Rationalization

The danger of AI systems that justify their decisions.

Alain Tapp
Université de Montréal, DIRO
Proud associate of MILA, RALI, CRM, IVADO

Causes, Justifications and Explanations

Why?

Why is the apple falling?

Aristotle: The nature of the apple

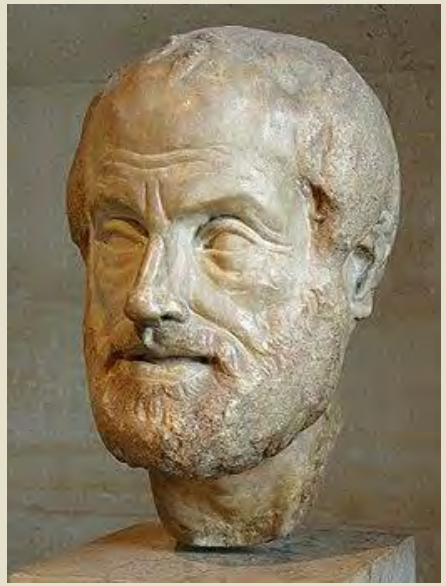
Newton: Gravitational force

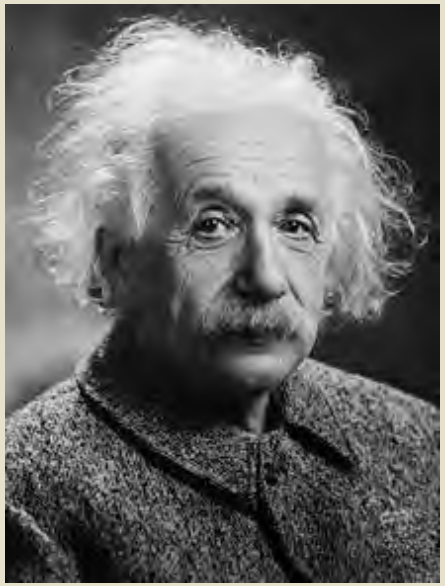
Einstein: Curvature of space

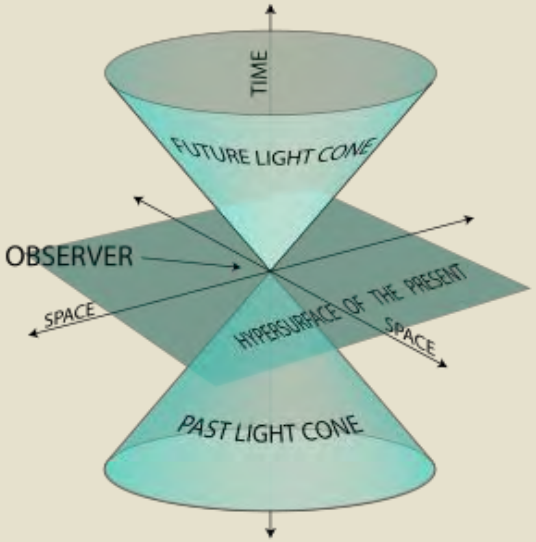
Better prediction but… why is it falling?

Butterfly effect

According to Aristotle,
God is the first cause, the unmoved mover.
Nowadays we believe that *some* events have no cause.







What is the cause of the fire?
Lightning?
Negligence?

Causes, Justifications and Explanations

Why ask why?

RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE

January 2015
Signed by dozens of scientists.

An Open Letter
RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents – systems that perceive and act in some environment. In this context, "intelligence" is related to statistical and economic notions of rationality – colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretical representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.

As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. **There is now a broad consensus that AI research is progressing steadily**, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008-09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. **We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do.** The attached research priorities document gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law and philosophy to computer security, formal methods and, of course, various branches of AI itself.

In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.

An Open Letter

RESEARCH PRIORITIES FOR ROBUST AND BENEFICIAL ARTIFICIAL INTELLIGENCE

There is now a broad consensus that AI research is progressing steadily…

We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial…

…our AI systems must do what we want them to do.

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

People are freaking out!

- Algorithms are incompetent and unreliable
- Algorithms are racist
- Algorithms are sexist
- Algorithms are self-fulfilling prophets
- Algorithms increase polarization
- Algorithms show us only more of the same
- Algorithms are spying on us, using us, and they stink….

Well… people decisions are often worst!

Please DO NOT site Newsweek!

Executive summary

- New techniques in machine learning are **opaque**.
- An entity affected by a decision is **entitled to an explanation**.
- Greater expectations if the decision is made by a machine.
- **Good definitions** are important.
- A justification can easily turn out to be **rationalization**.

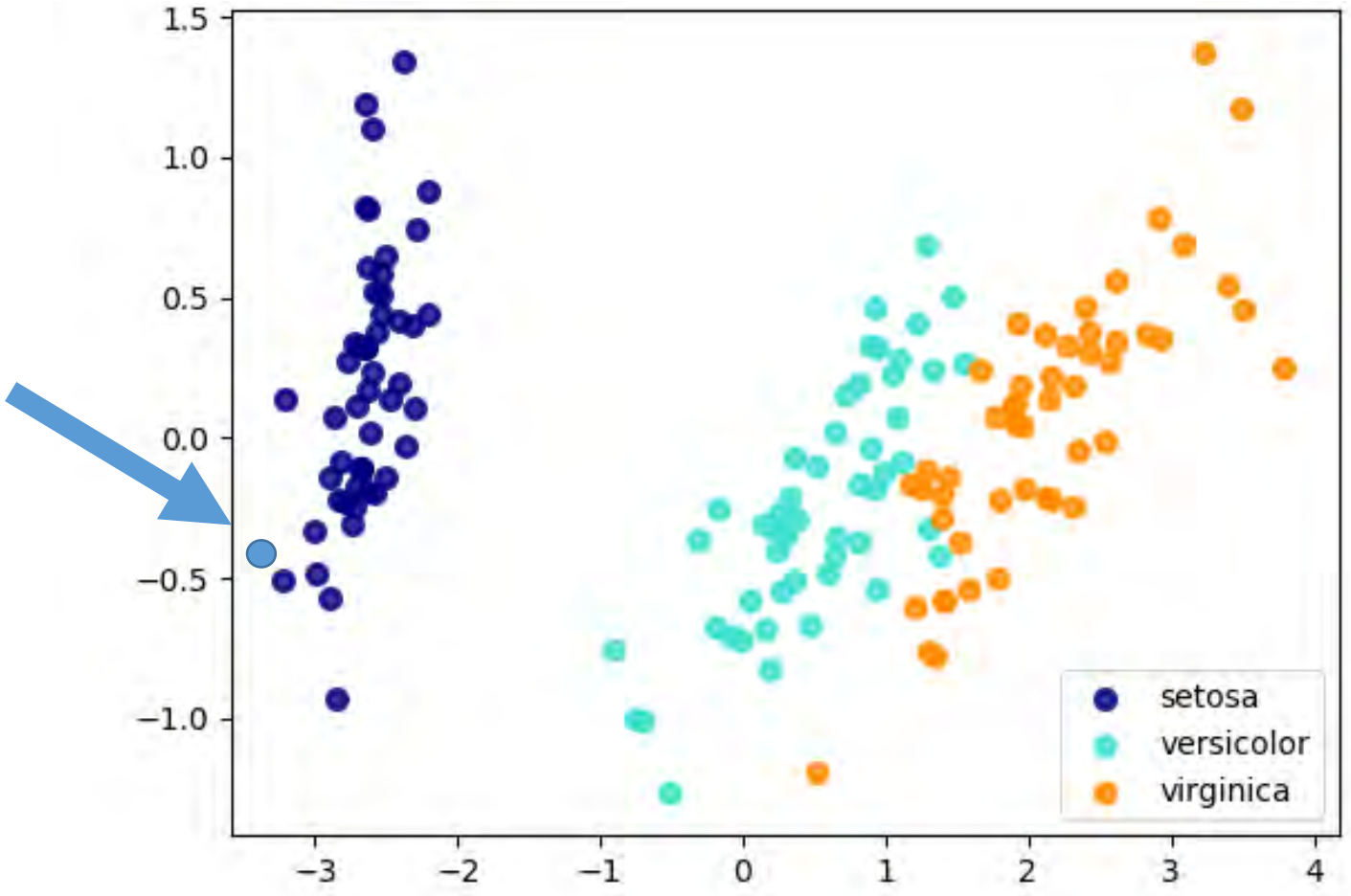
Causes, Justifications and Explanations

A good definition is most of the solution!

Justification

Some possible definitions:

- Proof (Seldom applies)
- Covering-law model (Epistemology)
- Visualization (Naïve)
- Explanation (Not a definition)
- Counterfactual argument (Very clever)
- Simplification (Subject to rationalization)



You have been rejected because you are in the dark blue cloud.
Look, it is quite clear!

Proofs

- Use a formal proof like one does in math
- Very transparent (if understood)
- Based on clear rules laws or conventions
- Logical and objective

You have been rejected because it is illegal to hire someone younger than 18.

Aquin to **covering law** justification principle in science.

Counterfactual approach

Assume a natural definition of distance or cost (C) that measure how difficult it is to transform a situation (data point) into another one.

Why is x not in $\mathbf{A}$?
Answer x', where x' minimize $C(x,x')$ with x in $\mathbf{A}$.

- Mortgage loan rejection: Reduce debt by 5K and have a collateral
- I propose *Blade Runner 2049:* Would not if you did not say you have liked *Incendi.*
- I propose *Across the universe*: Not possible, I propose that to everyone.

Which definition?

My conclusion is that talking about justification and explanation in general is a mistake and that one has to explicitly specify the real intent of the justification.

For example: conformity, legality, reliability, fairness, robustness, psychological, …

It might even be more reasonable to address those demand directly.

Causes, justifications and explanations

Rationalization!

Rationalization

In psychology

- Unconscious self-deception to avoid being self critical.

Here

- Intentional dishonest justification.

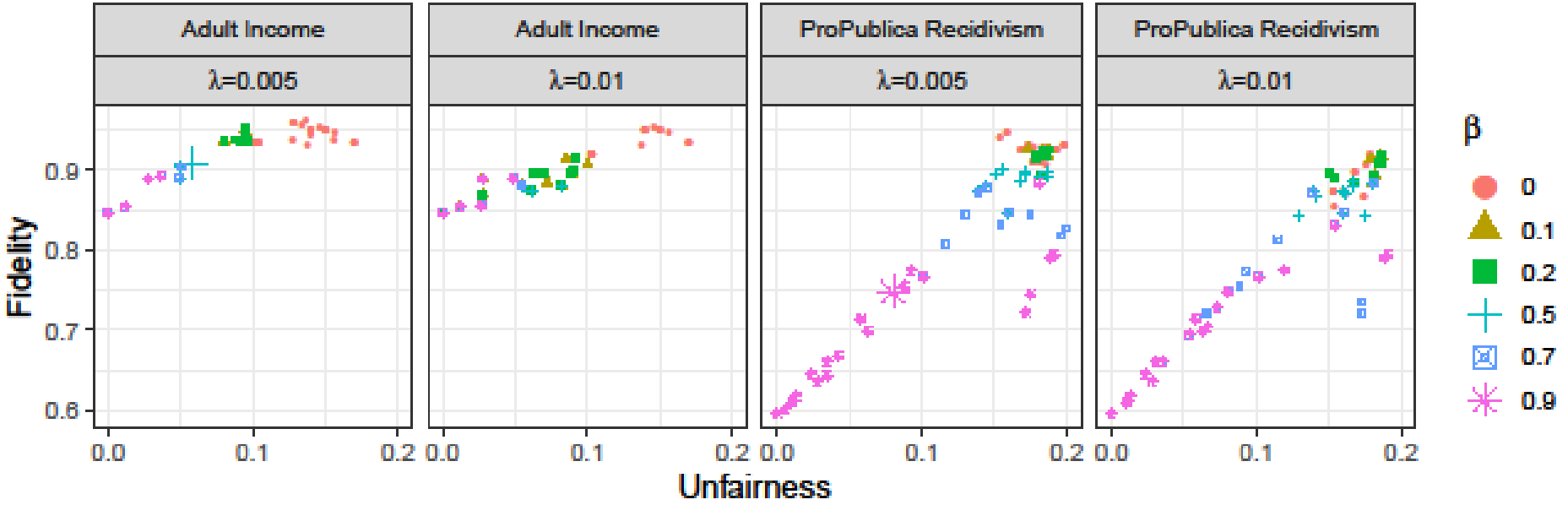
Example:

- Cherry picking
- Lawyer vs Judge

Method vs Decision

It is one thing to **justify a decision** to a user, for example answering the question: Why have I been rejected?

It is another thing to **justify a model** to a government so that we certify it is safe or fair.



Fairwashing

- Adult Income (protected attribute: gender)
- ProPublica Recidism (protected attribute: race)
- Black-box classiers: **random forests**
- Unfairness of the Black-box classiers: **0.13, 0.17**

Fairwashing in Machine Learning The risk of black-box explanation, Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sebastien Gambs, Satoshi Hara, Alain Tapp, ICML 2019.