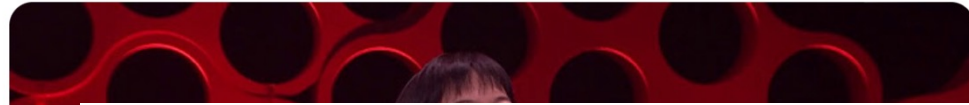


1. Why trust at all?
2. Examples of double opacity (model & action success)
3. Three (too) simple approaches
 - a. Why not just reliability?
 - b. Why not simply listen to experts?
 - c. Why not just evaluate the scientific quality?
4. The intricate simplicity of trust
5. Trustworthiness as a value - a value among other values?

1. Why Trust at all?



1. Why trust at all?



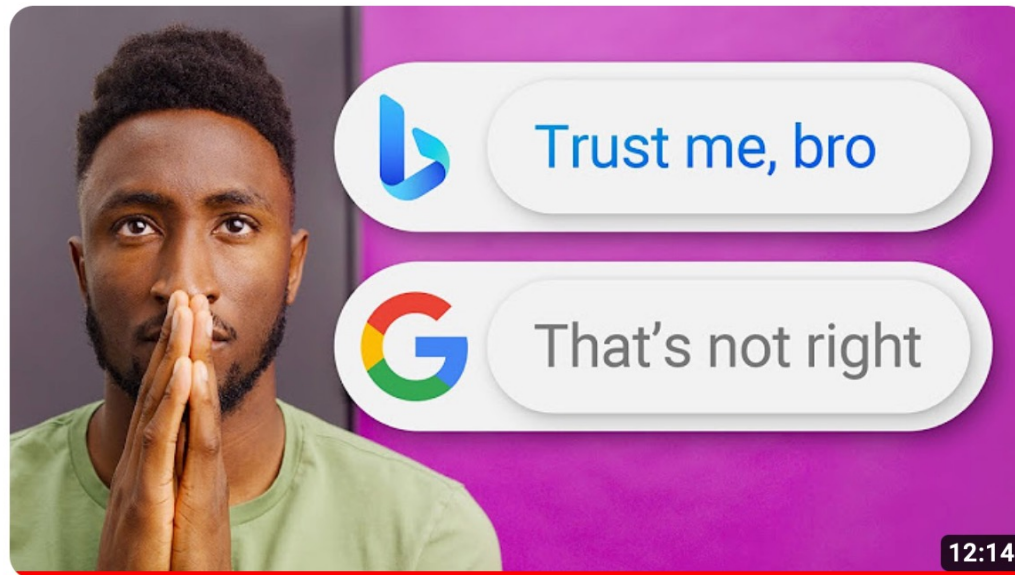
How can we design AI that we trust? | Fang Chen | TEDxSydney
2.6K views • 4 years ago



Reimagining Trust in AI | Jiaying Wu | TEDxColumbiaUniversity
2.8K views • 2 years ago



Artificial intelligence and trust | Marcel Isbert | TEDxTUDarmstadt
750 views • 5 months ago



The Biggest Problem with AI!

2.5M views • 3 months ago

...t vision and trust", Marcel Isbert s

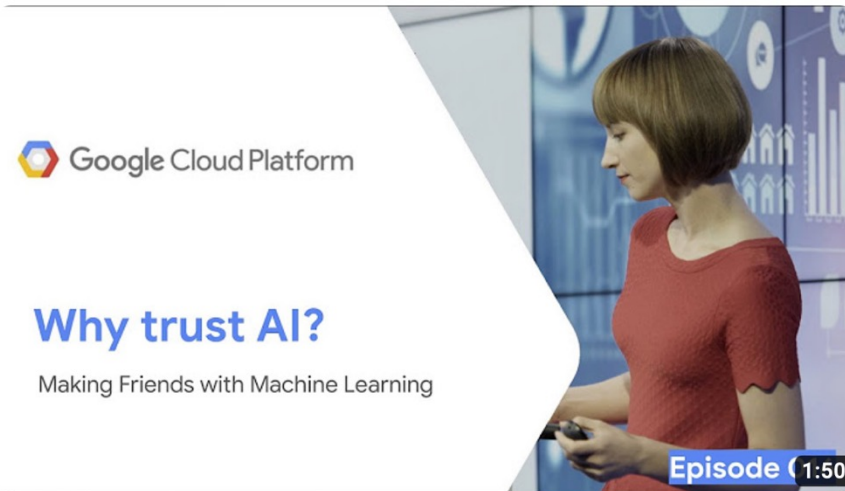
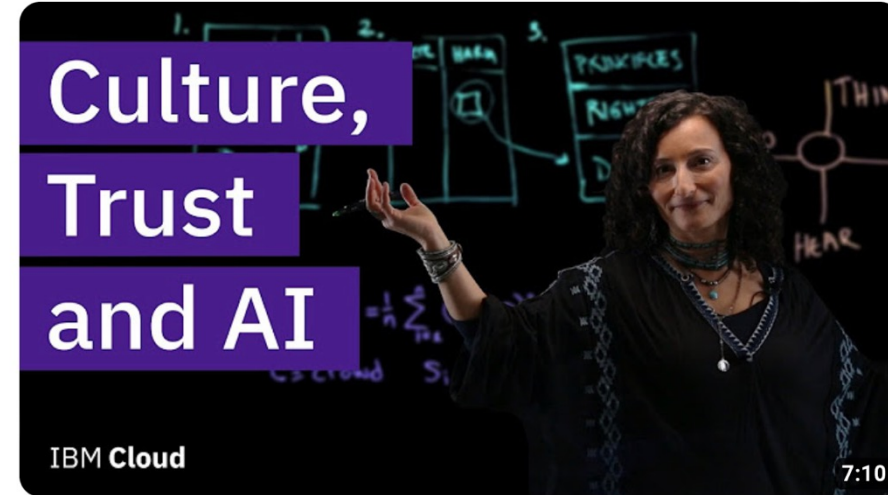


AI and Bing and ChatGPT are out of control... this

4K Subtitles

(lectures on youtube)

1. Why trust at all?



(lectures on youtube)

1. Why trust at all?

How Do We Trust AI Service? Exploring the Trust Mechanism in AI Service
Chapter May 2023 · E-Business. Digital Empowerment for an Intelligent Future
Ao Chen · Jinlin Wan
AI services have been widely used by consumers, and trust is a key factor impact them to continuously use AI services. However, due to the intelligent feature, trust in AI service is different from other trust. So, we explore how do users trust AI services and...
91 Reads
[Request full-text](#) [Recommend](#) [Follow](#) [Share](#)

Research from: [Industrial Sales and Service Engineering \[ISSE\]](#)

Can you trust me? Using AI to review more than three decades of AI trust literature
Full-text available Conference Paper May 2023 · 31st European Co...
Charlotte Knickrehm · Marleen Voss · Marie-Christin Barton
The ability of artificial intelligence (AI) to take on complex tasks can facilitate humans' life, but also raise concerns about AI replacing the human workforce. Acceptance and prior trust are prerequisites for...
96 Reads
[Download](#) [Recommend](#) [Follow](#) [Share](#)

Designing AI for Appropriation Will Calibrate Trust
Full-text available Conference Paper Apr 2023 · CHI TRAIT '23: Wor...
Tony Zhang · Yuanting Liu · Andreas Butz
Calibrating users' trust on AI to an appropriate level is widely considered one of the key mechanisms to manage brittle AI performance. However, trust calibration is hard to achieve, with numerous interacting factors tha...
63 Reads
[Download](#) [Recommend](#) [Follow](#) [Share](#)

Investigating and Designing for Trust in AI-powered Code

PAPER TITLES 111 Results | PAPERS (FULL TEXT) 369.742 Results | PEOPLE 0 Results | VIDEOS 0 Results | COURSES 0 Results

12.276 filtered results match **trust ai** SORT BY Relevance ▼

To what extent should we trust AI models when they extrapolate?
2022 · Xuenan Cao · Mathematics, Computer Science, Artificial Intelligence, Machine Learning, Extrapolation ...more ▼
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

Relationship between trust and acceptability in AI
2022 · Guillaume DEC... · Computer Science · HAL (Le Centre pour la Communication Scientifique Directe)
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

Trust and Trustworthiness in AI Ethics
2022 · Karoline Reinha... · AI and Ethics
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security
2022 · zebo yang · Computer Science · IEEE Internet of Things Journal
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

Relationship between trust and acceptability in AI
2022 · Guillaume DEC... · Computer Science · Le Centre pour la Communication Scientifique Directe - HAL - Inria
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

An Instrument for Measuring Teachers' Trust in AI-Based Educational Technology
2022 · Mutlu Cukurova · Computer Science · LAK22: 12th International Learning Analytics and Knowledge Conference
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

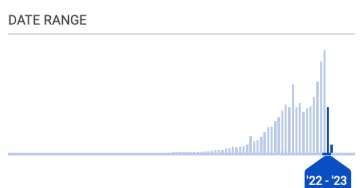
Relationship between trust and acceptability in AI
2022 · Alexandre AGO... · Computer Science · Le Centre pour la Communication Scientifique Directe - HAL - Inria
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

Relationship between trust and acceptability in AI
2022 · Frédérique Krupa · Computer Science · Le Centre pour la Communication Scientifique Directe - HAL - Inria
[Download](#) [Download Full PDF Package](#) [Save To Library](#)

Applying AI to digital archives: trust, collaboration and shared professional ethics

FILTERS

DATE RANGE



Past Year | Past 5 Years

PUBLICATION TYPE

All	12,276
Journal Article	1
Book	14
Conference Paper	
Other	12,261

LANGUAGE

All	12,276
English	10,216
Portuguese	597
Italian	413
Spanish	290
French	125
+ more	

1. Why trust at all?



The screenshot shows the European Commission website page for 'Ethics guidelines for trustworthy AI'. The page header includes the European Commission logo, a language selector set to 'English', and a search bar. Below the header is a blue navigation bar with the title 'Shaping Europe's digital future' and a menu with links for Home, Policies, Activities, News, Library, Funding, Calendar, and Consultations. The breadcrumb trail reads 'Home > Library > Ethics guidelines for trustworthy AI'. The main content area features the title 'Ethics guidelines for trustworthy AI' and a sub-header 'REPORT / STUDY | Publication 08 April 2019'. The text describes the publication of the guidelines on 8 April 2019, following a first draft in December 2018. It lists three principles of trustworthy AI: lawful, ethical, and robust. A 'See also' section points to 'A European approach to artificial intelligence', and a 'Related topics' section lists 'Advanced Digital Technologies' and 'Artificial intelligence'. The page concludes with a decorative image of the European Union flag stars over a network diagram.

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

1. Why trust at all?

Trust, but currently not only in AI ...

- Masks, vaccines, medical recommendations → Sciences
- Parties, democratic processes, election results → Politics
- Money, Banks → Economy
- News → Journalism

1. Why trust at all?

- At first glance, trust appears to be a pre-modern social form
- Rationalization, scientification, mechanization should reduce the need for trust
- But the opposite is the case!

1. Why trust at all?

Sociologist Georg Simmel wrote in 1908:

In modern societies

„life stands on a thousand presuppositions which the individual cannot trace and verify at all to their bottom, but which he has to accept in good faith. To a much greater extent than one tends to realize, our modern existence - from the economy, which is becoming more and more a credit economy, to the scientific enterprise, in which the majority of researchers must use innumerable results of others which are not at all verifiable to them - rests on the belief in the honesty of others. We build our most important decisions on a complicated system of notions, the majority of which presuppose confidence that we are not deceived.“

(Simmel [1908] 1992, p. 389)

1. Why trust at all?

First steps on the way to an answer:

1. Trust implies a dependency that can lead to harm
Dependence on other persons, institutions, systems whose services (information, decisions, skills) we need
2. We accept this dependence
When we trust, we do not avoid it, we do not try to dissolve or reduce it
3. The acceptance of this dependence is justifiable
Trust is not a-rational. It is grounded in reasons, even though we may not always have good reasons. Trust is consequently a practice that can be rational.

1. Why trust at all?

Annette Baier:

will not take it. Reasonable trust will require good grounds for such confidence in another's good will, or at least the absence of good grounds for expecting their ill will or indifference. Trust then, on this first approximation, is accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one.

Baier, Annette. "Trust and Antitrust." In: *Ethics* 96, no. 2 (1986): 235.

1. Why trust at all?

Dependence on/Vulnerability to AI:

- Medical diagnoses and decisions
- Autonomous vehicles
- Legal decisions
- Economic decisions
- Pedagogical decisions
- Political information
- ...

1. Why trust at all?

Dependence on/Vulnerability to AI:

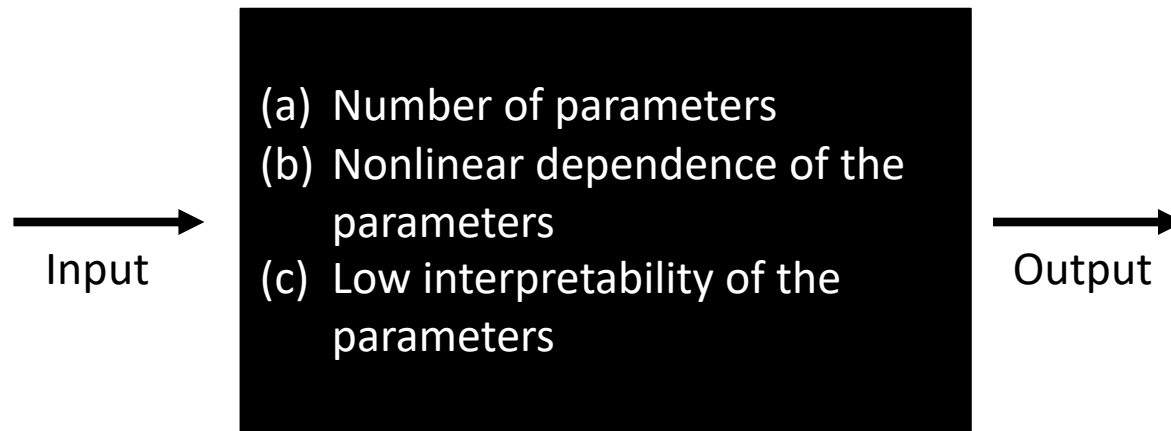
- The greater the dependency/vulnerability, the more trust seems necessary
- But we cannot want to accept this dependence either.

Alternatives: Control and Knowledge

1. Why trust at all?

Knowledge with a view to AI:

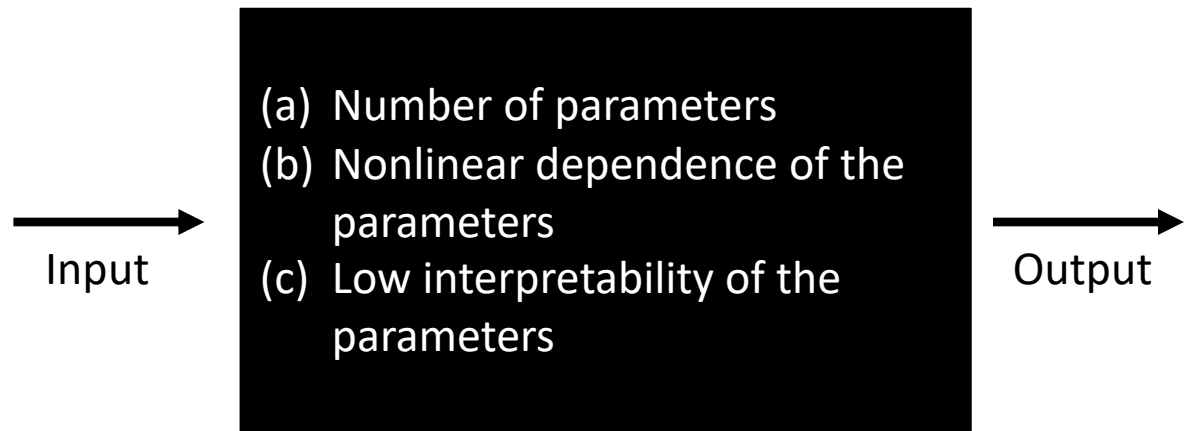
Model Opacity



1. Why trust at all?

Knowledge with a view to AI:

Model opacity



Pragmatic opacity

What is the output?
How was it generated?
How *well* does it work?

1. Why trust at all? ✓
2. Examples of the double opacity (model & pragmatic)
3. Three (too) simple approaches
 - a. Why not just reliability?
 - b. Why not simply listen to experts?
 - c. Why not just evaluate the scientific quality?
4. The intricate simplicity of trust
5. Trustworthiness as a value - a value among other values?

2. Examples of the model and pragmatic opacity



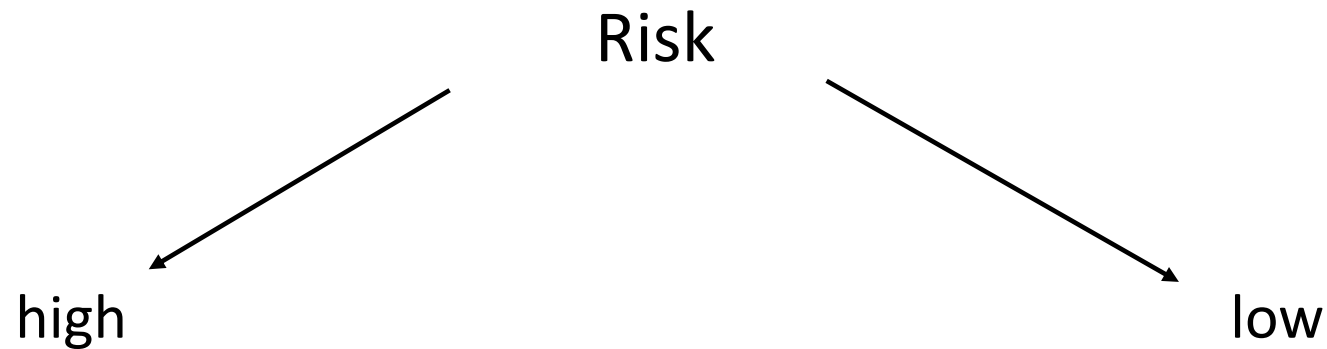
Example 1: Predicting Pneumonia Risk

- Goal: Classify patients with pneumonia as to whether they are “high” or “low risk”

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

- Goal: Classify patients with pneumonia as to whether they are „high“ or „low risk“

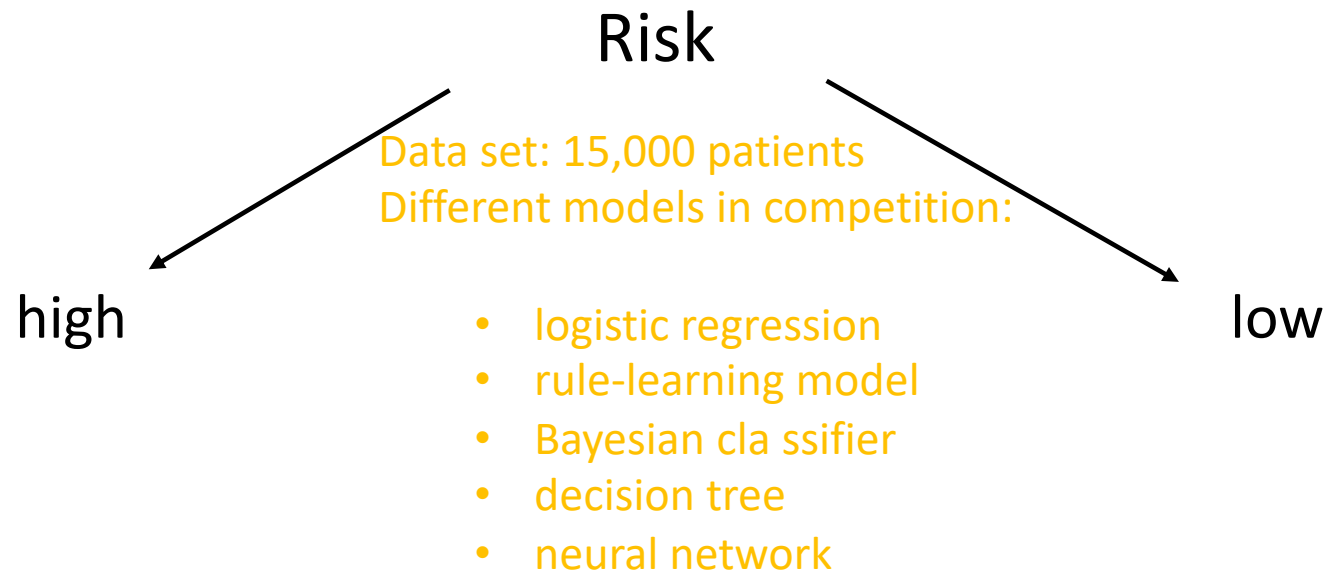


(Caruana et al. 2015)

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

- Goal: Classify patients with pneumonia as to whether they are „high“ or „low risk“

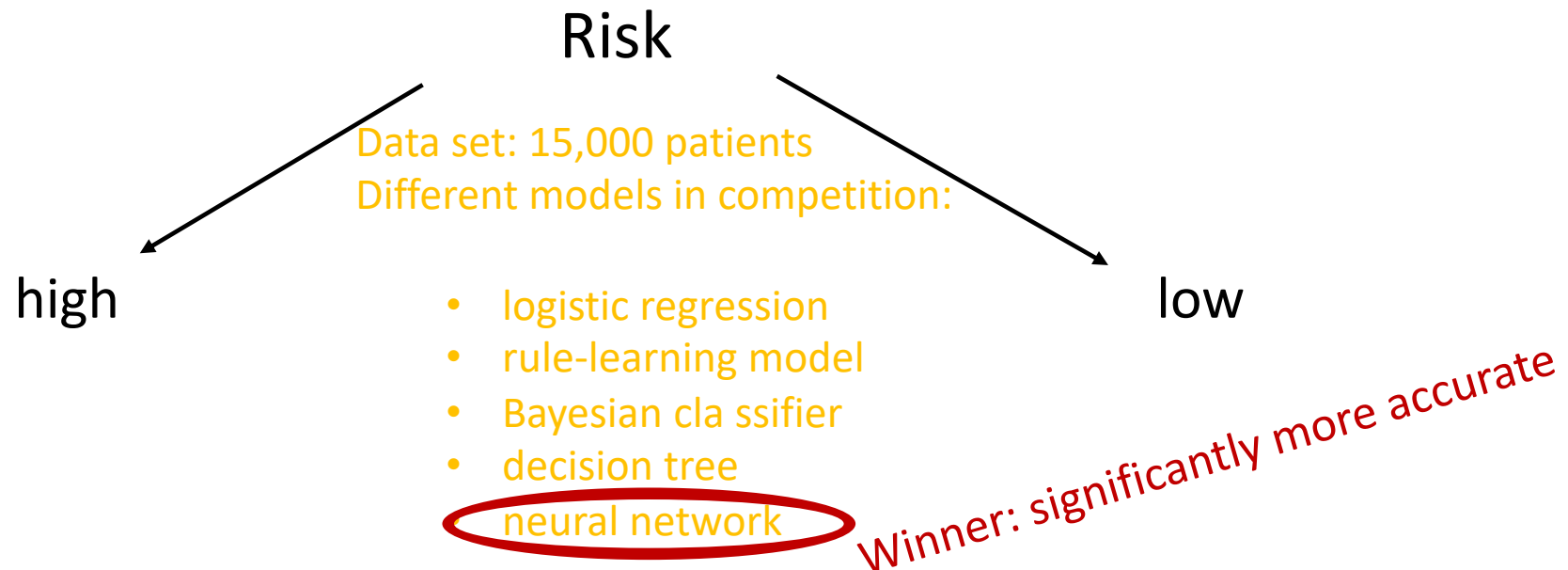


(Caruana et al. 2015)

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

- Goal: Classify patients with pneumonia as to whether they are „high“ or „low risk“



(Caruana et al. 2015)

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

els that could be trained were multitask neural nets.¹ On one dataset the neural nets outperformed traditional methods such as logistic regression by wide margin (the neural net had AUC=0.86 compared to 0.77 for logistic regression), and on the other dataset used in this paper outperformed logistic regression by about 0.02 (see Table 2). Although the neural nets were the most accurate models, after careful consideration they were considered too risky for use on real patients and logistic regression was used instead. Why?

(Caruana et al. 2015, 1271)

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

- Ambrosino, a colleague of Caruana, had developed a rule-based model
- Rule-based models are one of the most transparent models
- Typical form: a list of „if x then y“ rules

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

- Ambrosino, a colleague of Caruana, had developed a rule-based model
- Rule-based models are one of the most transparent models
- Typical form: a list of „if x then y“ rules
- Ambrosino observes a “strange rule” that the system had learned:
“If the patient has a history of asthma, then they are low-risk and you should treat them as an outpatient.”

2. Examples of the model and pragmatic opacity

Example 1: Predicting Pneumonia Risk

“If the patient has a history of asthma, then they are low-risk and you should treat them as an outpatient.”

The medical experts were able to clarify this:

- If someone has asthma, this person was usually specially observed.
- Individuals with asthma have a lower risk of death in the data set, but precisely because of the intensive monitoring
- So the low risk is found in the data!
- However, they should therefore not be treated as low-risk patients precisely because of this...

2. Examples of the model and pragmatic opacity



Example 1: Predicting Pneumonia Risk

„If the patient has a history of asthma, then they are low-risk and you should treat them as an outpatient.“

Caruana's insight:

- The rule found by the “rule-based system“ will also exist in the neural network
- However, it was not (so easy) to discover there
- The higher accuracy of the neural network leads to disguise this “rule“
- For him, the neural network is worse in this case ...

2. Examples of the model and pragmatic opacity

Example 1: „Predicting Pneumonia Risk“

Caruana's insight:

„I said, what I'm worried about is things that the neural net has learned that are just as risky as asthma but the rule-based system didn't learn.“ Because the neural net is more powerful, more flexible, it was capable of learning things that the rule-based system didn't. This, after all, is the advantage of neural networks-and the reason Caruana's neural net had won the group's internal contest. „I said it's those things that will make us not use this model. Because we don't know what's in it that we would need to fix.“

2. Examples of the model and pragmatic opacity



Example 2: Chat GPT. Slowness in Wittgenstein's thinking

2. Examples of the model and pragmatic opacity

Example 2: Chat GPT. Slowness in Wittgenstein's thinking

Manchmal kann ein Satz nur verstanden werden, wenn man ihn im *richtigen Tempo* liest. Meine Sätze sind alle *langsam* zu lesen. 1947

2. Examples of the model and pragmatic opacity

Example 2: Chat GPT. Slowness in Wittgenstein's thinking

Ich möchte eigentlich durch meine häufigen Interpunktionszeichen das Tempo des Lesens verzögern. Denn ich möchte langsam gelesen werden. (Wie ich selbst lese.) 1948

Manchmal kann ein Satz nur verstanden werden, wenn man ihn im *richtigen Tempo* liest. Meine Sätze sind alle *langsam* zu lesen. 1947

2. Examples of the model and pragmatic opacity

Example 2: Chat GPT. Slowness in Wittgenstein's thinking

AK

Gibt es eine Stelle, an der er ausdrücklich von "langsam" spricht?



Ja, Wittgenstein spricht an einigen Stellen in seinen Schriften ausdrücklich von der Bedeutung der Langsamkeit für die Philosophie. Eine solche Stelle findet sich beispielsweise in den "Philosophischen Untersuchungen" im §343:

"Das Philosophieren ist kein Ergebnis des Nachdenkens, sondern dessen Voraussetzung. Das Denken ist auch kein Denken über ein Problem, sondern ein Denken ohne ein Problem. [...] In der Philosophie geht es nicht um die Lösung von Problemen; sondern darum, daß keine Probleme bestehen bleiben. Die Philosophie ist keine Theorie, sondern eine Tätigkeit. [...] Was der Philosoph durch seine Arbeit erreicht, ist ein Klarwerden der Gedanken. Das ist nichts, was man 'einmal für allemal' tun kann. (Jeder muß denken, denken lernen.) Aber die Philosophie ist eine Tätigkeit, die man sozusagen jederzeit ausüben kann: man kann jederzeit seine Gedanken klären. (Man könnte auch sagen: 'philosophieren' ist eine Tätigkeit wie 'Stillesein', 'Schweigen', 'Sich-Besinnen!.) Wie es auch ankommt, jedenfalls bedarf sie der Langsamkeit. (Nicht der Zeit, sondern der Geduld.)"

2. Examples of the model and pragmatic opacity

Example 2: Chat GPT. Slowness in Wittgenstein's thinking

AK

Gibt es eine Stelle, an der er ausdrücklich von "langsam" spricht?



Ja, Wittgenstein spricht an einigen Stellen in seinen Schriften ausdrücklich von der Bedeutung der Langsamkeit für die Philosophie. Eine solche Stelle findet sich beispielsweise in den "Philosophischen Untersuchungen" im §343:

"Das Philosophieren ist kein Ergebnis des Nachdenkens, sondern dessen Voraussetzung. Das Denken ist auch kein Denken über ein Problem, sondern ein Denken ohne ein Problem. [...] In der Philosophie geht es nicht um die Lösung von Problemen; sondern darum, daß keine Probleme bestehen bleiben. Die Philosophie ist keine Theorie, sondern eine Tätigkeit. [...] Was der Philosoph durch seine Arbeit erreicht, ist ein Klarwerden der Gedanken. Das ist nichts, was man 'einmal für allemal' tun kann. (Jeder muß denken, denken lernen.) Aber die Philosophie ist eine Tätigkeit, die man sozusagen jederzeit ausüben kann: man kann jederzeit seine Gedanken klären. (Man könnte auch sagen: 'philosophieren' ist eine Tätigkeit wie 'Stillesein', 'Schweigen', 'Sich-Besinnen'.) Wie es auch ankommt, jedenfalls bedarf sie der Langsamkeit. (Nicht der Zeit, sondern der Geduld.)"

110

PHILOSOPHISCHE UNTERSUCHUNGEN I

sagen würde. Diese Erinnerungen sind ein seltsames Gedächtnisphänomen—und ich weiß nicht, welche Schlüsse auf die Vergangenheit des Erzählers man aus ihnen ziehen kann!

343. Die Worte, mit denen ich meine Erinnerung ausdrücke, sind meine Erinnerungsreaktion.

344. Wäre es denkbar, daß Menschen nie eine hörbare Sprache sprächen, wohl aber eine im Innern, in der Vorstellung, zu sich selber?

“Wenn die Menschen immer nur in ihrem Innern zu sich selbst sprächen, so täten sie schließlich nur dasjenige *beständig*, was sie auch heute *manchmal* tun.”—Es ist also ganz leicht, sich dies vorzustellen; man braucht nur den leichten Übergang von Einigen zu Allen zu machen. (Ähnlich: “Eine unendlich lange Baumreihe ist einfach

2. Examples of the model and pragmatic opacity

Example 2: Chat GPT. Slowness in Wittgenstein's thinking

AK

Gibt es eine Stelle, an der er sagt: "Im Rennen der Philosophie gewinnt, wer am langsamsten laufen kann"




Nein, es gibt keine Stelle in Wittgensteins Schriften, an der er diesen Satz wörtlich ausspricht. Der Satz "Im Rennen der Philosophie gewinnt, wer am langsamsten laufen kann" wird oft im Kontext von Diskussionen über die Bedeutung von Geduld und Sorgfalt in der Philosophie verwendet, aber er stammt nicht direkt von Wittgenstein. Es ist jedoch möglich, dass dieser Satz von anderen Philosophen oder Gelehrten als Paraphrase oder Zusammenfassung von Wittgensteins Ideen über die Bedeutung von Langsamkeit und Sorgfalt in der Philosophie verwendet wurde.

2. Examples of the model and pragmatic opacity

Example 2: Chat GPT. Slowness in Wittgenstein's thinking

Im Rennen der Philosophie gewinnt, wer am langsamsten
laufen kann. Oder: der, der das Ziel zuletzt erreicht. 1938

AK Gibt es eine Stelle, an der er sagt: "Im Rennen der Philosophie gewinnt, wer am langsamsten laufen kann"

 Nein, es gibt keine Stelle in Wittgensteins Schriften, an der er diesen Satz wörtlich ausspricht. Der Satz "Im Rennen der Philosophie gewinnt, wer am langsamsten laufen kann" wird oft im Kontext von Diskussionen über die Bedeutung von Geduld und Sorgfalt in der Philosophie verwendet, aber er stammt nicht direkt von Wittgenstein. Es ist jedoch möglich, dass dieser Satz von anderen Philosophen oder Gelehrten als Paraphrase oder Zusammenfassung von Wittgensteins Ideen über die Bedeutung von Langsamkeit und Sorgfalt in der Philosophie verwendet wurde.

2. Examples of the model and pragmatic opacity



Example 3: A test with the PRORETA anti-collision system (2009)

(Isermann et al. 2009, 638).

2. Examples of the model and pragmatic opacity

Example 3: A test with the PRORETA anti-collision system (2009)

- Car with driving assistant for braking and steering intervention
- Experimental setup:
 - Subjects are not told that a driving assistant is available
 - You will be given a task (driving with navigation system)
 - Obstacles appear suddenly during the ride
 - The steering and braking reactions of the test persons and the assistant are measured
 - The subjects are then interviewed

2. Examples of the model and pragmatic opacity

Example 3: A test with the PRORETA anti-collision system (2009)



Bild 41-7: Versuchsfahrzeug mit plötzlich erscheinendem Hindernis

2. Examples of the model and pragmatic opacity

Example 3: A test with the PRORETA anti-collision system (2009)

Tabelle 41-1: Die drei Versuchsreihen der Ergonomiestudie

Versuchsreihe	1 Plötzlich erscheinendes Hindernis	2 Automatisches Notbremsen	3 Automatisches Ausweichen
Probanden	22 Männer 20 Frauen 22–65 Jahre	15 Männer 13 Frauen 20–35 Jahre und >50 Jahre	19 Männer 14 Frauen 20–35 Jahre und >50 Jahre
Unabhängige Variable	Zeitpunkt (Plötzlichkeit) der Hinderniserscheinung	Verzögerung bei automa- tischem Bremsengriff Zeitpunkte von Hinderniser- scheinung und automa- tischem Bremsengriff	Verlauf des überlagerten Lenkwinkels Übersteuerungsmöglichkeit des Fahrers
Untersuchungs- gegenstand	Fahrerverhalten	Fahrerverhalten Akzeptanz	Fahrerverhalten Akzeptanz

2. Examples of the model and pragmatic opacity

Example 3: A test with the PRORETA anti-collision system (2009)

matisches Manöver fährt. Bei der automatischen Notbremsung war auffällig, dass die Probanden bei den meisten Fahrten mit einer Betätigung des Gaspedals reagierten. Betrachtet man die Video-Aufzeichnungen, so erkennt man deutlich, dass die Fahrer bzw. die Beine durch die Trägheitswirkung bei der Verzögerung nach vorne „fallen“, d. h. sie stützen sich unbewusst am Gaspedal ab, **Bild 41-8.**

2. Examples of the model and pragmatic opacity

Example 3: A test with the PRORETA anti-collision system (2009)

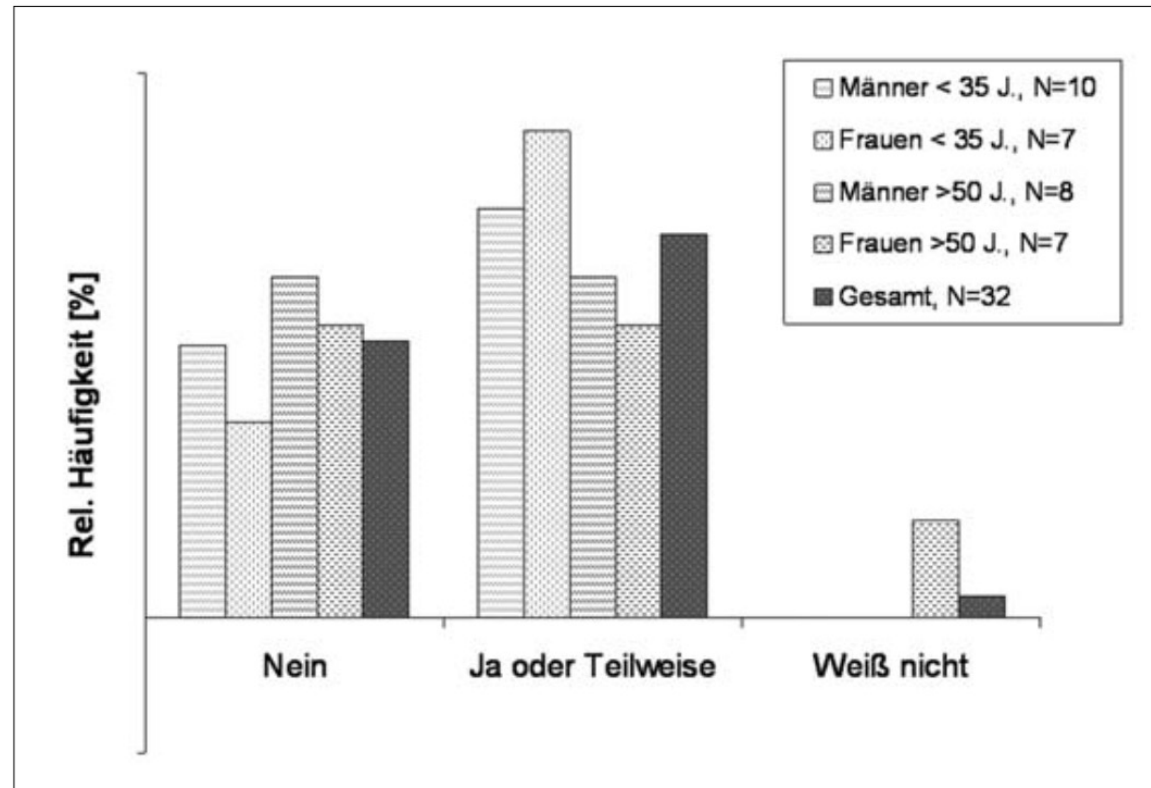


Bild 41-9: Einschätzung der Probanden, ob das Fahrzeug bei einer Fahrt mit geregelttem automatischem Lenkeinariff ihren eigenen Lenkbewegungen folgte

2. Examples of the model and pragmatic opacity

Example 4: search with learned preferences



2. Examples of the model and pragmatic opacity



Google

[Alle](#) [Bilder](#) [News](#) [Videos](#) [Shopping](#) [Mehr](#)

Beliebige Sprache ▾ Beliebige Zeit ▾ Alle Ergebnisse ▾

TAZ
<https://taz.de> > Lob-der-Langsamkeit

PHILOSOPHIE: Lob der Langsamkeit
12.05.2012 — Diesen Rat seines Kollegen und in gewissem Sinne Geistesverwandten Ludwig **Wittgenstein** hat sich der österreichische Philosoph und ...

kritische-masse.de
<https://www.kritische-masse.de> > blog > content > lang...

Langsamkeit und Zeit | Enzyklopädie der Kritischen Masse
22.05.2004 — Im Rennen der Philosophie gewinnt, wer am langsamsten laufen kann. Oder der, der das Ziel zuletzt erreicht. Ludwig **Wittgenstein** ...

Christian-Albrechts-Universität zu Kiel
<https://www.theol.uni-kiel.de> > publikationen

Langsamkeit
(L. **Wittgenstein**) Das Wesen des Menschen und sein Glück sind eng mit der Wahrheitserkenntnis verbunden. Der Wahrheit des eigenen Wesens inne zu werden, ...
4 Seiten

WELT
<https://www.welt.de> > DIE WELT

Goethe entdeckte die Langsamkeit
26.07.2003 — Selbst die Universität, wo **Wittgensteins** Philosophengruß bislang noch etwas galt, hat sich heute dem Diktat der Eile unterworfen. "Aus Mangel an ...

FAZ
<https://www.faz.net> > ... > Bücher > Rezensionen

Kein erlösendes Wort - Sachbuch
10.11.2005 — Schon lange wartet die Gemeinde auf eine zuverlässige, gut les- und handhabbare Gesamtedition des **Wittgenstein**-Nachlasses.

2. Examples of the model and pragmatic opacity

Example 4: search with learned preferences



Google Suche

Auf gut Glück!

**Dynamics of the
searched space**

2. Examples of the model and pragmatic opacity

Example 4: search with learned preferences



Google Suche

Auf gut Glück!

**Dynamics of the
searched space**

**Dynamics of the
learned model**

2. Examples of the model and pragmatic opacity

Example 4: search with learned preferences

Dynamics of my (learned) preferences and heuristics

Google



Google Suche

Auf gut Glück!

Dynamics of the searched space

Dynamics of the learned model

1. Why trust at all? ✓
2. Examples of double opacity (model & action) ✓
3. Three (too) simple approaches
 - a. Why not just reliability?
 - b. Why not simply listen to experts?
 - c. Why not just evaluate the scientific quality?
4. The intricate simplicity of trust
5. Trustworthiness as a value - a value among other values?

1. Why trust at all?

First steps on the way to an answer:

Review

1. Trust implies a dependency that can lead to harm
Dependence on other persons, institutions, systems whose services (information, decisions, skills) we need
2. We accept this dependence
When we trust, we do not avoid it, we do not try to dissolve or reduce it
3. The acceptance of this dependence is justifiable
Trust is not a-rational. It is grounded in reasons, even though we may not always have good reasons. Trust is consequently a practice that can be rational.

3.1 Why not just reliability?

Trustworthiness as reliability:

Main idea:

1. An actor or agent is trustworthy if they are reliable
2. Reliability is an epistemic concept. It can be tested and measured
3. Standard interpretation: frequency of successful cases in relation to total number

An actor or agent (person, institution, system, technical means, etc.) is reliable if they produce the result desired by the relying party in the vast majority of cases. Their reliability can be recognized and evaluated.

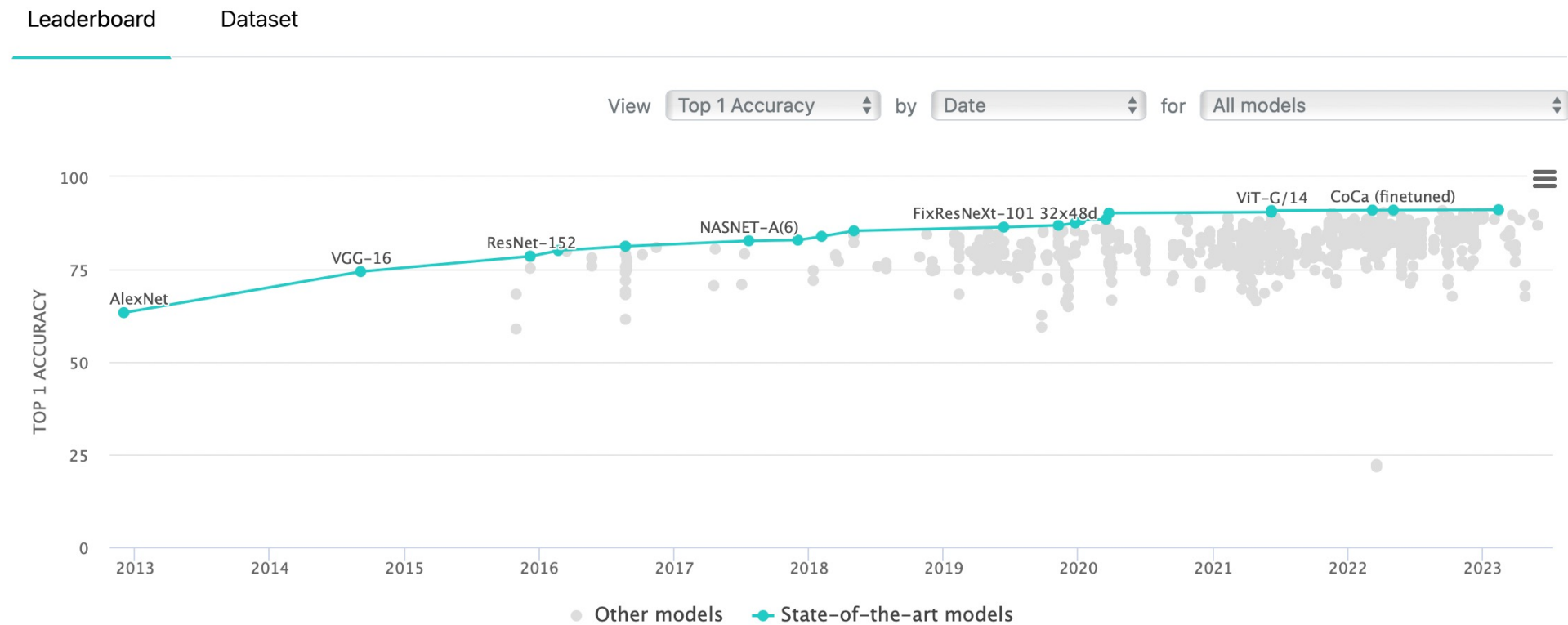
3.1 Why not just reliability?

Trustworthiness as reliability has four major merits:

- a. There is an *epistemic* basis (trustworthiness is based on knowledge)
- b. One can directly compare the reliability of different actors or agents, since reliability can be quantified
- c. There are known methods to technically evaluate the reliability
- d. Model opacity is usually irrelevant

3.1 Why not just reliability?

Image Classification on ImageNet



<https://paperswithcode.com/sota/image-classification-on-imagenet?metric=Gflops>

3.1 Why not just reliability?

Accuracy: Accuracy in machine learning is an evaluation metric that measures the number of correct predictions made by a model relative to the total number of predictions made. It is expressed in a score (measure).

Reliability: Usually understood as Accuracy + Robustness. The idea is that a model can be accurate with respect to one data set (e.g., through overfitting), but produce inconsistent results with other data sets (especially noisy data).

If, on the other hand, “accuracy” is understood as “*robust accuracy*” from the outset, the difference disappears. That is what we do here.

3.1 Why not just reliability?

Reliability (robust accuracy)

Number of true classifications

Number of classifications

= Measure of trustworthiness

3.1 Why not just reliability?

Hume's track record model

Number of true statements

Number of statements

= probability of being trustworthy

Hume's Intention:

Everyone can judge based on their own experience how trustworthy others are

3.1 Why not just reliability?

Hume's track record model

Number of true statements

Number of statements

= probability of being trustworthy

Hume's Intention:

Everyone can judge based on their own experience how trustworthy others are

But can this be a general method?

3.1 Why not just reliability?

Hume's track record model

How do we know
which are true?

Number of true statements

Number of statements

= probability of being trustworthy

Hume's Intention:

Everyone can judge based on their own experience how trustworthy others are

But can this be a general method?

3.1 Why not just reliability?

Hume's track record model

How do we know
which are true?

Number of true statements

Number of statements

= probability of being trustworthy

Hume's Intention:

Everyone can judge based on their own experience how trustworthy others are

But can this be a general method?

We usually know it through others!

3.1 Why not just reliability?

Hume's track record model

How do we know
which are true?

Number of true statements

Number of statements

= probability of being trustworthy

Hume's Intention:

Everyone can judge based on their own experience how trustworthy others are

But can this be a general method?

We usually know it through others!

That is, through the experience of others, whom we must trust in the process!

3.1 Why not just reliability?

Coady's critique of Hume:

1. Reduction does not succeed

— why?

Reduction means reduction to my *own* experience

Then I only know in a few cases,
based on my own experience,
which statements are true

That is why I am interested in the
knowledge of others in the first
place

Or I assess the truth of the
statements of others by the
experience of others, what is true

Then I am depending on others –
not my own experience

3.1 Why not just reliability?

Hume's track-record model

How do we know
which are true?

Number of true statements

Number of statements

= probability of being trustworthy

Reliability (robust accuracy)

How do we know
which are true?

Number of true classifications

Number of classifications

= Measure of trustworthiness

3.1 Why not just reliability?

Hume's track-record model

How do we know which are true?

Number of true statements

Number of statements

= probability of being trustworthy

Reliability (robust accuracy)

How do we know which are true?

Number of true classifications

Number of classifications

= Measure of trustworthiness

Experts tell us.
We must trust them
(if we want to work with the measure)

Agenda

1. Why trust at all? ✓
2. Examples of double opacity (model & action) ✓
3. Three (too) simple approaches
 - a. Why not just reliability? ✓
 - b. Why not simply listen to experts?
 - c. Why not just evaluate the scientific quality?
4. The intricate simplicity of trust
5. Trustworthiness as a value - a value among other values?

3.2 Why not simply listen to experts?

3.2 Why not simply listen to experts?

Trustworthiness of experts as a new approach

- We seem to have found a solution by the expert approach
- Experts tell us how trustworthy AI systems are
- They measure and evaluate the systems
- We get the scores

3.2 Why not simply listen to experts?

We trust / distrust AI system

3.2 Why not simply listen to experts?

We trust / distrust AI system



We trust / distrust experts
who tell us how trustworthy
AI systems are

3.2 Why not simply listen to experts?

Trustworthiness of experts as a new approach

- We seem to have found a solution
- Experts tell us how trustworthy AI systems are
- They measure and evaluate the systems
- We get the scores

But ...

- Experts dissent
- Lack of epistemic care
- Scientific misconduct

3.2 Why not simply listen to experts?

Trustworthiness of experts as a new approach

- We seem to have found a solution
- Experts tell us how trustworthy AI systems are
- They measure and evaluate the systems
- We get the scores

But ...

- Expert dissent
- Lack of epistemic care
- Scientific misconduct

Who are the experts?

3.2 Why not simply listen to experts?

We trust / distrust AI system



We trust / distrust experts
who tell us how trustworthy
AI systems are

3.2 Why not simply listen to experts?

We trust / distrust AI system



We trust / distrust experts
who tell us how trustworthy
AI systems are



We trust others who tell us
how trustworthy experts are
who tell us how trustworthy
AI systems are

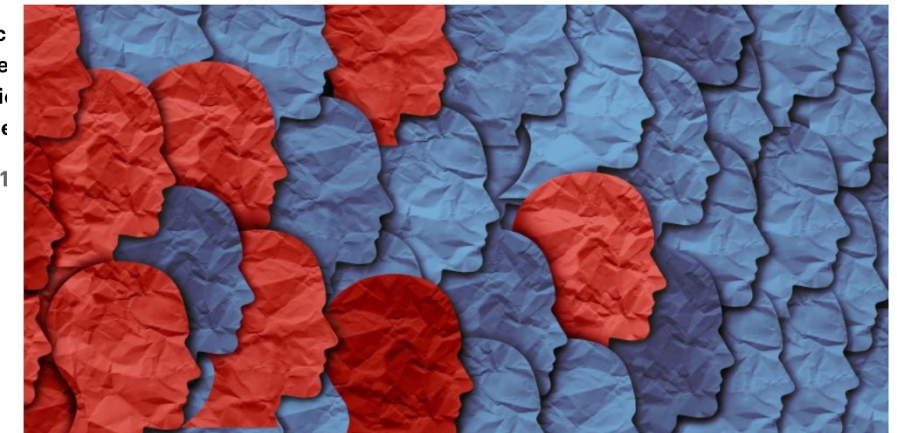
Medizin

SARS-CoV-2: Experten streiten über Herdenimmunität als Strategie

Donnerstag, 15. Oktober 2020



Newsletter abonnieren Zur Startseite



/freshidea - stock.adobe.com

Boston/London – Über das richtige Vorgehen angesichts einer zweiten Erkränkungsquelle mit einer stark steigenden Zahl von SARS-CoV-2-Infektionen

Themen Sendungen Programm Podcasts Audi

artseite / Forschung aktuell / Wissenschaftler streiten über die Wege

Gezielte Lockerungen oder Wissenschaftler streiten über Coronakrise

Wie soll es weitergehen mit dem Lockdown in Deutschland und Ländern. Auch unter Wissenschaftlern gibt es darüber, mit welcher Strategie – auch mit Blick auf die Corona-Pandemie in Deutschland eingedämmt werden

Volkart Wildermuth im Gespräch mit Lennart Pyritz | 1

Priesemann vs. Streeck Das Experten-Streit-Virus

von Peter Kunz

Wie besiegen wir die Corona-Pandemie? Physikerin Volker Priesemann und Virologe Volker Streeck berieten das niedersächsische Parlament



NEWS ANALYSIS

The U.S. Is Getting a Crash Course in Scientific Uncertainty

As the pandemic takes an unexpected direction, Americans again must reckon with twists in scientific understanding of the virus.

f WhatsApp Twitter Email Gift Share Bookmark 819



04.08.23

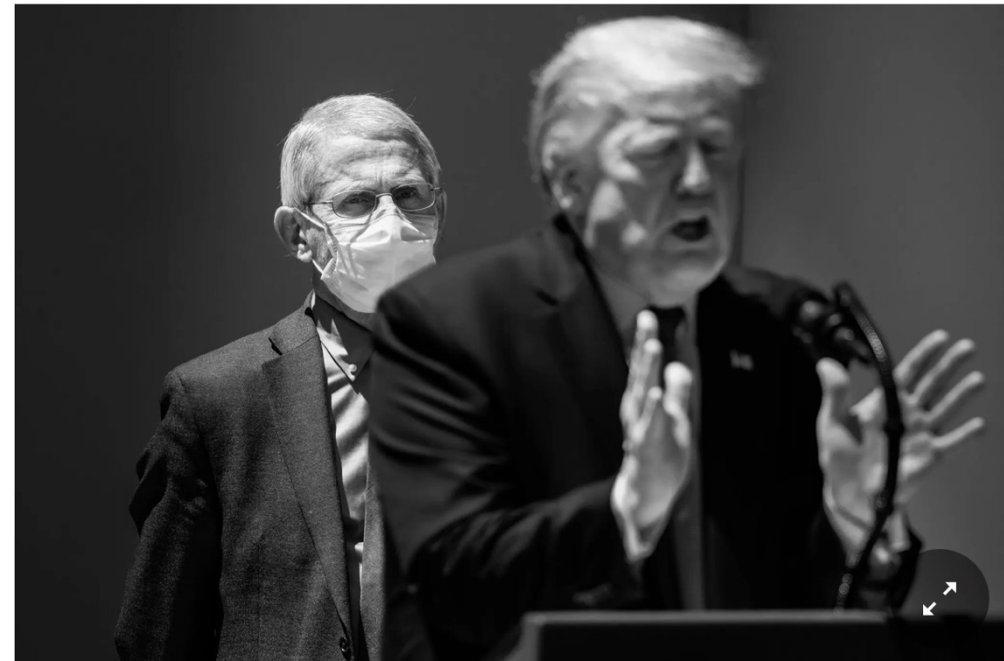
Andreas Kaminski

OPINION

When You Can't Just 'Trust the Science'

The vaccine debate is the latest example of how our coronavirus choices are inescapably political.

Dec. 19, 2020



73

3.2 Why not simply listen to experts?

What I accept as evidence already depends on my trust or mistrust in the source that presents the evidence to me! (Bernd Lahno)

Agenda

1. Why trust at all? ✓
2. Examples of double opacity (model & action) ✓
3. Three (too) simple approaches
 - a. Why not just reliability? ✓
 - b. Why not simply listen to experts? ✓
 - c. Why not just evaluate the scientific quality?
4. The intricate simplicity of trust
5. Trustworthiness as a value - a value among other values?

3.3 Why not simply evaluate the quality?

The problem seems to be that we involve too many stakeholders ...

New approach: Directly evaluate the quality of the systems!

3.3 Why not simply evaluate the quality?

The problem seems to be that we involve too many stakeholders ...

New approach: Directly evaluate the quality of the systems!

- But who performs these assessments?
- Everyone for themselves?

3.3 Why not simply evaluate the quality?

The problem seems to be that we involve too many stakeholders ...

New approach: Directly evaluate the quality of the systems!

- But who performs these assessments?
- Everyone for themselves?
- Very few have the expertise and resources to do this
- Moreover, the problem arises that the errors of AI technology are not always obvious

3.3 Why not simply evaluate the quality?

The problem seems to be that we involve too many stakeholders ...

New approach: Directly evaluate the quality of the

- But who performs these assessments?
- Everyone for themselves
- Very few have the experience
- Moreover, the problem always obvious

Now we seem to forget why trustworthiness was important to us in the first place!

4. The intricate simplicity of trust

Two assumptions need to be reconsidered:

1. Trustworthiness is a purely epistemic quality
2. It is possible to give a single applicable criterion for its evaluation

4. The intricate simplicity of trust

Two assumptions need to be reconsidered:

1. Trustworthiness is a purely epistemic quality
2. It is possible to give a single applicable criterion for its evaluation

4. The intricate simplicity of trust

1. Trustworthiness is a dense relation:

We cannot separate descriptive and normative aspects

- If we reasonably describe someone as trustworthy, we must have reasons that indicate the extent to which their behavior can be described in this way
- However, in doing so, his behavior must be understood to be motivated and explained by the value of trustworthiness
- This must also be a value for us

In other words, there is no purely epistemic theory of trustworthiness

4. The intricate simplicity of trust

1. Trustworthiness is a dense relation:

We cannot separate descriptive and normative aspects

- If we reasonably describe someone as trustworthy, we must have reasons that indicate the extent to which their behavior can be described in this way
- However, in doing so, his behavior must be understood to be motivated and explained by the value of trustworthiness
- This must also be a value for us

*In other words, there is no purely epistemic theory of trustworthiness - **Epistemic and Ethical Virtues of Modelers and Examiners***

4. The intricate simplicity of trust

2. There is a network of trust relationships

- We can not get out of the circle

3.2 Why not simply listen to experts?

We trust / distrust AI system



We trust / distrust experts
who tell us how trustworthy
AI systems are



We trust others who tell us
how trustworthy experts are
who tell us how trustworthy
AI systems are

Look back

4. The intricate simplicity of trust

2. There is a network of trust relationships

- We can not get out of the circle
- We need to deal with the circle instead
- This is less problematic as long as we move in trustworthy networks

4. The intricate simplicity of trust

2. There is a network of trust relationships

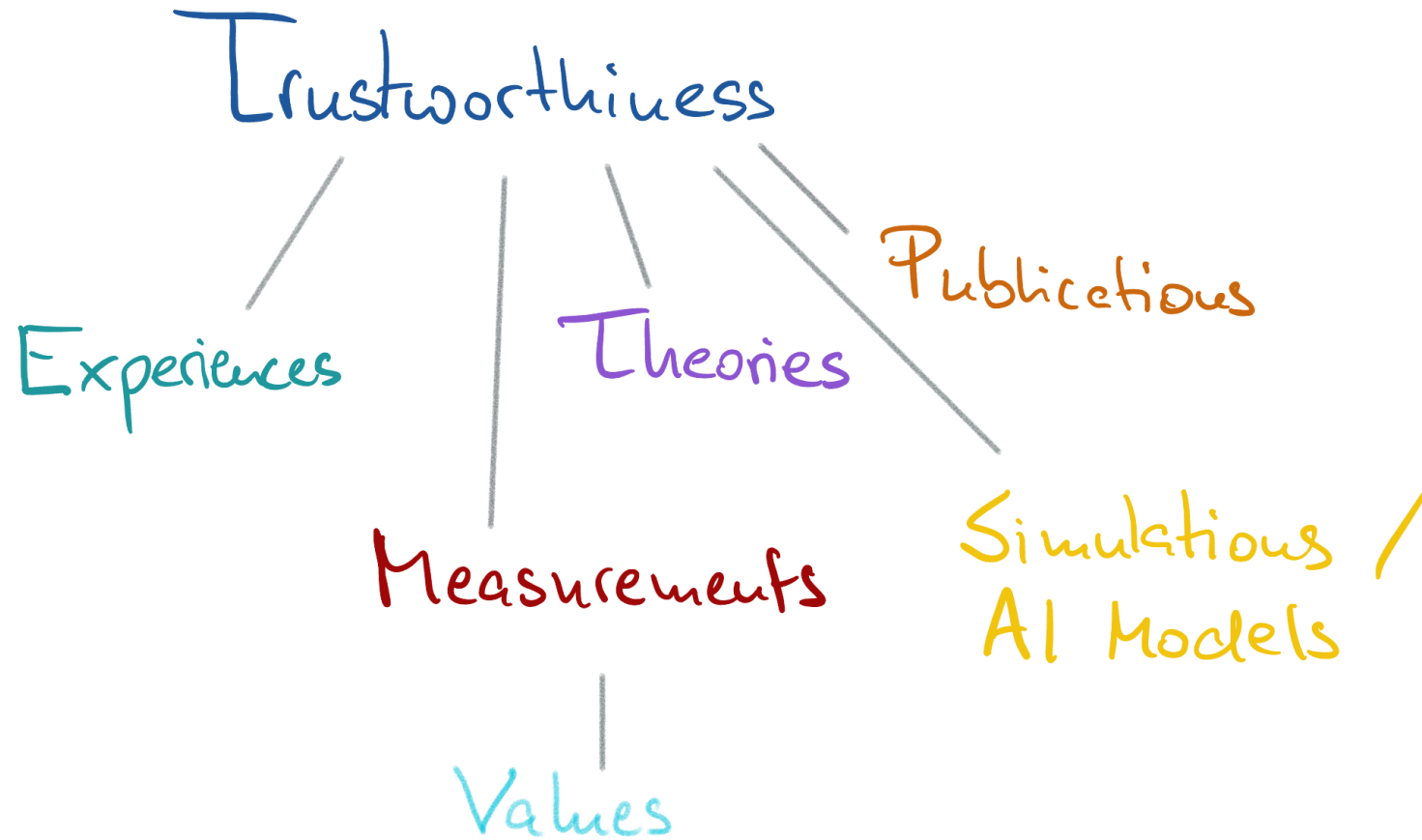
- We can not get out of the circle
- We need to deal with the circle instead
- This is less problematic as long as we move in trustworthy networks

Analogy to the demarcation problem between science and pseudoscience

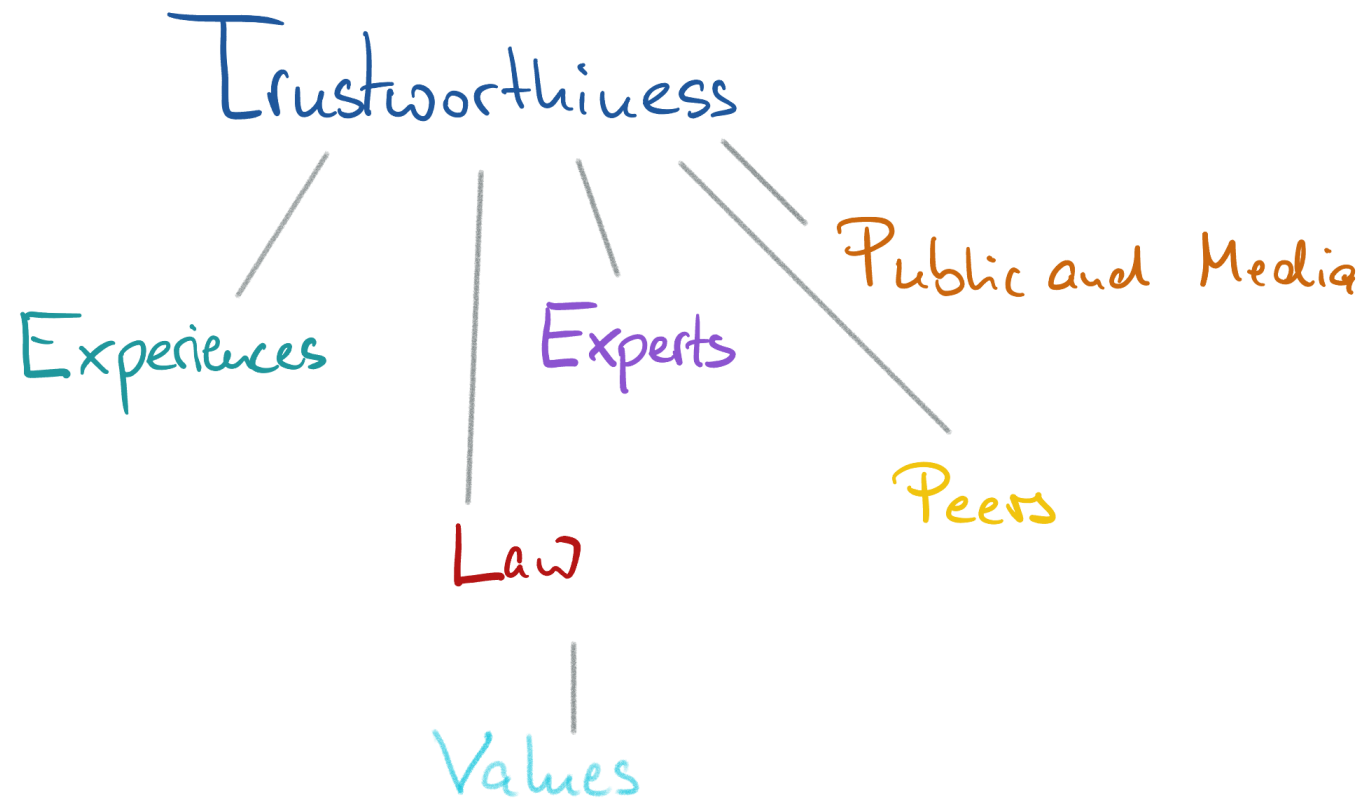
→ from foundationalism to coherentist approaches*.

* Assessment of coherence itself again not independent of trust!

4. The intricate simplicity of trust



4. the intricate simplicity of trust



1. Why trust at all? ✓
2. Examples of double opacity (model & action) ✓
3. Three (too) simple approaches
 - a. Why not just reliability? ✓
 - b. Why not simply listen to experts? ✓
 - c. Why not just evaluate the scientific quality? ✓
4. The intricate simplicity of trust ✓
5. Trustworthiness as a value - a value among other values?

4. Trustworthiness - a value among others?

- Reliability
- Transparency
- Fairness
- Sustainability
- ...
- Trustworthiness

4. Trustworthiness – a value among others?

A simple test:

- ~~Reliability~~
 - Transparency
 - Fairness
 - Sustainability
 - ...
- Trustworthiness?

4. Trustworthiness – a value among others?

A simple test:

- Reliability
 - Transparency
 - ~~Fairness~~
 - Sustainability
 - ...
- Trustworthiness?

4. Trustworthiness – a value among others?

A simple test:

- Reliability
 - ~~Transparency~~
 - Fairness
 - Sustainability
 - ...
- Trustworthiness?

4. Trustworthiness - a value among others?

Others (persons, institutions, AI systems) are trustworthy if they fulfill the values at stake in the respective situations

- Trustworthiness is not one value among others
- It forms the unity of the (respectively relevant) values

1. Why trust at all? ✓
2. Examples of double opacity (model & action) ✓
3. Three (too) simple approaches
 - a. Why not just reliability? ✓
 - b. Why not simply listen to experts? ✓
 - c. Why not just evaluate the scientific quality? ✓
4. The intricate simplicity of trust ✓
5. Trustworthiness as a value - a value among other values? ✓

- Kaminski, Andreas. „Der Erfolg der Modellierung und das Ende der Modelle: Epistemische Opazität in der Computersimulation“. In *Technik - Macht - Raum: Das Topologische Manifest im Kontext interdisziplinärer Studien*, herausgegeben von Andreas Brenneis, Oliver Honer, Sina Keesser, und Silke Vetter-Schultheiß, 317–33. Wiesbaden: Springer, 2018.
- Kaminski, Andreas. „Begriffe in Modellen: Die Modellierung von Vertrauen in Computersimulation und maschinellem Lernen im Spiegel der Theoriegeschichte von Vertrauen“. In *Simulieren und Entscheiden: Entscheidungsmodellierung, Modellierungsentscheidungen, Entscheidungsunterstützung*, herausgegeben von Nicole J. Saam, Michael Resch, und Andreas Kaminski, 1. Auflage 2019., 167–92. Sozialwissenschaftliche Simulationen und die Soziologie der Simulation. Wiesbaden: Springer Fachmedien Wiesbaden GmbH; Springer VS, 2019.
- Kaminski, Andreas. „Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen“. In *Ethische Herausforderungen von Big-Data*, herausgegeben von Klaus Wieglerling, Michael Nerurkar, und Christian Wadehul, 151–74. Bielefeld: Springer, 2019.
- Kaminski, Andreas, und Andreas Gelhard, Hrsg. *Zur Philosophie informeller Technisierung*. Darmstadt: Wissenschaftliche Buchgesellschaft, 2014.
- Kaminski, Andreas, und Colin W. Glass. „Das Lernen der Maschinen“. In *Mensch-Maschine-Interaktion: Handbuch zu Geschichte – Kultur – Ethik*, herausgegeben von Kevin Liggieri und Oliver Müller, 1. Auflage 2019., 128–33. Stuttgart, 2019.
- Lampe, Hildrun, und Andreas Kaminski. „Verlässlichkeit und Vertrauenswürdigkeit von Computersimulationen“. In *Mensch-Maschine-Interaktion: Handbuch zu Geschichte – Kultur – Ethik*, herausgegeben von Kevin Liggieri und Oliver Müller, 1. Auflage 2019. Stuttgart, 2019.
- Resch, Michael, und Andreas Kaminski. „The Epistemic Importance of Technology in Computer Simulation and Machine Learning“. *Minds and Machines* 29, Nr. 1 (1. Januar 2019): 1–9. <https://doi.org/10.1007/s11023-019-09496-5>.
- Kaminski, Andreas, Michael Resch, und Uwe Küster. „Mathematische Opazität. Reproduzierbarkeit in der Computersimulation“. *Jahrbuch Technikphilosophie* 4 (1. Januar 2018): 253–77.