EXPLAINABILITY THOUGHT EXPERIMENT #1

Machine or human?
(What is the threshold number..?)

EXPLAINABILITY THOUGHT EXPERIMENT #2

1, 2, 3, 4?

Why for *developer?*
Why for *user?*

Is anyone else involved?

Virginia Dignum
@vdignum

Follow

Dilemma from my talk to top executives in finance:

Your organisation is considering to choose an AI system to support decision-making for credit applications.

- A well-known vendor, used by many other banks, can provide a highly accurate algorithm (90% accuracy), but this algorithm is not able to provide an explanation for the decisions.
- Your own AI team can develop an algorithm that provides reliable explanations, but can only guarantee 75% accuracy.

Your choice:

1. Yes, great track record and used by competitors
2. Yes, algorithm is accurate and most clients will not ask
3. No, our clients have a right to explanation
4. No, don't want to be accountable for any unlikely mistakes of the algorithm

EXPLAINABILITY THOUGHT EXPERIMENTS

What is the *value* of explainability (on developer and user sides)?

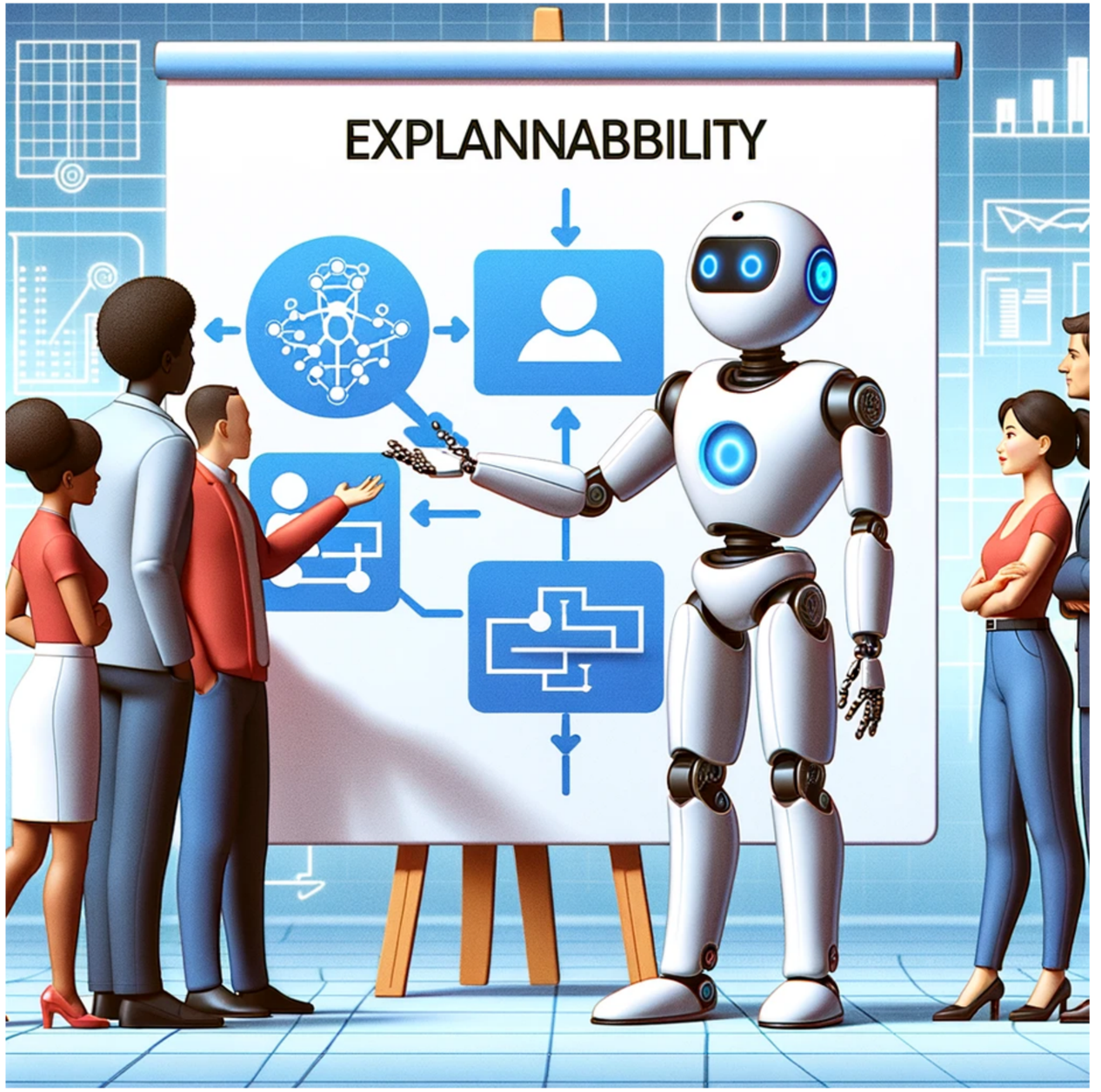
Why do we want it?
(*If the machine works at 100%, is the need for explainability 0?*)

How can it be described, or measured?

Definition explainability (in AI)

1. ChatGPT and the need to explain
2. Explainability theory
3. Explainability practice

Depict the concept of "explainability" in AI with an image. (ChatGPT 2023)

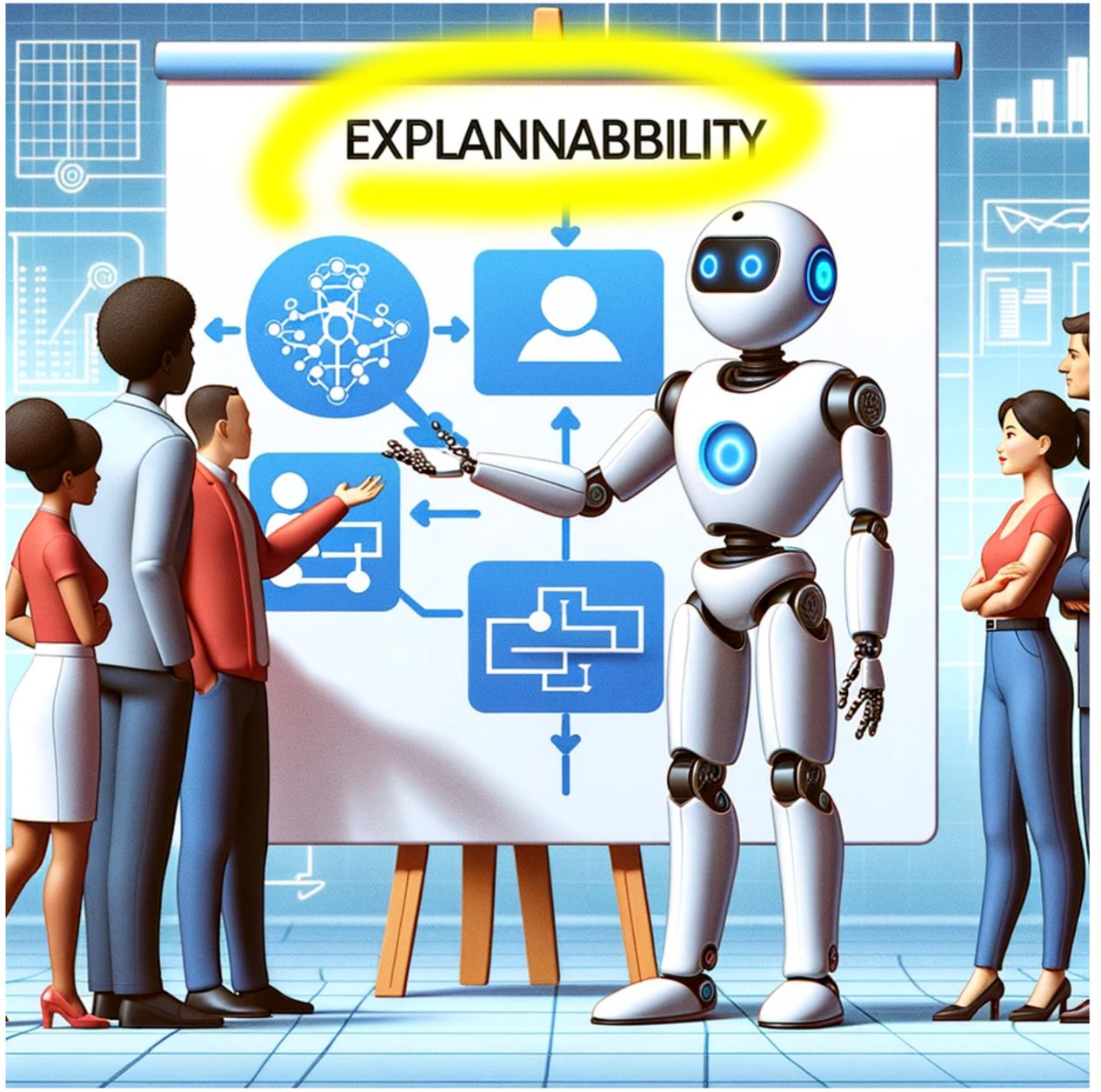


Definition explainability (in AI)

Need for explianability:

Reflected in the *kinds* of errors machines make

"How can the machine be so right and efficient, AND so wrong…?"



Depict the concept of "explainability" in AI with an image. (ChatGPT 2023)

Definition explainability (in AI)

Explainability Theory

Describe why/how output was produced
in the languages of the various stakeholders

AI Scoring of Loan Application

Stakeholders: Users, developers,
company lawyers…

Definition explainability (in AI)

Explainability in Practice

User: interpretability ("What do I need to change about myself?")
Developer: transparency ("What do I need to change in the AI?")
Lawyer: accountability ("Who can I credit/blame for AI performance?")

*Interpretability as "why" without "how" = which inputs create change in outputs...
*Interpretability presumably easier to achieve

Outline

AI doctor and loan officer

What is explainability?

Describe why the machine did what it did - for different audiences

Explainability and Accountability: 2 Cases

Forget explainability!

Impediments to explainability

Redress with blurry explainability (Compensating errors)

Explainability and Accountability

Explainability as ethical imperative (not just technical) because allows accountability

Accountability gap: Who do police blame…?

Explainability and Accountability

"Many hands problem," (industrial reality) *versus* "Accountability gap" (Digital reality)

Explainability and Accountability: Case 1

Does this justify the accountability gap?

Why, why not?

Dilemma

How many lives do we need to save to justify zero accountability for the ones who die...?

TESLA

A Tragic Loss

The Tesla Team •
June 30, 2016

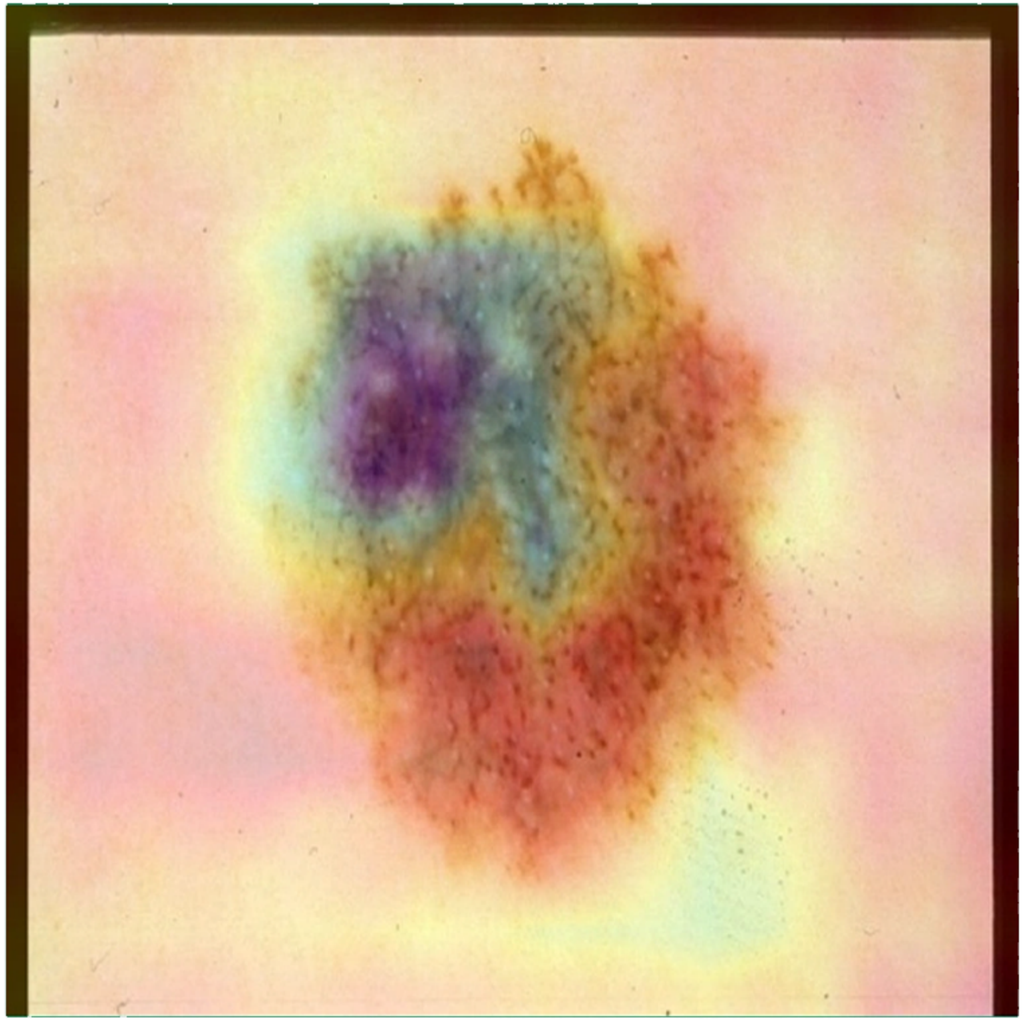
We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

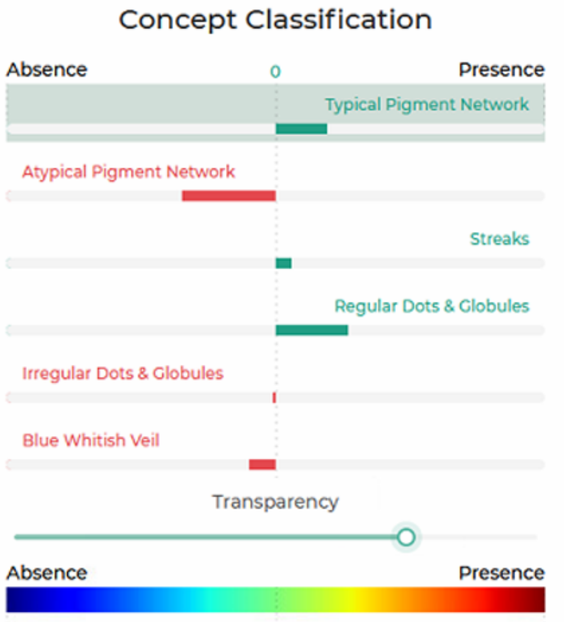
Explainability case 2: high accuracy / low explainability

Technical reality:
AI skin lesion analysis for cancer =
Blackbox + Accurate

Human problem:
Doctors ignore accurate AI image analysis since their patient could be 1 in 100 mis-diagnosed with no explanation

Accountability gap obstructs life-saving AI use...





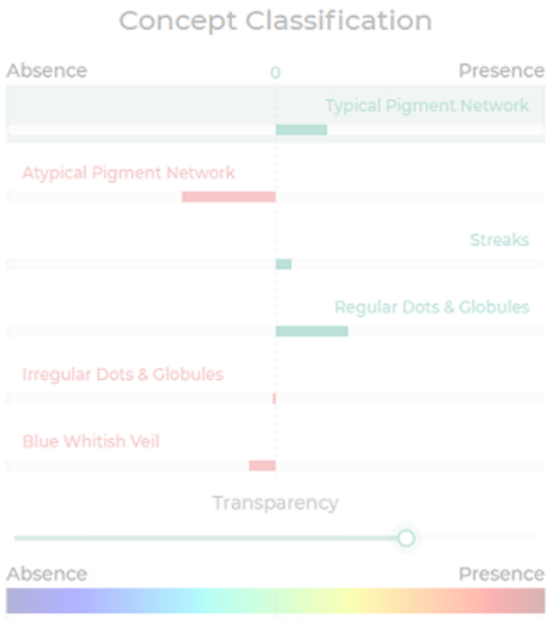
Explainability case 2: high accuracy / low explainability

Three Solutions:

1) Eliminate machine
   Famous accuracy/explainability tradeoff

2) Eliminate human
   Require doctors to defer to machine
   (Precedent of seatbelts…)

3) Explainable wrapper
   AI German Research Center (DFKI), exAID product
   "Heatmap" of suspicion



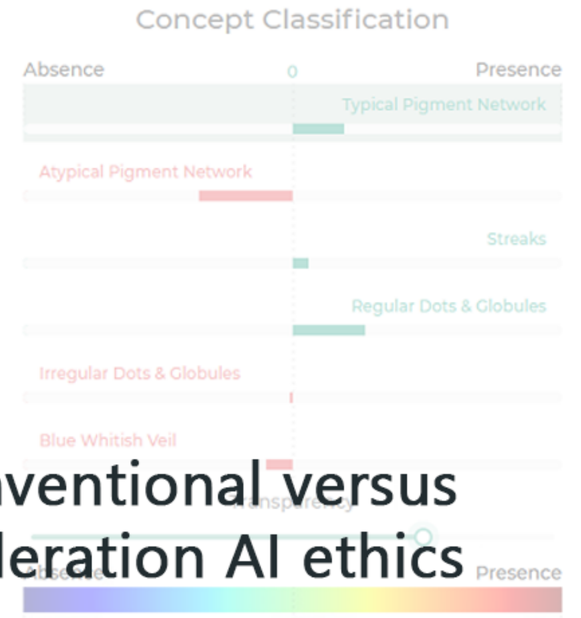
Explainability case 2: high accuracy / low explainability

Three Solutions:

1) Eliminate machine
   Famous accuracy/explainability tradeoff

2) Eliminate human
   Require doctors to defer to machine
   (Precedent of seatbelts…)

3) Explainable wrapper
   AI German Research Center (DFKI), exAID product
   "Heatmap" of suspicion



***Conventional versus acceleration AI ethics**

Outline

AI doctor and loan officer

What is explainability?
- Describe why the machine did what it did - for different audiences

Explainability and Accountability: 2 Cases
- Accountability gap, and skin lesion case

Forget explainability!

Impediments to explainability

Redress with blurry explainability (Compensating errors)

Forget explainability!

Question: Does it matter that "correct way"?

Vectors (CAV). Our results on independent evaluation sets clearly show that the classifier learns and encodes human understandable concepts in its latent representation. Additionally, TCAV scores (Testing with CAV) suggest that the neural network indeed makes use of disease-related concepts in the correct way when making predictions. We anticipate that this work can not only increase confidence of medical practitioners on CAD but also serve as a stepping stone for further development of CAV-based neural network interpretation methods.

Forget explainability!

2 Arguments against the value of explainability

Practical

Epistemological

Forget explainability! Practical argument

Carnegie Mellon University

CMU.EDU

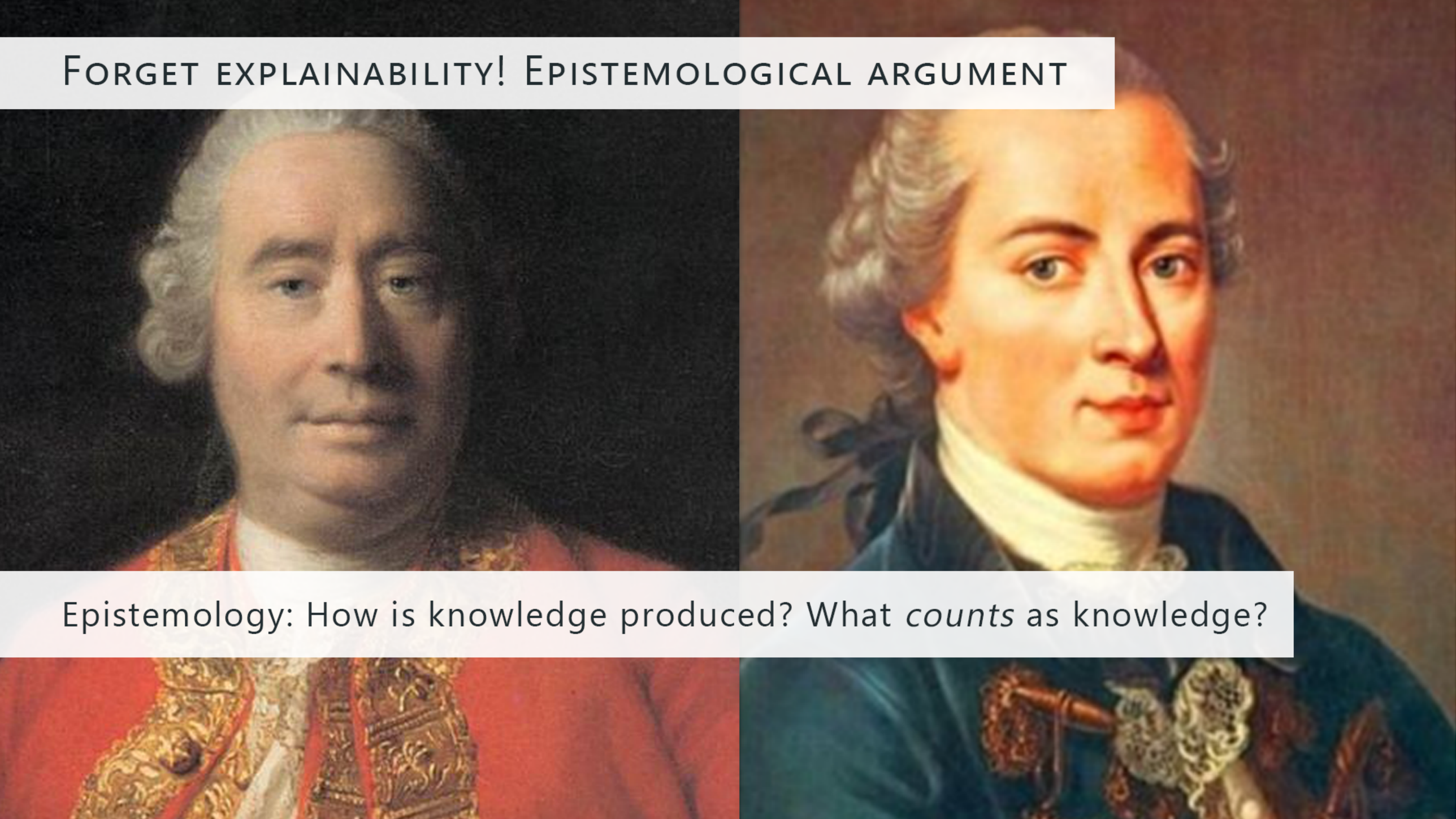
Manuela Veloso and microwave ovens...

...and aspirin

Forget explainability! Epistemological argument

Chris Anderson
and the end of theory...

Forget explainability! Epistemological argument



Epistemology: How is knowledge produced? What *counts* as knowledge?

Forget explainability! Epistemological argument

*Constant correspondence versus causality

Hume

- Knowledge production from constant correspondence & pattern recognition (Never see a cause)

versus **Kant**

- Knowledge produced from reasons, cause/effect as condition of experience (This is what it *means* to have knowledge, and only humans have it…)

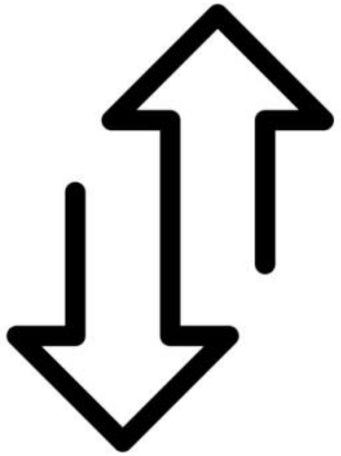
Forget explainability! Epistemological argument

Clouds and rain...

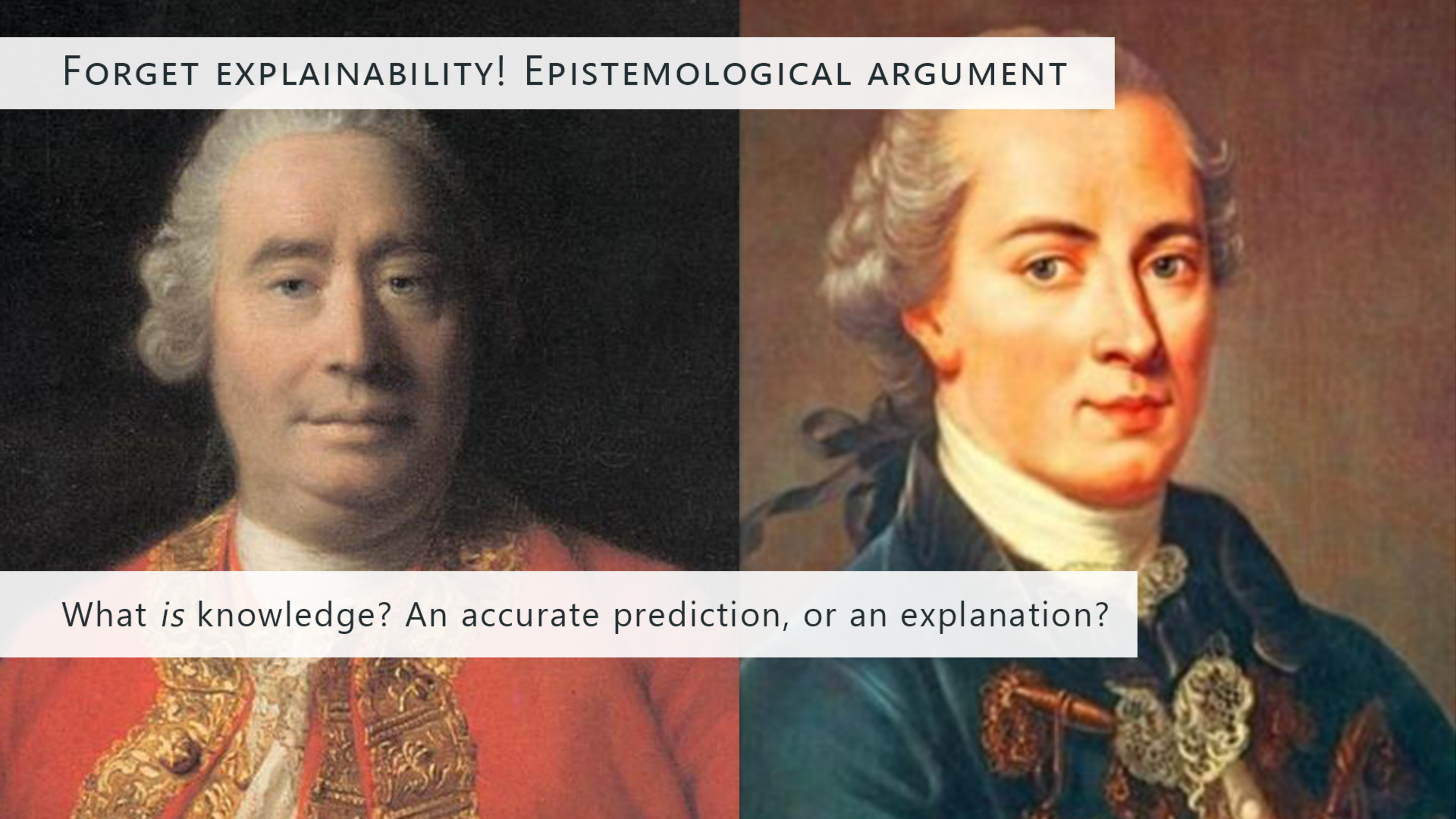
Forget explainability! Epistemological argument

Do humans reduce to thinking at machine level,
or do machines try to think up to human level?

(Hume: Humans as crude as machines
versus
Kant: Humans above machines by definition)

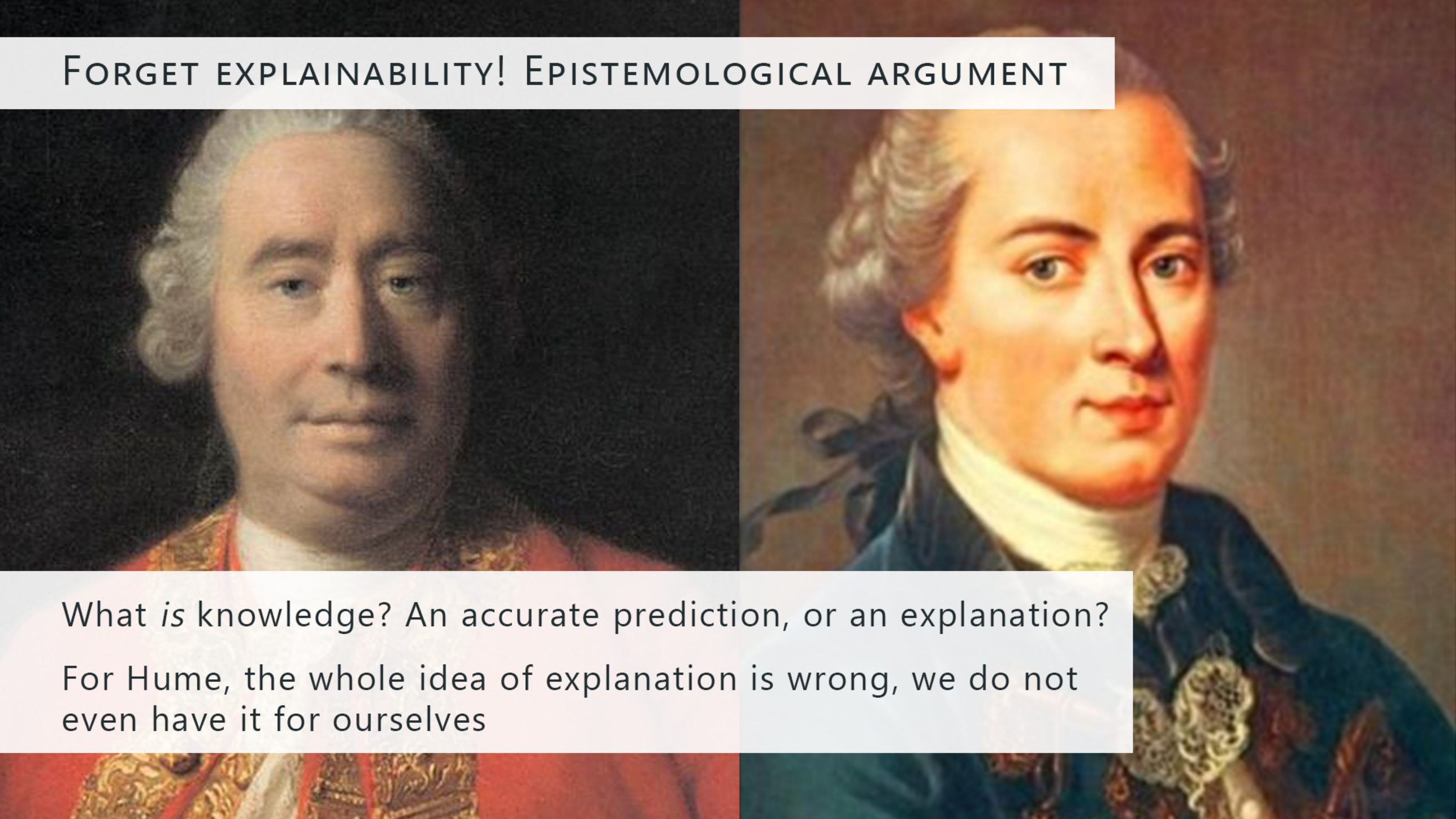


Forget explainability! Epistemological argument



What *is* knowledge? An accurate prediction, or an explanation?

Forget explainability! Epistemological argument



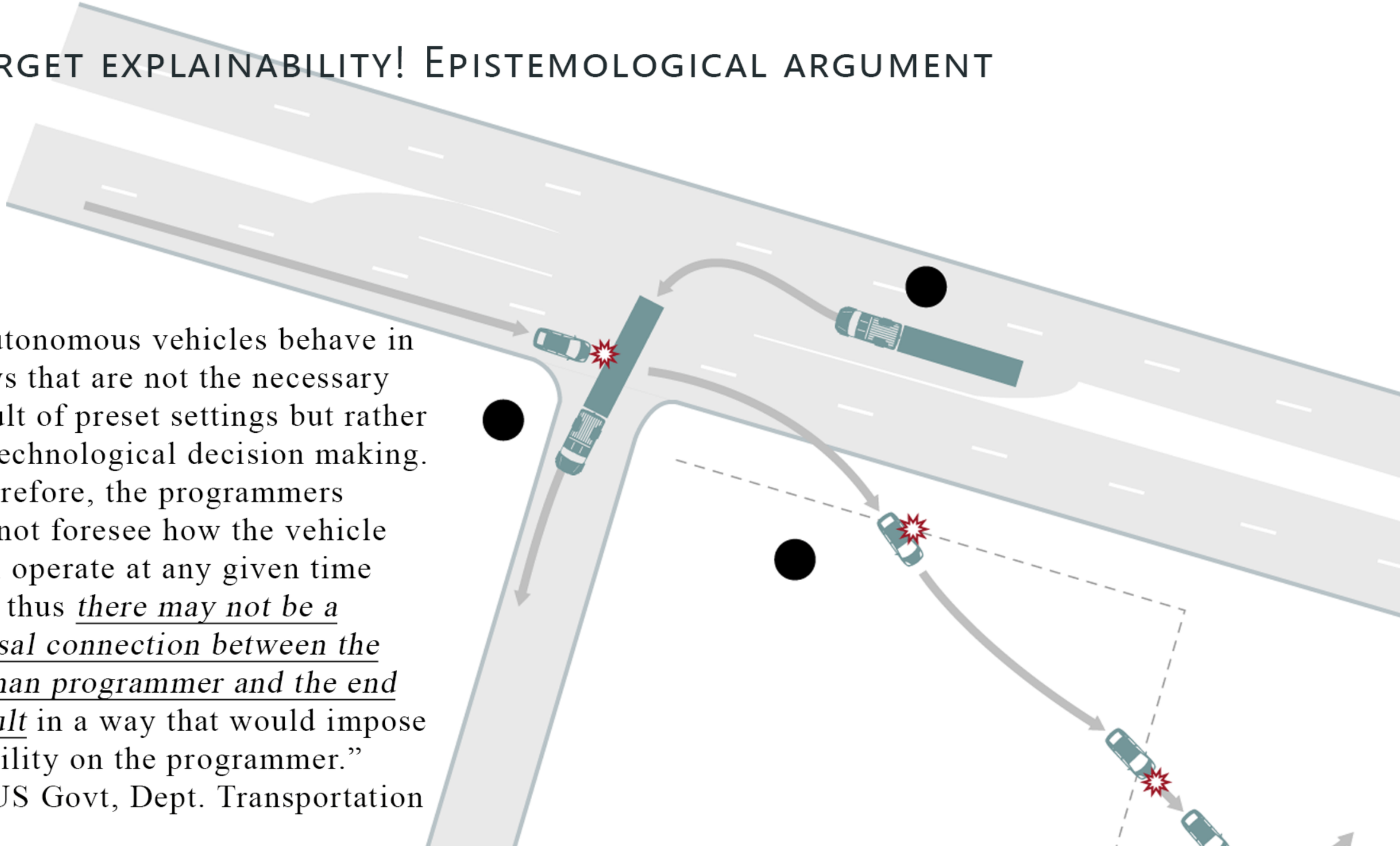
What *is* knowledge? An accurate prediction, or an explanation?

For Hume, the whole idea of explanation is wrong, we do not even have it for ourselves

Forget explainability! Epistemological argument

Chris Anderson
and the end of theory =
pure efficiency, pure
predictive knowledge

Forget explainability! Epistemological argument



"Autonomous vehicles behave in ways that are not the necessary result of preset settings but rather of technological decision making. Therefore, the programmers cannot foresee how the vehicle will operate at any given time and thus *there may not be a causal connection between the human programmer and the end result* in a way that would impose liability on the programmer."

-US Govt, Dept. Transportation

Case: Forget explainability?

Carnegie Mellon University

CMU.EDU

One approach to question:

Is AI like a microwave oven?
A pharmaceutical drug?
Something else?

What is it *like?*...

Outline

AI doctor and loan officer

What is explainability?
- Describe why the machine did what it did - for different audiences

Explainability and Accountability: 2 Cases
- Accountability gap, and skin lesion case

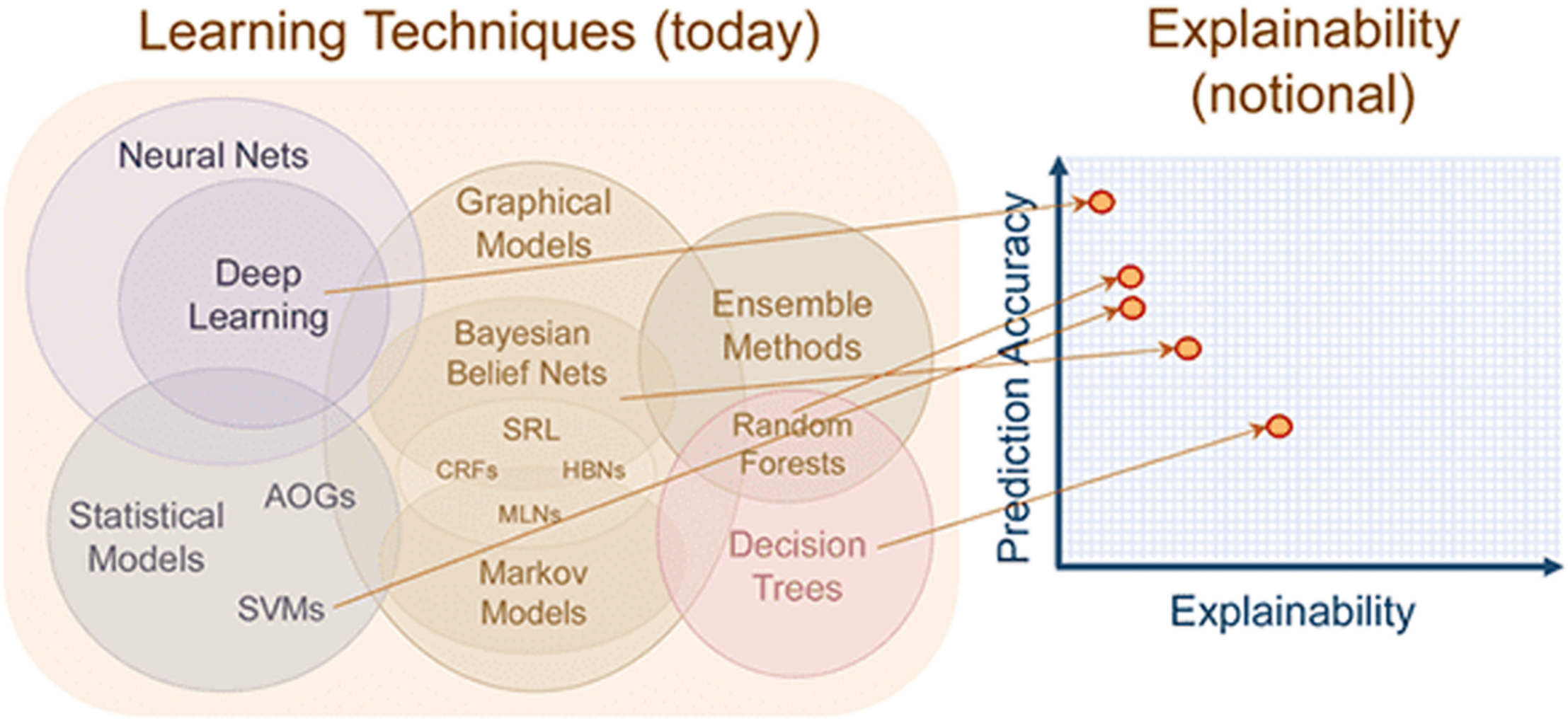
Forget explainability!
- Practical and epistemological arguments

Impediments to explainability

Redress with blurry explainability (Compensating errors)

Impediments to explainability: Practical (1 of 2)

Performance sacrifice: How much? In which contexts?



Impediments to explainability: Practical (2 of 2)

Black box nature (Practically impossible bc inherent in system/self-learning models)

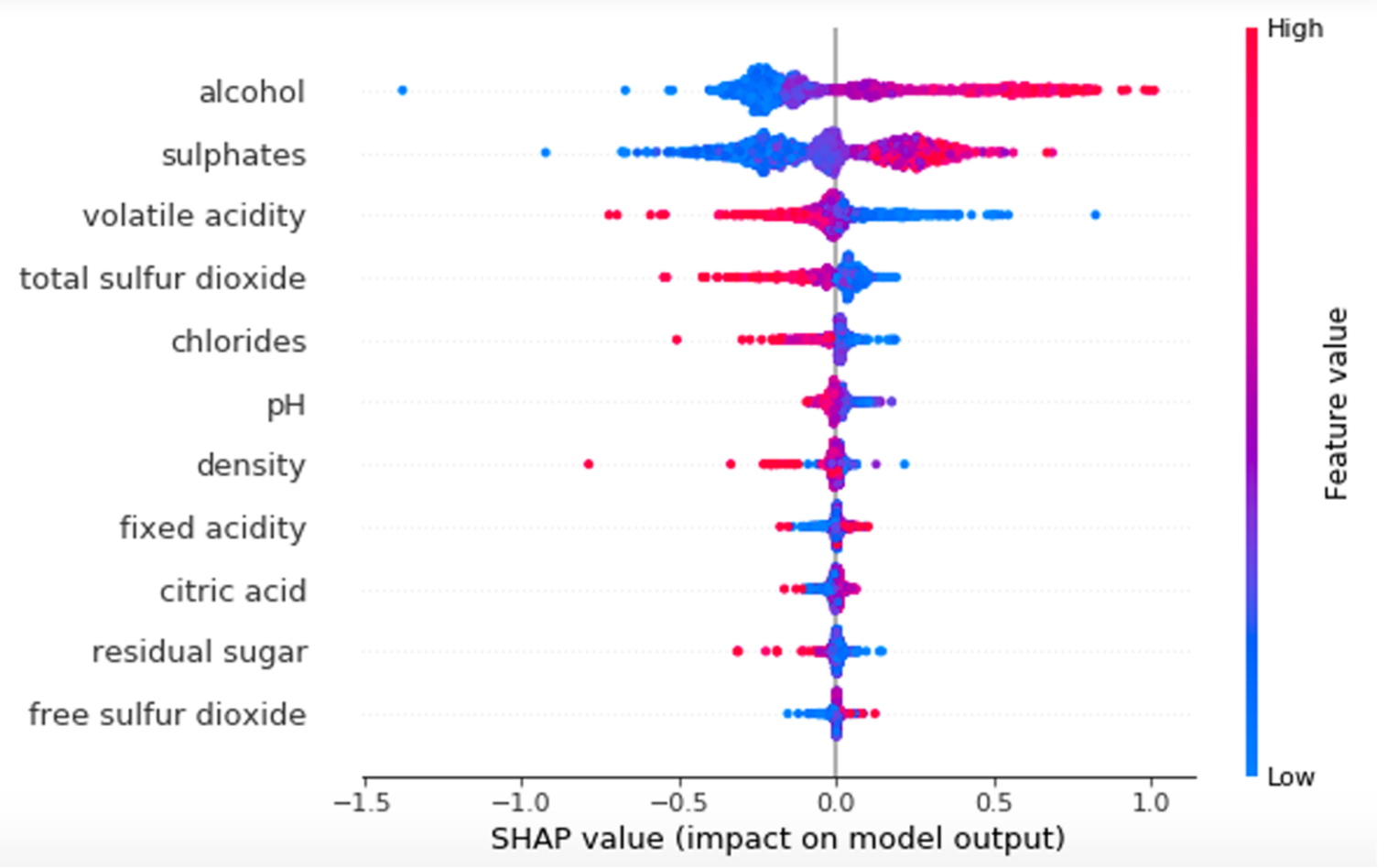
Intellectual property rights

Use/integration of third-party products

Impediments to explainability: Theoretical (Parts and wholes)

Shapley values.

But, is AI like cooking?



Outline

AI doctor and loan officer

What is explainability?
- Describe why the machine did what it did - for different audiences

Explainability and Accountability: 2 Cases
- Accountability gap, and skin lesion case

Forget explainability!
- Practical and epistemological arguments

Impediments to explainability
- Practical and theoretical

Redress with blurry explainability (Compensating errors)

Redress

(Applying accountability/responsibility across accountability gap)

4 Strategies

1. Doctrine of learned intermediary
2. Human creator
3. Legal personhood for AI
4. Redress-by-design (Form of explainability…)

Redress strategies

1. **Doctrine of learned intermediary**
   Human operator accountable, regardless of AI operation

Problem: Tesla develops autopiloting not so drivers can ride gripping wheel, but to nap on way home from work

ML advance = further extend from learned intermediary: Better AI means it is harder to pin blame on human overseer

TESLA

A Tragic Loss

The Tesla Team •
June 30, 2016

It is important to note that Tesla disables Autopilot by default and requires explicit acknowledgement that the system is new technology and still in a public beta phase before it can be enabled. When drivers activate Autopilot, the acknowledgment box explains, among other things, that Autopilot "is an assist feature that requires you to keep your hands on the steering wheel at all times," and that "you need to maintain control and responsibility for your vehicle" while using it. Additionally, every time that Autopilot is engaged, the car reminds the driver to "Always keep your hands on the wheel. Be prepared to take over at any time." The system also makes frequent checks to ensure that the driver's hands remain on the wheel and provides visual and audible alerts if hands-on is not detected. It then gradually slows down the car until hands-on is detected again.

Redress strategies

2. Human creator

Problems:

Multiple creators in each component
(Which human?)

"Organic nature" of AI versus
separable components of
traditional engineering
(Which human?)

Machine independent learning
AV feedback/reinforcement example

Real-Time Interaction: In some advanced systems, AI can adapt to the driving style of the human driver in real-time.

For example, if a vehicle is in a semi-autonomous mode, the AI might learn from the human's driving patterns and preferences, adjusting its own behavior accordingly.

-ChatGPT, no source..

"I don't have a specific source for this statement as it's based on general knowledge up to my last training data in April 2023. The field of AI in autonomous vehicles is rapidly evolving..."

Redress strategies

3. Legal personhood for AI

AI as a corporation = the “business” is responsible

Advantage: Well established by industrial economy/allows innovation

Problem: No human accountability

Redress strategies

4. Redress-by-design (Form of explainability…)
Formulating AIs so that outputs can be contested

When a loan is denied, redress-by-design helps applicants understand specific data leading to rejection, and enables opportunity to object effectively.

Problem: Loss of accuracy (We understand wrong answers better, but there are more of them…)

Redress strategies

4. Redress-by-design example #2

As an example of redress-by-design implementation, consider the recent EU Directive on copyright: an AI system will decide if a content is purportedly legitimate or is in violation of someone's copyright.

Although overall the number of false positives may be small, the damage caused to these persons' right to freedom of speech, for those persons will be total.

The details of the appeal procedure in the Directive are very scarce.... We are therefore not going to have an effective way of timely and freely correct wrong decisions and, even more, we will not gain a precise view of the actual precision of the systems.

A Redress-by-design alternative, for example, could have been granting uploading users the possibility of immediately opposing a decision to block a publication of a content, forcing its publication upon provision by the user of an indirect verification mechanism usable by a court. (pretty much like phone number verification by a messaging app or wifi splash page).

https://blog.quintarelli.it/2019/04/we-need-redress-by-design-for-ai-systems/

Problem: Loss of the efficiency AI exists to create

Redress and EC *Ethics Guidelines*

Elements for adequate redress in AI systems:

- a way for users to contest decisions
- a way for users to make a claim for harm done
- available information about how to make the claim
- available information about the circumstances that may occasion such a claim

Outline

AI doctor and loan officer

What is explainability?
- Describe why the machine did what it did - for different audiences

Explainability and Accountability: 2 Cases
- Accountability gap, and skin lesion case

Forget explainability!
- Practical and epistemological arguments

Impediments to explainability
- Practical and theoretical

Redress with blurry explainability (Compensating errors): 4 strategies