Link Uncertainty and ML

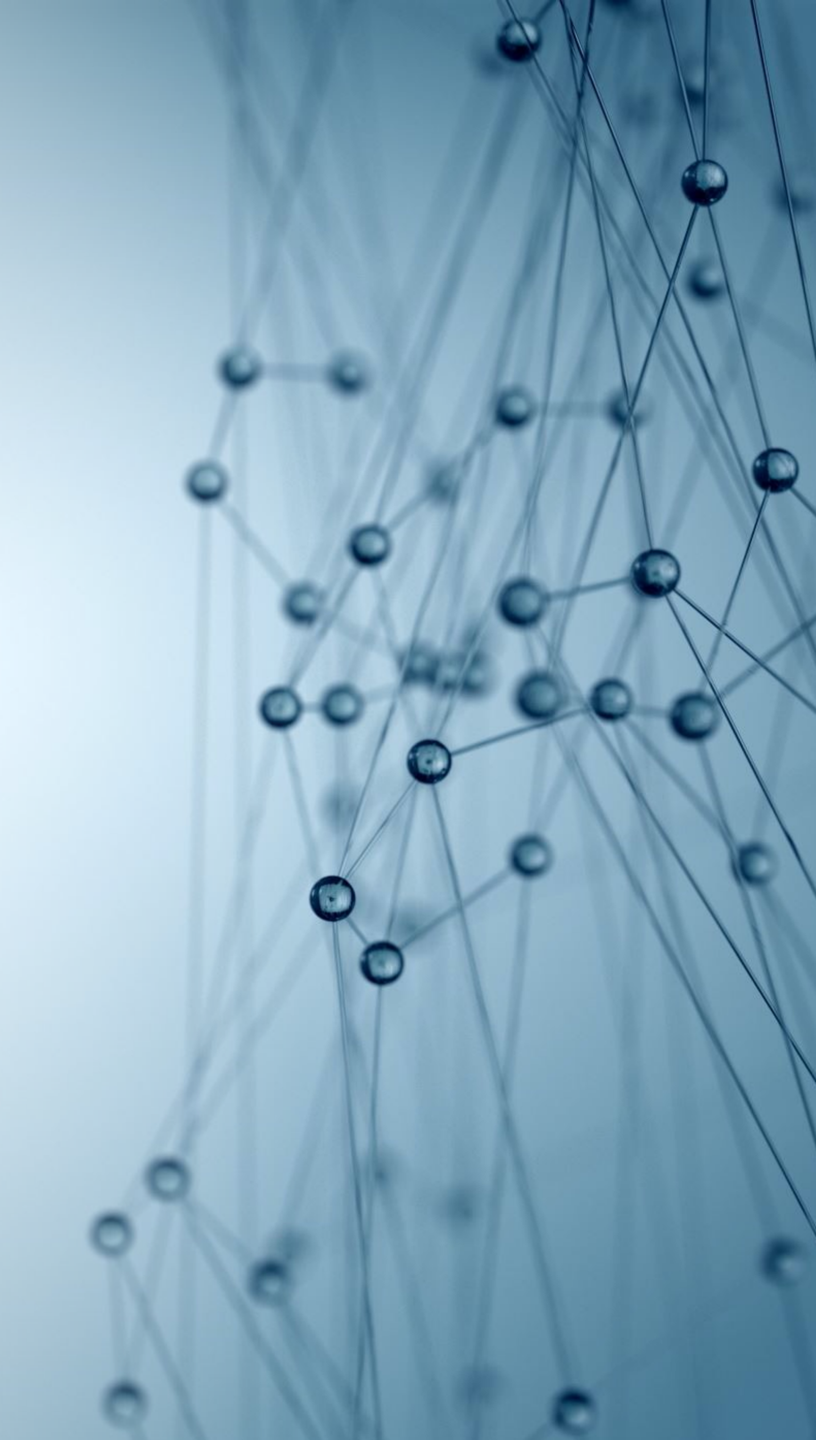
Emily Sullivan
Philosophy and Ethics
Eindhoven University of Technology
Eindhoven Artificial Intelligence Systems Institute

esdit
ethics of socially disruptive technologies

NWO

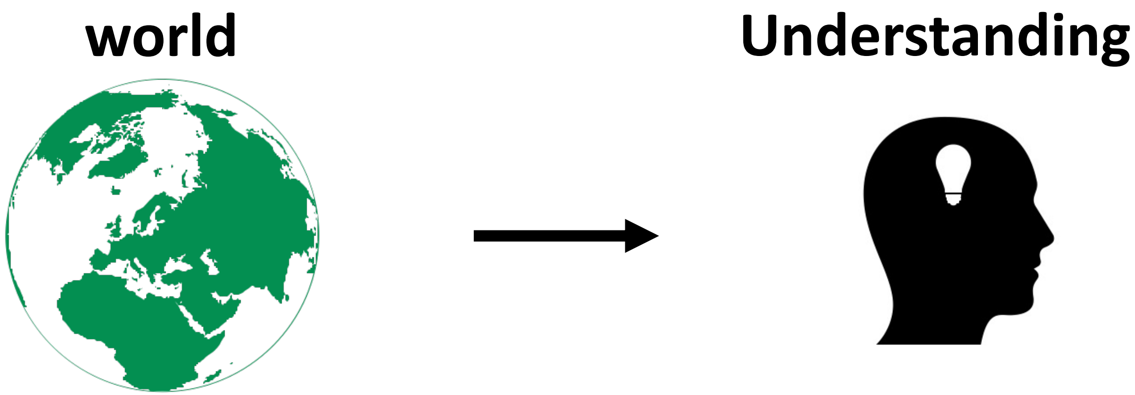
TU/e
EAISI

ϕAI/e



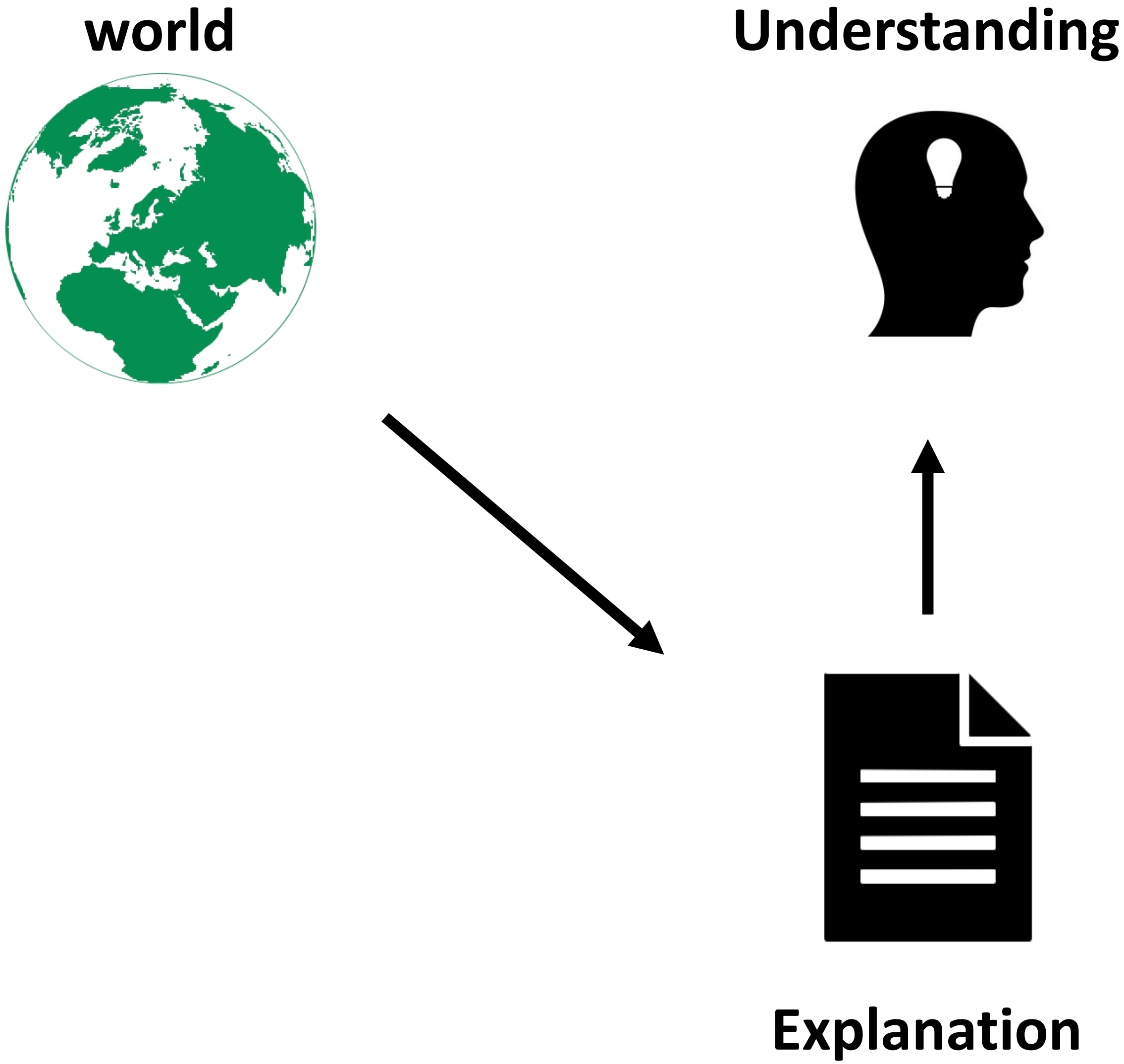
world

Understanding



world

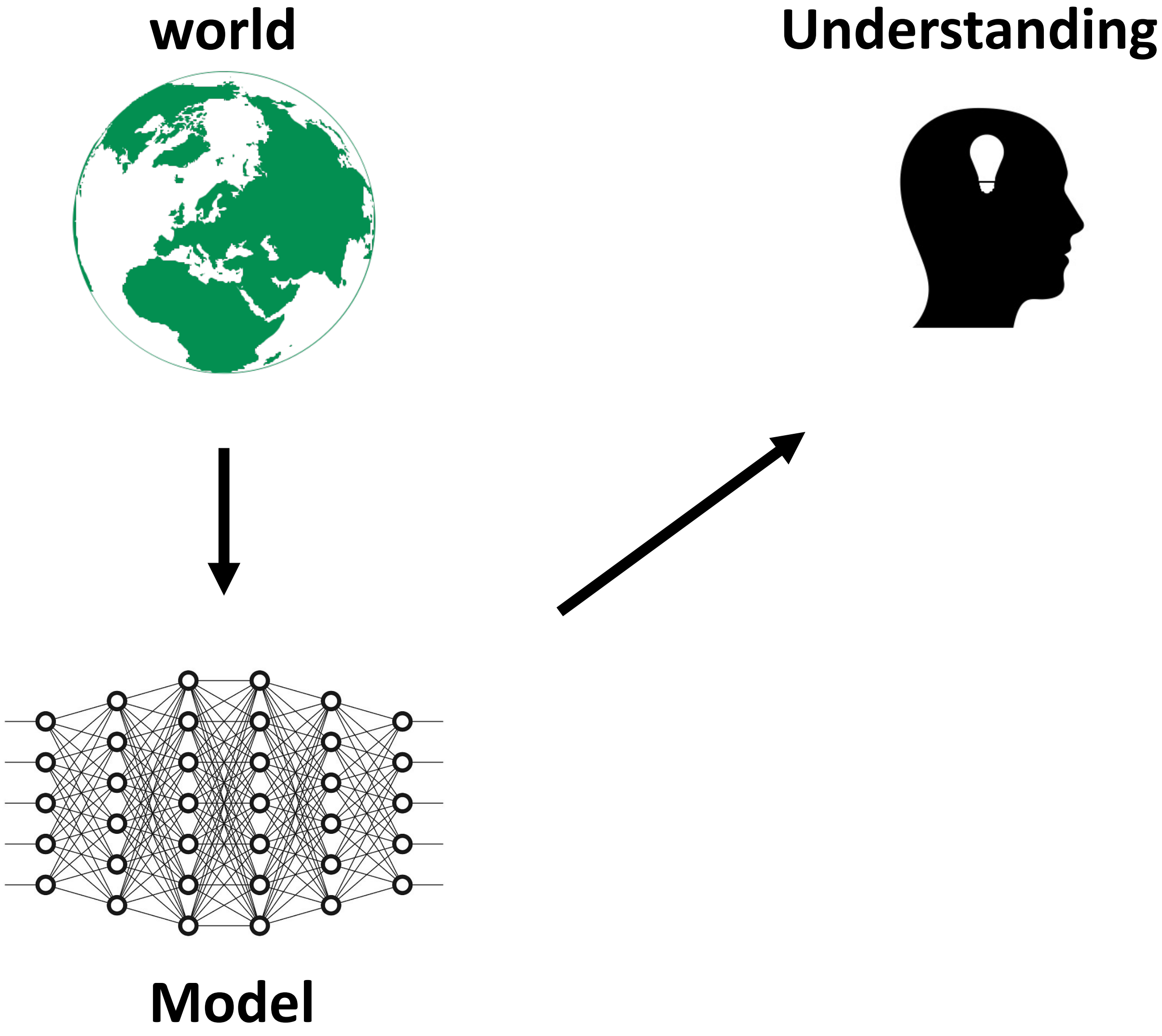
Understanding



world

Understanding

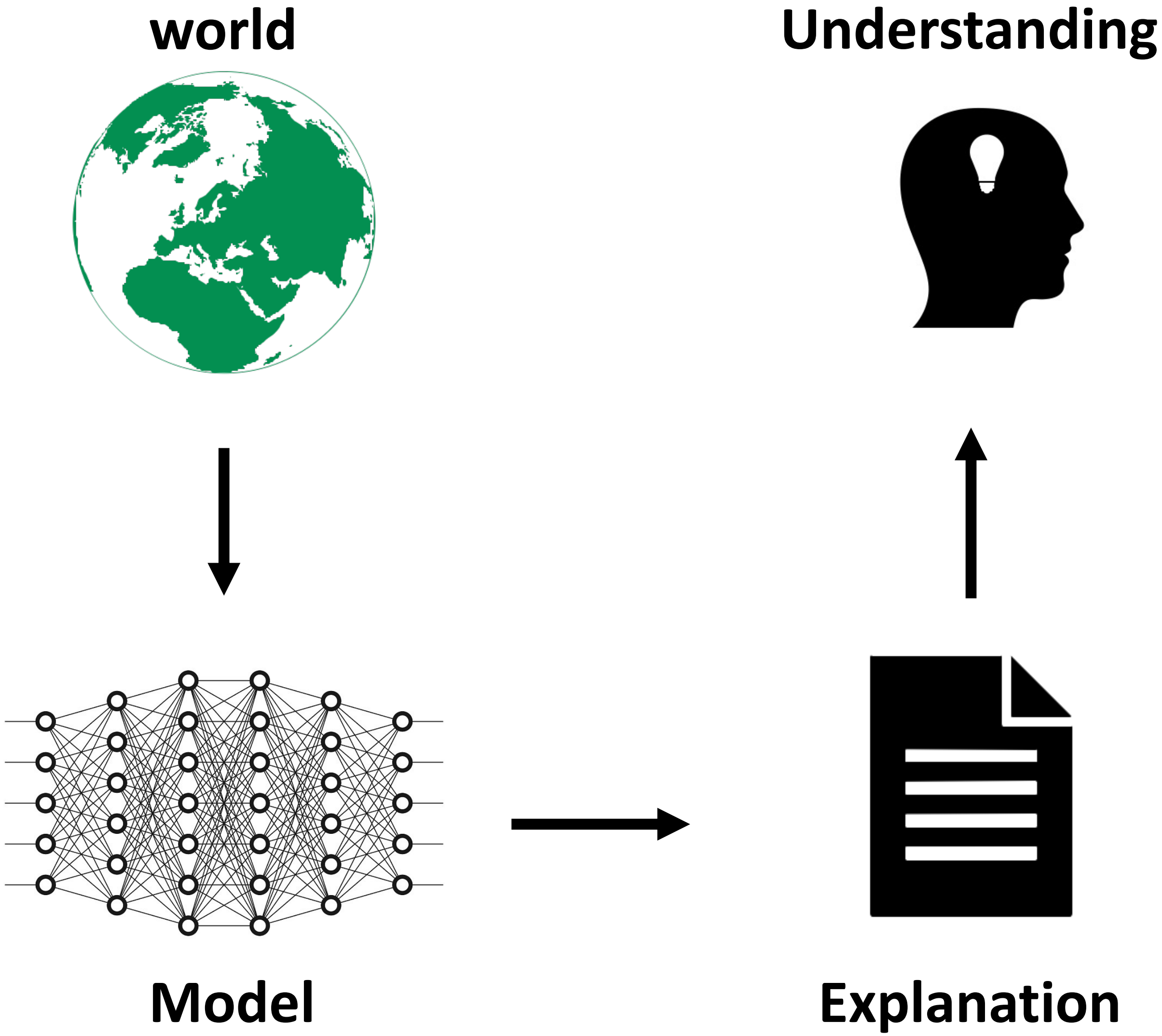
Explanation



world

Understanding

Model

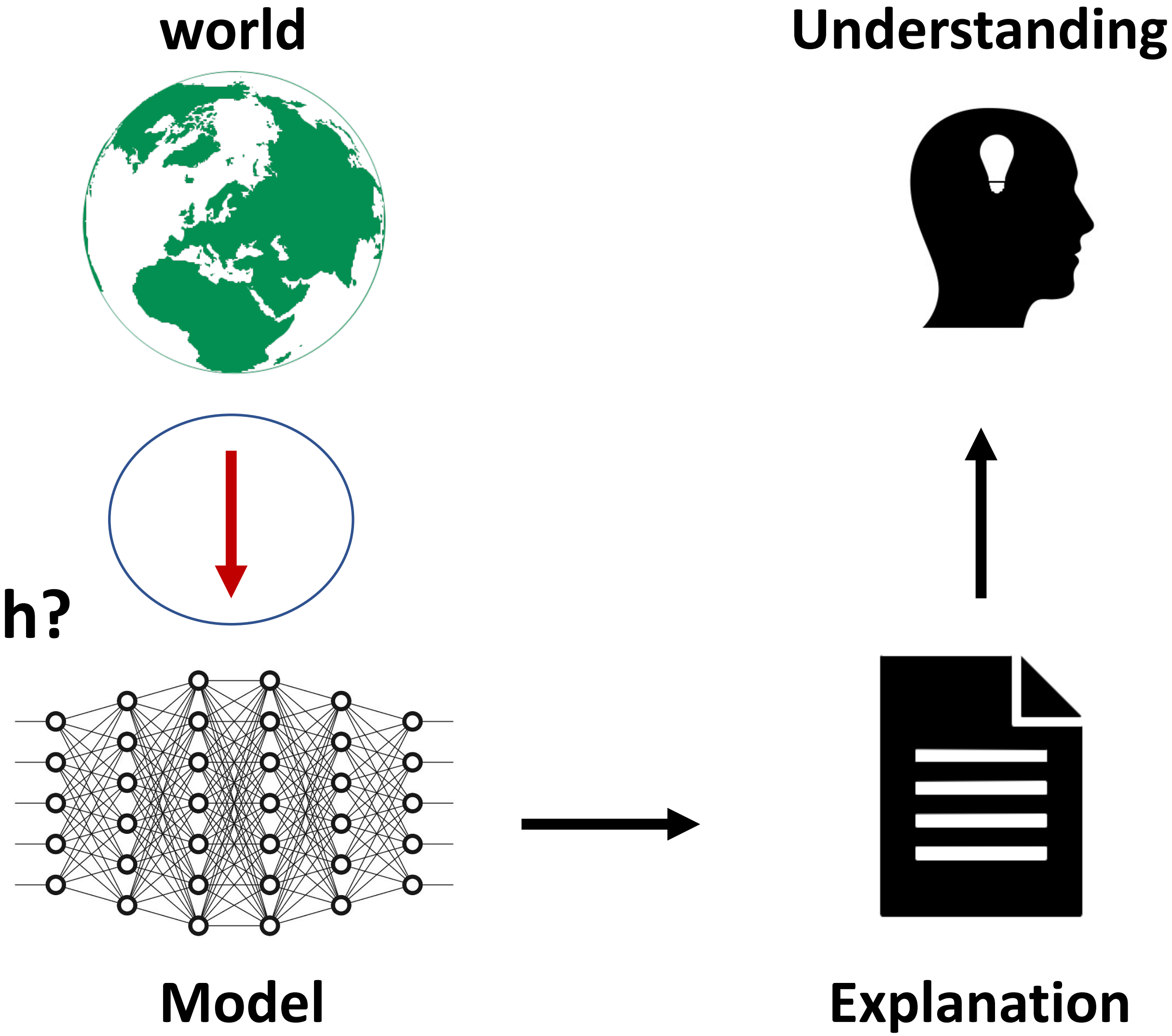


world

Understanding

Model

Explanation



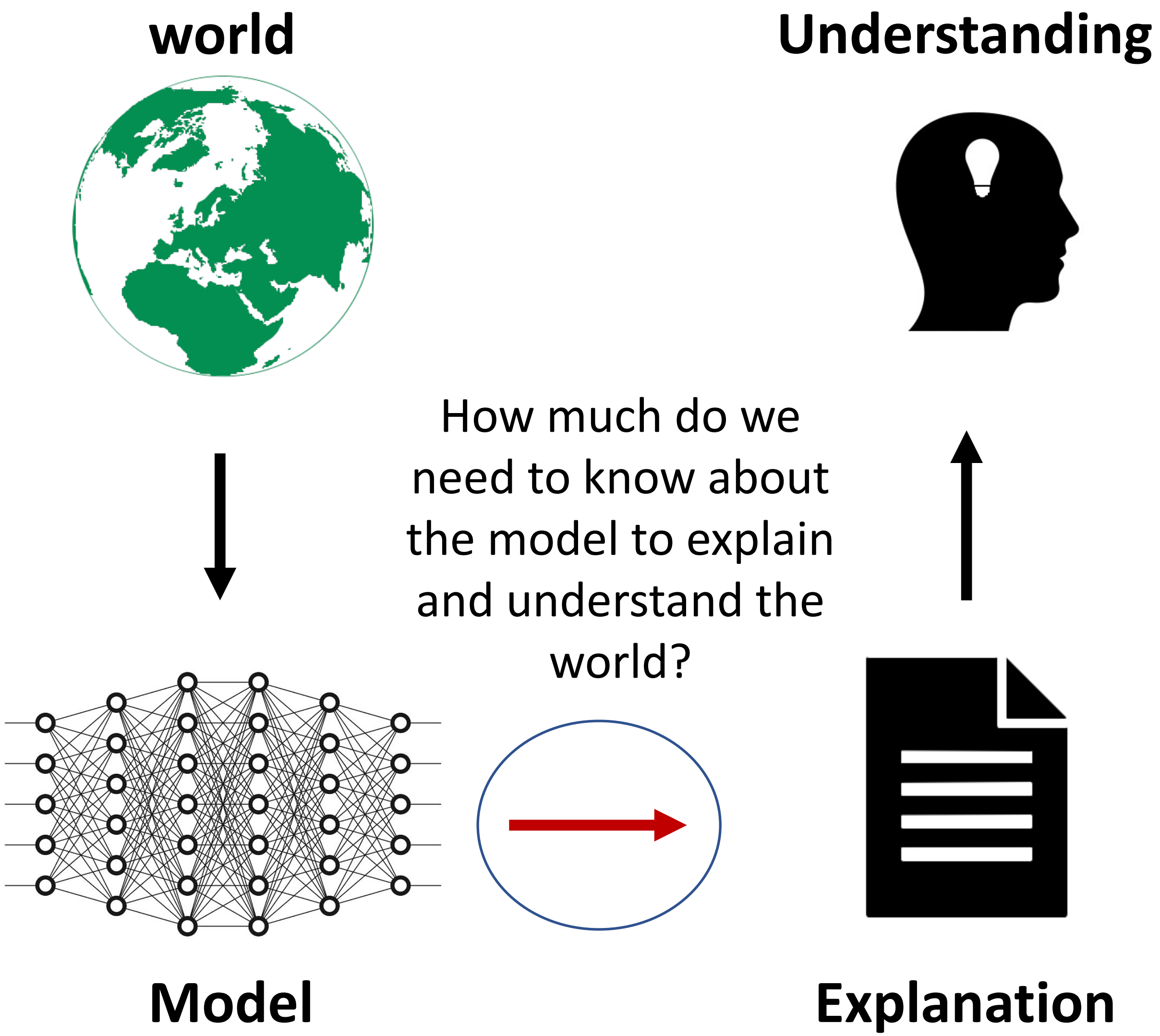
world

Understanding

Model

Explanation

When is this connection strong enough?



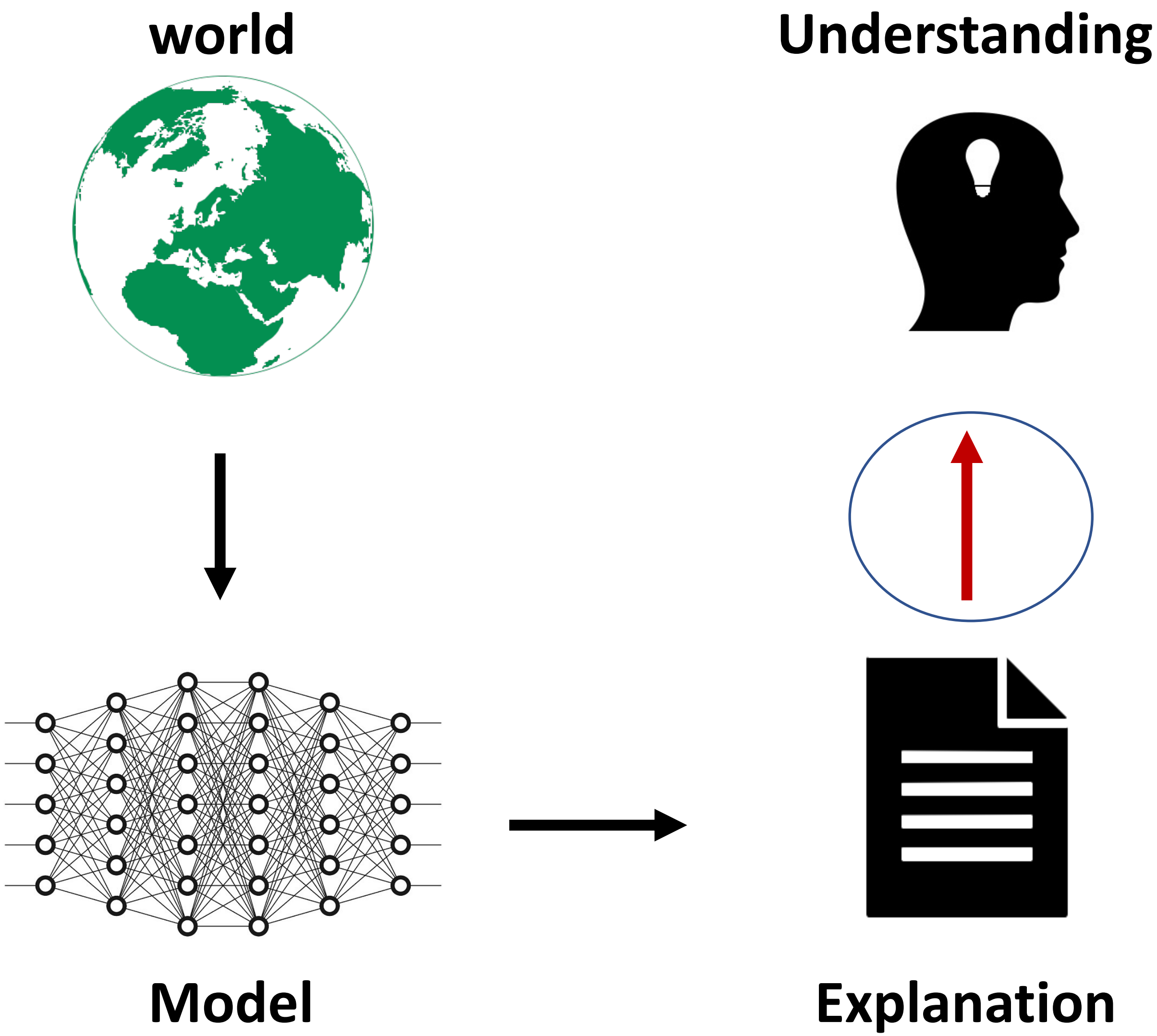
world

Understanding

How much do we need to know about the model to explain and understand the world?

Model

Explanation



world

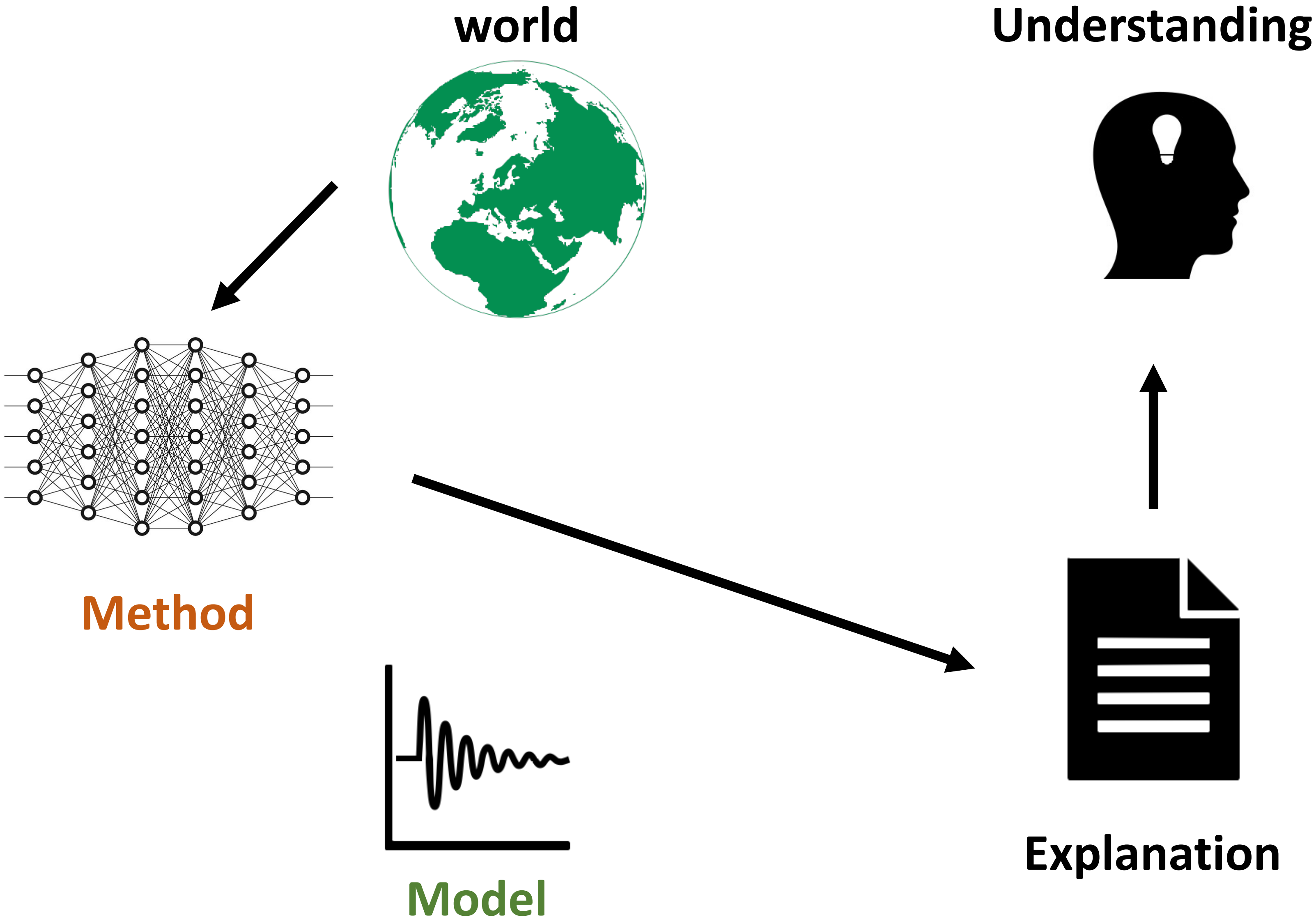
Understanding

Model

Explanation

What are the norms of explanation such that they can enable understanding?

Model independence?



world

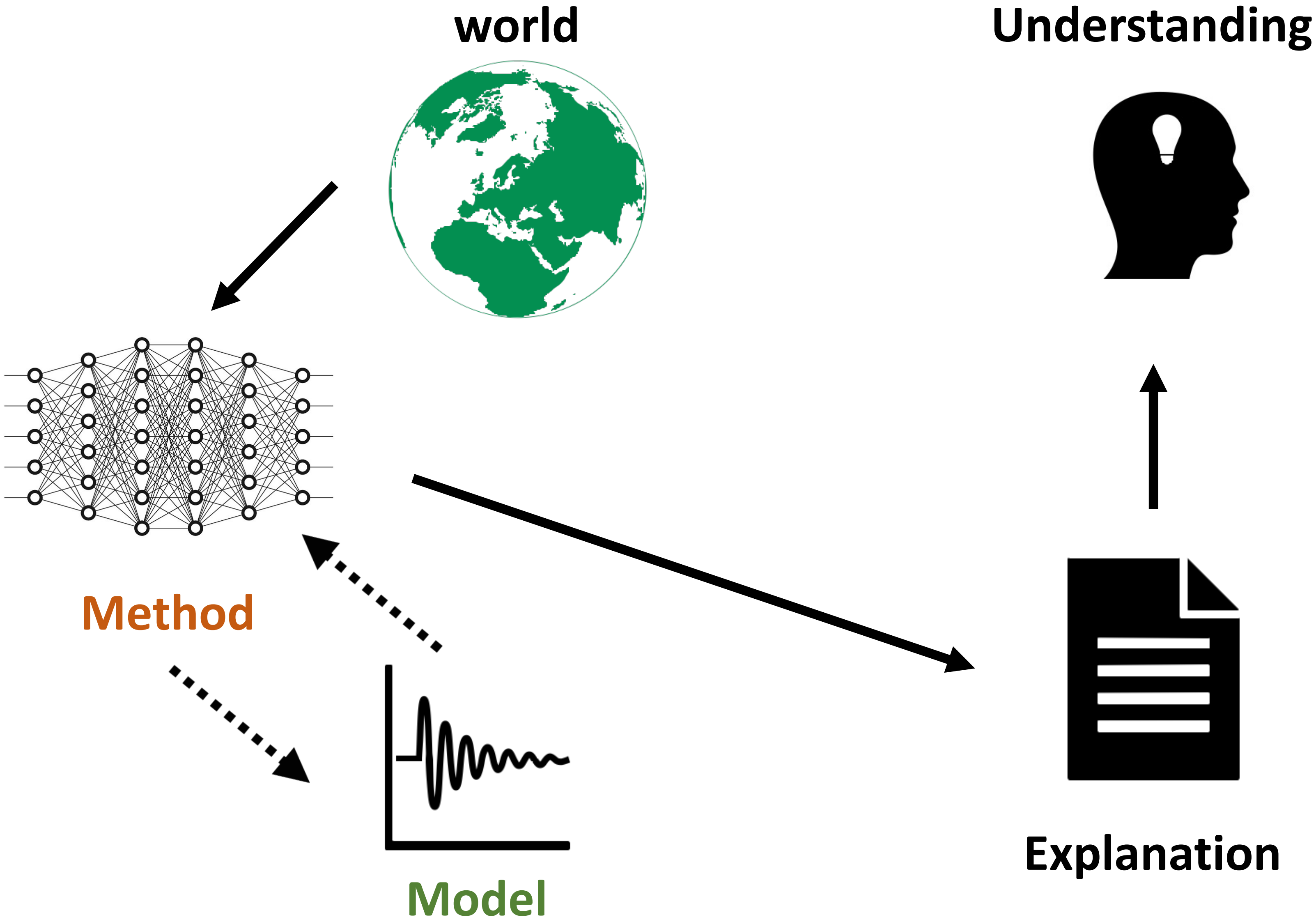
Understanding

Method

Model

Explanation

Model independence?



world

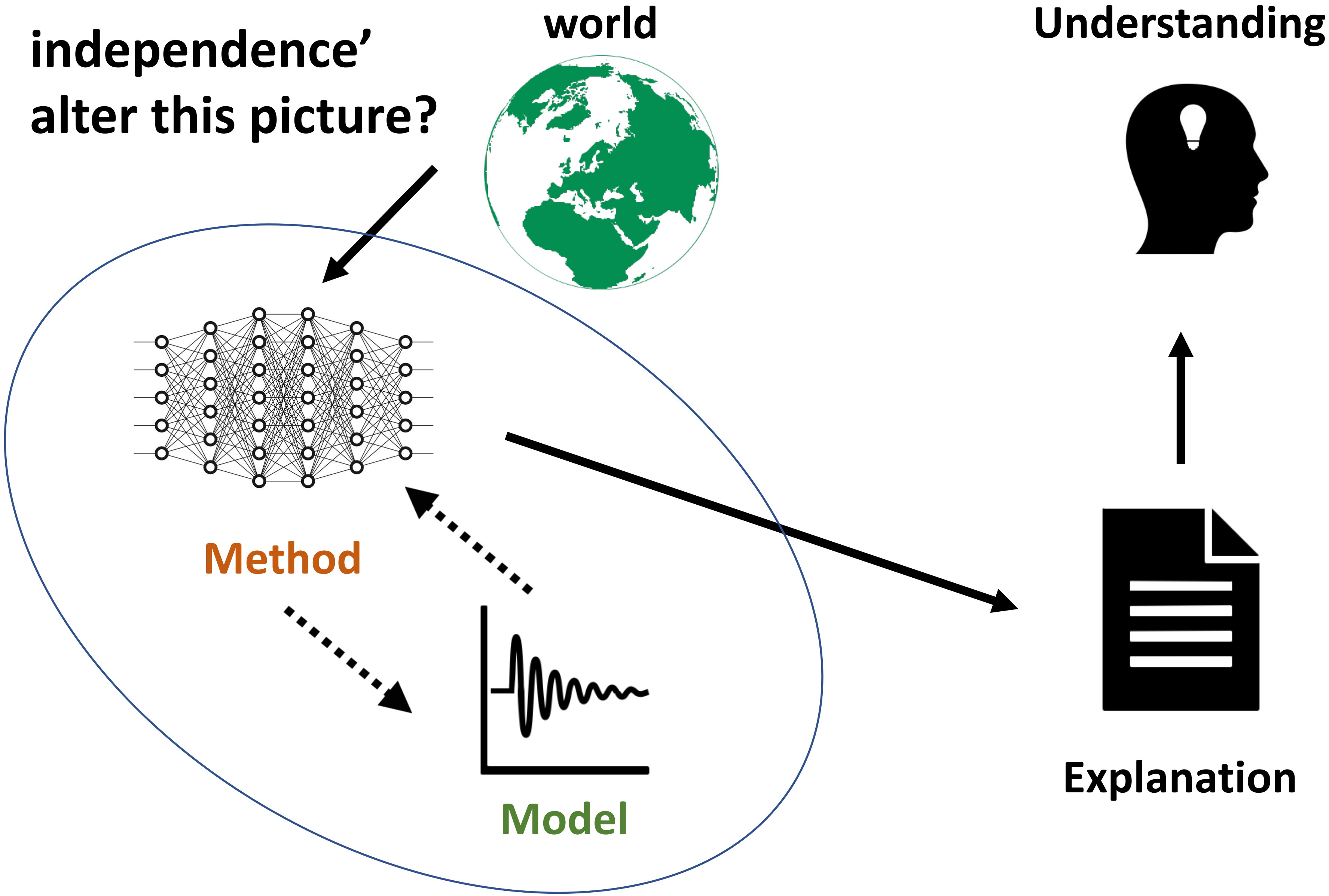
Understanding

Method

Model

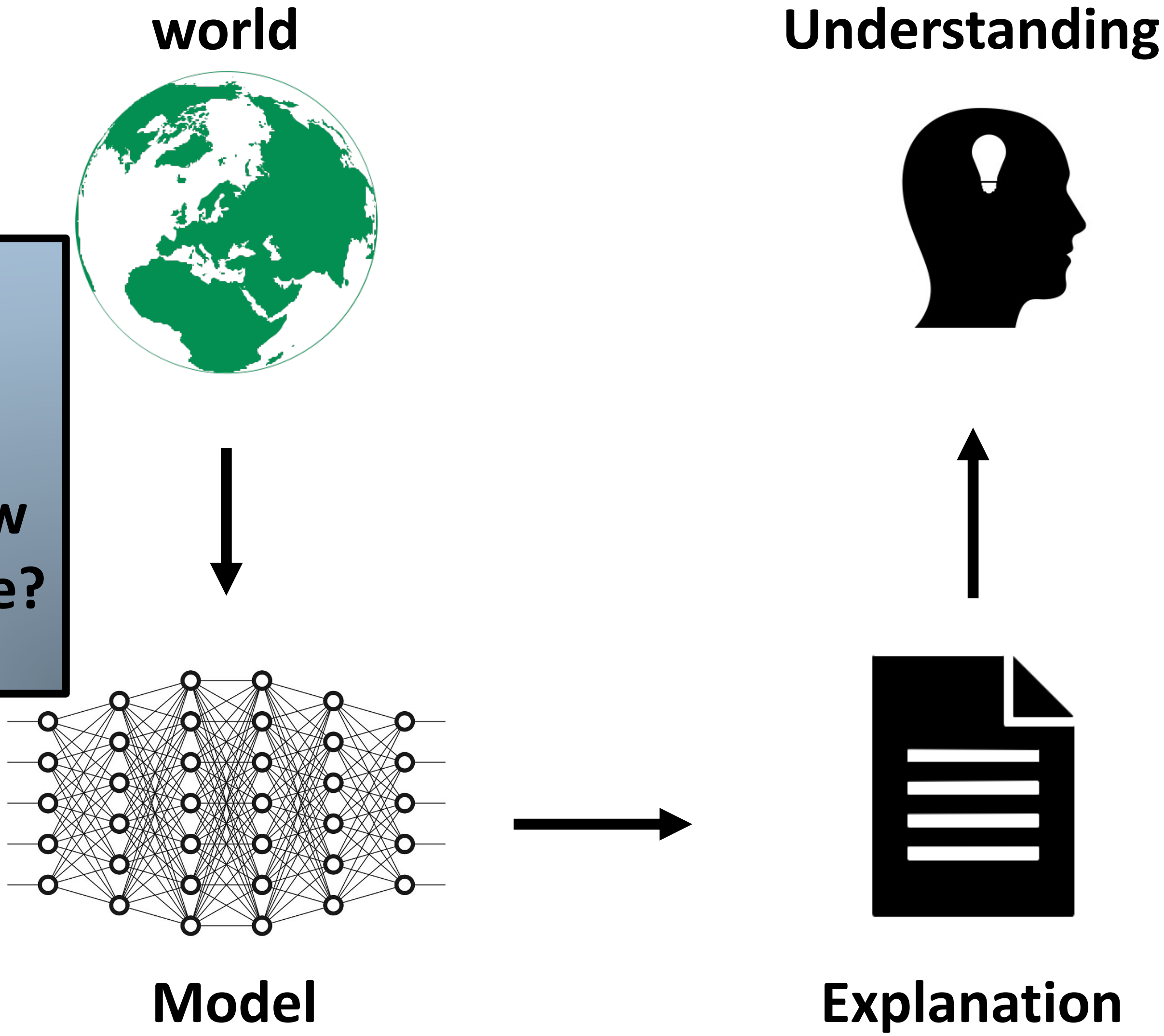
Explanation

How does ‘model independence’ alter this picture?



ML and XAI

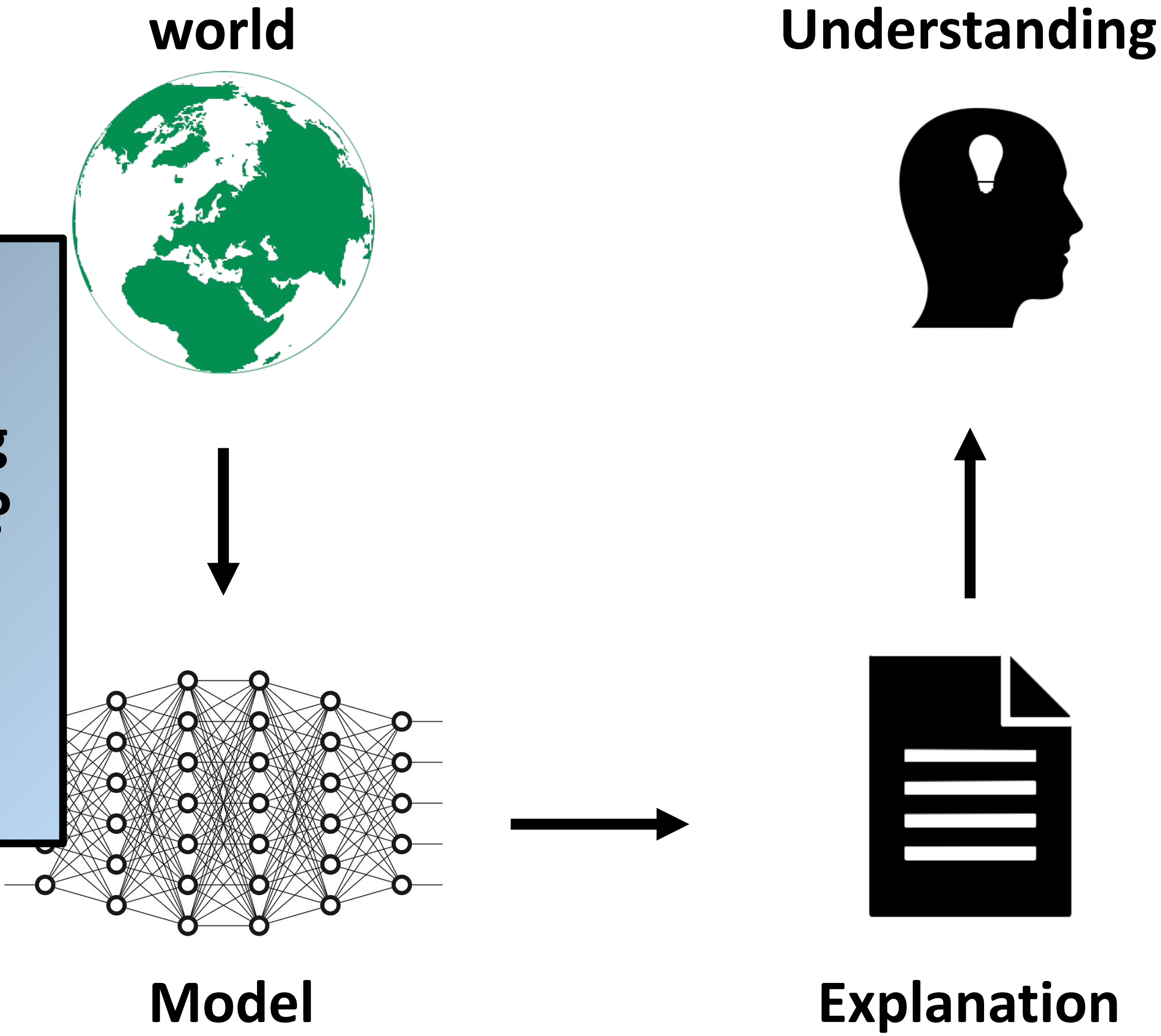
Is there anything philosophically new and interesting here?



ML and XAI

How can modeling practices improve?

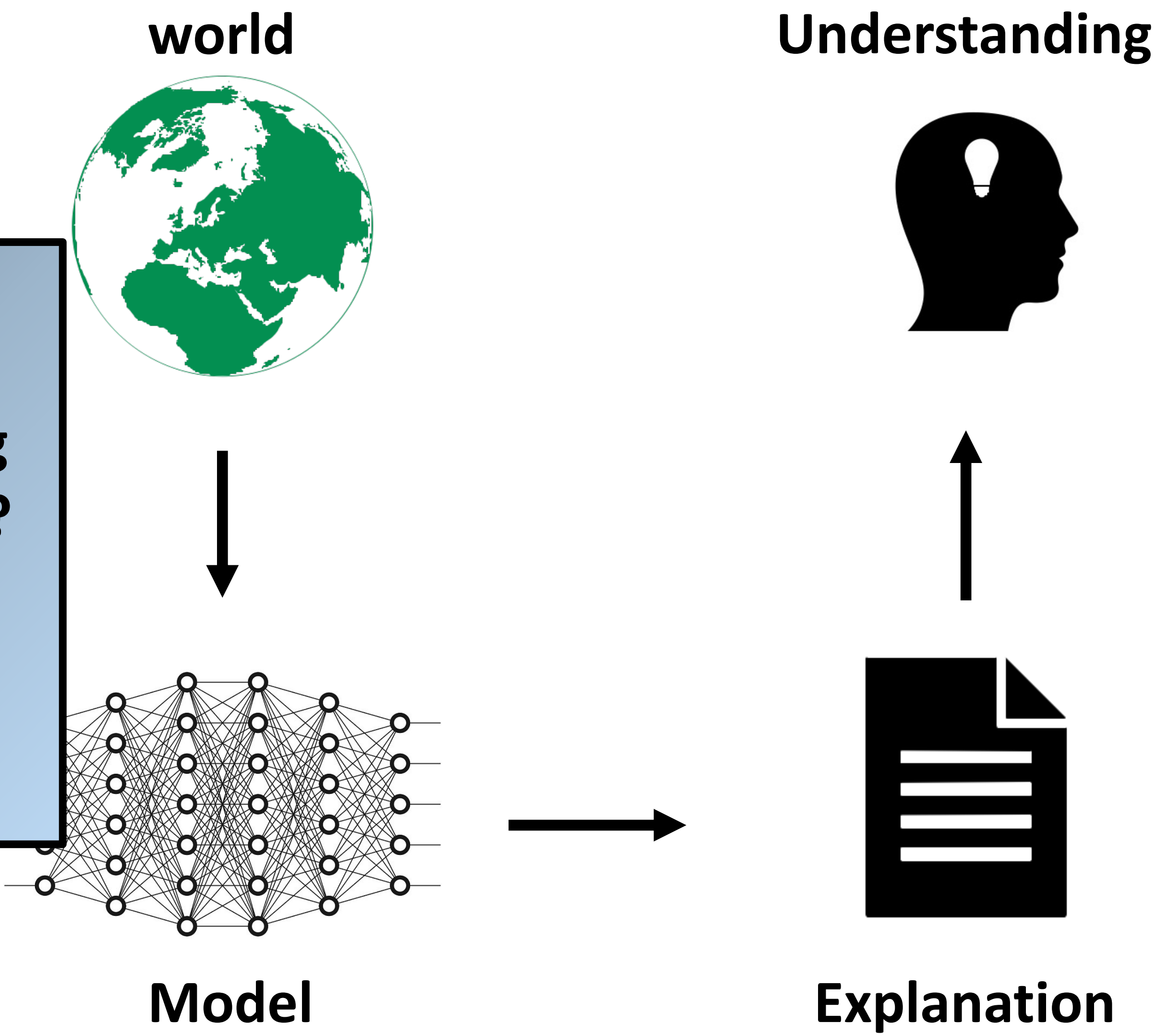
How should we evaluate models?



ML and XAI

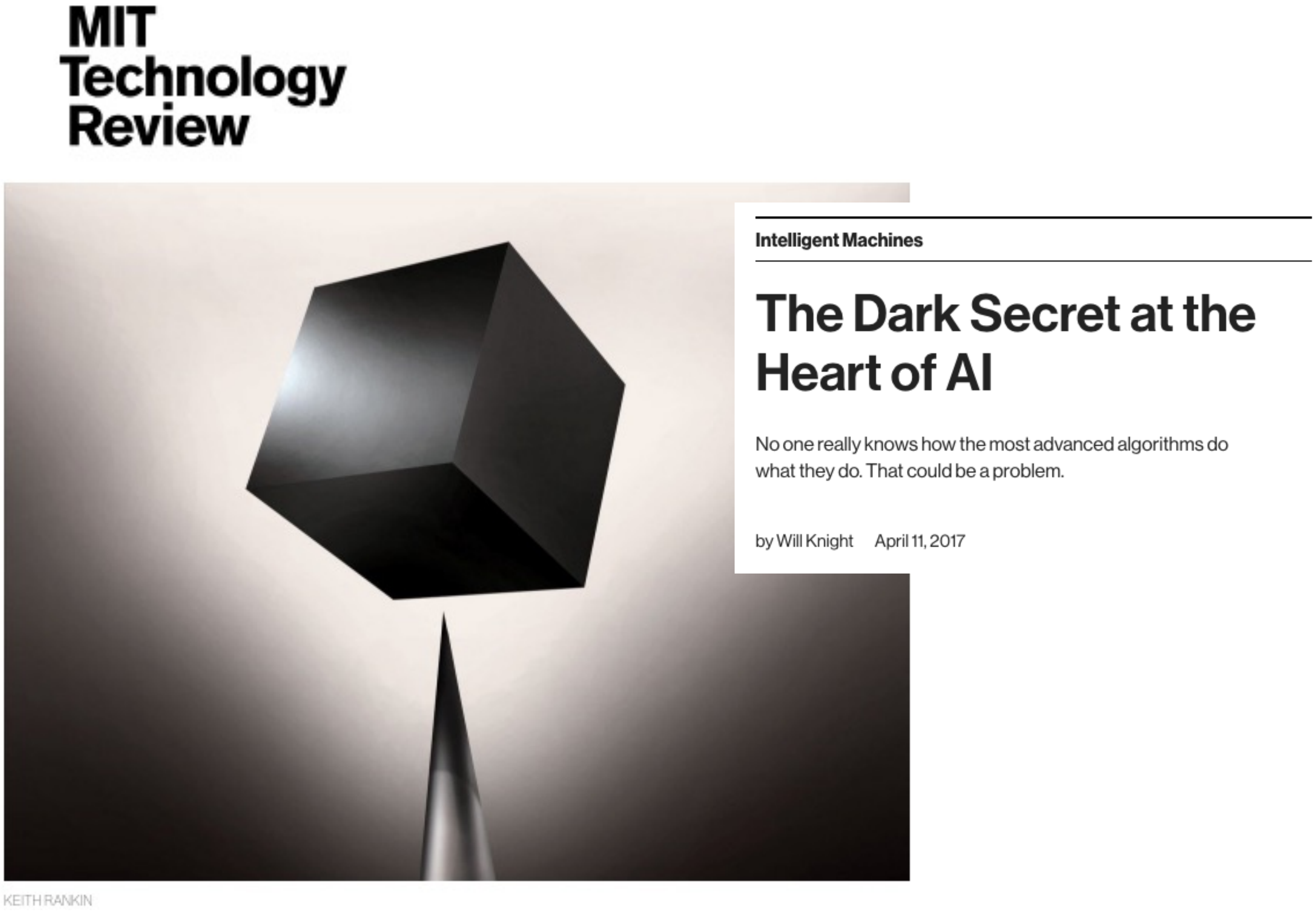
How can modeling practices improve?

How should we evaluate models?



Understanding from ML Models

The British Journal for the Philosophy of Science
Volume 72 / Issue 4
December 2021
Published by the University of Chicago Press
on behalf of the British Society for the Philosophy of Science



MIT Technology Review

Intelligent Machines

The Dark Secret at the Heart of AI

No one really knows how the most advanced algorithms do what they do. That could be a problem.

by Will Knight April 11, 2017

KEITH RANKIN

MIT
Technology
Review

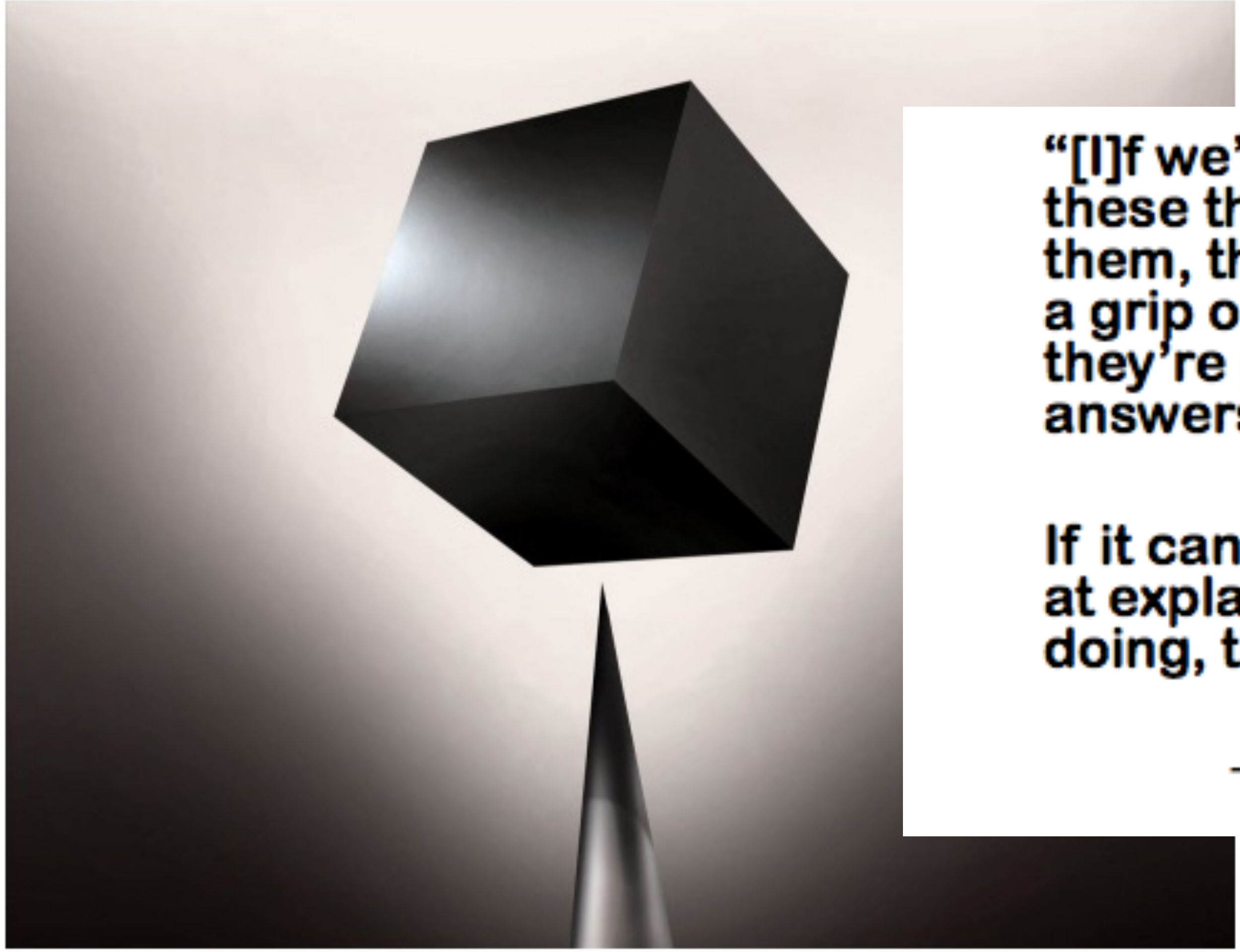
KEITH RANKIN

> “We can build these models, but we don’t know how they work.”

Joel Dudly

Deep Patient Project
Mount Sinai Hospital,
New York

MIT
Technology
Review

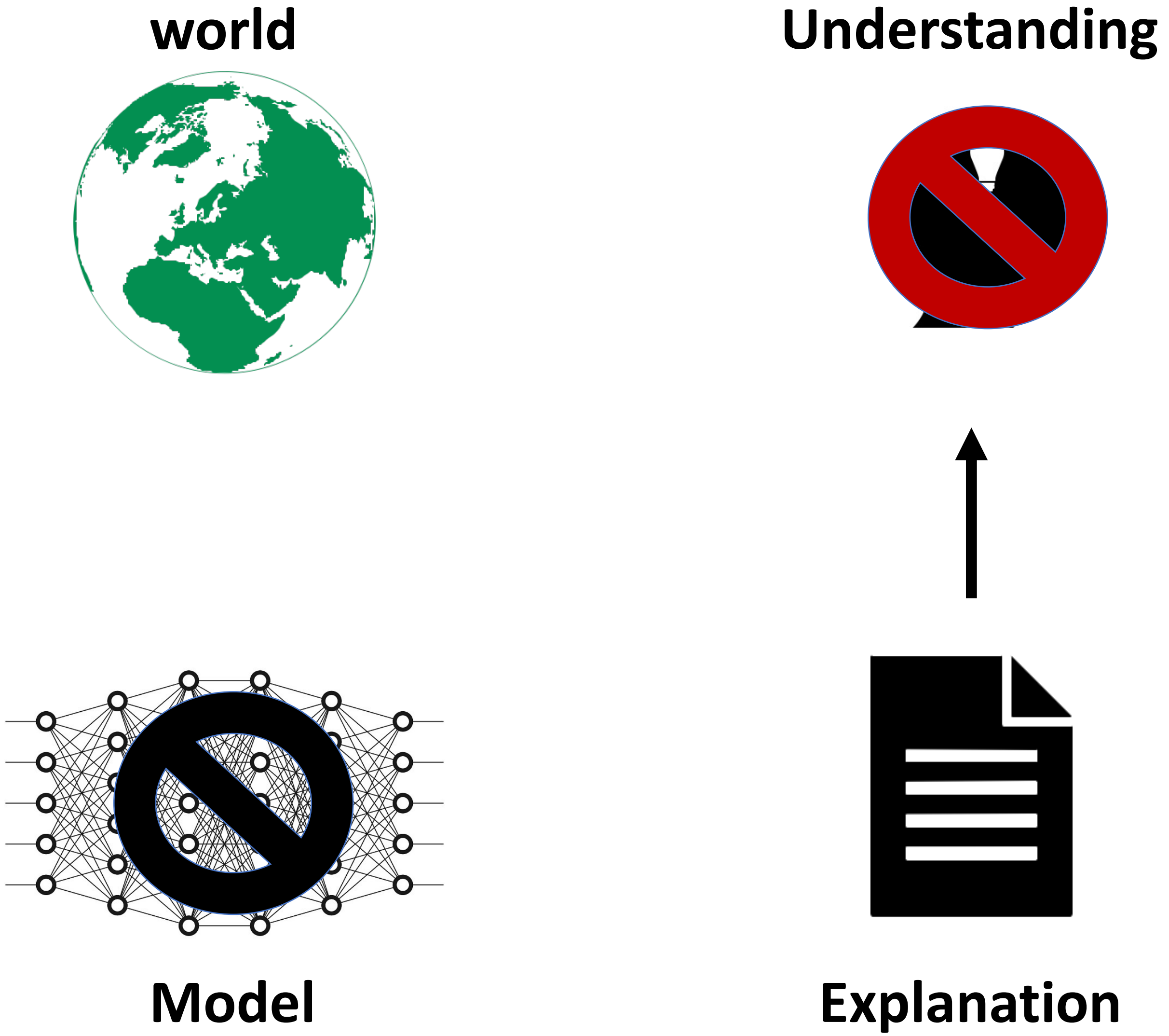


KEITH RANKIN

> “[I]f we’re going to use these things and rely on them, then let’s get as firm a grip on how and why they’re giving us the answers as possible…
>
> If it can’t do better than us at explaining what it’s doing, then don’t trust it.”
>
> - Daniel Dennett

Opacity Hypothesis

Complex and opaque models cannot enable understanding of phenomena because the inner workings of the model are opaque, black-boxed, or unintelligible.



world

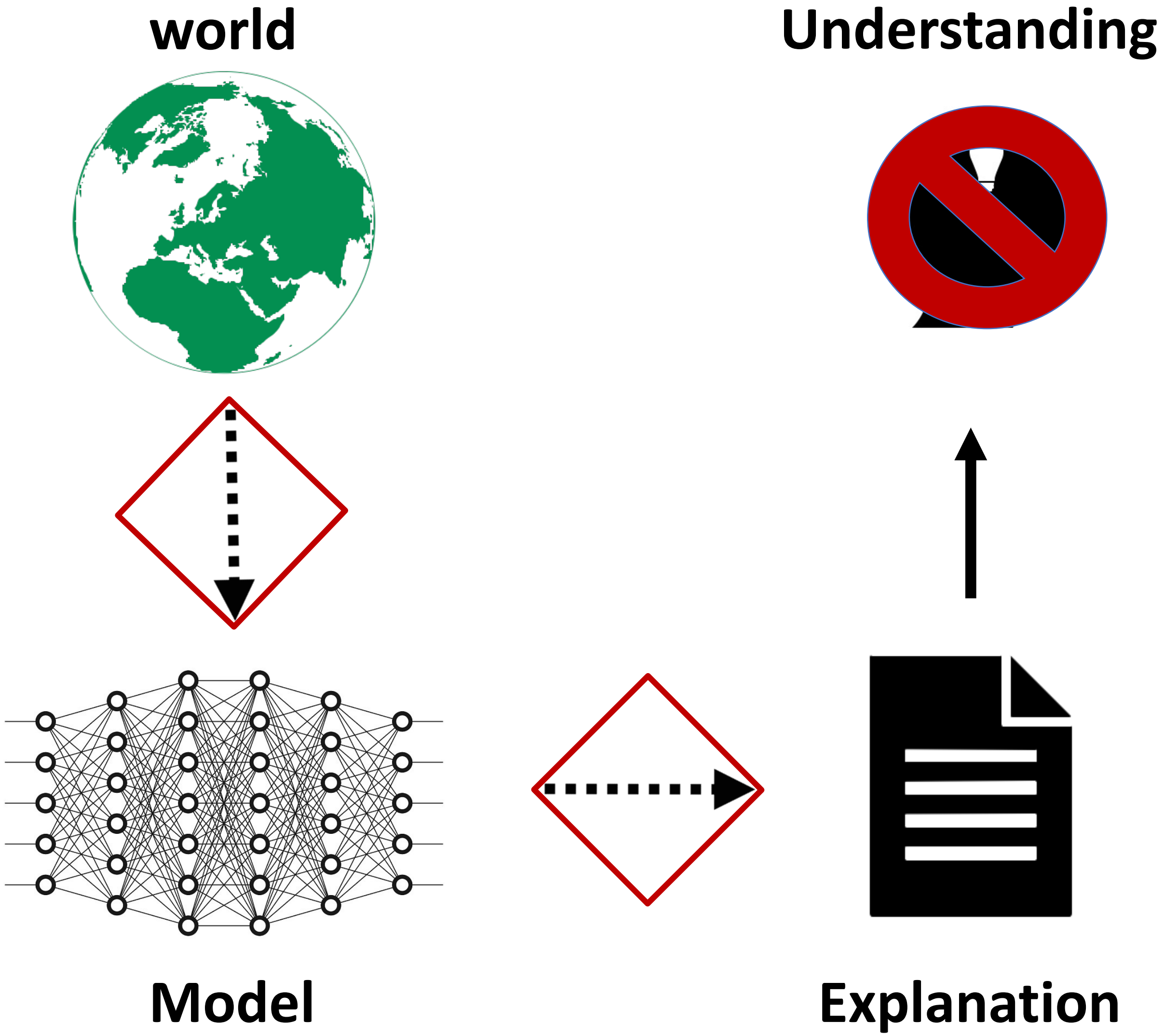
Understanding

Model

Explanation

Link Uncertainty Hypothesis

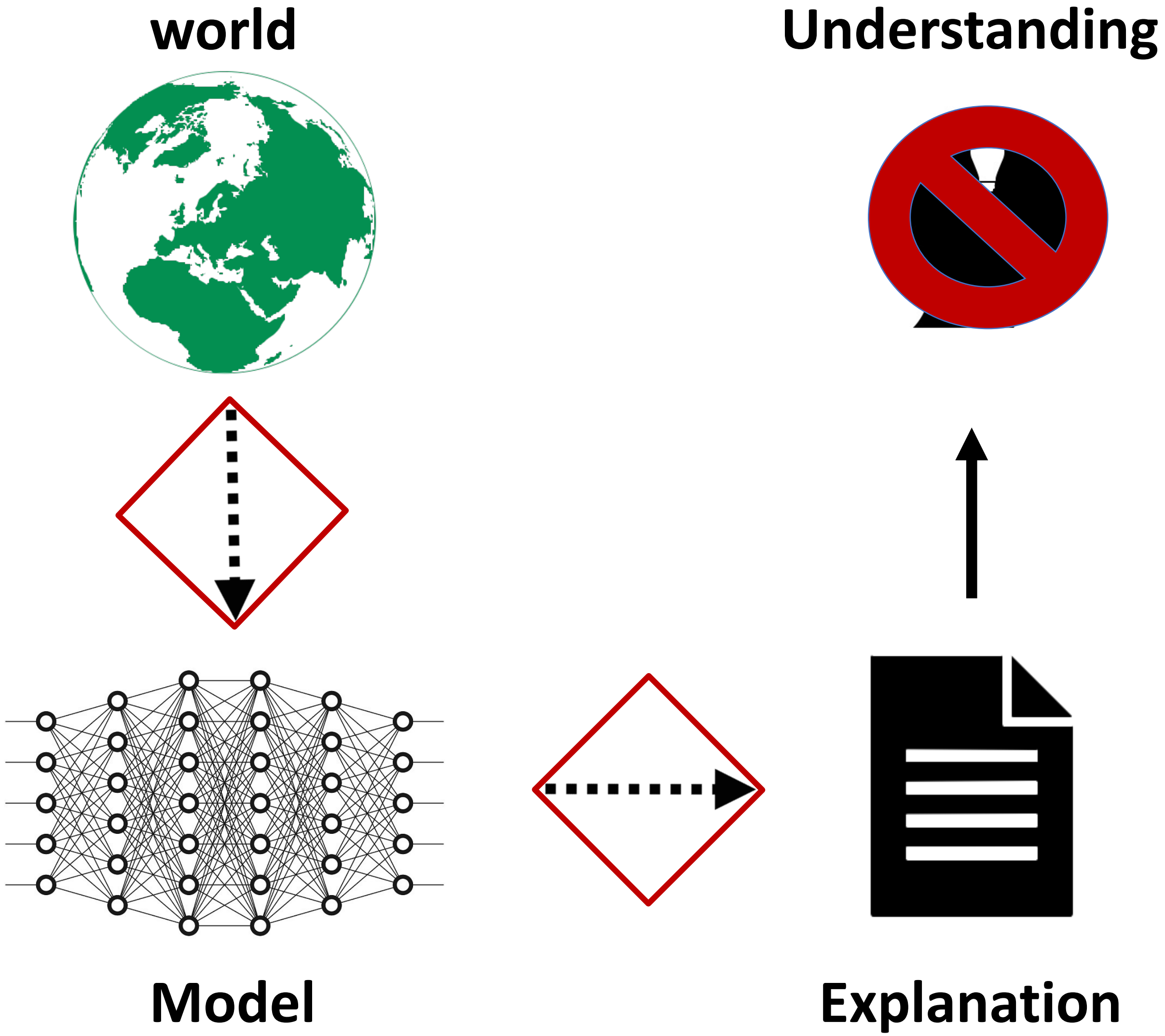
Complex or opaque models fail to explain or enable understanding when the link between the phenomenon and the model is uncertain.



Link Uncertainty Hypothesis

Complex or opaque models fail to explain or enable understanding when the link between the phenomenon and the model is uncertain.

External problem



world

Understanding

Model

Explanation

Outline

Explanation for understanding phenomena

Simple models

ML models

LU and model independence in physics?

Explanation for understanding phenomena

Explanation

Explanation starts with a question:

Why-questions, how-possibly questions, what-if questions …

Explanations are a **type of answer** to a question.

Understanding is knowing a correct explanation

Explanation

Why did the window break?

Glass has x and y physical properties that under great force causes it to break.

Sally threw a rock at a glass window that exhibited great force.

Thus, the window broke

Explanation

Models are not explanations

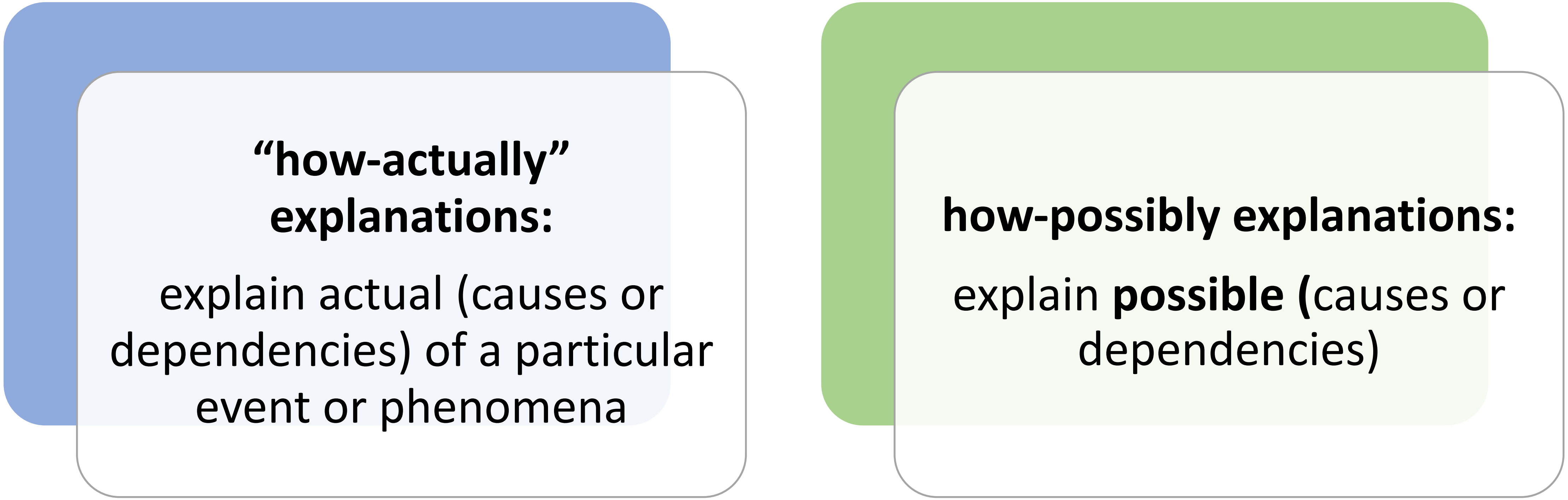
The target of understanding is need not be the model (how it works)

When models help answer ‘why questions’ they explain

XAI methods allows researchers to **discover an explanation** for the phenomenon of interest.

XAI methods only need to **reveal** aspects of a model that **help to induce an explanation of phenomena**.

Explanation



“how-actually” explanations:

explain actual (causes or dependencies) of a particular event or phenomena

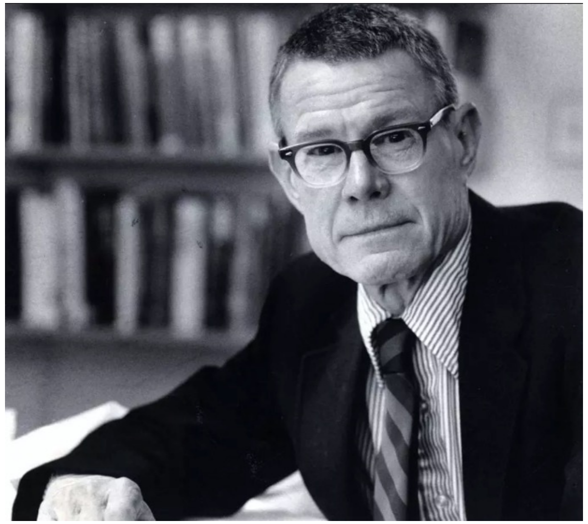
how-possibly explanations:

explain **possible (**causes or dependencies)

Simple models

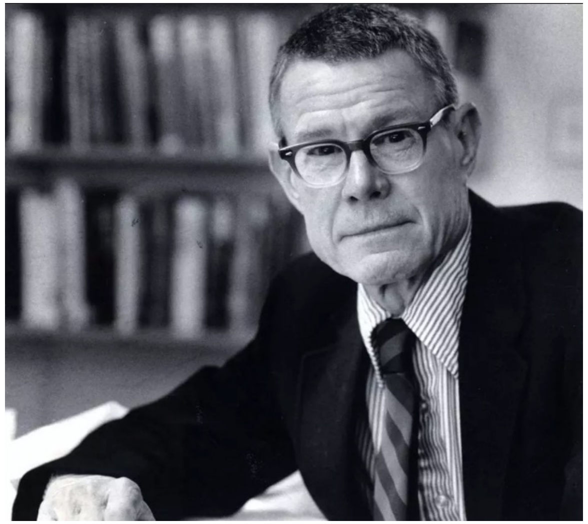
Why are so many real-world populations segregated?

Schelling's
Checkerboard Model
(1971)



Schelling’s Checkerboard Model (1971)

Importantly Schelling’s model provides insight by help of a simple algorithm.



Schelling’s Model

Coins of (two) different types are placed randomly on a board.

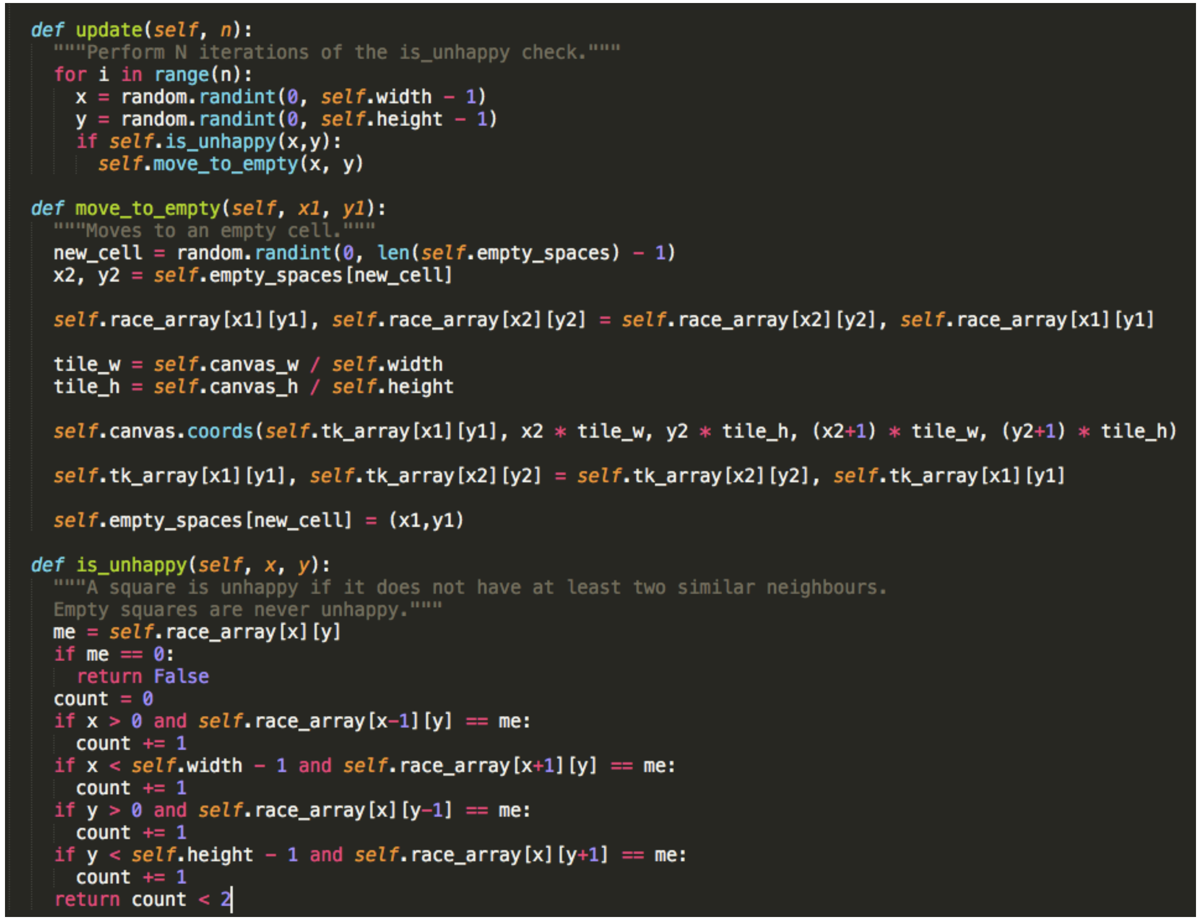
Schelling's Model

If a coin is adjacent to too many coins of the other type, then that coin is moved to closest empty space.

Schelling’s Model

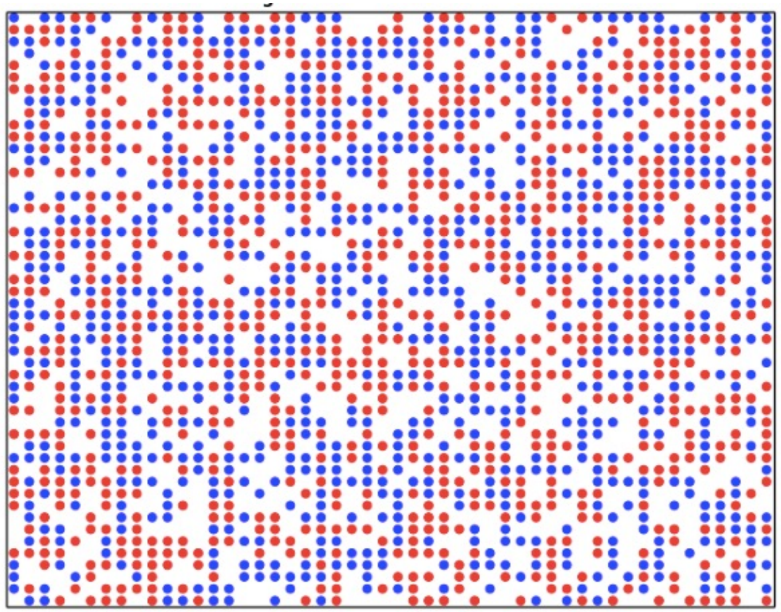
This is repeated until no more changes are made (reaches equilibrium).

Schelling’s Model

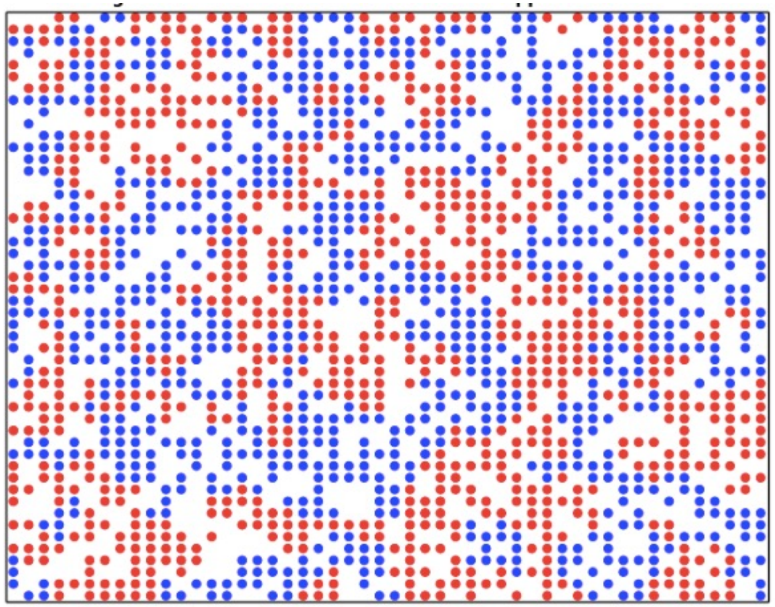


```python
def update(self, n):
    """Perform N iterations of the is_unhappy check."""
    for i in range(n):
        x = random.randint(0, self.width - 1)
        y = random.randint(0, self.height - 1)
        if self.is_unhappy(x,y):
            self.move_to_empty(x, y)

def move_to_empty(self, x1, y1):
    """Moves to an empty cell."""
    new_cell = random.randint(0, len(self.empty_spaces) - 1)
    x2, y2 = self.empty_spaces[new_cell]

    self.race_array[x1][y1], self.race_array[x2][y2] = self.race_array[x2][y2], self.race_array[x1][y1]

    tile_w = self.canvas_w / self.width
    tile_h = self.canvas_h / self.height

    self.canvas.coords(self.tk_array[x1][y1], x2 * tile_w, y2 * tile_h, (x2+1) * tile_w, (y2+1) * tile_h)

    self.tk_array[x1][y1], self.tk_array[x2][y2] = self.tk_array[x2][y2], self.tk_array[x1][y1]

    self.empty_spaces[new_cell] = (x1,y1)

def is_unhappy(self, x, y):
    """A square is unhappy if it does not have at least two similar neighbours.
    Empty squares are never unhappy."""
    me = self.race_array[x][y]
    if me == 0:
        return False
    count = 0
    if x > 0 and self.race_array[x-1][y] == me:
        count += 1
    if x < self.width - 1 and self.race_array[x+1][y] == me:
        count += 1
    if y > 0 and self.race_array[x][y-1] == me:
        count += 1
    if y < self.height - 1 and self.race_array[x][y+1] == me:
        count += 1
    return count < 2
```

Schelling's Model

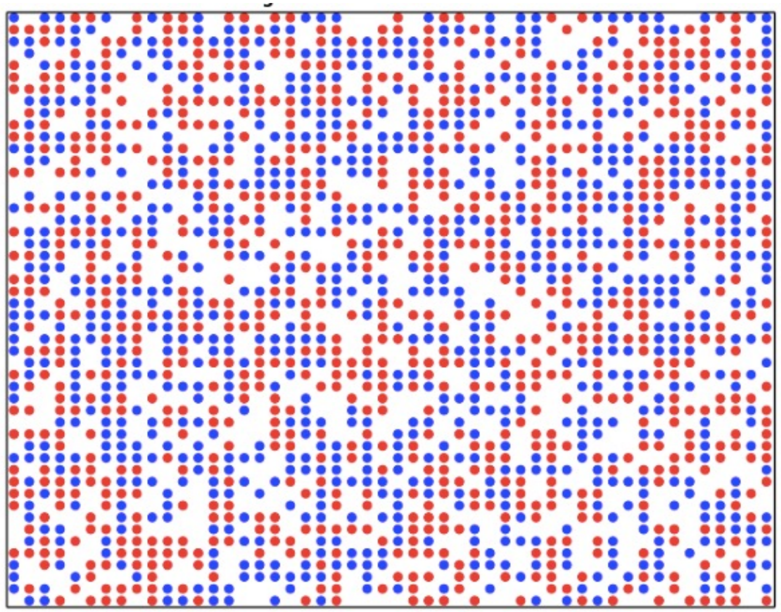


Original State

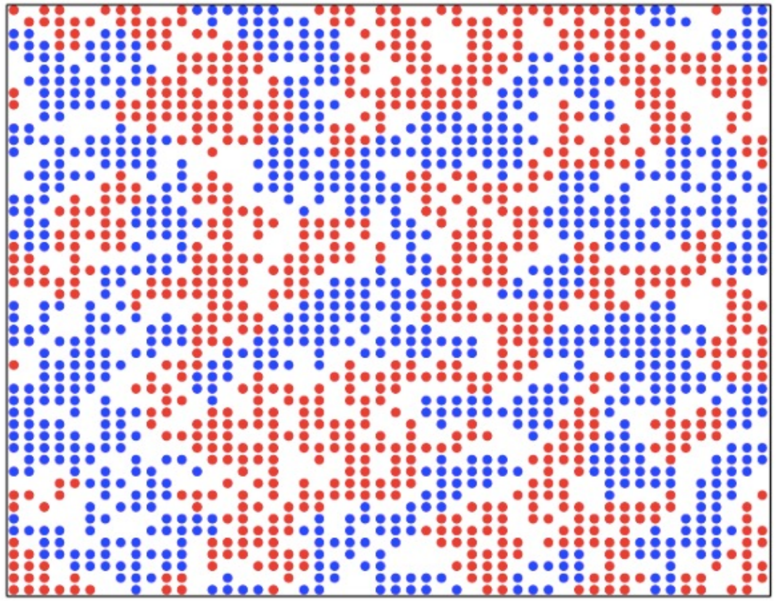


30% preference

Schelling's Model

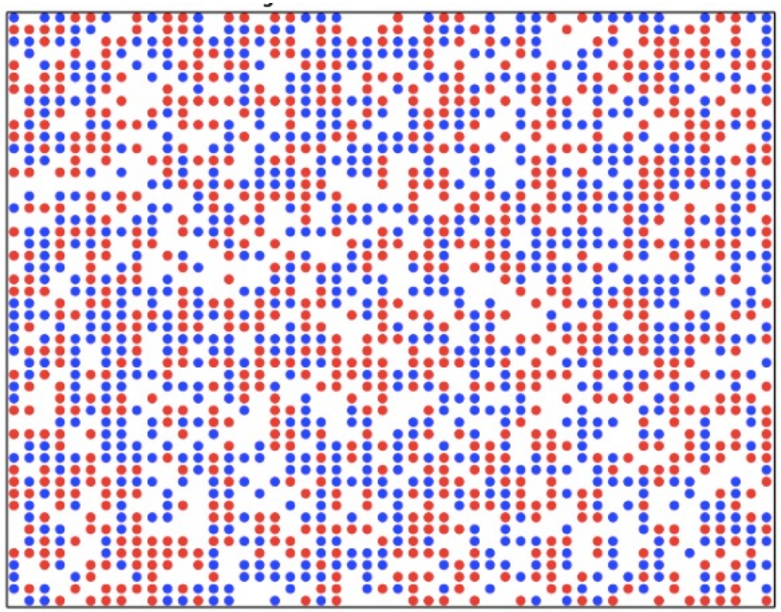


Original State

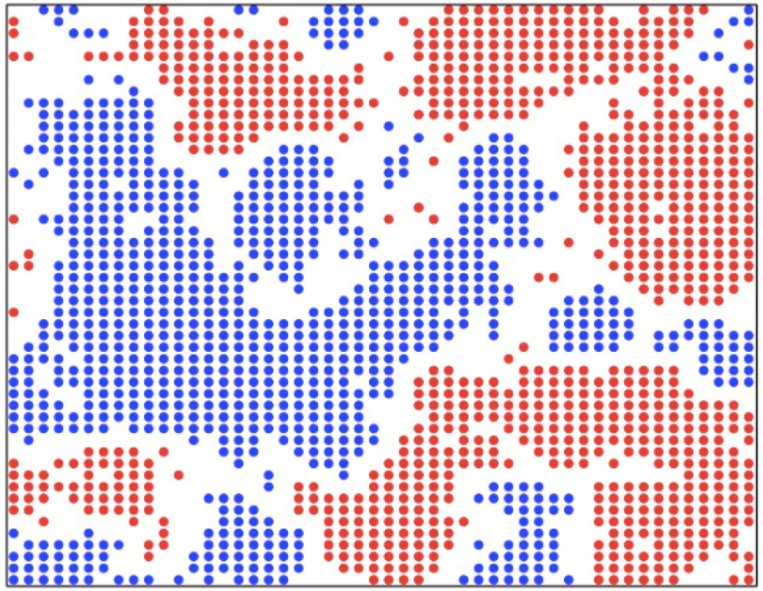


50% preference

Schelling's Model



Original State



80% preference

Explanatory Interests

How does Schelling's Model explain segregation?

How does the algorithm work?

How is it possible that segregation could occur without institutional racism?

Why are so many real-world populations segregated?

Explanatory Interests

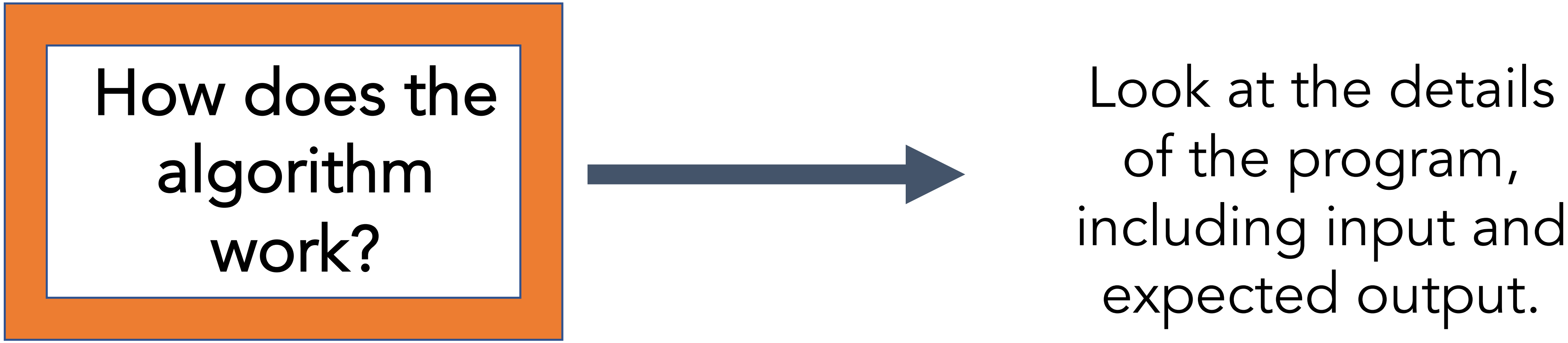
Depending on the question, the algorithm will play a different role in the explanation and understanding.

How does the algorithm work?

How is it possible that segregation could occur without institutional racism?

Why are so many real-world populations segregated?

Explanatory Interests



How does the algorithm work?

Look at the details of the program, including input and expected output.

Explanatory Interests

How is it possible that segregation could occur without institutional racism?

- Look how the algorithm could be used to simulate a possible population.

- The dots represent people of different races; the empty spaces represent possible houses.

- Identify the key feature behind the algorithm and how it maps onto a possible population.

- Need some external support to motivate

Explanatory Interests

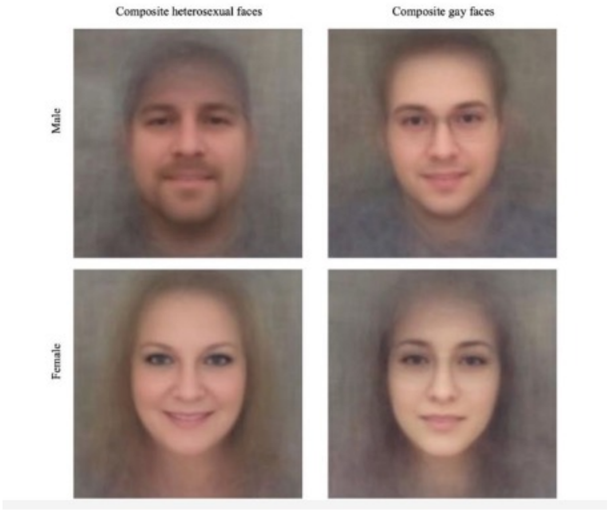
Why are so many real-world populations segregated?

- how the algorithm simulates a real population.
- What is the key feature of the algorithm and how does it map onto real-world populations.
- Must go beyond the model
- Need external evidence that people's preferences primarily determined housing choices
- The appropriate link between the phenomenon and the model must be established.
- In Schelling's case, it is that individual preferences alone can cause segregation in real-world populations.

ML models

Explanation: DNN

SO Model

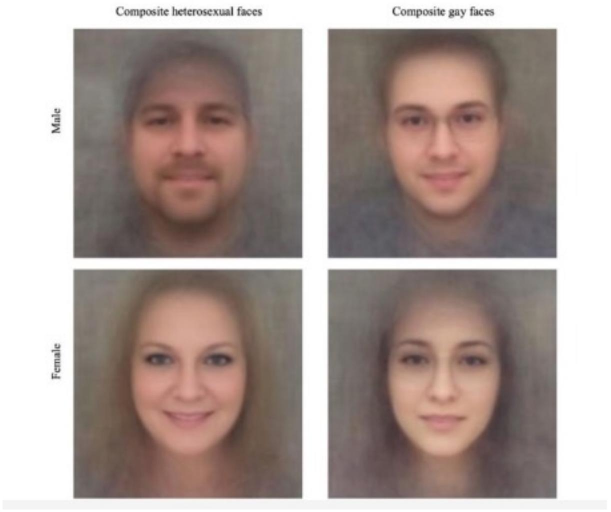


Identify if an individual is gay or straight through facial recognition.

(Wang, Y. and Kosinski, M., 2018)

Explanation: DNN

SO Model



(Wang, Y. and Kosinski, M., 2018)

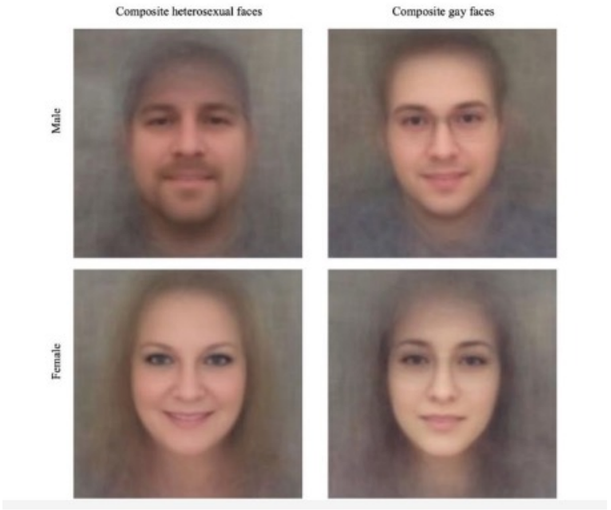
Identify if an individual is gay or straight through facial recognition.

Possible Questions

Is it possible to identify one's sexual orientation based on facial features?

Explanation: DNN

SO Model



(Wang, Y. and Kosinski, M., 2018)

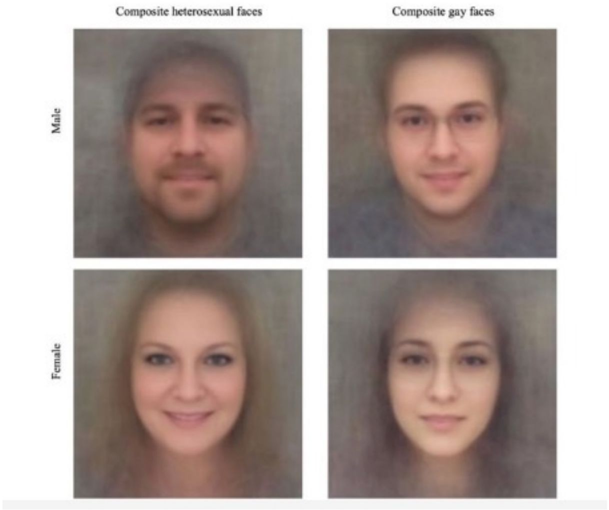
Identify if an individual is gay or straight through facial recognition.

Possible Questions

What would cause facial features to depend on sexual orientation?

Explanation: DNN

SO Model



(Wang, Y. and Kosinski, M., 2018)

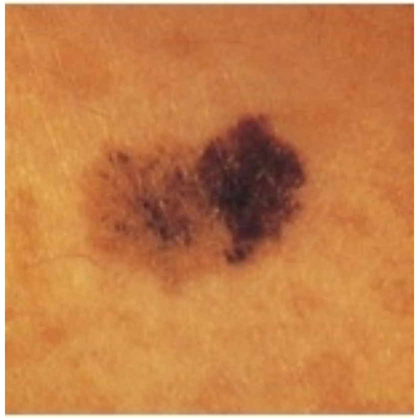
Identify if an individual is gay or straight through facial recognition.

Possible Questions

Why is the model able to with a high accuracy classify sexual orientation through facial images?

Explanation: DNN

Melanoma Model

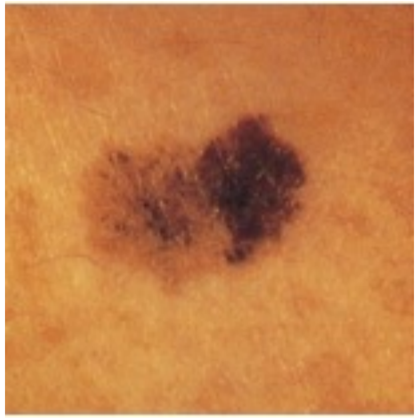


(Esteva, A., et. al, 2017)

Identify if a mole is likely to be a melanoma.

Explanation: DNN

Melanoma Model



(Esteva, A., et. al, 2017)

Identify if a mole is likely to be a melanoma.

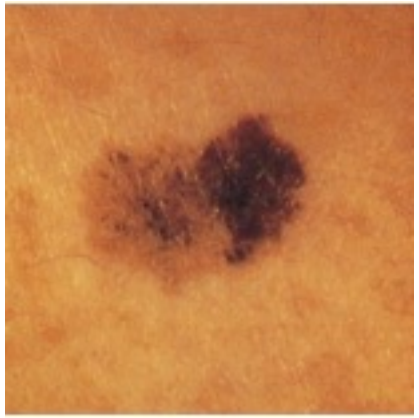
→

Possible Questions

What are the visual signs of melanoma?

Explanation: DNN

Melanoma Model



(Esteva, A., et. al, 2017)

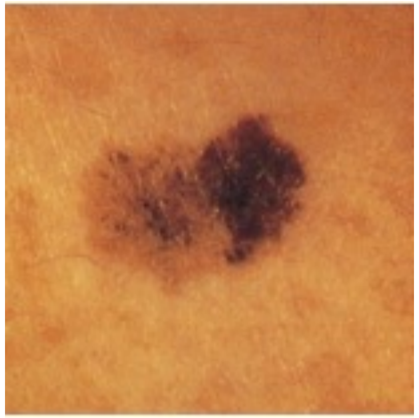
Identify if a mole is likely to be a melanoma.

Possible Questions

How does a melanoma differ from a mole?

Explanation: DNN

Melanoma Model



(Esteva, A., et. al, 2017)

Identify if a mole is likely to be a melanoma.

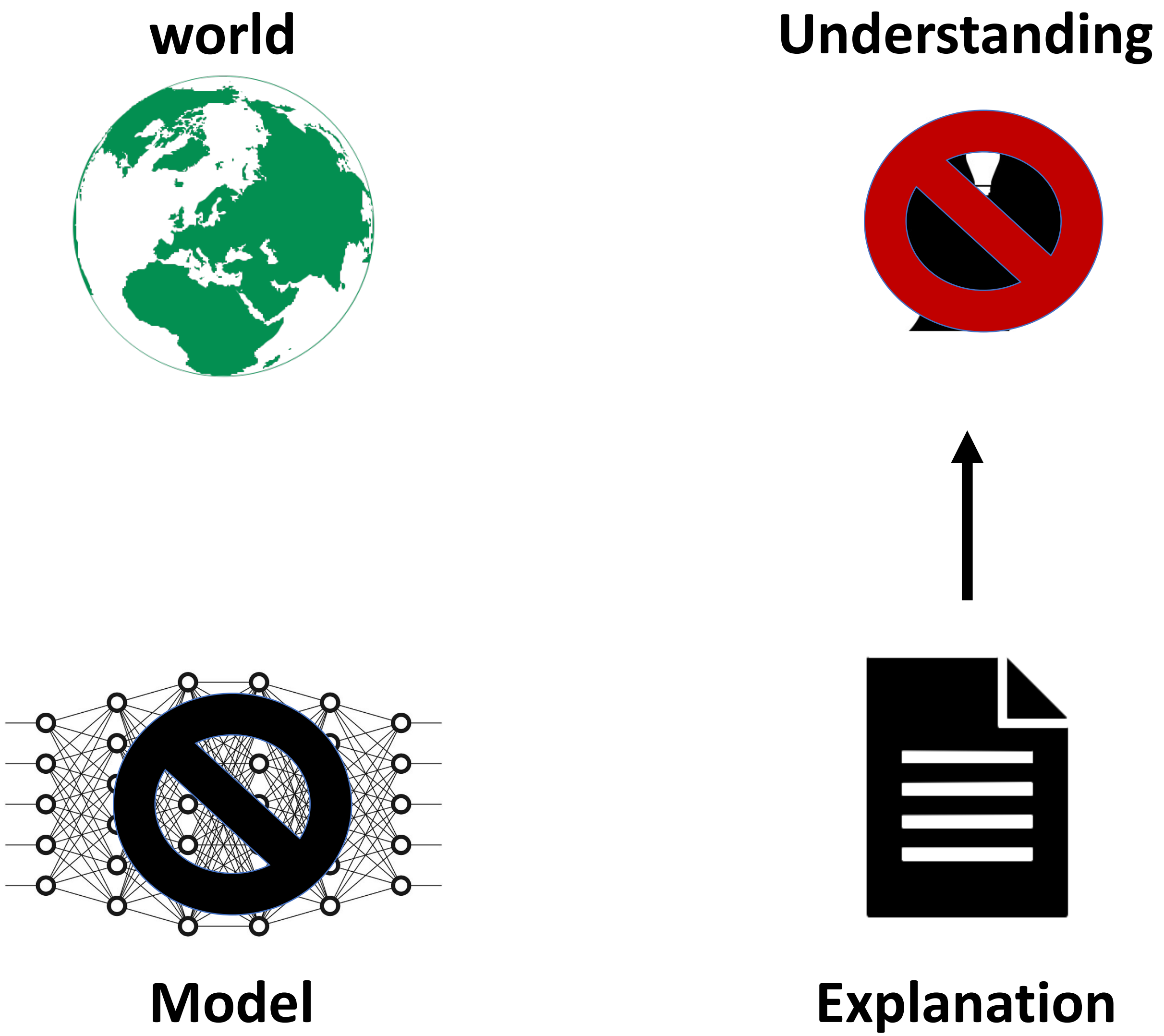
→

Possible Questions

Why should a particular patient's mole be biopsied for melanoma?

Opacity Hypothesis

Complex and opaque models cannot enable understanding of phenomena because the inner workings of the model are opaque, black-boxed, or unintelligible.



world

Understanding

Model

Explanation

Black-Boxes

Implementation Black-Box

Low-level details of algorithm implementation is obscured, unknown, or illegible.

Example:
Computing factorials

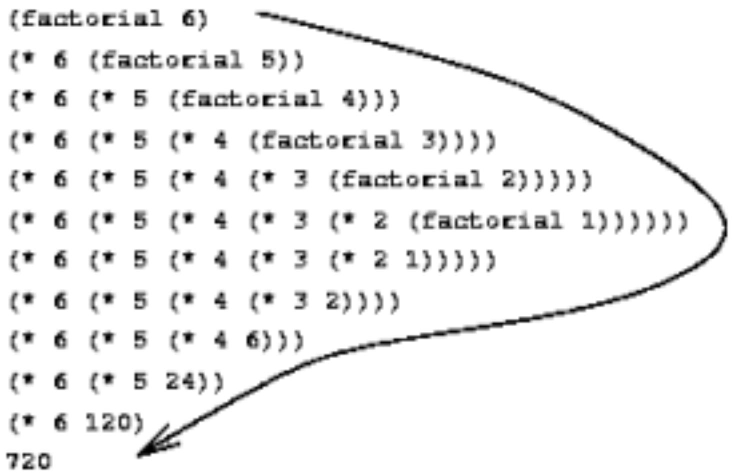
(factorial 6)

There is a black-box around how (factorial 6) is implemented.

Black-Boxes

Recursive process

```
(define (factorial n)
  (if (= n 1)
      1
      (* n (factorial (- n 1)))))
```



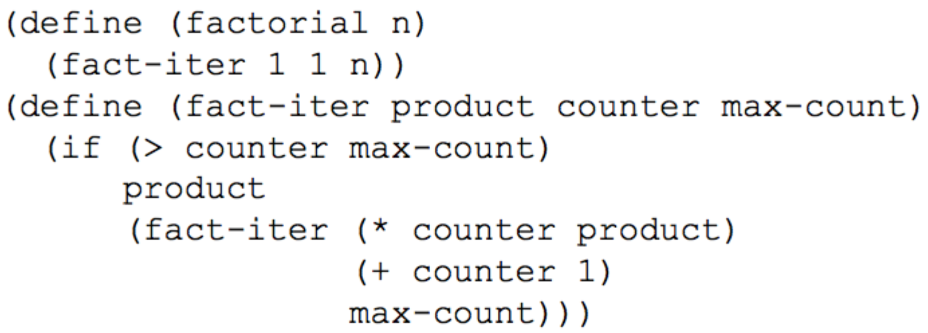
```
(factorial 6)
(* 6 (factorial 5))
(* 6 (* 5 (factorial 4)))
(* 6 (* 5 (* 4 (factorial 3))))
(* 6 (* 5 (* 4 (* 3 (factorial 2)))))
(* 6 (* 5 (* 4 (* 3 (* 2 (factorial 1))))))
(* 6 (* 5 (* 4 (* 3 (* 2 1)))))
(* 6 (* 5 (* 4 (* 3 2))))
(* 6 (* 5 (* 4 6)))
(* 6 (* 5 24))
(* 6 120)
720
```

Figure 1.3: A linear recursive process for computing 6!.

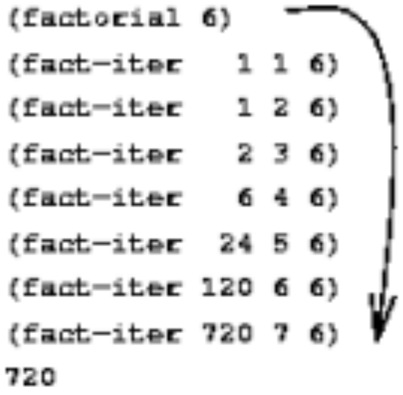
Example from *SICP*
Sussman and Abelson

Black-Boxes

Iterative process



```
(define (factorial n)
  (fact-iter 1 1 n))
(define (fact-iter product counter max-count)
  (if (> counter max-count)
      product
      (fact-iter (* counter product)
                 (+ counter 1)
                 max-count)))
```



```
(factorial 6)
(fact-iter   1 1 6)
(fact-iter   1 2 6)
(fact-iter   2 3 6)
(fact-iter   6 4 6)
(fact-iter  24 5 6)
(fact-iter 120 6 6)
(fact-iter 720 7 6)
720
```

Figure 1.4: A linear iterative process for computing 6!.

Example from *SICP*
Sussman and Abelson

Black-Boxes

Explanation / Understanding?

Implementation Black-Box

Low-level details of algorithm implementation is obscured, unknown, or illegible.

Black boxes do not prohibit understanding in virtue of abstracted ***implementation*.**

Don't need to know how (factorial 6) is implemented to understand, say a climate model, that utilizes factorials.

For higher level questions the exact implementation does not need to be known in order to enable understanding.

Black-Boxes

When implementation matters

Implementation Black-Box

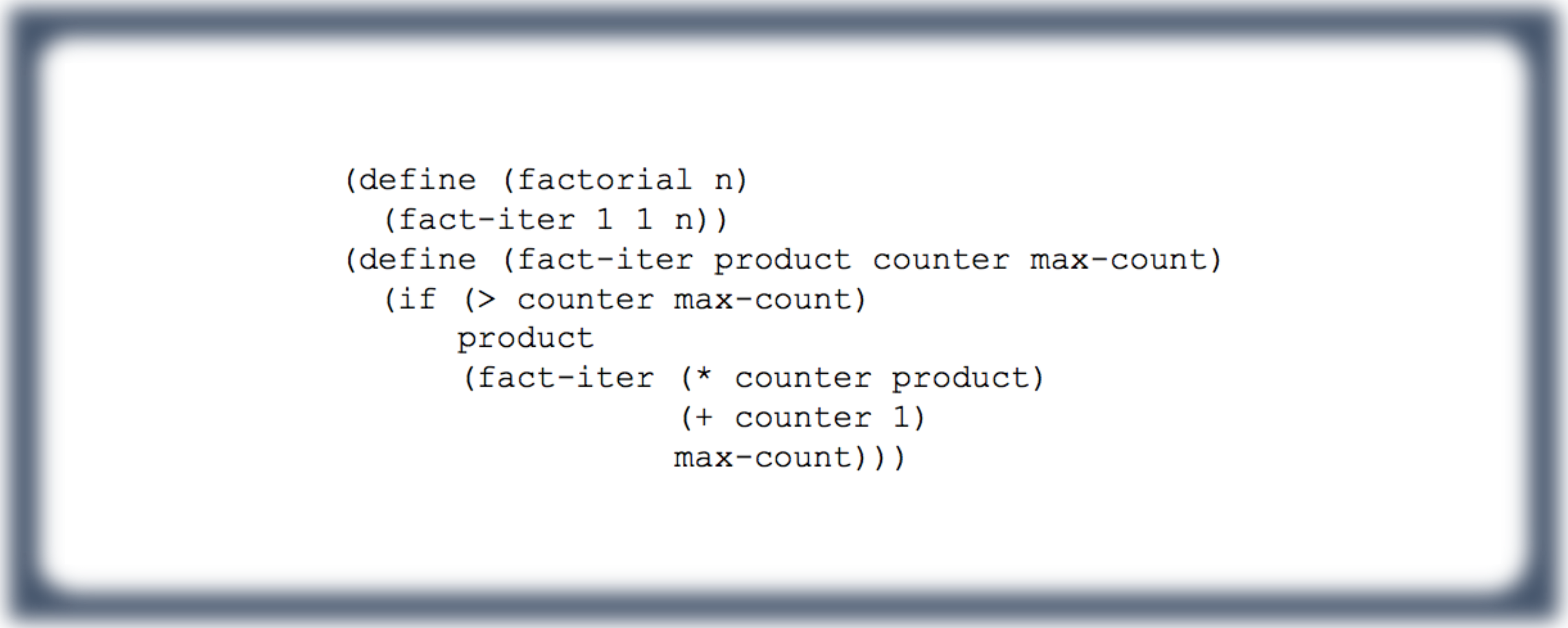
Low-level details of algorithm implementation is obscured, unknown, or illegible.

How is this feature of the algorithm implemented?

Why is this implementation better (or faster) than this other implementation?

Black-Boxes

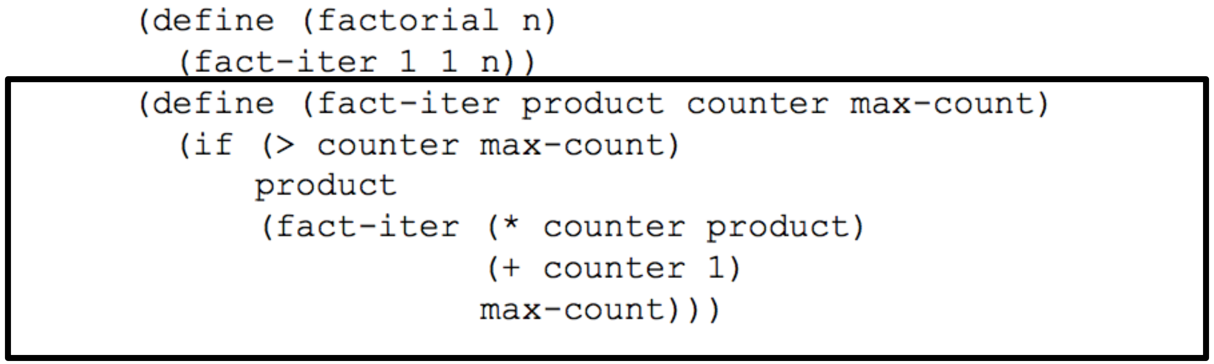
Levels of black-boxes



```
(define (factorial n)
  (fact-iter 1 1 n))
(define (fact-iter product counter max-count)
  (if (> counter max-count)
      product
      (fact-iter (* counter product)
                 (+ counter 1)
                 max-count)))
```

Black-Boxes

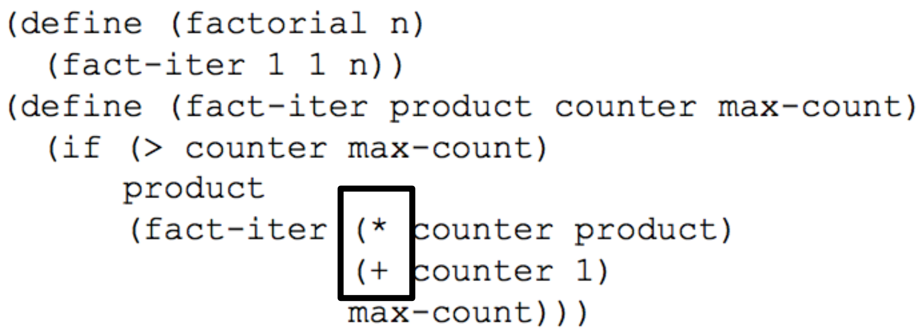
Levels of black-boxes



```
(define (factorial n)
  (fact-iter 1 1 n))
(define (fact-iter product counter max-count)
  (if (> counter max-count)
      product
      (fact-iter (* counter product)
                 (+ counter 1)
                 max-count)))
```

Black-Boxes

Levels of black-boxes



```
(define (factorial n)
  (fact-iter 1 1 n))
(define (fact-iter product counter max-count)
  (if (> counter max-count)
      product
      (fact-iter (* counter product)
                 (+ counter 1)
                 max-count)))
```

Black-Boxes

Levels of black-boxes

What levels of implementation black boxes undermine explanation and understanding?

If the algorithm is indeterminate?

If it changes or updates while running it?

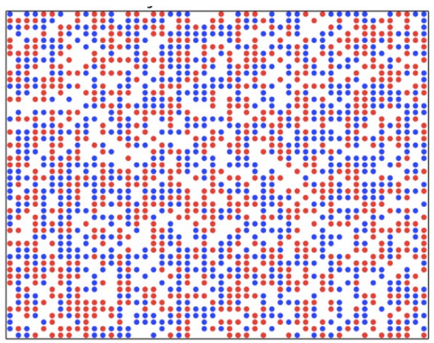
Black-Boxes

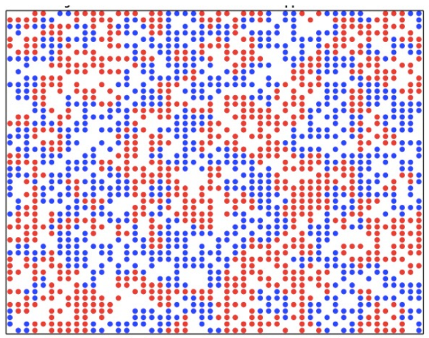
Levels of black-boxes

What levels of implementation black boxes undermine explanation and understanding?

If the algorithm is indeterminate?

If it changes or updates while running it?





Black-Boxes

Levels of black-boxes

Highest-level of black-box

input

output

The entire algorithm is obscured.

Black-Boxes

Levels of black-boxes

Highest-level of black-box

input
output

Goal of algorithm (model)

Way (model / algorithm) achieves goal

Black-Boxes

Levels of black-boxes

Highest-level of black-box

input

output

~~Goal of algorithm (model)~~

Way (model / algorithm) achieves goal

Black-Boxes

Levels of black-boxes

Highest-level of black-box

input

output

Goal of algorithm (model)

~~Way (model / algorithm) achieves goal~~

Black-Boxes

Levels of black-boxes

What about DNN models?

Highest-level of black-box

input

output

Goal of algorithm (model) ✓

Way (model / algorithm) achieves goal

Black-Boxes

Levels of black-boxes

What about DNN models?

Highest-level of black-box

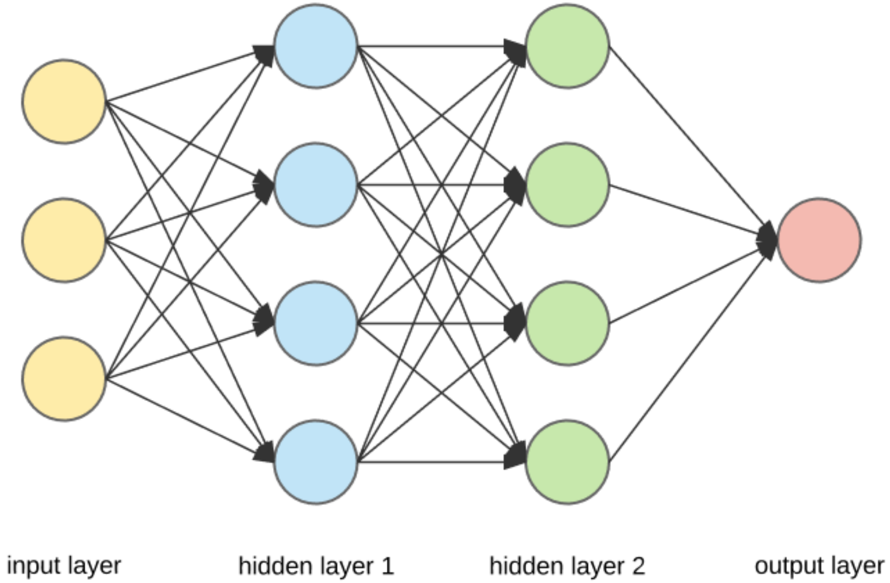
input

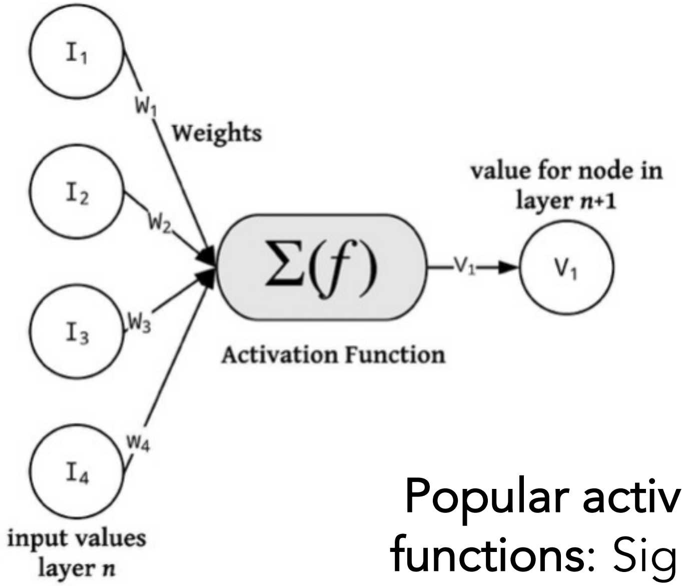
output

Goal of algorithm (model) ✓

Way (model / algorithm) achieves goal ✓

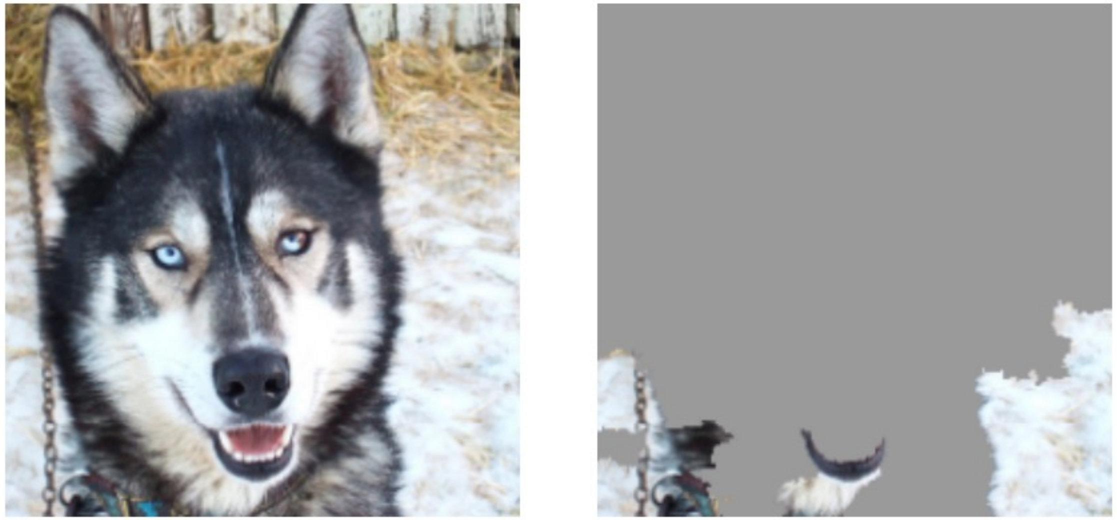
We know the high-level algorithmic structures of DNNs.





Popular activation functions: Sigmoid, hyperbolic tangent function (tanh), rectified linear units (ReLu)

XAI



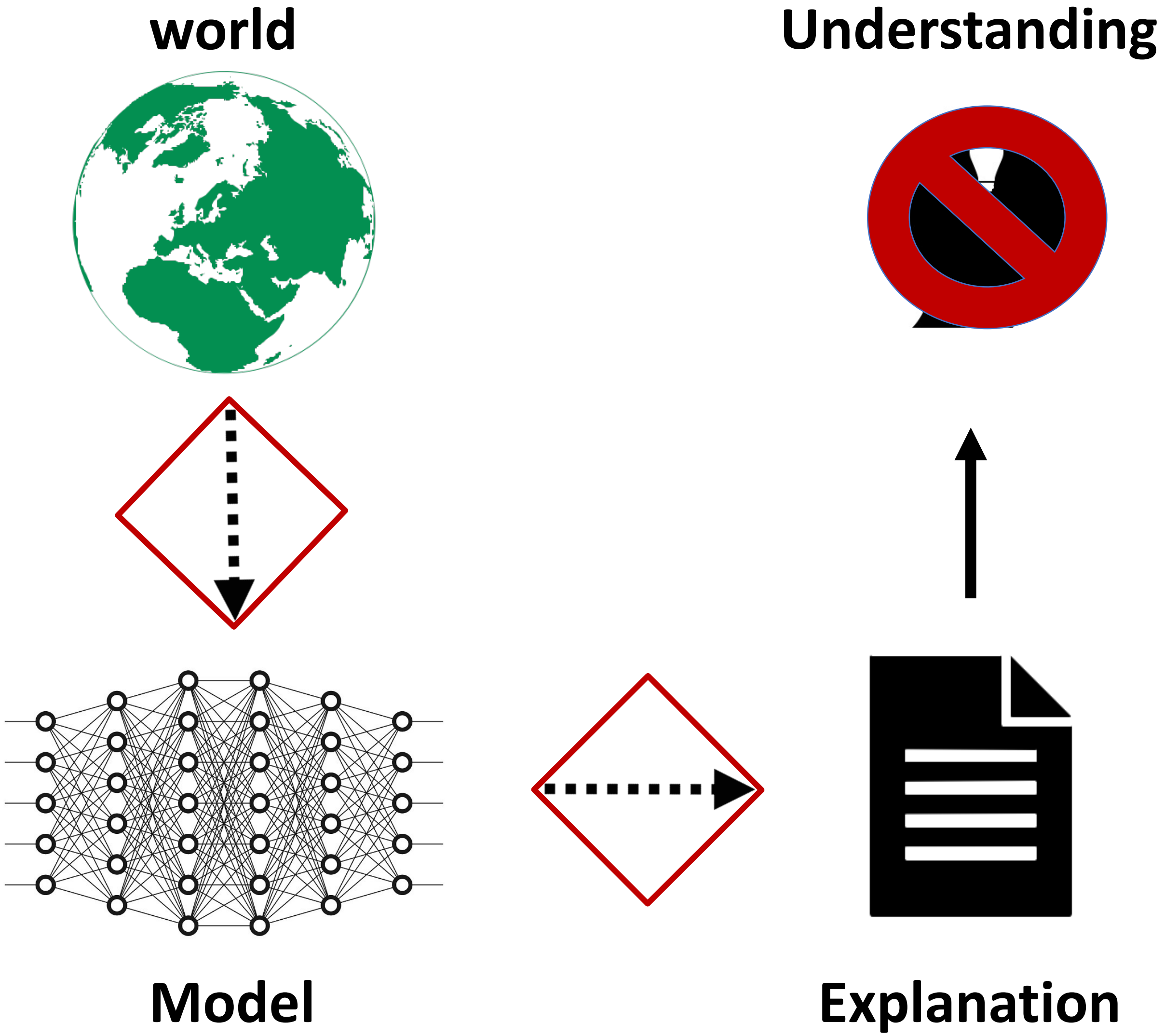
(a) Husky classified as wolf (b) Explanation

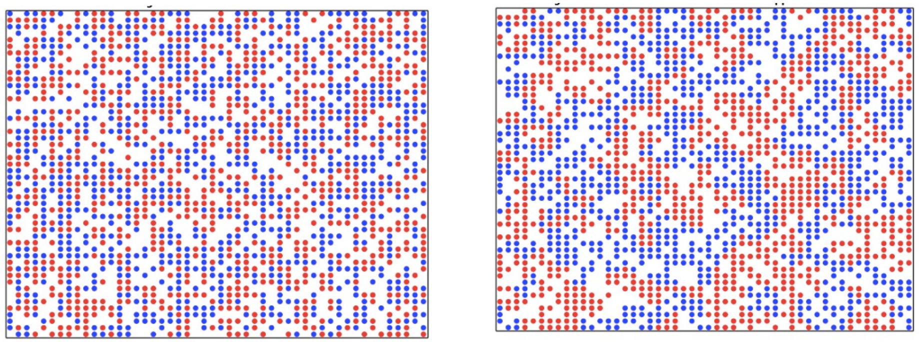
Ribeiro et al. 2016

LIME

Link Uncertainty Hypothesis

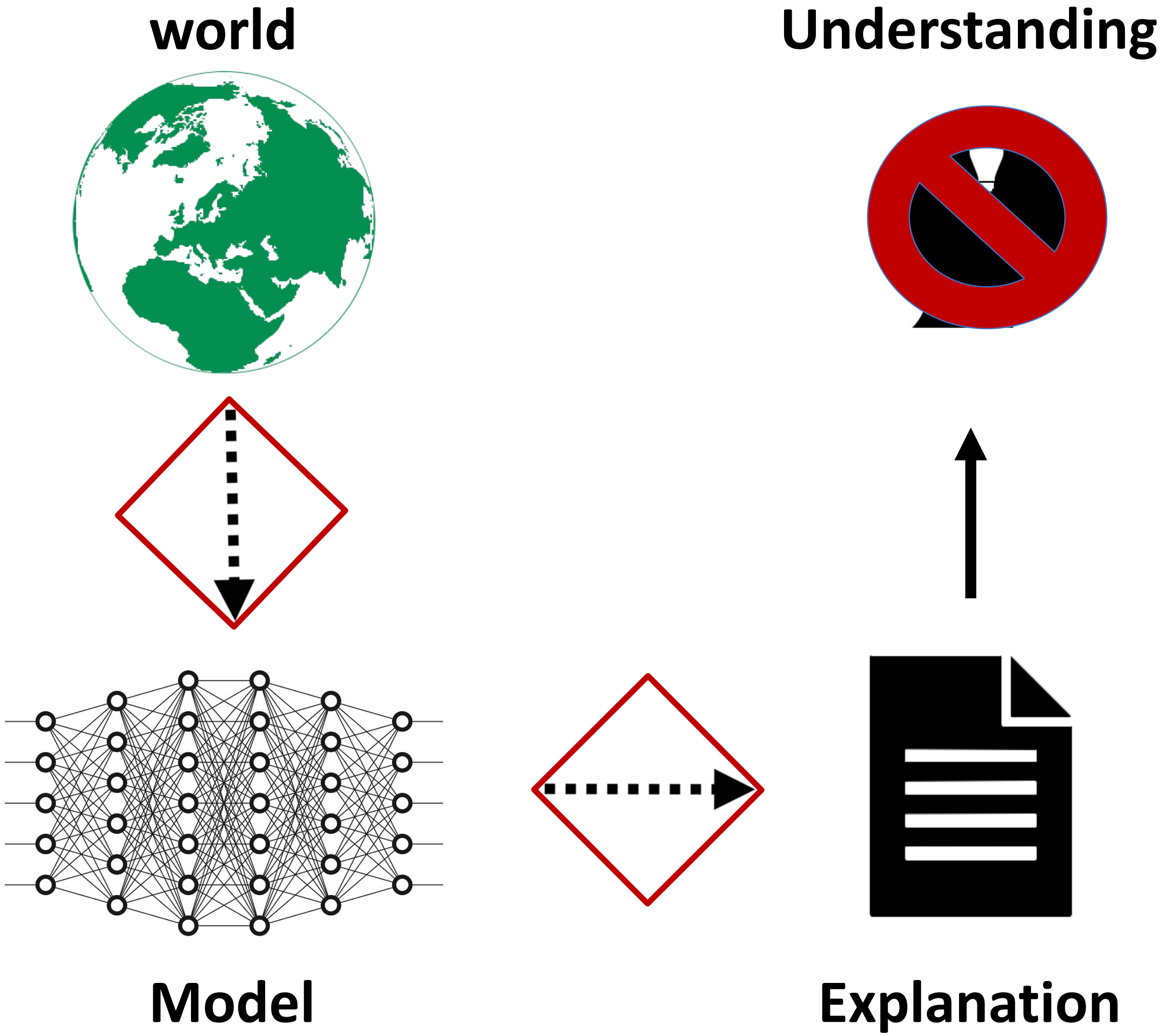
Complex or opaque models fail to explain or enable understanding when the link between the phenomenon and the model is uncertain.





How-possibly?

How-actually?

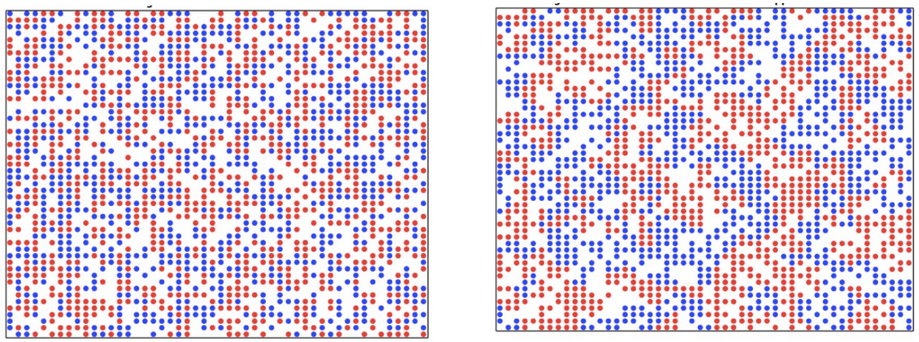


world

Understanding

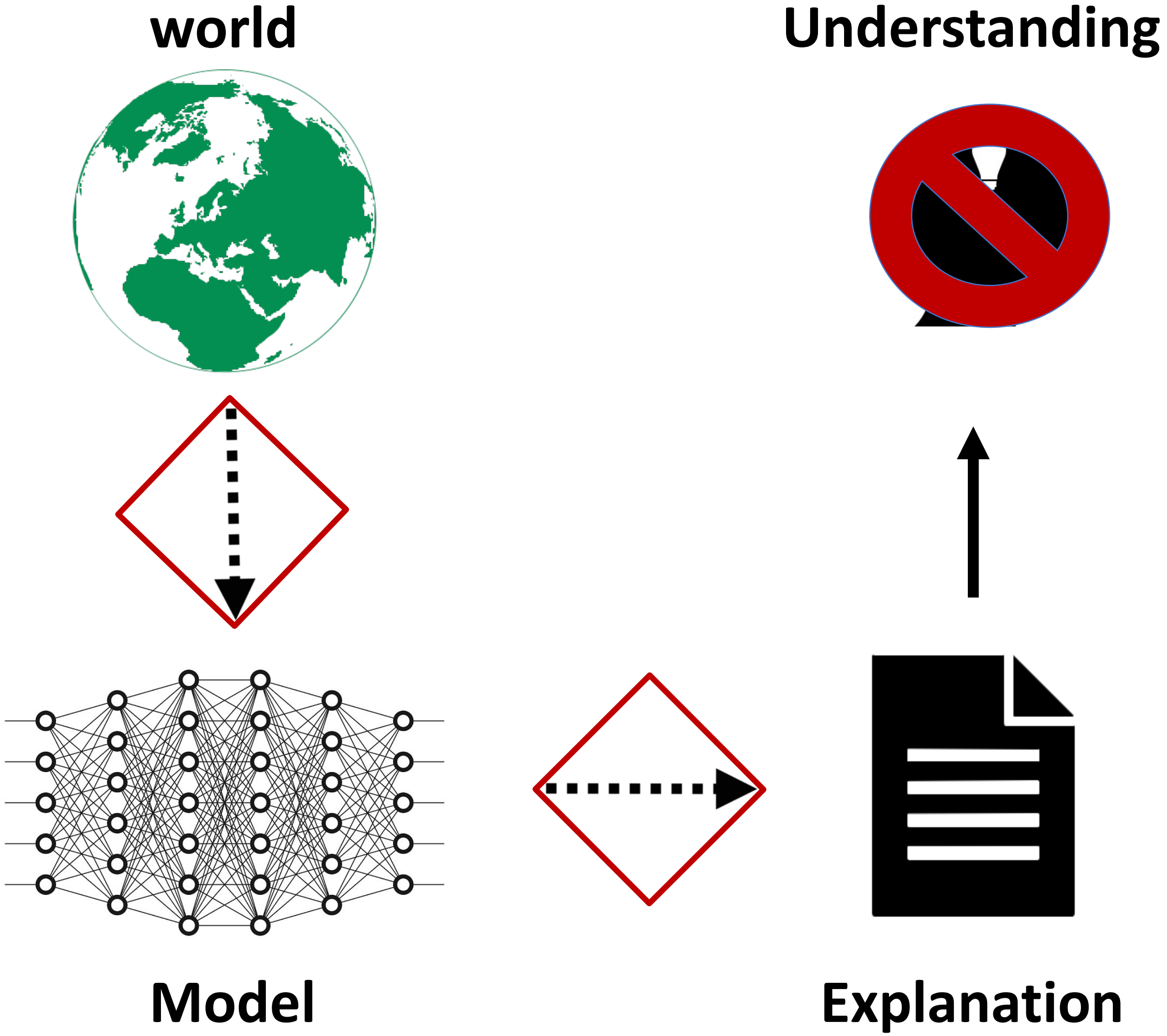
Model

Explanation



How-possibly? ✓

How-actually?

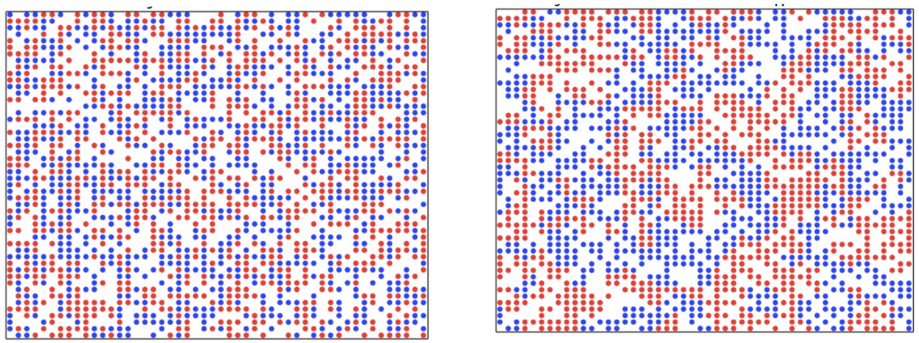


world

Understanding

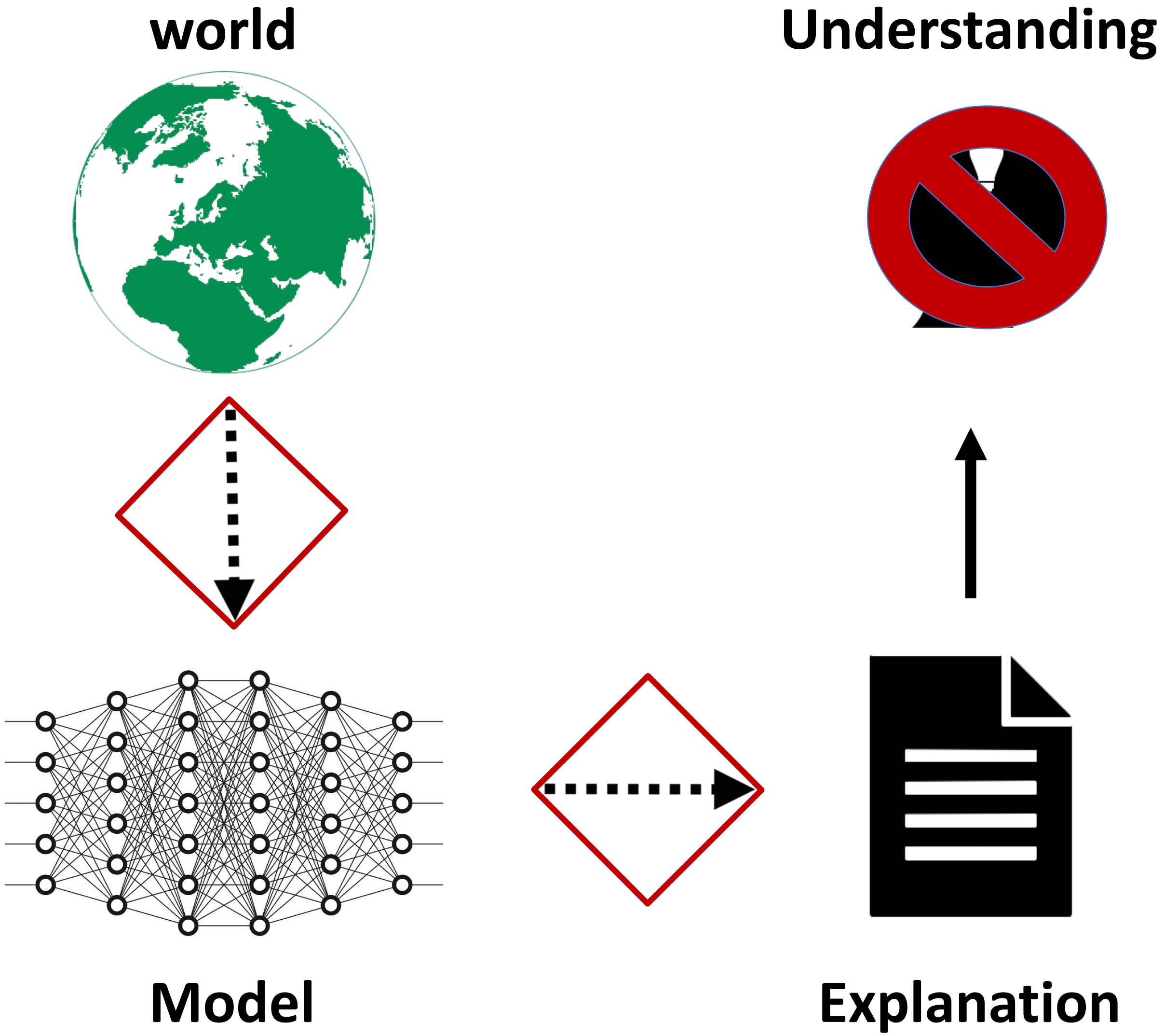
Model

Explanation



How-possibly? ✓

~~How-actually?~~

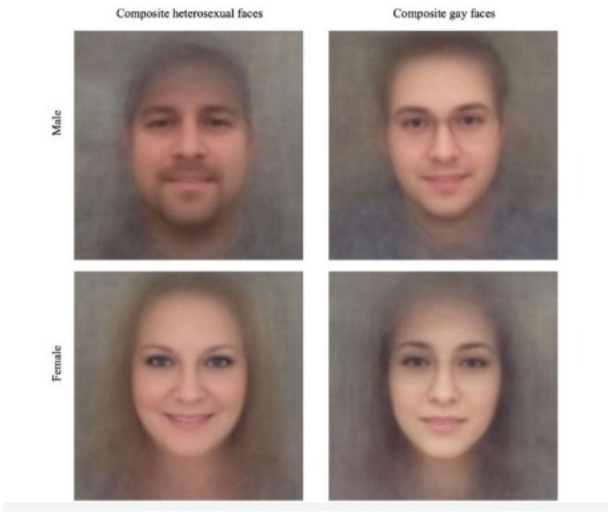


world

Understanding

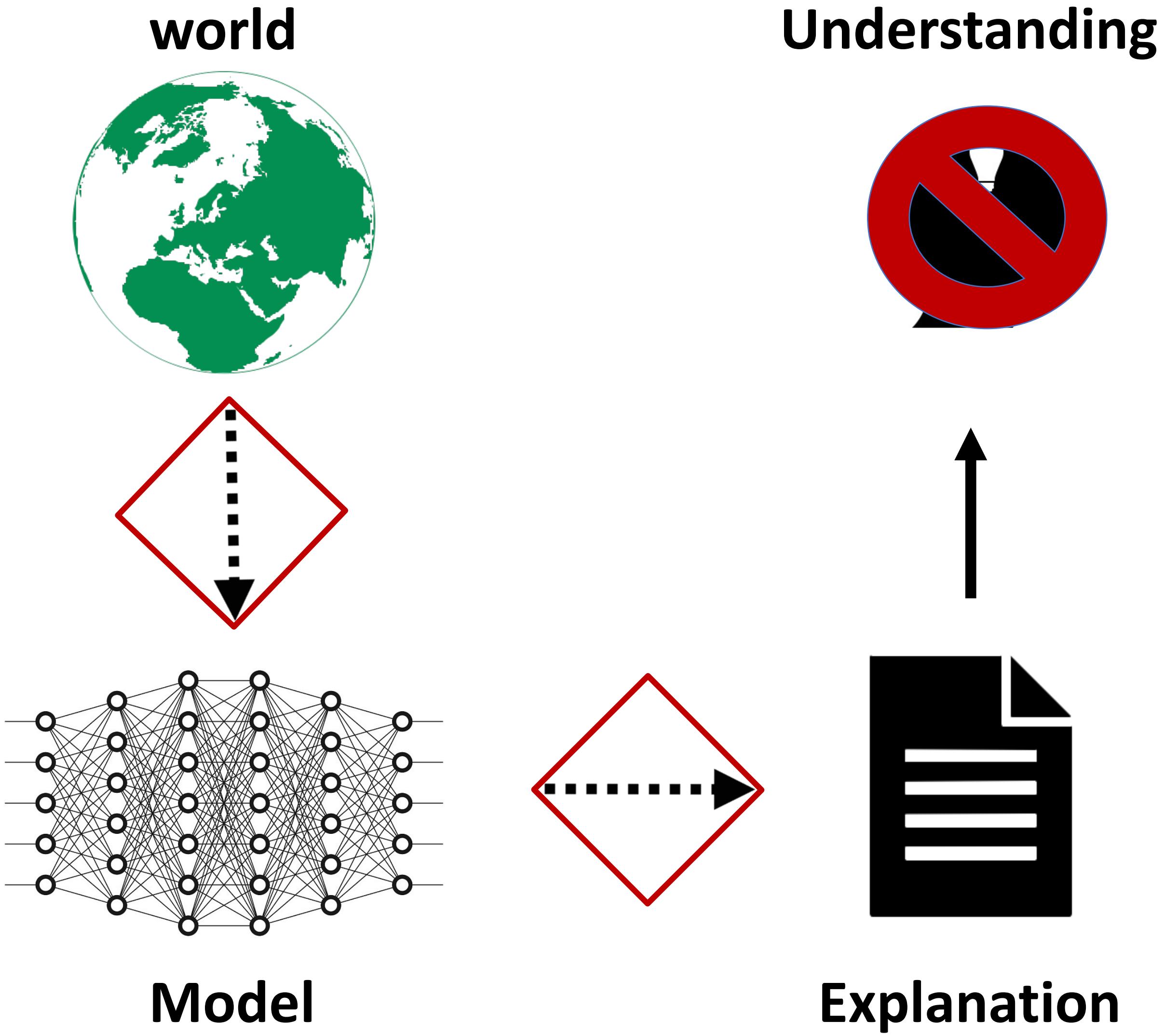
Model

Explanation



How-possibly?

How-actually?

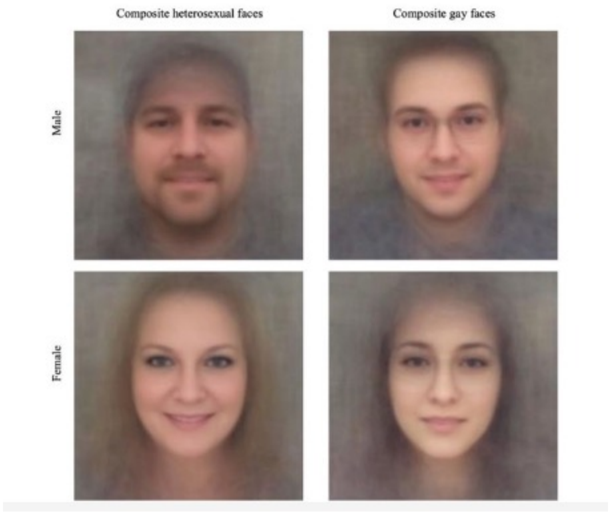


world

Understanding

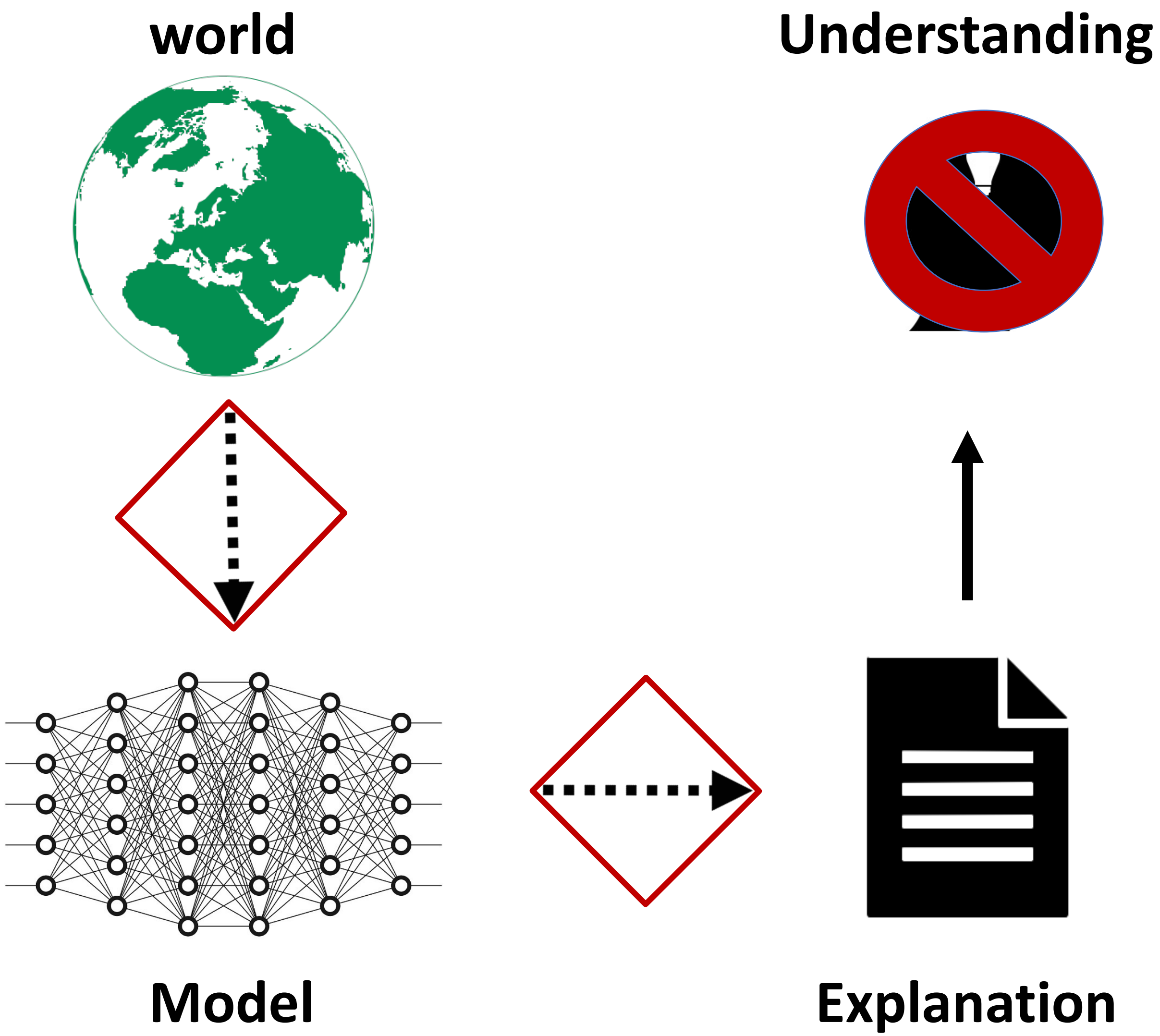
Model

Explanation



How-possibly? ✓

How-actually?

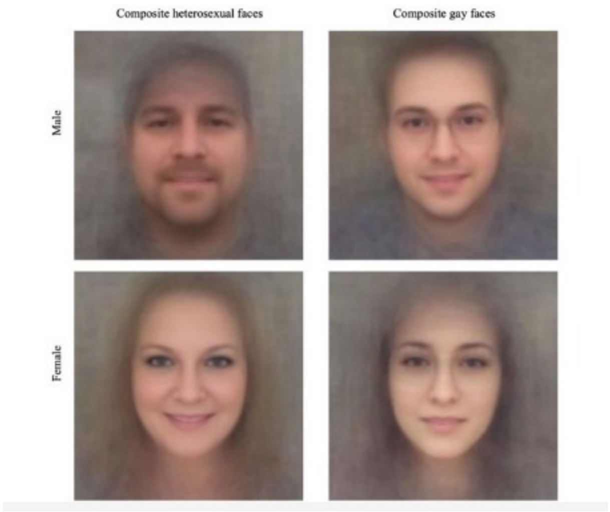


world

Understanding

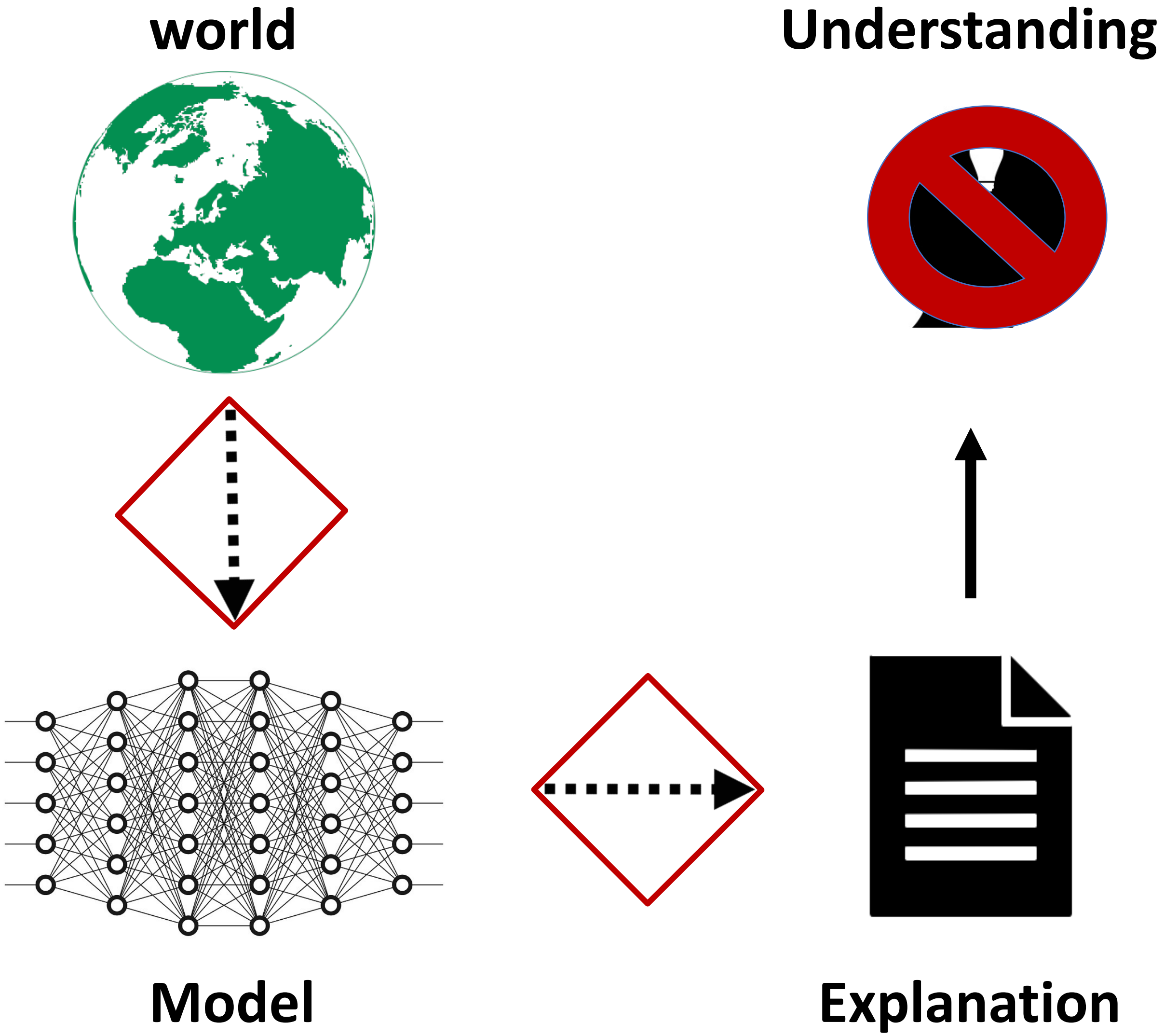
Model

Explanation



How-possibly? ✓

~~How-actually?~~

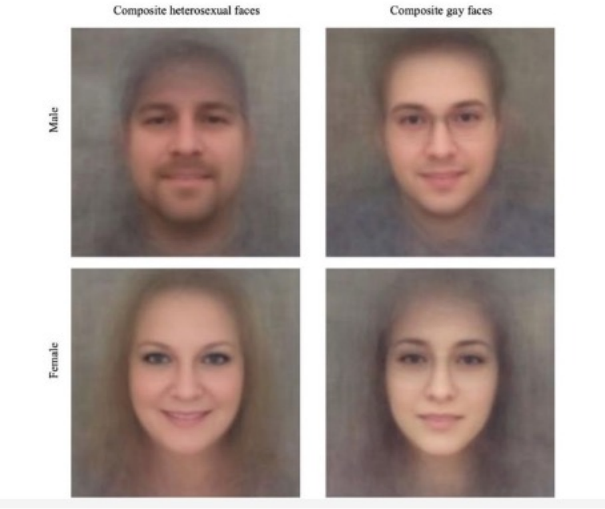


world

Understanding

Model

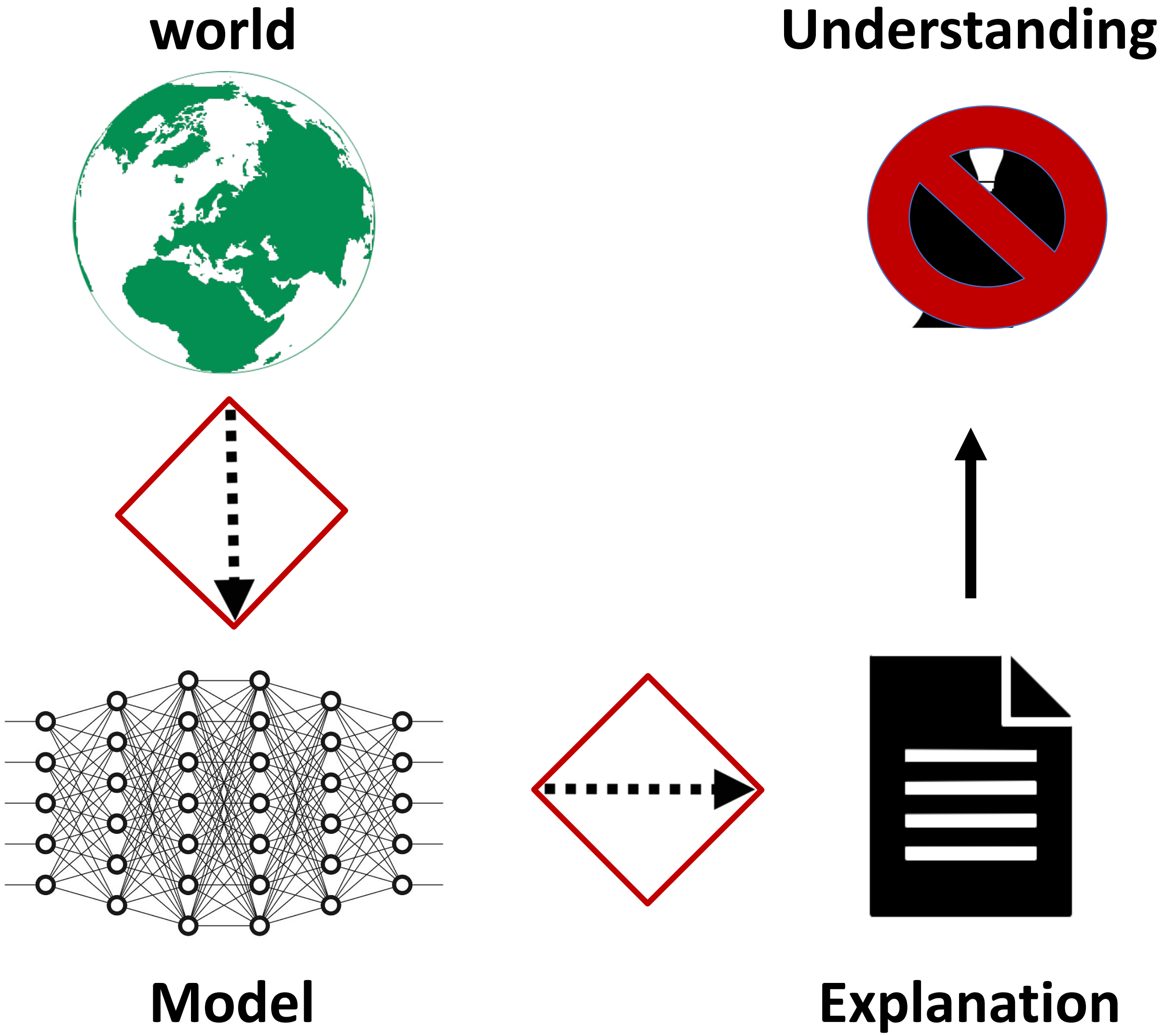
Explanation



The idealized assumptions underlying the model—

e.g. that sexual orientation is binary and static, that those who are openly gay on social media are representative of the whole gay population, and ignoring gender and racial variance—

distort important difference makers in real-world populations.

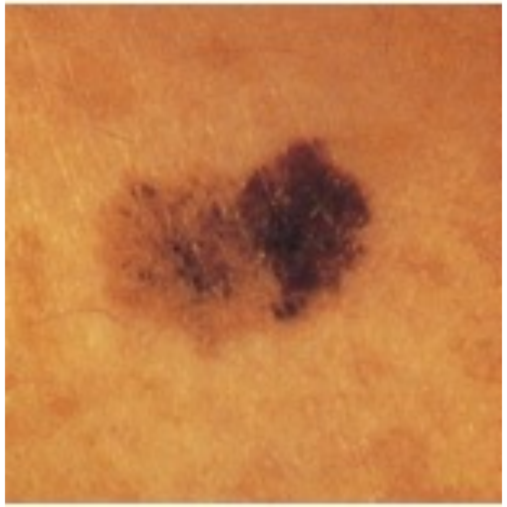


world

Understanding

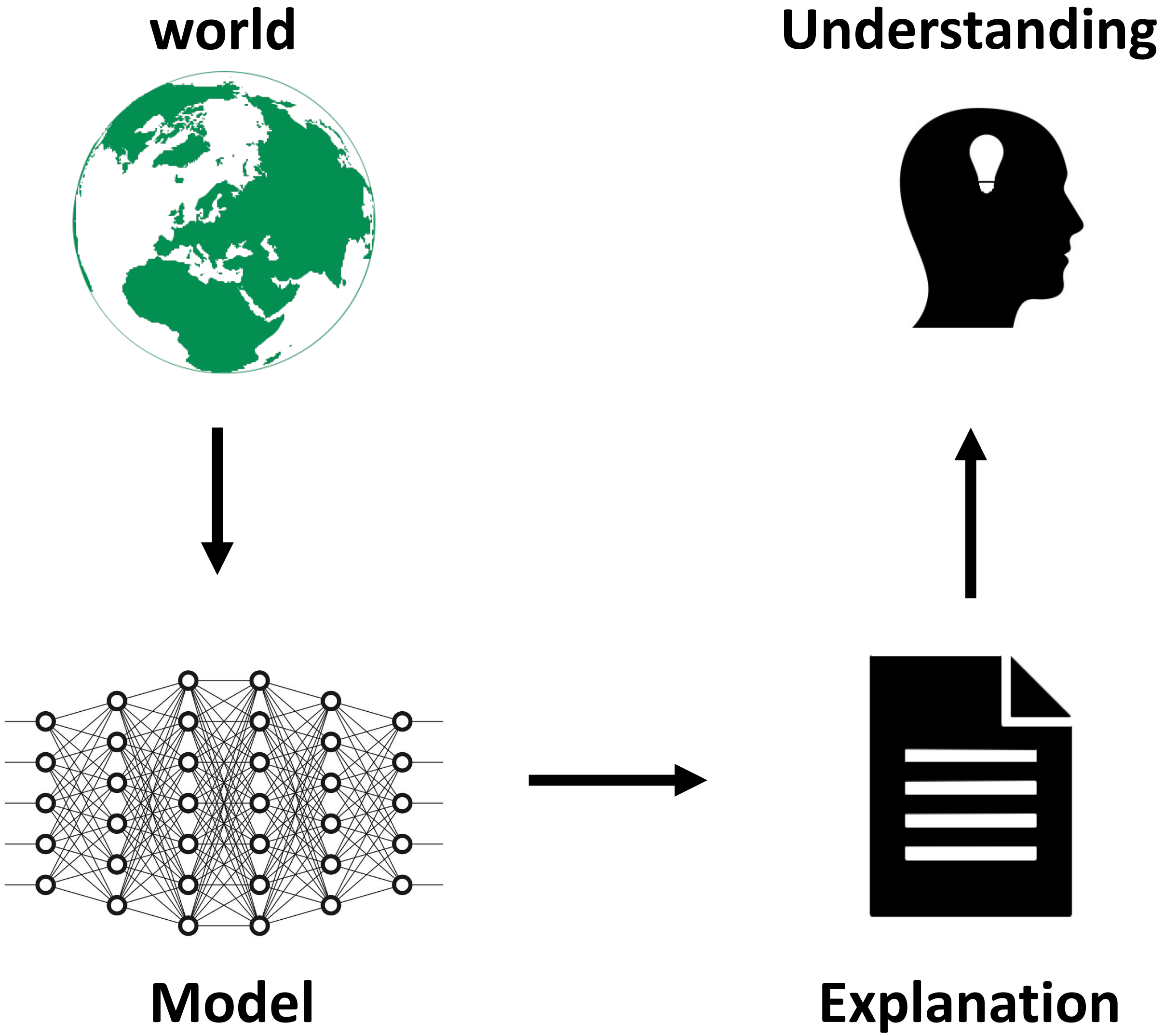
Model

Explanation



How-possibly?

How-actually?

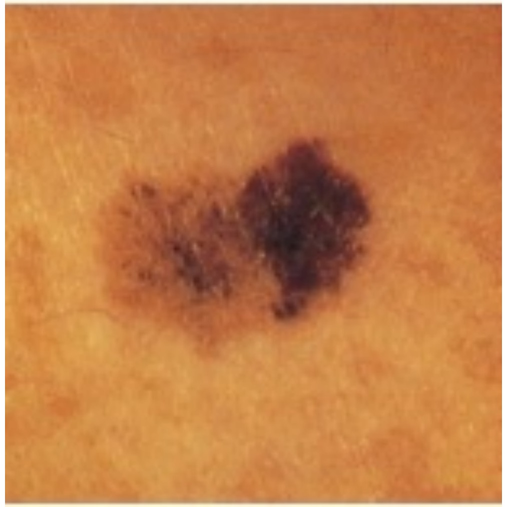


world

Understanding

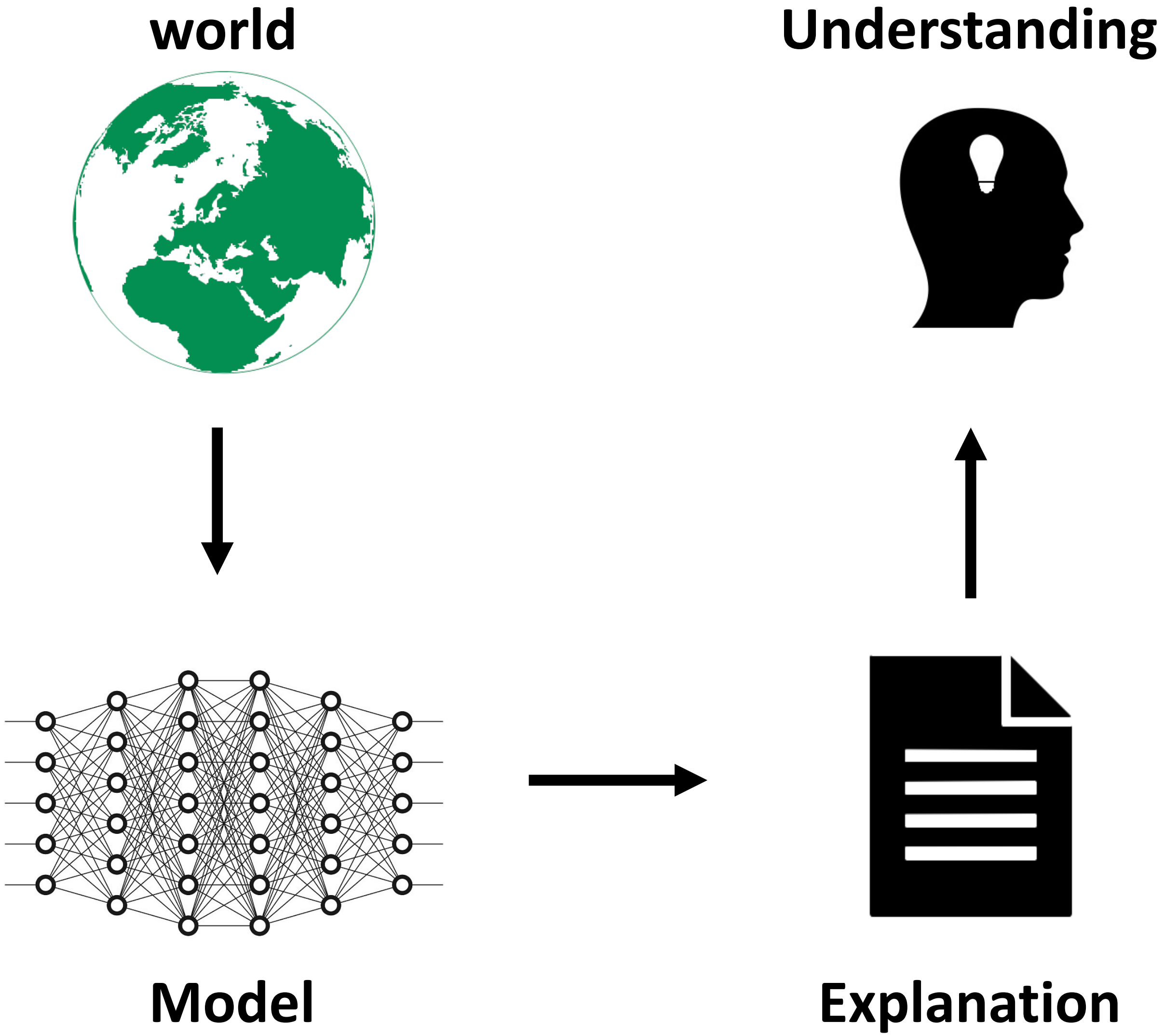
Model

Explanation



How-possibly? ✓

How-actually?

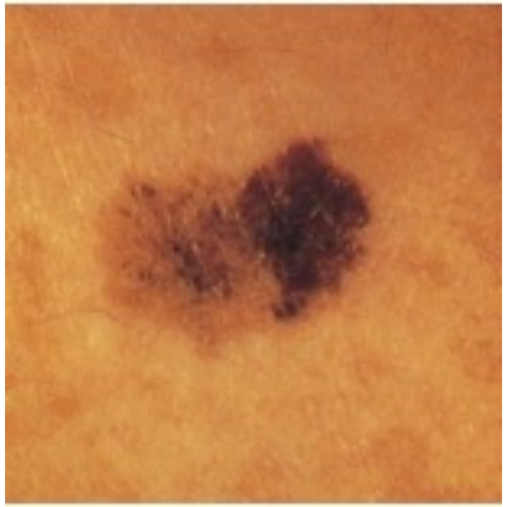


world

Model

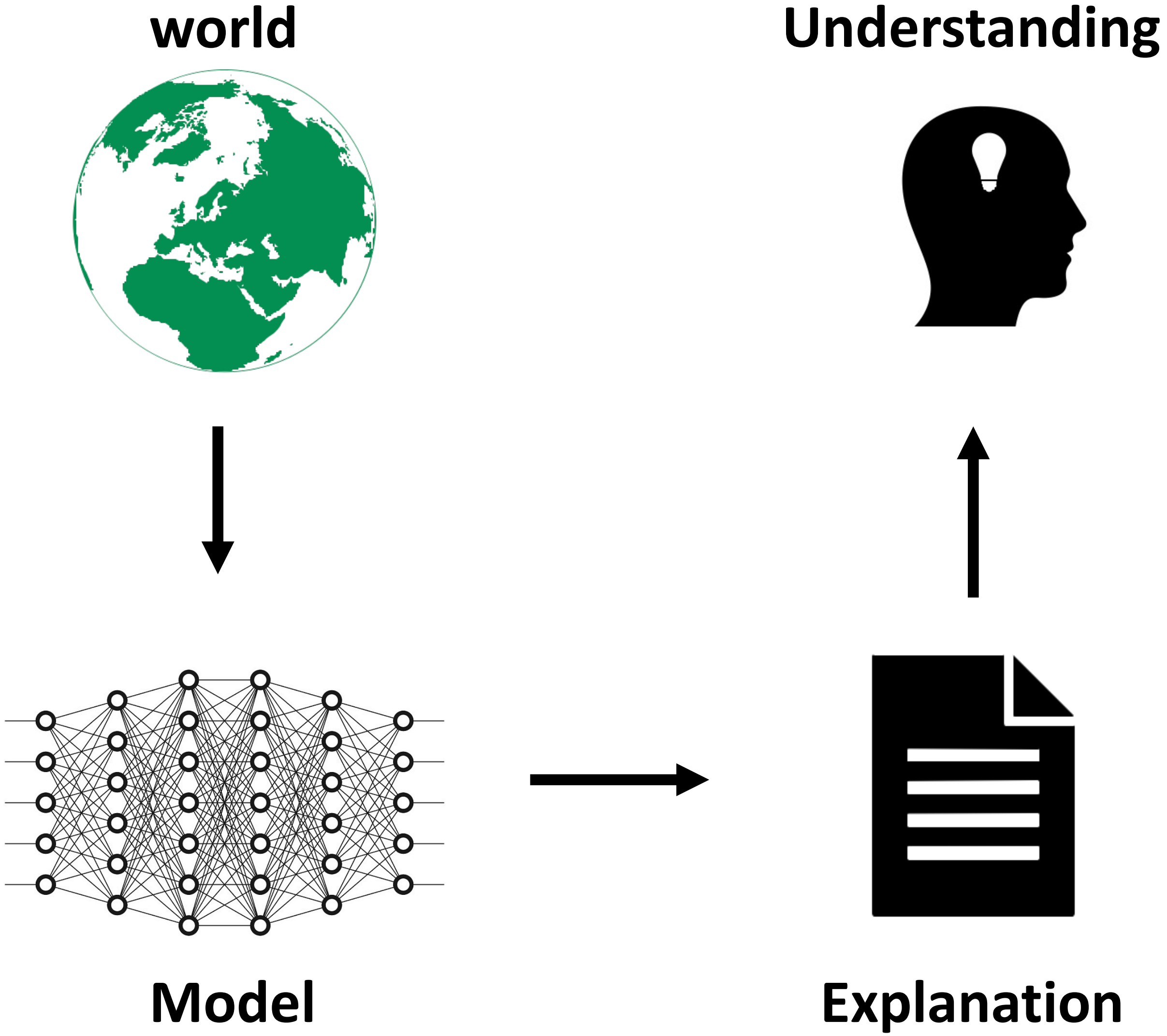
Explanation

Understanding



How-possibly? ✓

How-actually? ✓

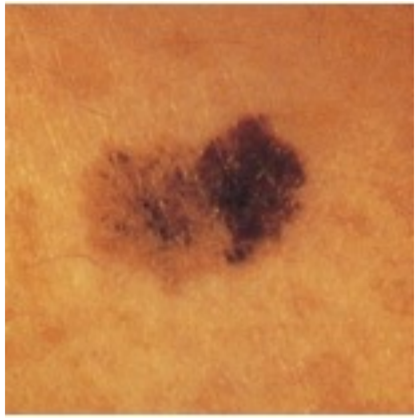


world

Understanding

Model

Explanation

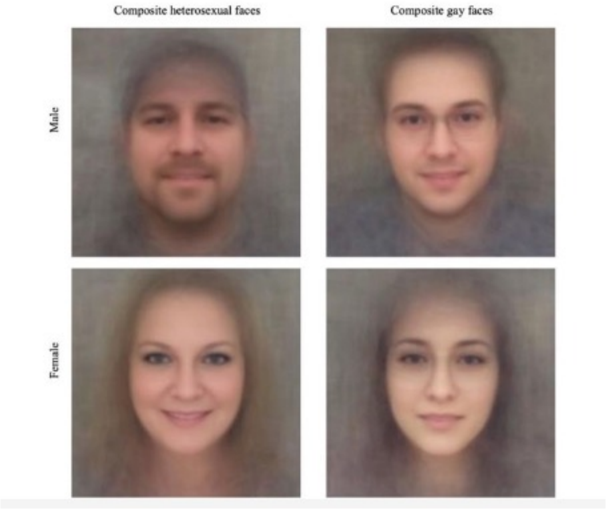


(Esteva, A., et. al, 2017)

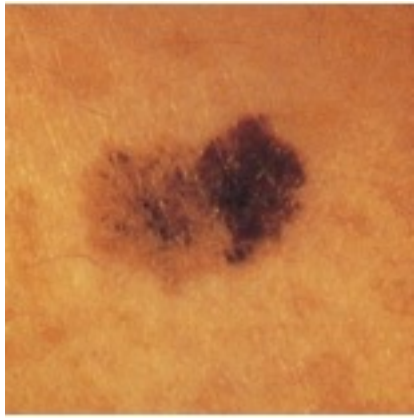
Black-box / Opacity

=

Any difference in the level of understanding we gain from these models must be due to something other than opacity.



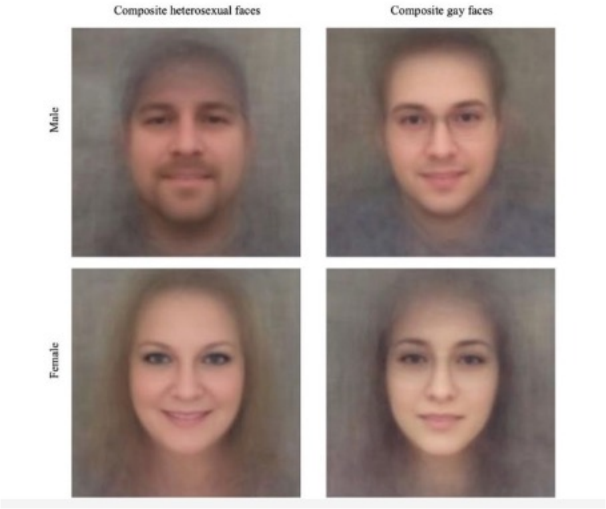
(Wang, Y. and Kosinski, M., 2018)



(Esteva, A., et. al, 2017)

≠

The level of link uncertainty between the phenomenon and the model differs.



(Wang, Y. and Kosinski, M., 2018)

Reducing Link Uncertainty

- requires connecting data, model architectures, and counterfactual the model makes inferences to the target phenomena
  - Robustness analysis (e.g. over different data distributions)
  - traditional empirical research
  - Improve ground truth methods for data labeling

There is a special worry with DNN models.

The power of these models mean that we need to take special care to make sure the models do not have high levels of link uncertainty before we rely on their results.

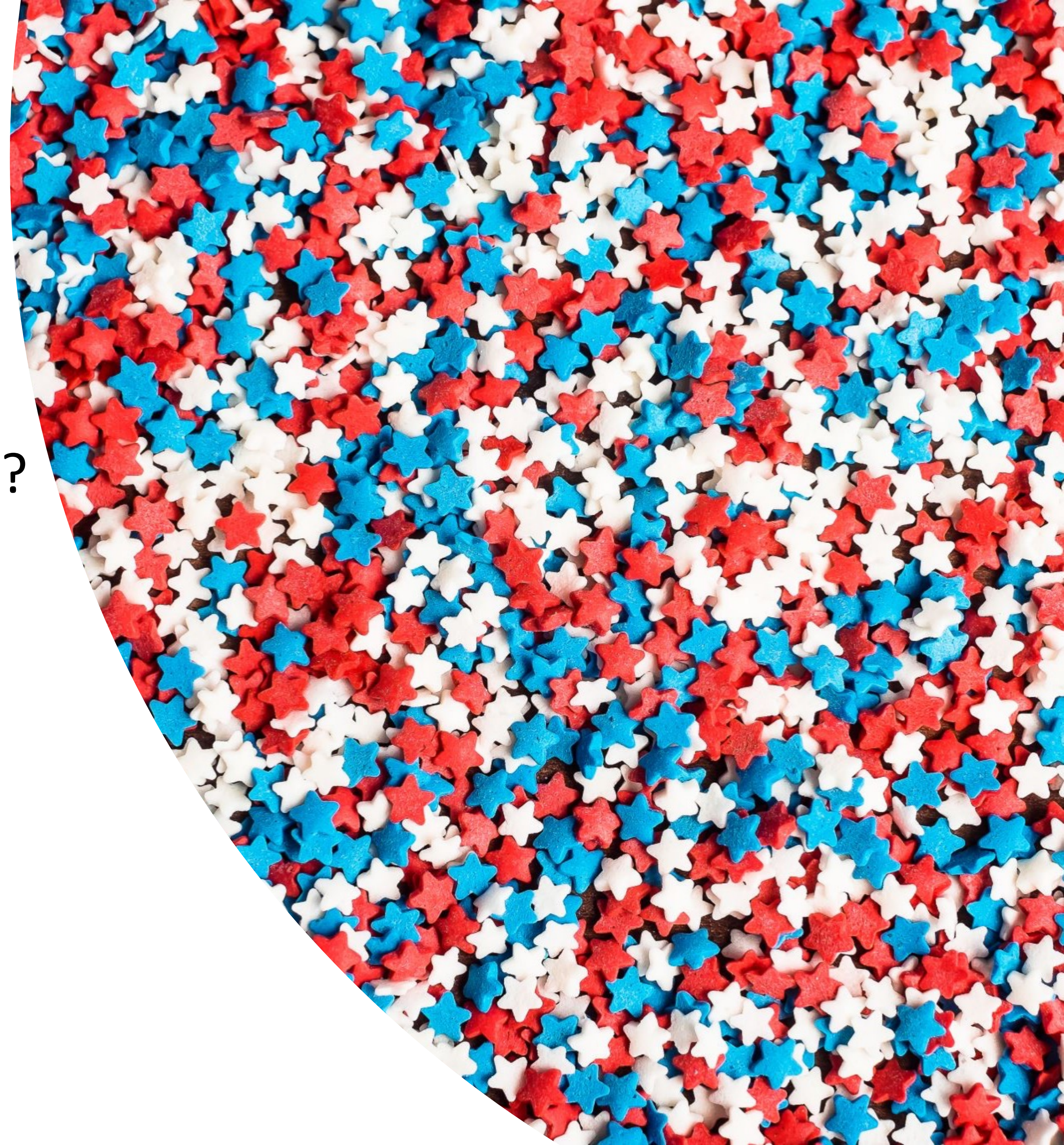
We need to be clear when a model is merely exploratory and not a new discovery.

ML wrap-up

- Difference between explaining the model and explaining phenomena with models.
- Model opacity is not in-principle a problem for explaining phenomena.
- For explaining phenomena, the problem of model opacity is an external problem of link uncertainty

Model independence?

How could the notion of LU be helpful?



Model independence

Disanalogies

- ”single class categorization”
- Don’t know what you are looking for
- Searching ‘without an alternative’
- Opacity a worry?

Analogies

- similar threat of treating data in non-realistic ways
- Searching in large space of parameters for a significant pattern
- Needing to know when findings are worth investigating further

Model independence

When could there high link uncertainty that would prevent explanation and understanding?

Some ideas I heard yesterday….

Misalignment (or uncertainty of alignment) between ML architectures and data --- (**Kyle Cranmer**)

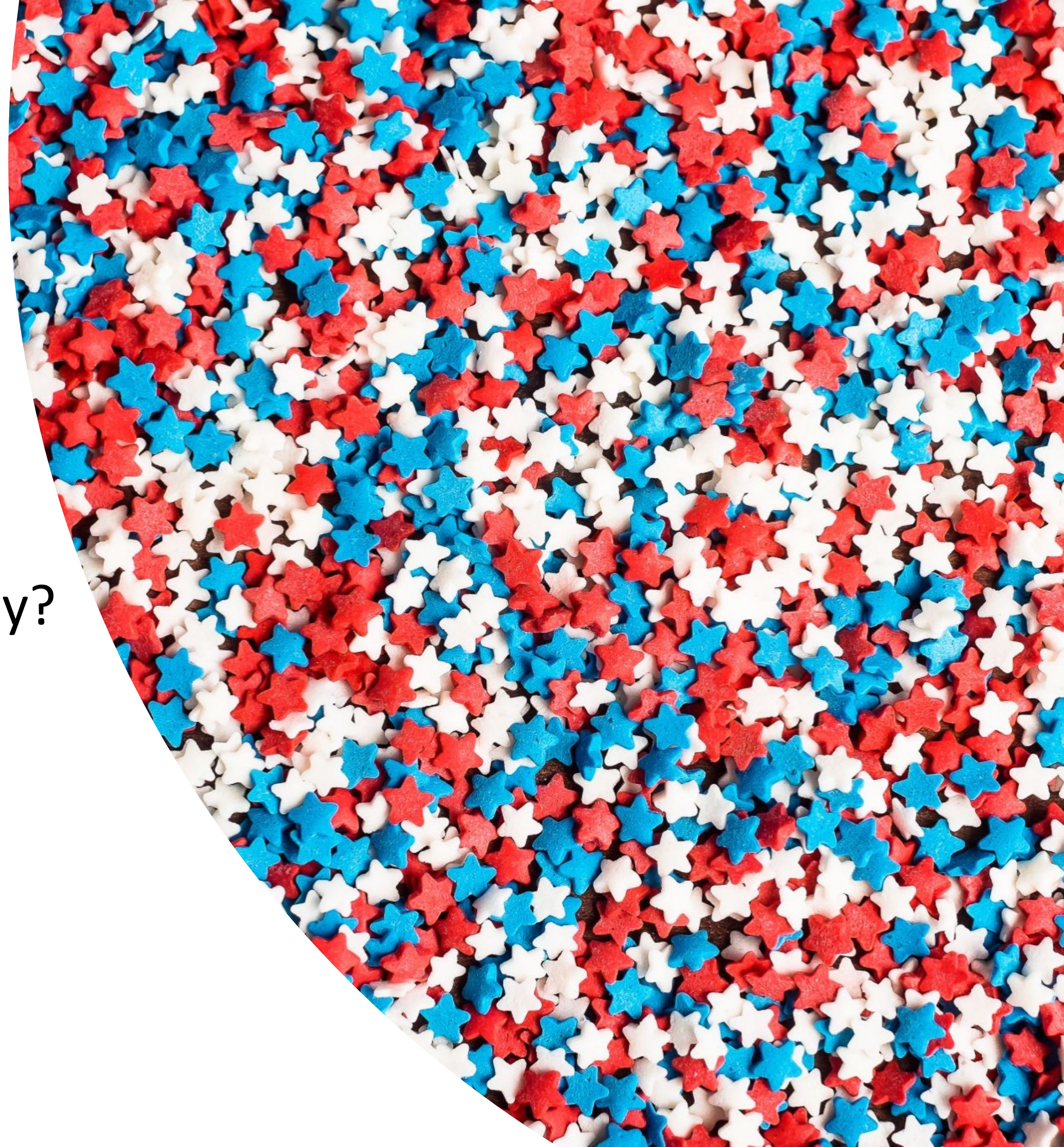
Uncertainty concerning inter-dependency between choice of operator bases --- (**Christophe Grojean**)

Bottom-up approaches to SMEFT *before* 'fit into global explanation theory'--- (**Martin King**)

Model independence?

What do you think?

Does this add more than EFT validity?



Link Uncertainty and ML

Emily Sullivan
Philosophy and Ethics
Eindhoven University of Technology
Eindhoven Artificial Intelligence Systems Institute

esdit ethics of socially disruptive technologies
NWO
TU/e EAISI
ϕAI/e