NOTTS - AI Monthly Catch-up - June 2019

Ethics Guidelines for Trustworthy AI

While we wait...are you on slack yet?
Yes, then good :)

No, here is a cryptic invite
- Head to https://notts.ai/
- Bottom right corner (footer)
- Find the invite

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Police facial recognition surveillance court case starts

By Clive Coleman
Legal correspondent, BBC News

Is Google Duplex ethical and moral?

Chris Butler Follow
May 10, 2018 · 5 min read

Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot

Company finally apologises after 'Tay' quickly learned to produce offensive posts, forcing the tech giant to shut it down after just 16 hours

The 'creepy Facebook AI' story that captivated the media

By Chris Baraniuk
Technology reporter

"Why Should I Trust You?"

“Its Magic…..
I Owe no
explanation"

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

Bias can occur when,

Framing the problem.
Collecting the data.
Preparing the data.

why it’s so hard to fix bias

Unknown unknowns.
Imperfect processes.
Lack of social context.
The definitions of fairness.

AI HLEG

Vision

Increasing public and private investments in AI to boost its uptake,

Preparing for socio-economic changes, and

Ensuring an appropriate ethical and legal framework to strengthen European values.

Needs to be

human-centric.

maximise the benefits of AI systems.

preventing and minimising their risks.

Trustworthy AI as foundational ambition.

Components

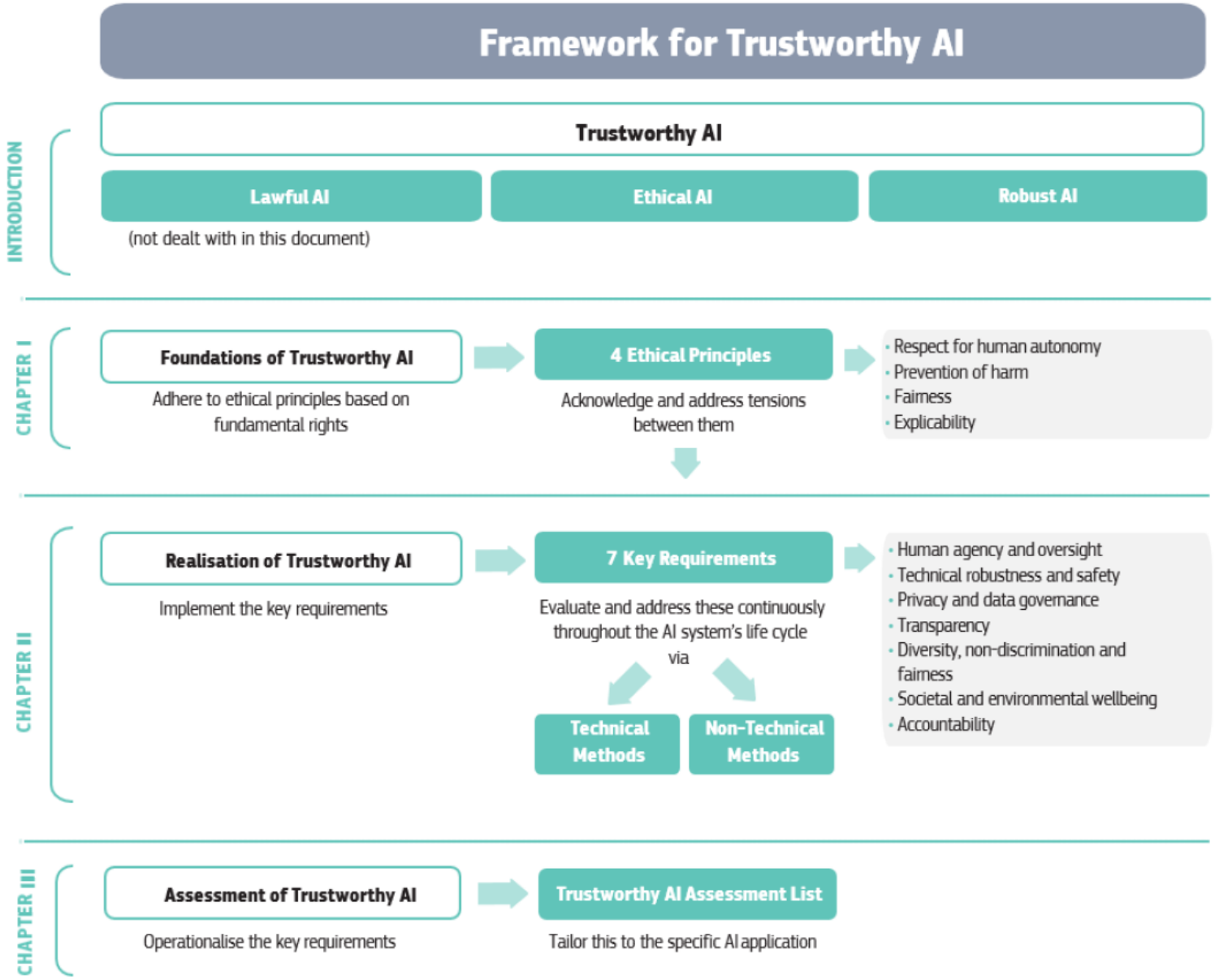
Lawful

Ethical

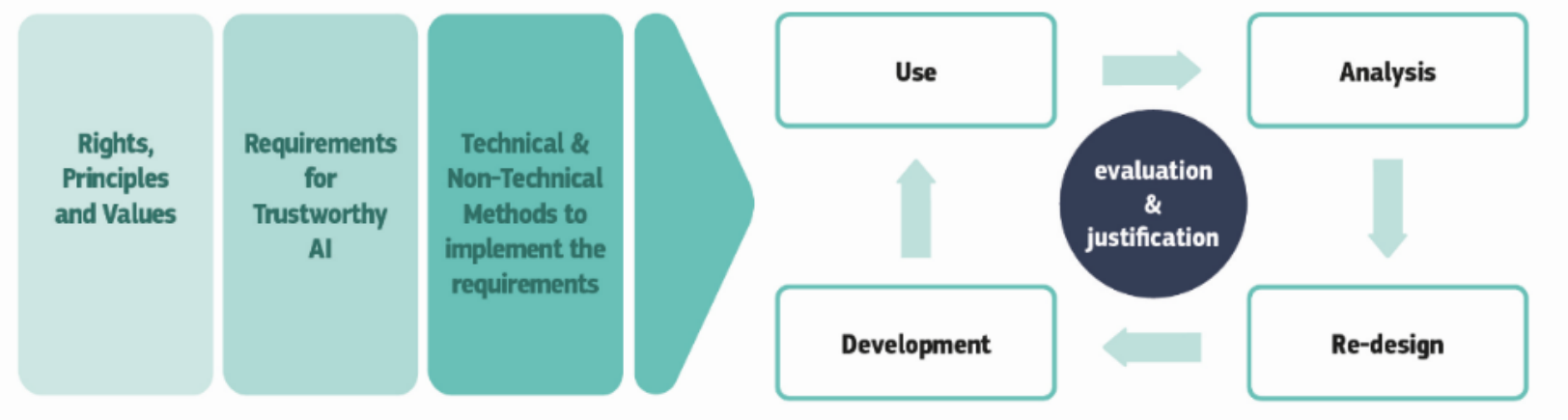
Robust

More information on the High-Level Expert Group on Artificial Intelligence is available online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

Guidelines as a Framework



Foundations of Trustworthy AI



The principle of respect for human autonomy

Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans

The principle of fairness

ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.

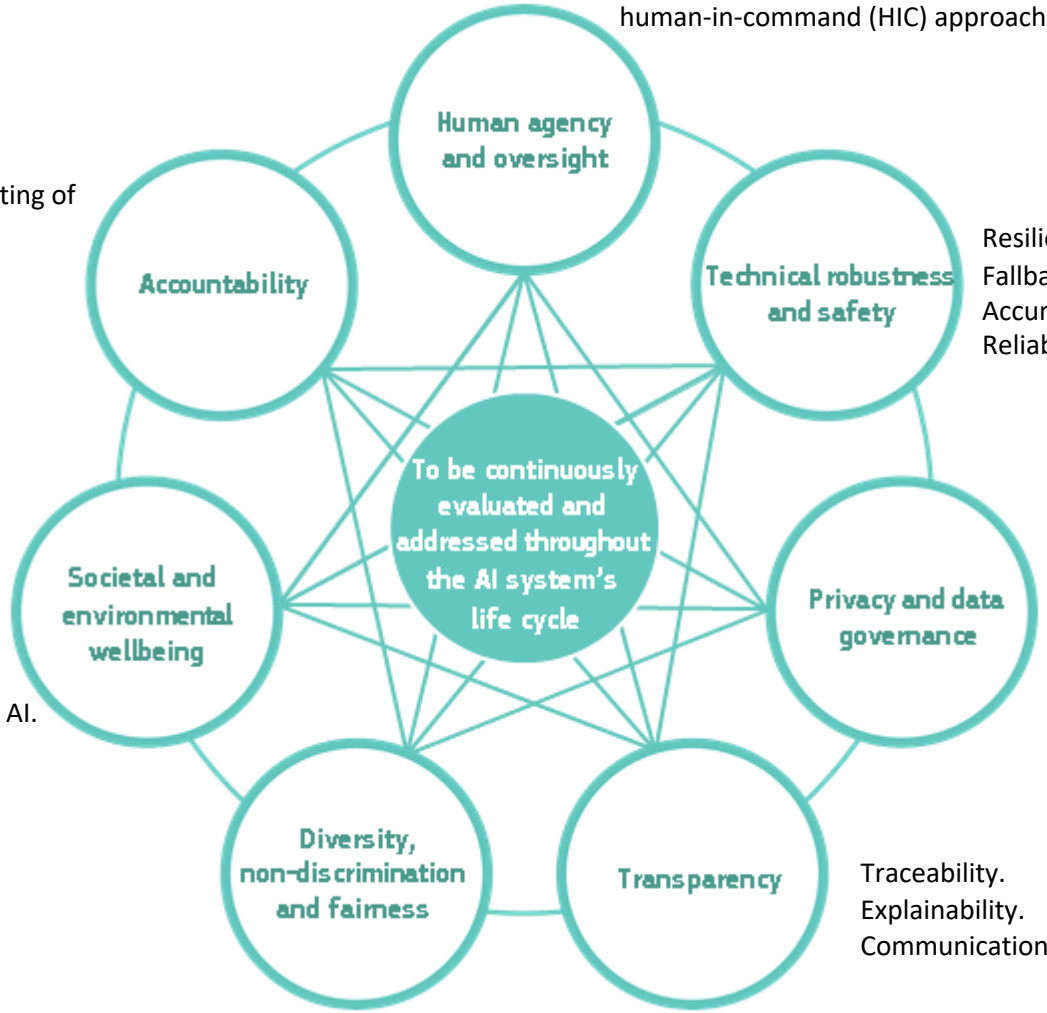
The principle of prevention of harm

AI systems should neither cause nor exacerbate harm. They must be technically robust and it should be ensured that they are not open to malicious use.

The principle of explicability

An explanation as to why a model has generated a particular output or decision. Cases where this is not possible (‘Black Box’) need special attention. Traceability, auditability and transparent communication on system capabilities required.

Requirements of Trustworthy AI



human-in-the-loop (HITL).
human-on-the-loop (HOTL).
human-in-command (HIC) approach.

Resilience to attack and security.
Fallback plan and general safety.
Accuracy.
Reliability and Reproducibility.

Data Protection.
Quality and integrity of data.
Access to data.

Traceability.
Explainability.
Communication.

Avoidance of unfair bias.
Accessibility and universal design.
Stakeholder Participation.

Sustainable and environmentally friendly AI.
Social impact.
Society and Democracy.

Auditability.
Minimisation and reporting of negative impacts.
Trade-offs.
Redress.

Assessing Trustworthy AI

1. Human agency and oversight

- ✓ Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?
  - Which detection and response mechanisms did you establish to assess whether something could go wrong?

2. Technical robustness and safety

Resilience to attack and security:

- ✓ Did you assess potential forms of attacks to which the AI system could be vulnerable?
  - Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?

7. Accountability

Auditability:

- ✓ Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?

Register for the Piloting Process

https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/register-piloting-process-0

Feedback will be gathered through an online survey, which will be launched in June 2019.

Based on this feedback, the High-Level Expert Group on AI (AI HLEG) will propose a revised version of the assessment list to the Commission in early 2020.

What tools can help me out

IBM's AI Fairness 360 (Open Source)

A comprehensive set of fairness metrics for datasets and machine learning models, explanations for these metrics, and algorithms to mitigate bias in datasets and models

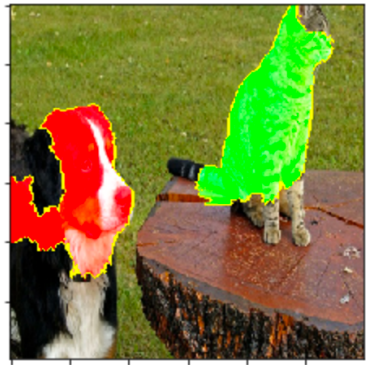
http://aif360.mybluemix.net/

LIME (Open Source)

Explaining the predictions of any machine learning classifier

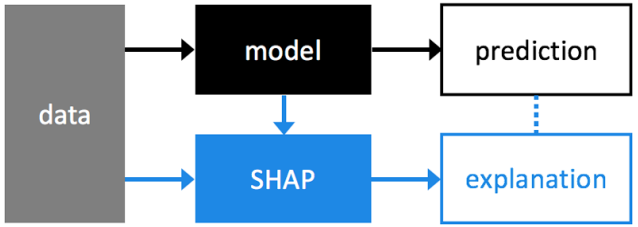
https://www.youtube.com/watch?v=hUnRCxnydCc

https://github.com/marcotcr/lime



SHAP (Open Source)

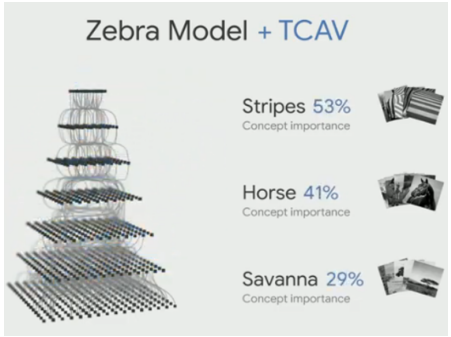
A unified approach to explain the output of any machine learning model.
https://github.com/slundberg/shap



Google's TCAV

Testing with Concept Activation Vectors (TCAV) is a new interpretability method to understand what signals your neural networks models uses for prediction.

https://github.com/tensorflow/tcav



Monthly Challenge (June 2019) – Data Visualisation

Idea: ‘World Cloud’ displaying interests of all the members.

Details: Read Meetup profile data. Lookup Interests. Generate a word cloud.

Dataset: Meetup profiles

Elsewhere on the Internet

- Samsung AI Makes the Mona Lisa 'Speak' https://www.youtube.com/watch?v=P2uZF-5F1wI

- Adversarial patch to fool an AI system https://youtu.be/MIbFvK2S9g8
  https://arxiv.org/abs/1904.08653