The meaningful AI Transparency Research Project

Ramak Molavi Vasse'i

The New York Times | https://www.nytimes.com/2016/09/02/technology/artificial-intelligence-ethics.html

How Tech Giants Are Devising Real Ethics for Artificial Intelligence

By John Markoff

Sept. 1, 2016

Now five of the world's largest tech companies are trying to create a standard of ethics around the creation of artificial intelligence. While science fiction has focused on the existential threat of A.I. to humans, researchers at Google's parent company, Alphabet, and those from Amazon, Facebook, IBM and Microsoft have been meeting to discuss more tangible issues, such as the impact of A.I. on jobs, transportation and even warfare.

> The authors of the Stanford report, which is titled "Artificial Intelligence and Life in 2030," argue that it will be impossible to regulate A.I. "The study panel's consensus is that attempts to regulate A.I. in general would be misguided, since there is no clear definition of A.I. (it isn't any one thing), and the risks and considerations are very different in different domains," the report says.

One main concern for people in the tech industry would be if regulators jumped in to create rules around their A.I. work. So they are trying to create a framework for a self-policing organization, though it is not clear yet how that will function.

Ethical guidelines, Ethics teams...

Since 2016:

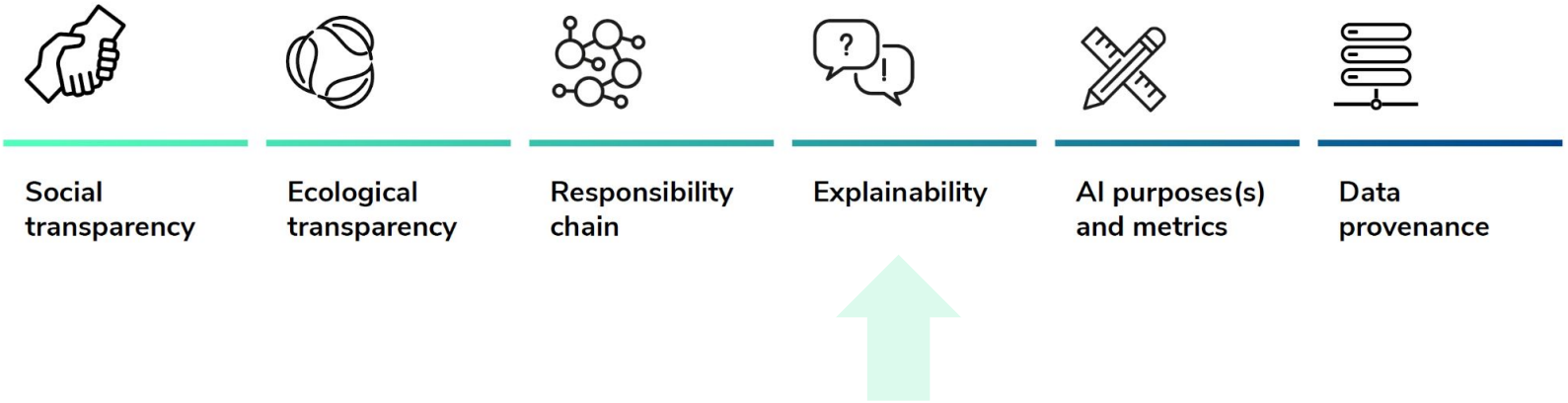
- More than 350 international ethical guidelines, most of them from industry. Almost all of them emphasise the **need for AI transparency**.
- Many ethics teams have been established.

2023:

- Twitch closed its ethics team last month. This follows similar moves at Twitter, Meta and Microsoft in the last six months.
- **How about applied AI Transparency?**
  Our research shows that AI transparency is still in its infancy.

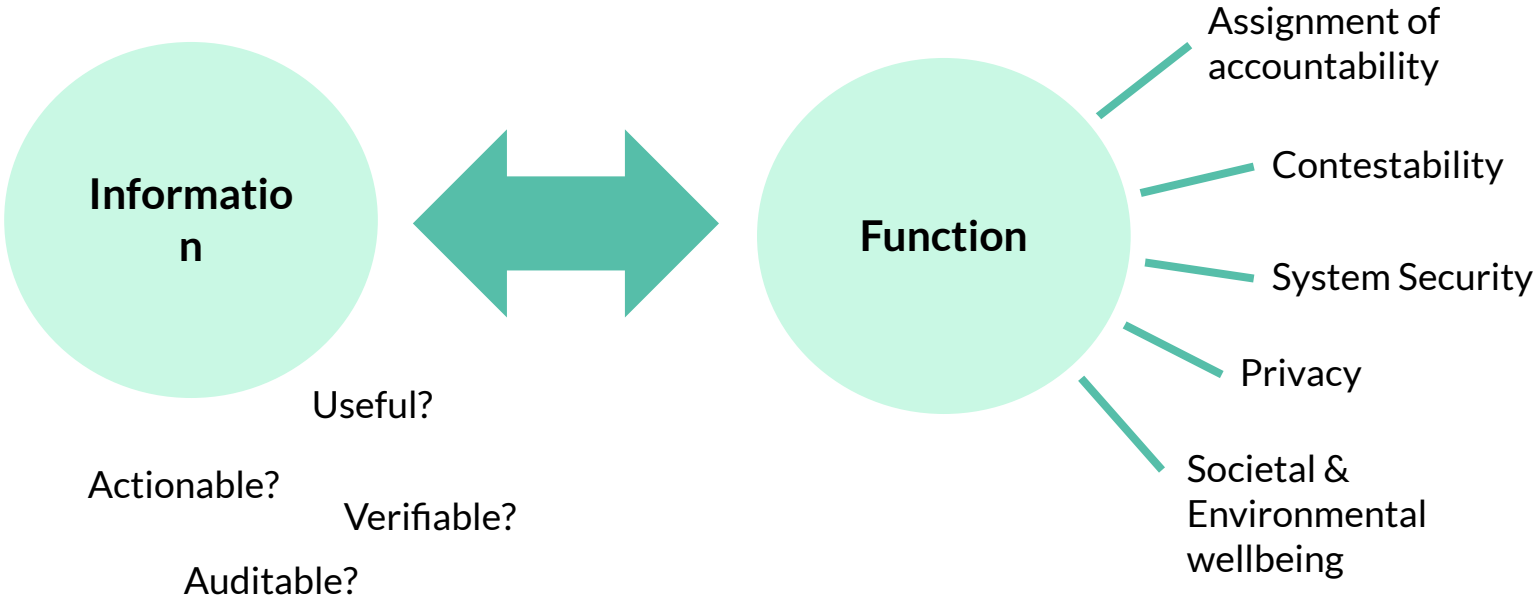
The Concept: I. Thematic Dimension

Range of Information that should be provided



The Concept: II. Functional Dimension

Transparency as a means to an end



The Research

- The Mission: XAI reality check, understand the challenges of the builders and translate them into action, help bridge the gap.

- 52 survey participants
  Machine learning Engineers, Software and automation engineers, Developers, Data scientists, Product Designer, Product Manager, QA tester.
  7 Interviews

What did we learn?

Motivating Factors

1. Accuracy and target goal achievement
2. Gain new insights by investigating learned prediction strategies
3. Impact assessment to avoid unwanted outcomes
4. Avoid bias
5. Justify decisions to subjects and other stakeholders
6. Enable user control
7. Increase security
8. Verify generalizability of the model
9. Disclose knowable information
10. Improve system robustness
11. Compliance with ethical guidelines / internal code of conduct
12. Legal compliance / audits

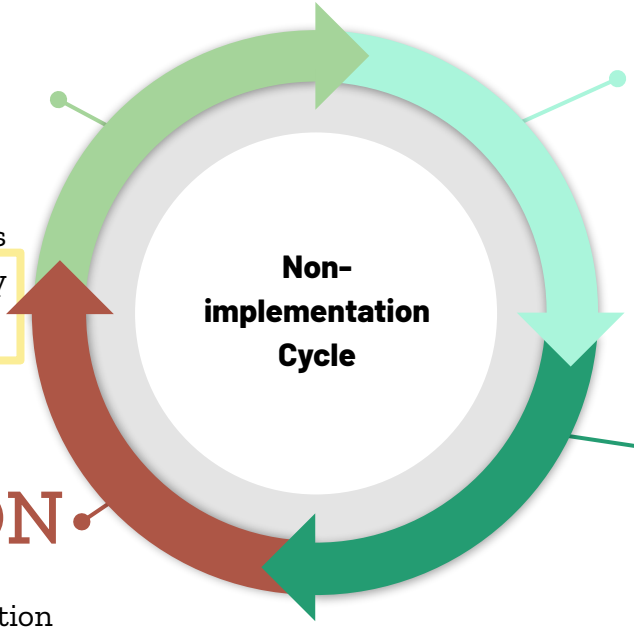
90%

Of respondents ranked ethical guidelines 11th on a 12 point scale

Mapping the Challenges

METHODS

- Lack of fitting tools
- Lack of resources
- Lack of reliable explanations
- How to deliver transparency to different stakeholders



ETHICS

- Lack of intrinsic motivation
- Lack of education /awareness /maturity
- Lack of work ethics/ oaths

INCENTIVES

- Lack of business incentives
- Lack of public interest & pressure

REGULATION

- (Perceived?) Lack of regulation
- Lack of harmonized standards

Ranking of Challenges & Obstacles

1. No clear guidance how to choose among explainable AI methods
2. The outcomes (explanations) are too ambigious (unclear) and/or still too complex
3. Decrease of development speed
4. Explanations are not correct
4. Lack of clear objectives/KPIs to build an incentive around transparency deep dive
4. Explanation of AI is not part of the education of ML professionals

5

- Absence of strandardized evaluation methods
- Cost of transparency implementation
- Lack of buy-in from CEO/Lead
- Lack of clear accountability for the transparency topic
- Transparency measure could enable malicious users to increase capabilities and performance of undesirable systems

6

- Lack of resources
- Lack of whistleblower protection for employees

7

- Does not align with our work practices
- Out transparency efforts are pure ethics washing

8

- Lack of internal expertise on how to use explainability techniques

Interpretable-by-design models

- Explainability tools are useful in some cases, but are fundamentally limited.
- Guidance: Is interpretability is a design requirement? The use of interpretable models is recommended.
- In many cases, an interpretable model can be just as accurate as the best black box model.

What comes next?

Help to shift industry norms & inform enforcement bodies/ AI auditing

1

Status quo & gaps 2022/2023

2

Guidance on Transparency Requirements

Art 13 AI Act

The "form" of Transparency

Transparency delivery 2023

*"High-risk AI systems shall be designed and developed in such a way to ensure that their operation is **sufficiently transparent** to enable users to interpret the system's output and use it **appropriately**." Art 13 AIA*


E.M.A.K
DEEPFAKE

Help to shift industry norms & inform enforcement bodies/ AI auditing

1

Status quo & gaps
2022/2023

2

Guidance on Transparency Requirements

Art 13 AI Act

Transparency delivery
2023

3

Synthetic content labeling & disclosure

Art 52 AI Act

AI disclosure
2023

4

Thank you.

If you would like to connect
or be involved in our research,
get in touch with us!

ramak@mozillafoundation.org