Fighting fire with fire: responsible AI through regulation or innovation?

Frank van Harmelen
Knowledge Representation & Reasoning Group
Vrije Universiteit Amsterdam

Part 0:
Scope & Goal

Scope

Informatics for a sustainable future

AI for a **socially** sustainable future

AI for a **FAT** future

Fair, Accountable, Transparant

- fatml.org
- facctconference.org (ACM)
- FATE @ Microsoft

Goals

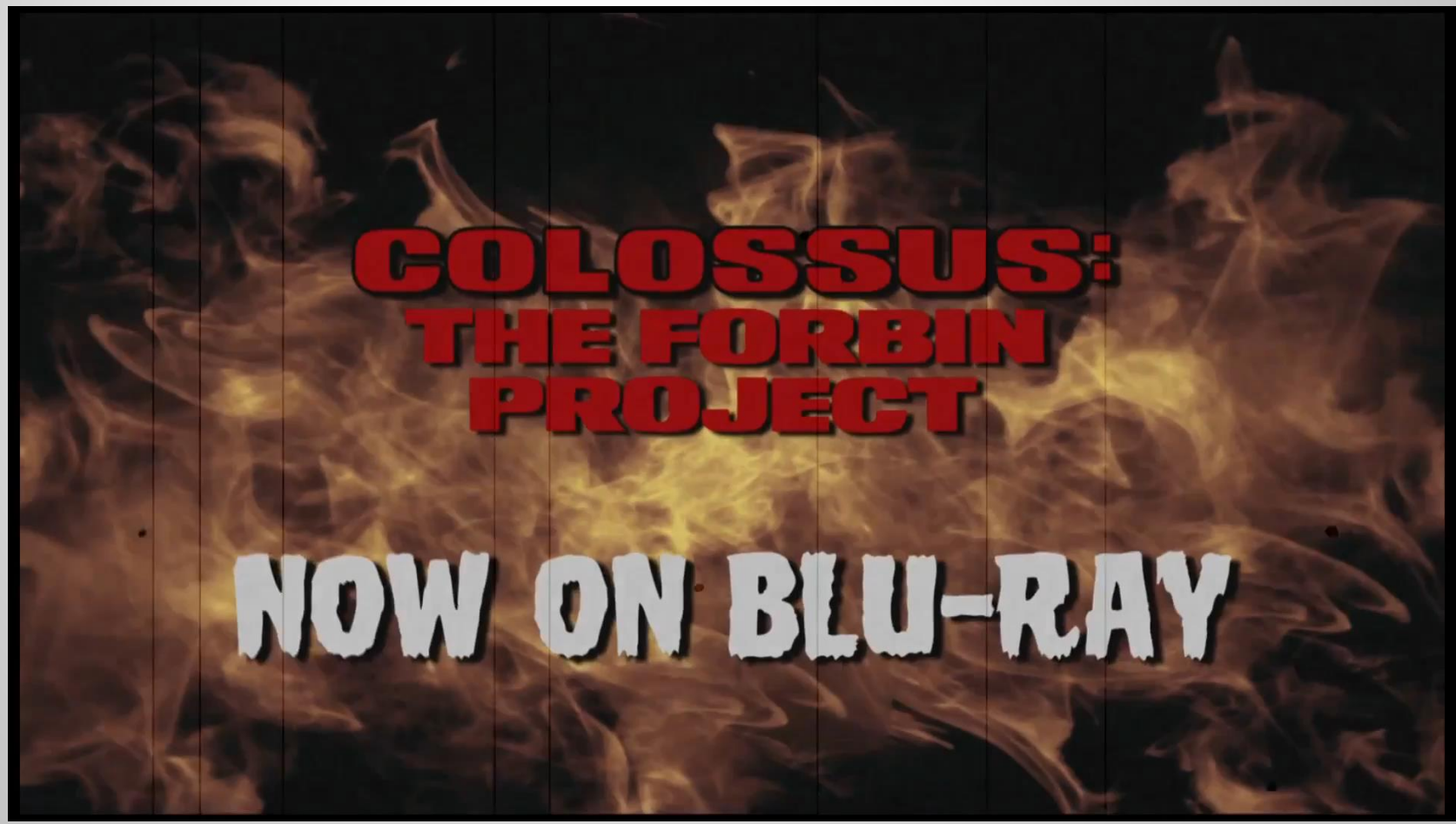
- Convince you there's a problem with (the public image of) AI
- Show you how lawmakers deal with this problem
- Show you how AI researchers deal with this problem
- Discuss with you lessons and recommendations

Part I:
Convince you there's a problem with (the public image) of AI

1. Different narratives about AI

(in the eyes of the public, politicians, other scientists)

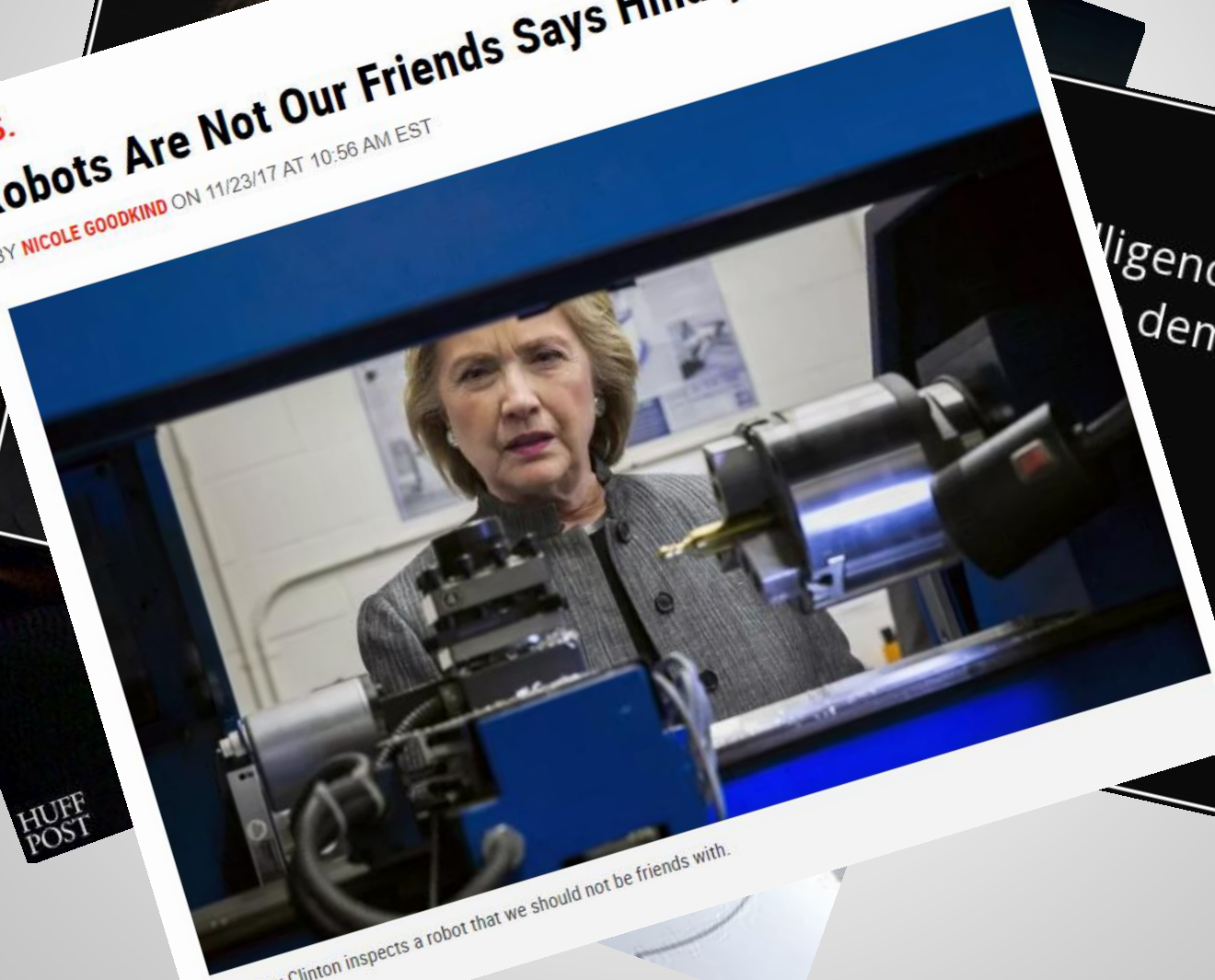
Narrative 1: AI is going to destroy the world



U.S.

Robots Are Not Our Friends Says Hillary Clinton

BY NICOLE GOODKIND ON 11/23/17 AT 10:56 AM EST



Hillary Clinton inspects a robot that we should not be friends with.

REUTERS

HUFF POST

ligence we are
demon.

1. Different narratives about AI

(in the eyes of the public, politicians, other scientists)

Narrative 2: AI is going to save the world

1. Different narratives about AI

(in the eyes of the public, politicians, other scientists)

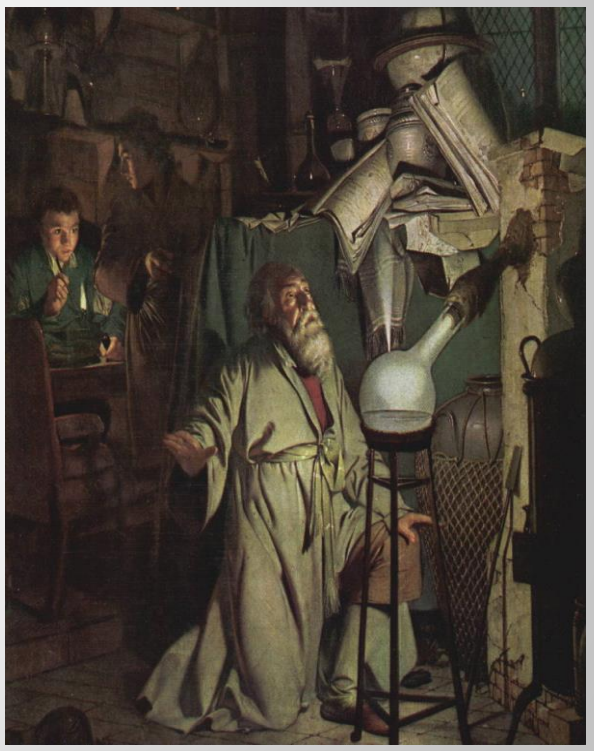
Narrative 3: AI is going to destroy science(!)

ML engineers assemble their codes with the same wishful thinking that the ancient alchemists had when mixing their magic potions.

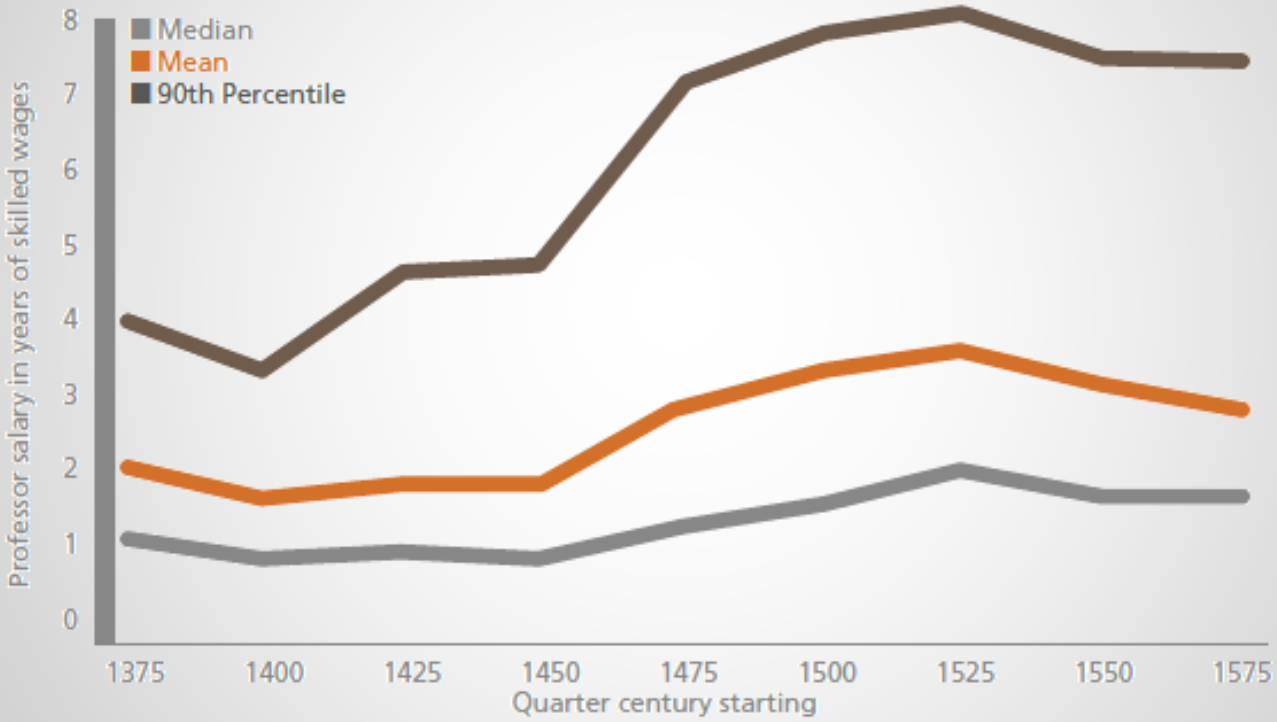
By deferring so much to machines, are we discarding the scientific method, and reverting to the dark practices of alchemy?

We should never forget the hard-won lessons of history. Alchemy was not [illegible] proto-science, but also a "hyper-science [illegible] overpromised and underdelivered.

Robbert Dijkgraaf
Quanta Mag. 2021

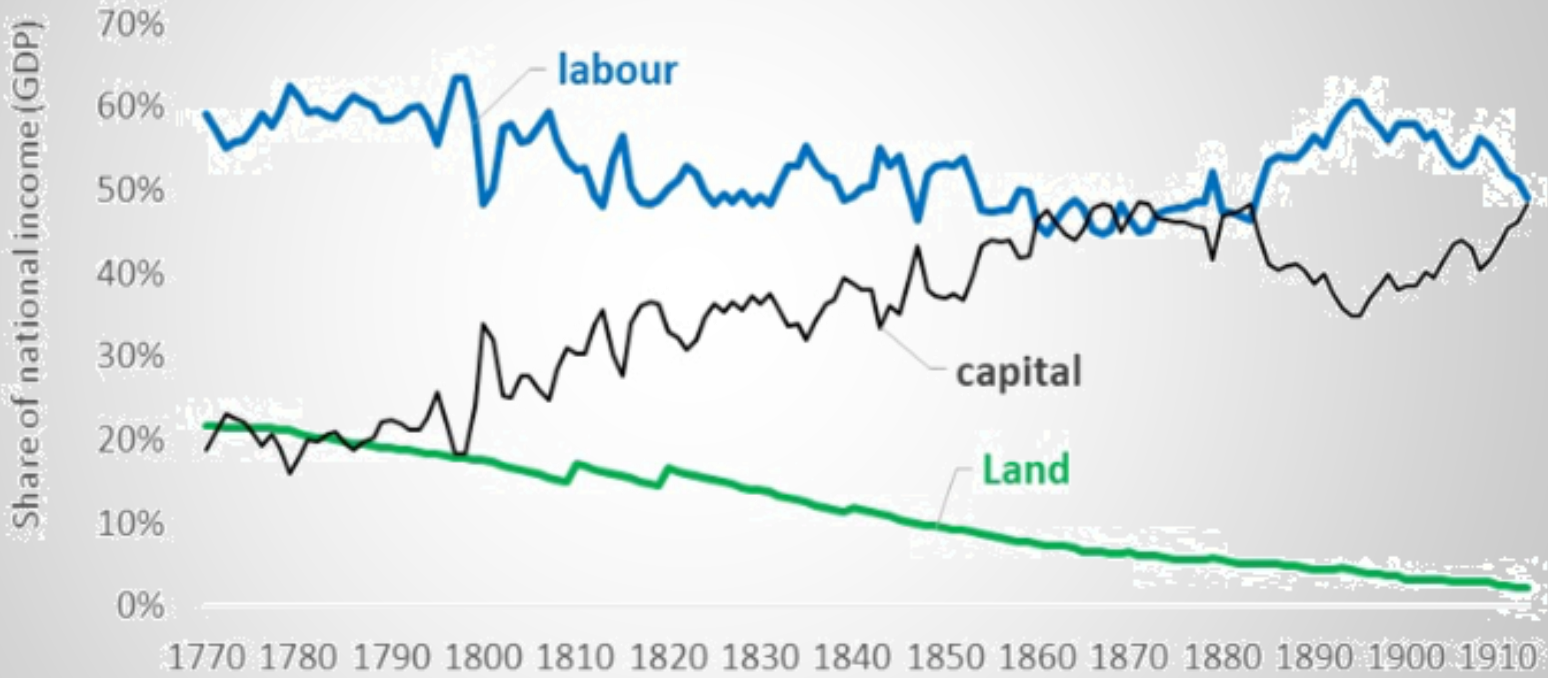


2. AI contributes socio-economic inequality
The printing press



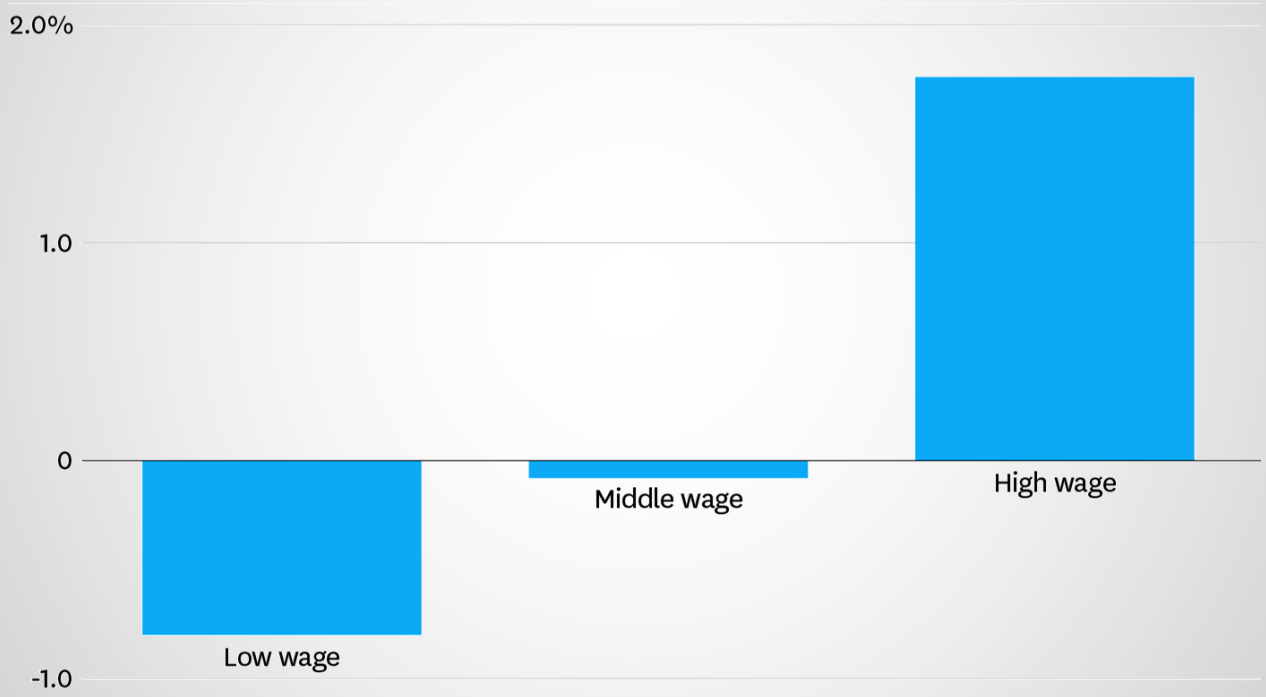
Salaries of professors at Italian universities (LSE)

2. AI contributes socio-economic inequality
The steam engine



Share of GDP (UNCTAD)

2. AI contributes socio-economic inequality
The computer



Annual job growth 1980-2000 (HBR)

3. AI contributes to unfairness

In NL AI algorithms used
- police records,
- education level,
- real-estate ownership,
- debts,
- citizenship status

to assess fraude risk
for daycare allowances

3. AI contributes to unfairness

Screenshot from 2020-04-03 09-51-57.png

Hand	77%
Gun	61%

Screenshot from 2020-04-02 11-51-45.png

Hand	72%
Monocular	60%

4. AI is non-transparent

Image labelling

?

Label: **shower cap**

Certainty: 99.7%

4. AI is non-transparent

Large language models

GPT-3: encoding of 500 **billion** words,
175 **billion** parameters

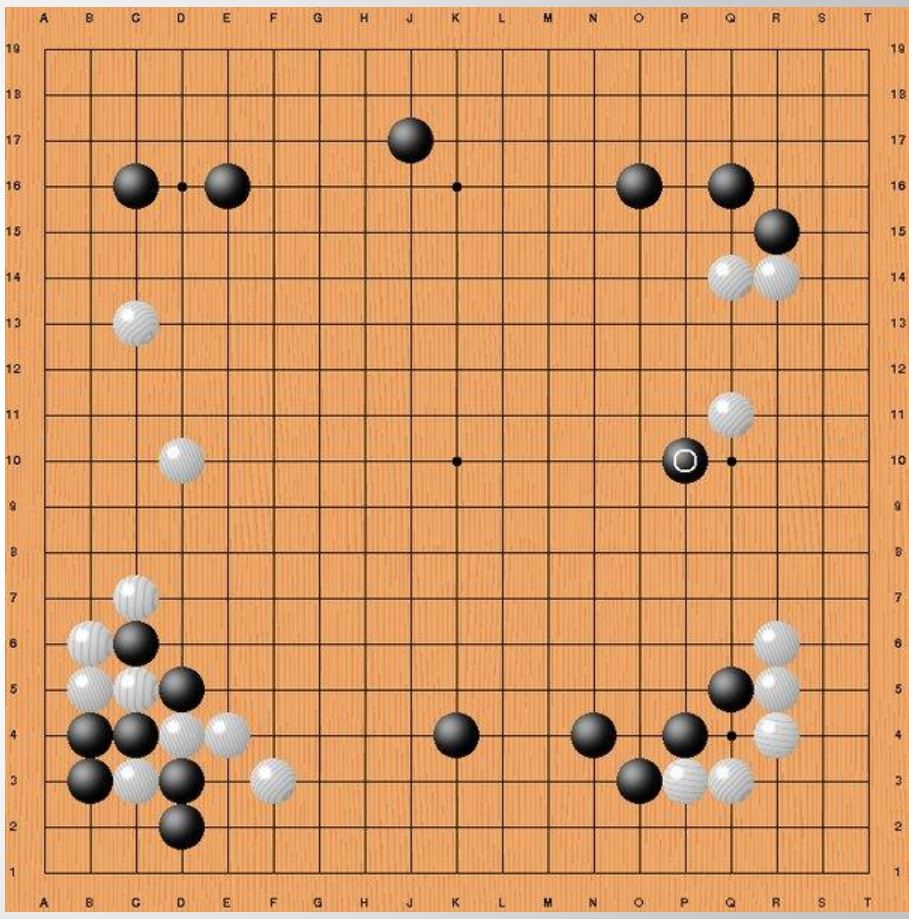
You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So …
you drink it. You are now dead.

4. AI is non-transparent

AlphaGo,

Game 2, Move 37

The Hand of God move



By now,
you should be convinced
there there is a real issue with FAT AI
that we cannot ignore
in our research and our teaching

Part II:
How do lawmakers try to solve this issue

1. Forbid the registration of sensitive data

Concern:

Makes it impossible to detect bias by proxy (“shortcut learning”)

- Postcode as proxy for ethnicity
- Name as proxy for gender:
  Anna, Lynda, Marja, Carla, Lisa, ….

In particular for Deep Learning

2. Introduce an algorithm register

Concern:

1. It's not the algorithm,
   it's the algorithm + the data + the application
2. Where to stop?
   The Dutch system for tax fraud detection used
   linear regression & decision trees,
   both are perfectly transparent and explainable.

3. Introduce guidelines

- OECD Principles on AI
- EU Ethics Guidelines for Trustworthy AI
- Chinese Government Ethical Norms for the New Generation AI
- UN framework for ethical AI
- Informatics Europe & EUACM Recommendations on Machine-Learned Automated Decision Making
- ….

Concern:
not sufficiently operational

(but see work by Richard Benjamins at Telefonica on operationalising them)

4. Introduce laws: EU AI Act

- AI = Machine Learning, Expert & Logic Systems, Baysian or statistical approaches
- Applies to: finance, education, human resources, law enforcement, industrial AI, medical devices, car industry, toys
- Three categories of AI uses:
  - Prohibited
  - High risk
  - Limited risk

4. Introduce laws: EU AI Act

Prohibited AI use =

- Harmful subliminal manipulation
- Harmful exploitation of age or disability
- Social credit scoring by governments
- Real-time remote biometric identification in public spaces by law enforcement agencies (except in limited cases)

4. Introduce laws: EU AI Act

High risk AI use =

- In one of 19 markets (aviation, cars, medical devices, ….)
- Critical infrastructure
- Access to education
- Worker management
- Essential services (including financial & credit scoring)
- Justice & law enforcement
- Migration, asylum, border control
- … (extendible list)

4. Introduce laws: EU AI Act

High risk AI use must:

- Have safeguards against biases in data sets
- Use prescribed data management practices
-  Be able to trace back outputs
-  Have acceptable levels of understandability for users
- Have human oversight

4. Introduce laws: EU AI Act

Limited risk AI use must inform users:

- Disclose“this is AI”
- Disclose which data for which purposes
- Disclose use of sensitive categories
- Disclose deep fakes

Fines up to 30m€ or 6% of turnover for prohibited AI
20m€ or 4% of turnover for high risk AI

4. Introduce laws: EU AI Act

Concerns: CLAIRE (largest AI network in the world)

This is too tough

- Unclear definitions (“AI”? “data quality”?)
- Regulation will impose burden
- These two will limit uptake of AI in Europe

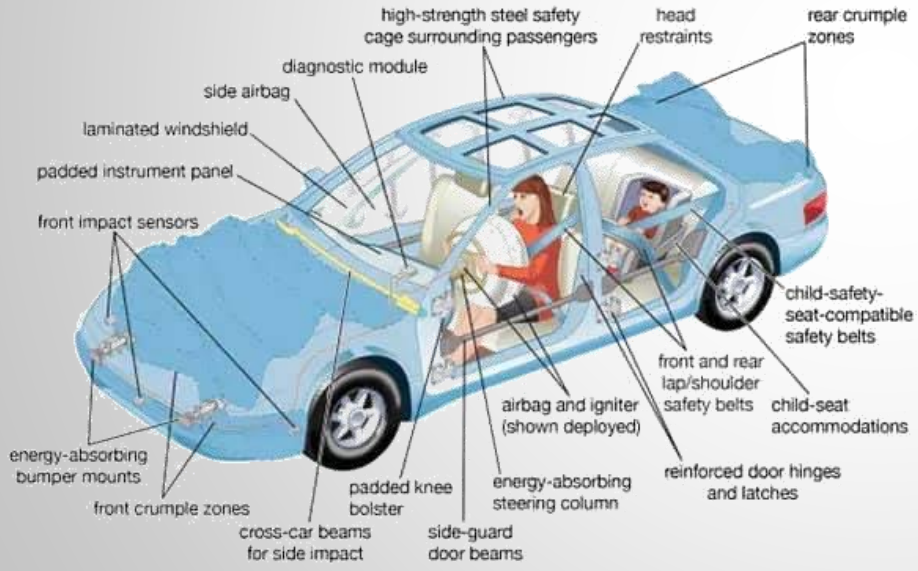
Concerns:

This is not tough enough

“*Prohibit* the use of *all* AI in

education, employment, law enforcement, biometric identification, banking, migration, justice”

Concern: red flag laws

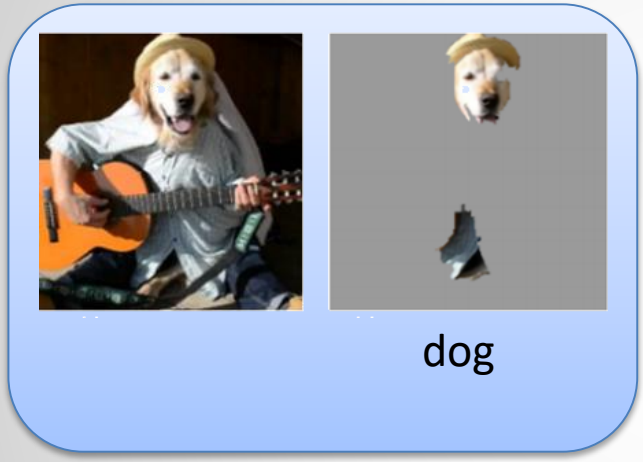


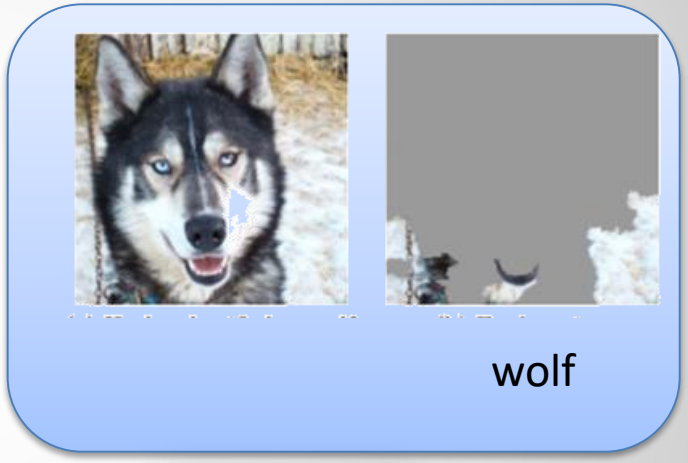
Cars got safer through more technology

Part III:
How do AI researchers try to solve this issue

1. Explanation by salience

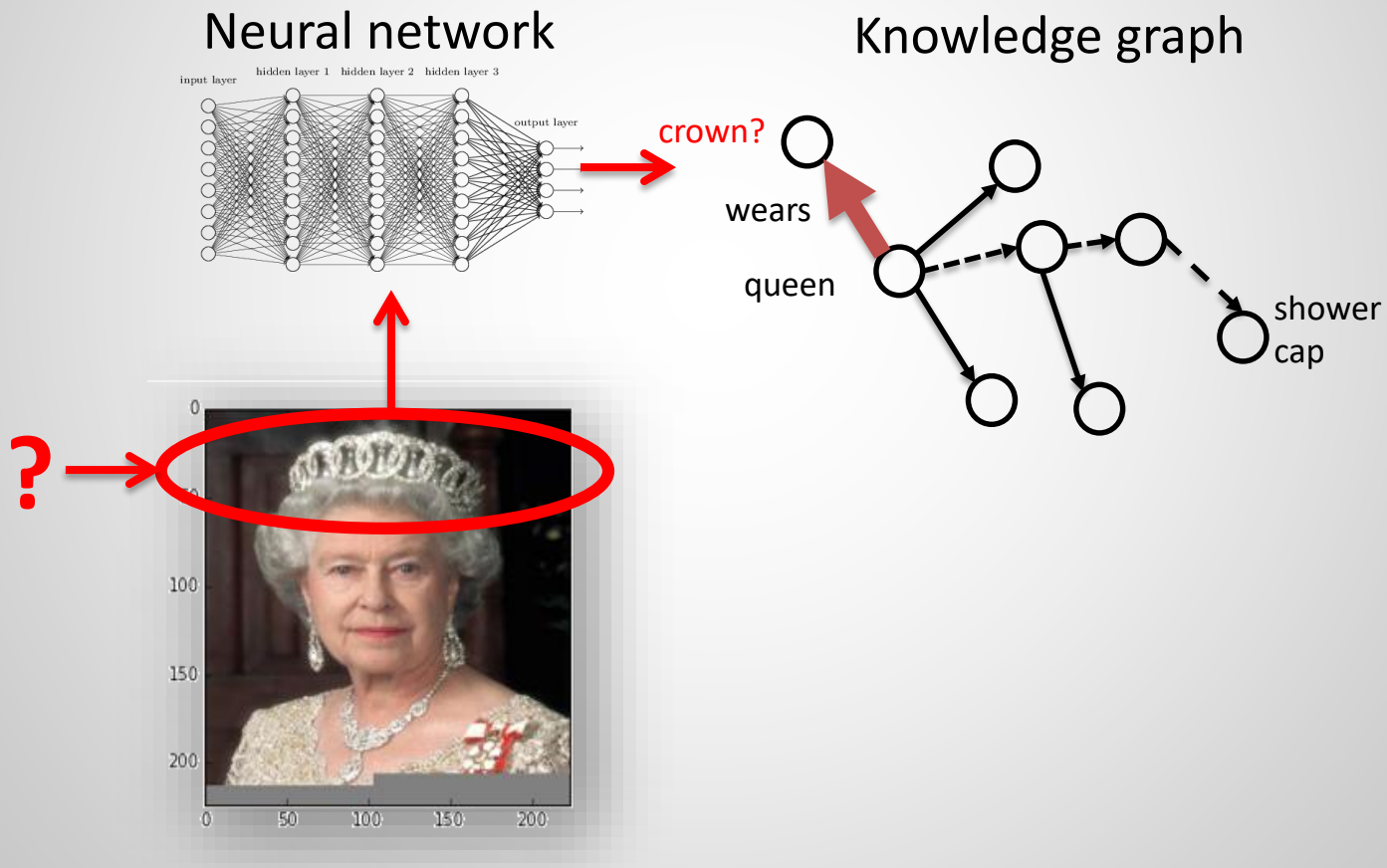
Which parts of the input contributed most to the output



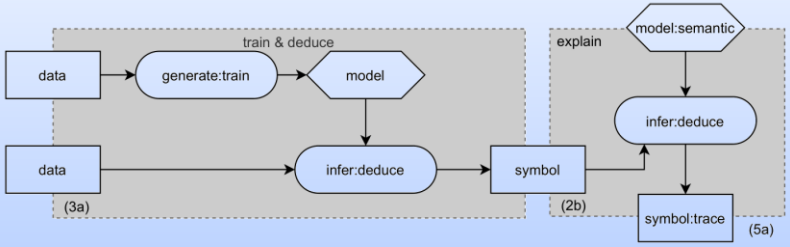


LIME (but now many others)
Exposes shortcut-learning
Would have explained the Google gun problem

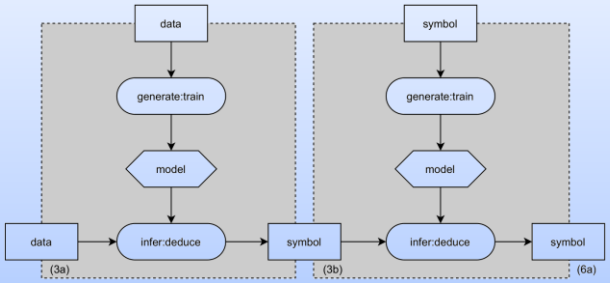
2. Explanation by rational justification



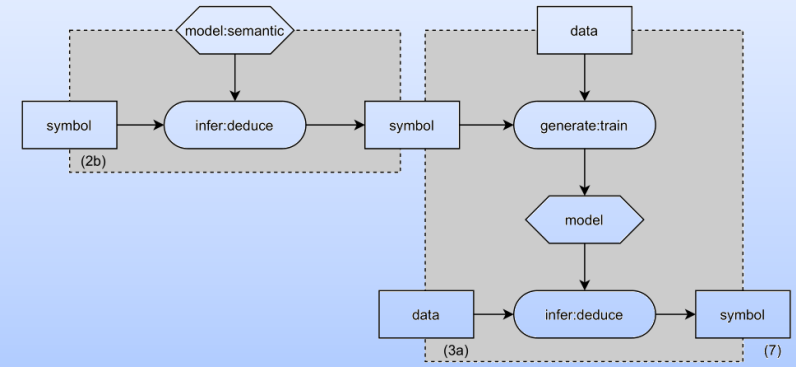
3. Trust by decomposition



[[(data+data) -> ML -> sym] -> KR -> sym]



[([(data+data) -> ML -> sym]+sym) -> ML -> sym]



[([symbol -> KR -> sym]+data+data) -> ML -> sym]

Good old program correctness:

Decompose AI system into components,

Proof properties about

components + their composition

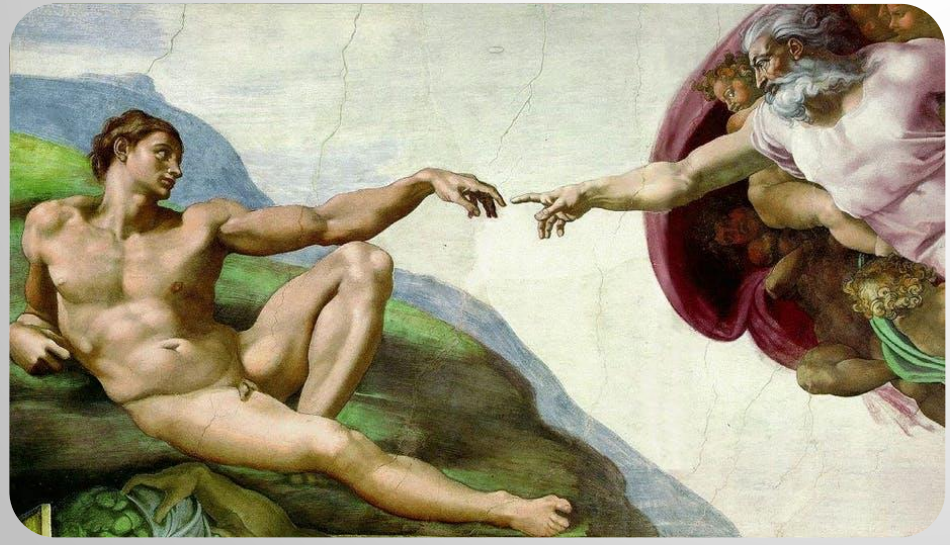
("boxology", van Harmelen et. al)

3. Trust by formal characterisation

Theorem 4.2: A logical classifier is captured by AC-GNNs
if and only if it can be expressed in graded modal logic
(or equivalently, in description logic $\mathcal{ALCQ}$)

THE LOGICAL EXPRESSIVENESS OF GRAPH NEURAL NETWORKS

Barceló et al, ICLR 2020



4. Trust & explanation by semantic loss function

flower?
cushion?

P(cushion|chair) >> P(flower|chair)

"Given the context of chair,
a cushion is much more likely
than a flower"

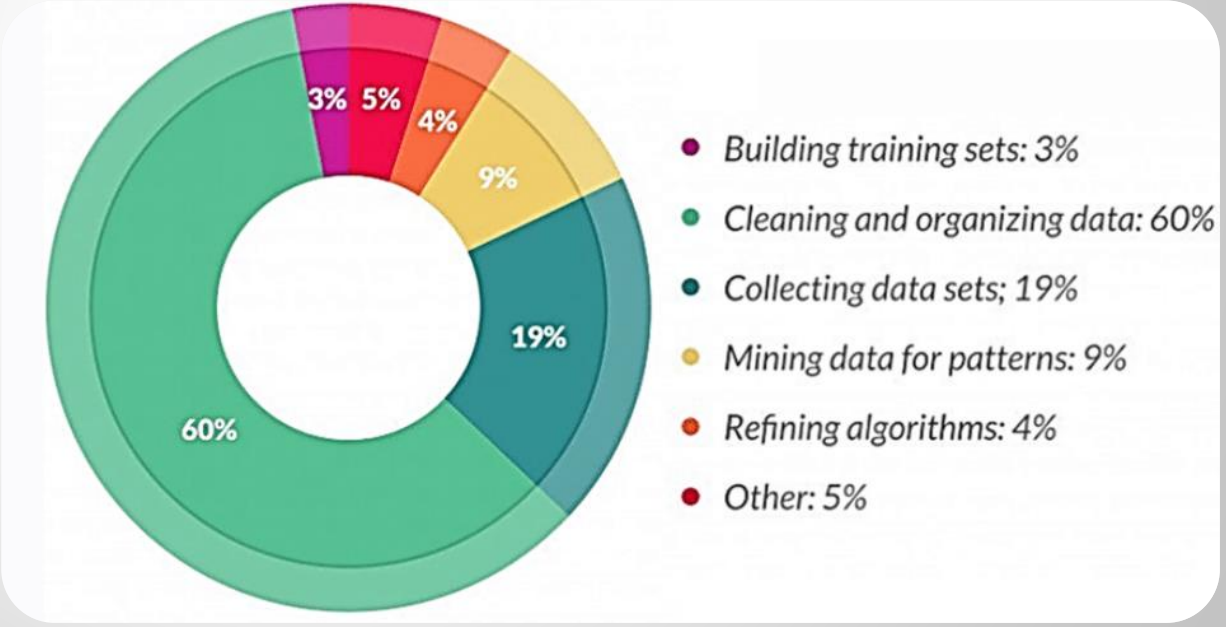
$\forall x, y$ chair$(x) \wedge$ partOf$(y, x) \rightarrow$
cushion$(y) \vee$ armRest(y)

"Parts of a chair are:
cushion and armrest"

= minimise the violation of knowledge about the world expressed in logical form

5. Trust by data provenance

The “dark 80%” of machine learning:
What do data scientists spend their time on?



Part IV:
Lessons & Recommendations for AI researchers and educators

Keep down the hype (remember the narratives?)

"a highly-trained and specialised radiologist may now be in greater danger of **being replaced by a machine** than his own executive assistant" (Andrew Ng, The Economist, 2016)

"People should stop training as radiologists now. It's just completely obvious that within 5 years, deep learning is **going to do better than radiologists**" (Geoffrey Hinton, The New Yorker, 2017)

Keep down the hype (remember the narratives?)

In research:

- A scientific paper is not a sales pitch
- Documented failures are import (but currently unpublishable)

In teaching:

- Teach the limitations as well as the successes
- Teach data science (80%), not just machine learning (20%)
- Teach all branches of AI, not just machine learning

As a community

Work *with* colleagues from humanities, social science, law before they start working *without you.*

(even employ them in your own department, eg. Nijmegen)

We should *innovate,* or else they will *legislate*