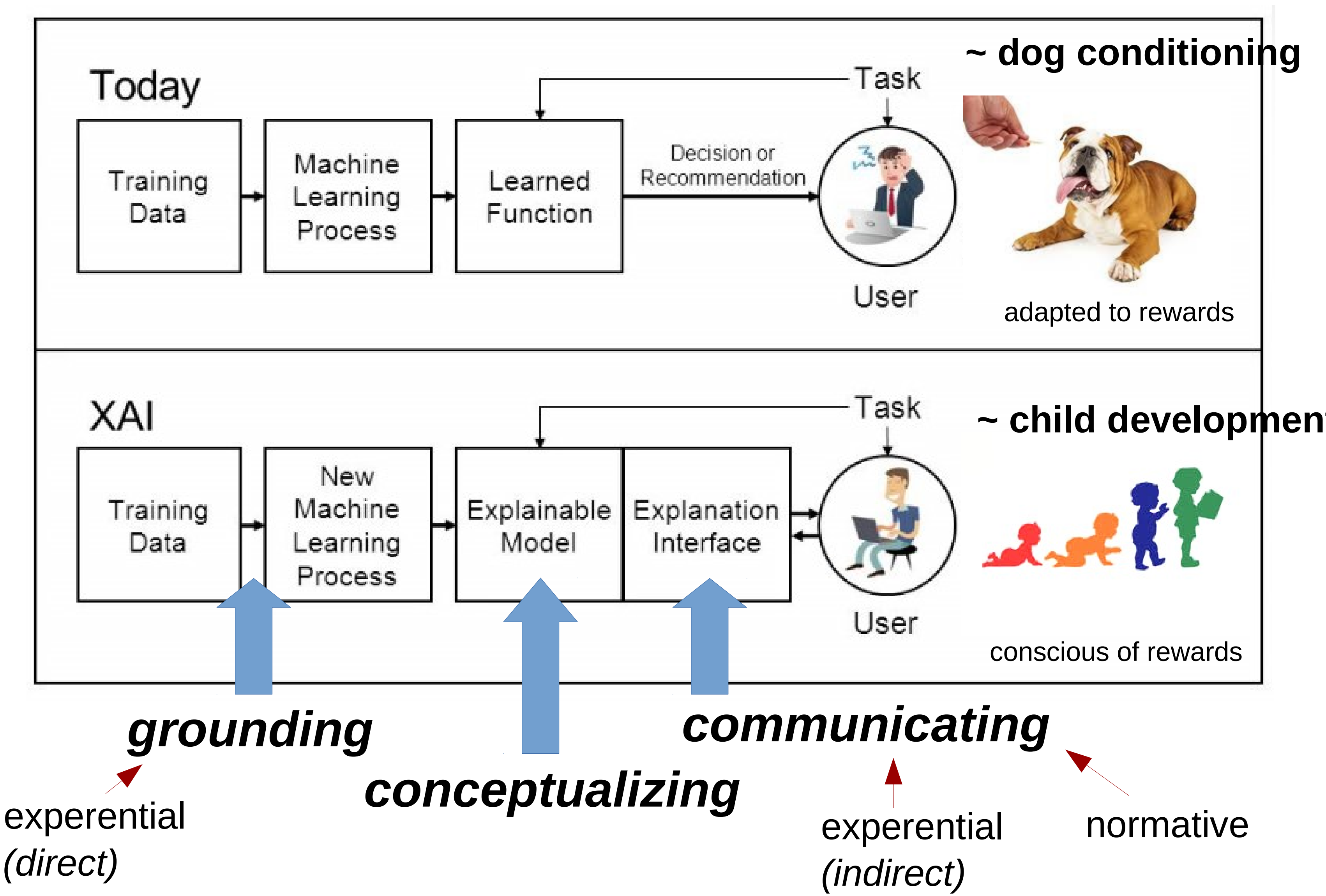
conceptual remoteness
for observer O
–

inadvertence
attributed to A
–

- From moral to legal responsibility:
  - **equity before the law**
  - law, as a reward system, defines emotion

Conclusion

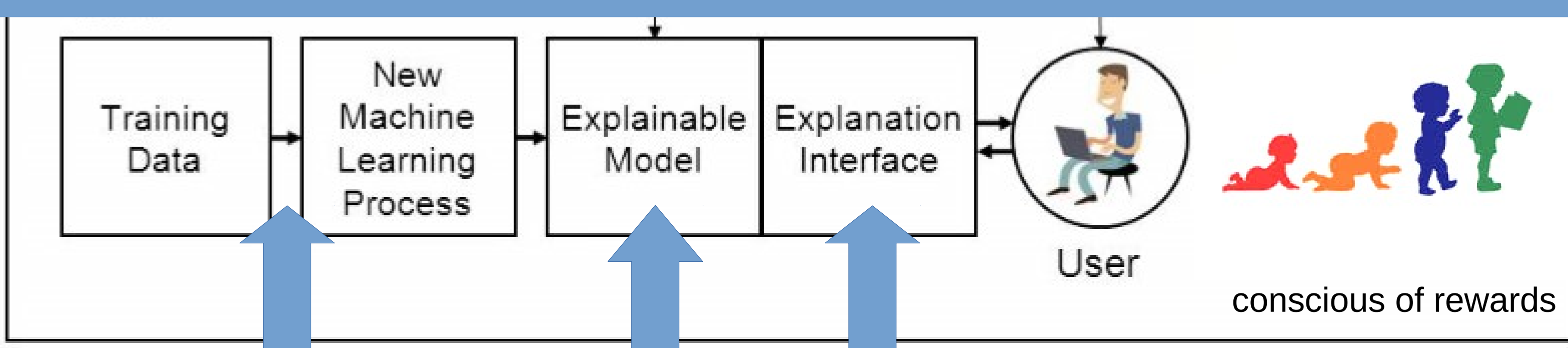
The call for *Explanaible AI* (XAI)



The call for *Explanaible AI* (XAI)

automated decision-making need to be:

- **non (primarily) statistical**
- **cognitively plausible**
- **linguistically competent**
- **able to take into account norms**



grounding

experiential
(direct)

conceptualizing

communicating

experiential
(indirect)

normative

Outlining the kernel of agency

- The core problem – of *normative, epistemic* and *ontological* **alignment** – is related to the different modalities that we, as agents, attribute to reality...

collective

individual

physical

Outlining the kernel of agency

- The core problem – of *normative, epistemic* and *ontological* **alignment** – is related to the different modalities that we, as agents, attribute to reality...

collective

individual

physical

This holds for humans, but also for artificial agents.