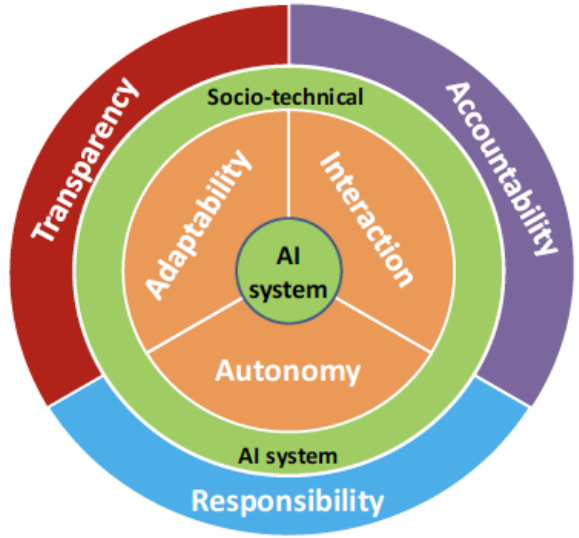
The ART Principles

Accountability, Responsibility, Transparency

Outline

- Beneficial/Harmful AI Systems
- Characteristics of Trustworthy Autonomous Systems
  - Autonomy, Adaptability, Interaction
- The ART Principles
  - Accountability
  - Responsibility
  - Transparency



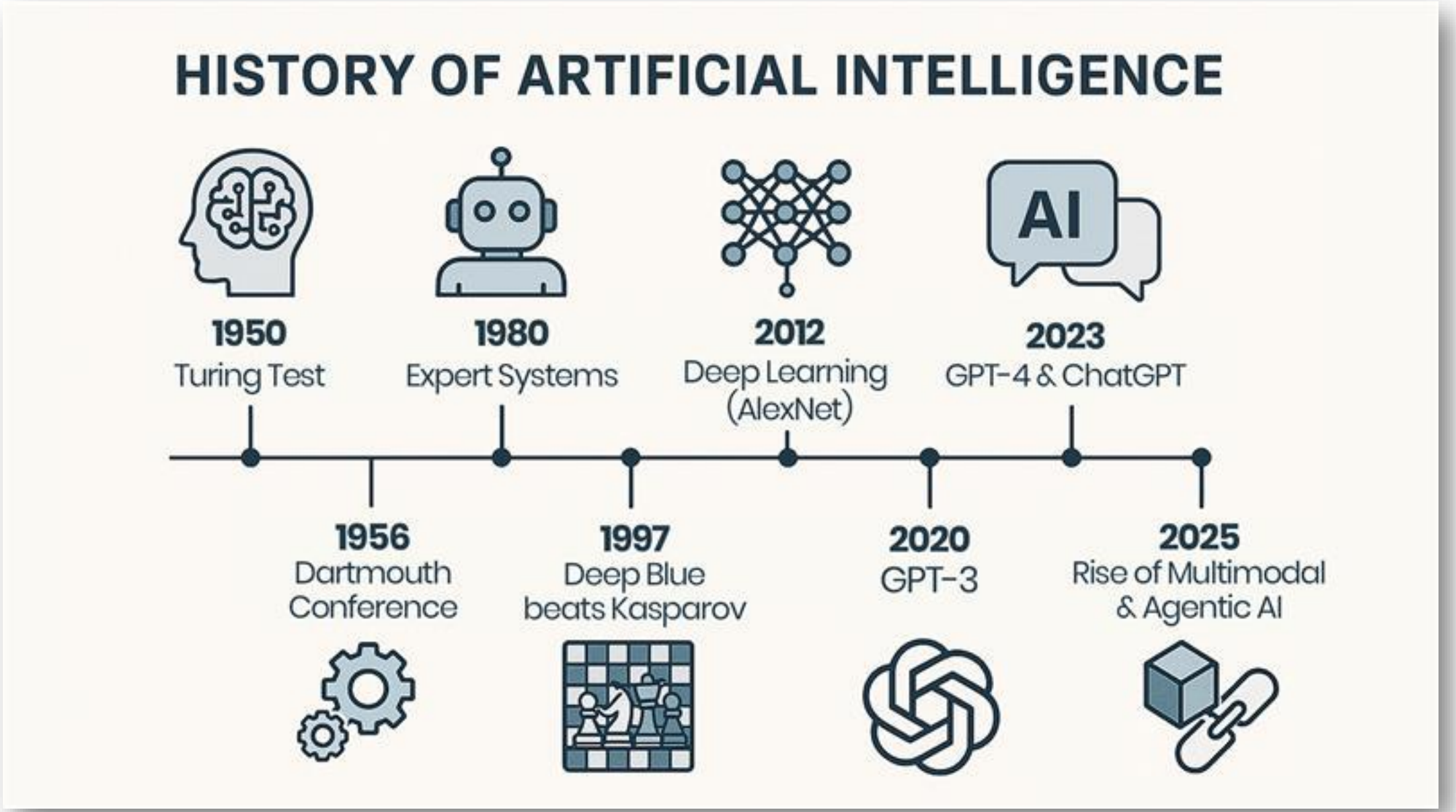
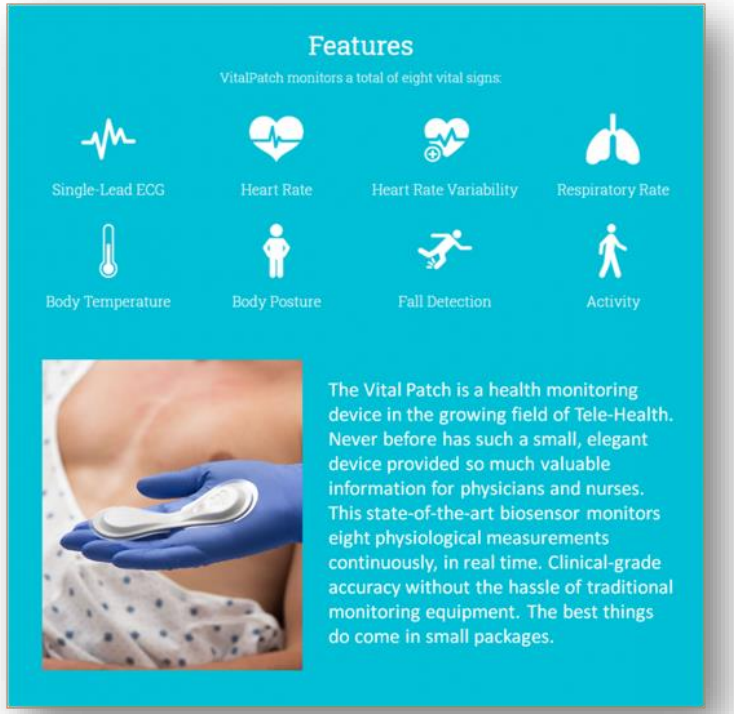
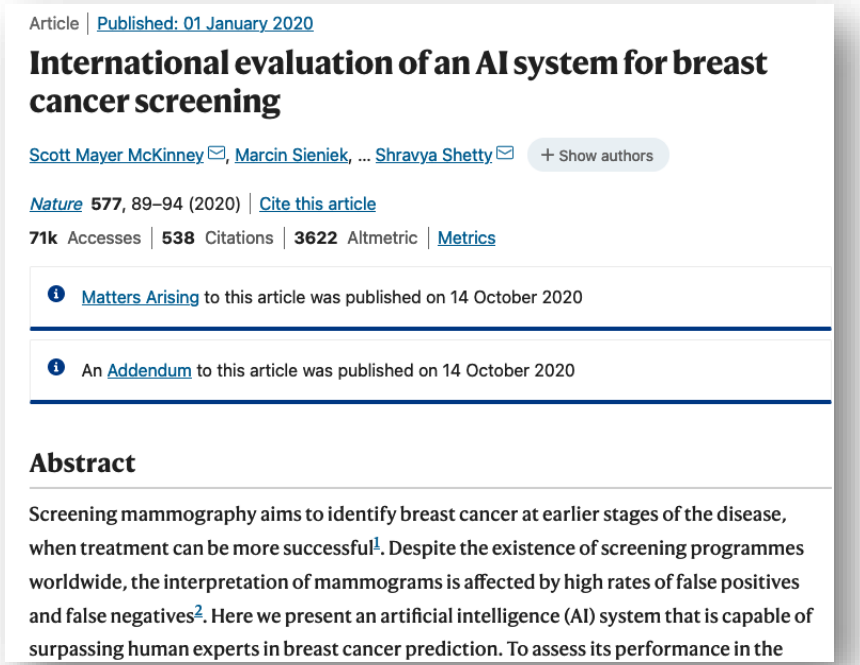


Image: https://www.linkedin.com/pulse/history-artificial-intelligence-timeline-innovation-from-bansal-cpa-asw6f

AI has great potential (if controlled)

- AI can bring significant benefits to society.
  - e.g., climate change, cure to diseases …


Features
VitalPatch monitors a total of eight vital signs:
Single-Lead ECG
Heart Rate
Heart Rate Variability
Respiratory Rate
Body Temperature
Body Posture
Fall Detection
Activity
The Vital Patch is a health monitoring device in the growing field of Tele-Health. Never before has such a small, elegant device provided so much valuable information for physicians and nurses. This state-of-the-art biosensor monitors eight physiological measurements continuously, in real time. Clinical-grade accuracy without the hassle of traditional monitoring equipment. The best things do come in small packages.


Article | Published: 01 January 2020
International evaluation of an AI system for breast cancer screening
Scott Mayer McKinney, Marcin Sieniek, ... Shravya Shetty + Show authors
Nature 577, 89–94 (2020) | Cite this article
71k Accesses | 538 Citations | 3622 Altmetric | Metrics
Matters Arising to this article was published on 14 October 2020
An Addendum to this article was published on 14 October 2020
Abstract
Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To assess its performance in the


Edinburgh
Holyrood Park
Blackford
Currie
Dalkeith
Swanston


Start learning
Scottish Gaelic!
balach
cù
cat
muc

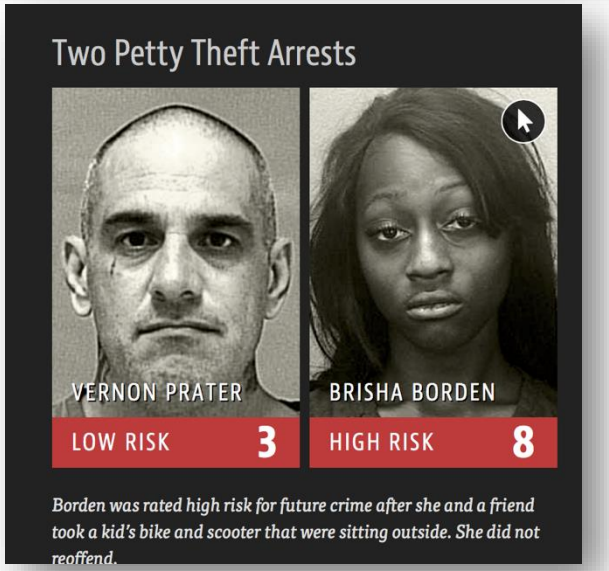
AI has great potential (if controlled)

- AI can bring significant benefits to society.
  - e.g., climate change, cure to diseases …
- AI can produce undesirable impacts.
  - e.g., amplifying biases, discrimination, misinformation, manipulation …

Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case

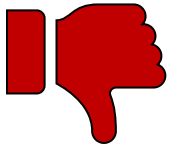
By Jay Stanley, Senior Policy Analyst, ACLU Speech, Privacy, and Technology Project
JUNE 2, 2017 | 1:30 PM

TAGS: Privacy & Technology



Generative AI

- Pattern Discovery
  - Original outputs
- Enhancing Learning
  - An assistant to help with writing
- Customer Engagement
  - Customized chatbots



- Hallucinations
  - Retrieval Augmented Generation
- Ethical concerns
  - Bias, Privacy, Trustworthiness
- Intellectual Property Issues

We need to find an ethically acceptable way of designing technology.

The Landscape of AI Ethics Principles

- A Google Scholar search reveals >2.5M results for "AI Ethics Principles" query.

- Jobin *et al.* Analyzed 84 papers to produce AI Ethics principles.

Perspective | Published: 02 September 2019

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca & Effy Vayena

Nature Machine Intelligence **1**, 389–399 (2019) | Cite this article

81k Accesses | **4305** Citations | **832** Altmetric | Metrics

Abstract

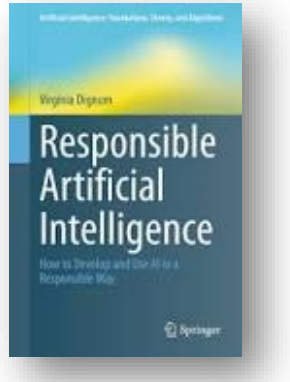
In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

Findings from Jobin *et al.*'s paper

- **Transparency** (appeared in 87% of the documents),
- **Justice and Fairness** (81%),
- **Non-maleficence** (71%),
- **Accountability/Responsibility** (71%),
- **Privacy** (56%),
- **Beneficence** (49%),
- **Freedom and Autonomy** (40%),
- **Trust** (33%),
- **Sustainability** (17%), **Dignity** (15%), and **Solidarity** (7%).

*Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell **1,** 389–399 (2019).*

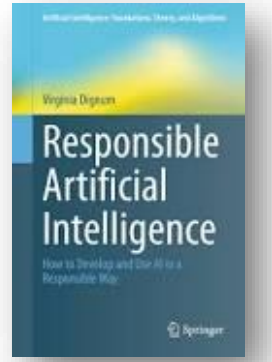
Characteristics of AI Systems



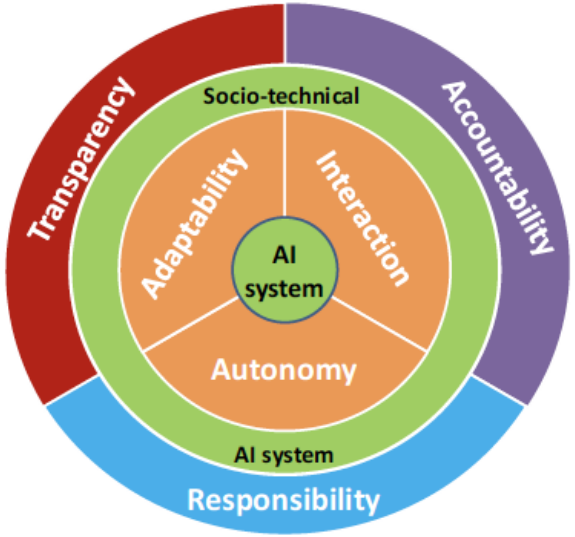
- Autonomy
  - deciding on an action
- Adaptability
  - learning from the environment, adapting its behavior
- Interaction
  - communicating with other agents in the environment

Floridi, L., and Sanders, J. On the morality of artificial agents. Minds and Machines 14, 3 (2004), 349-379.

The ART Principles for Trustworthy Autonomous Systems



- Accountability
- Responsibility
- Transparency



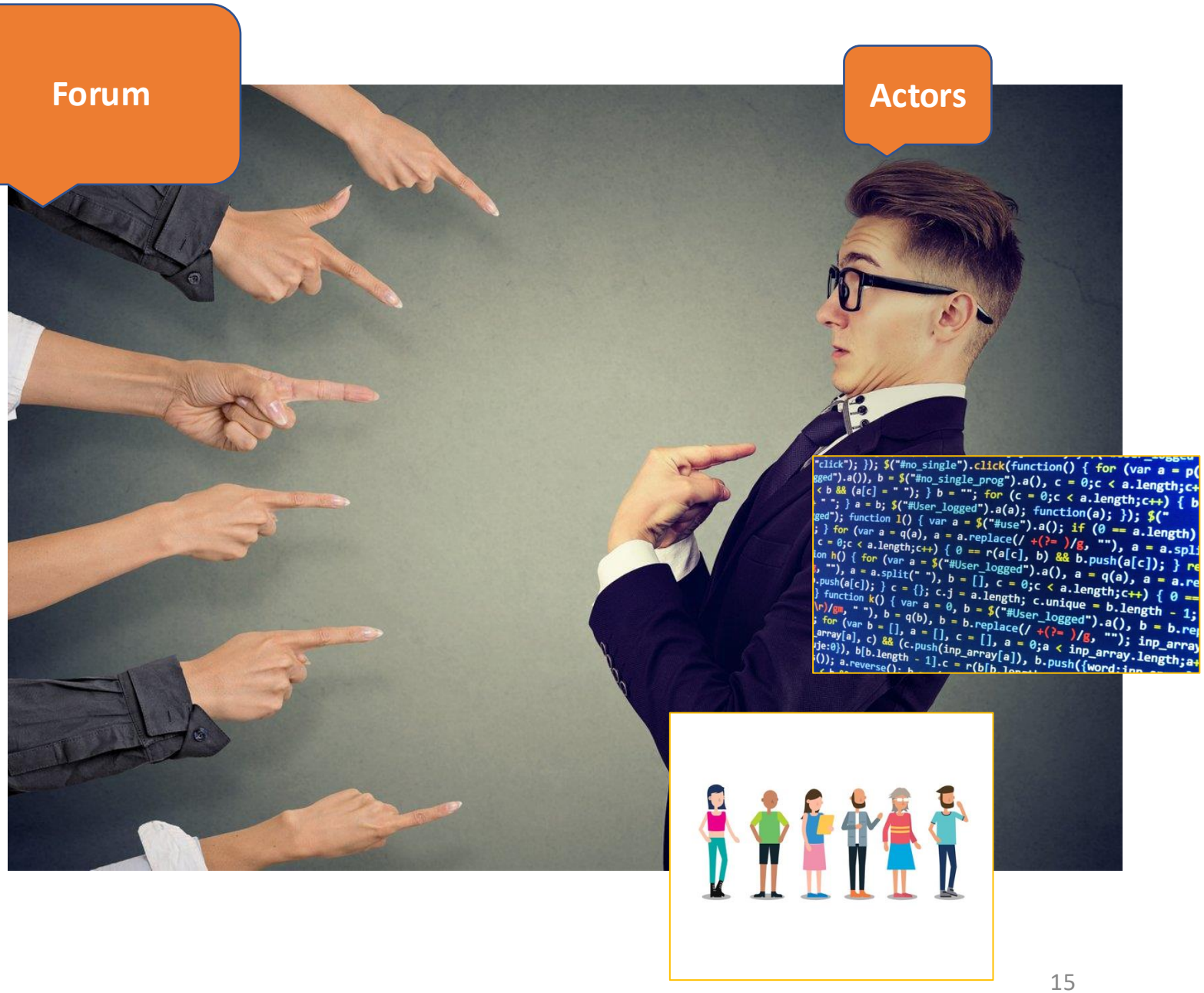
Required to build social trust

Accountability

ART

Accountability

- The actor has an obligation to explain.
- The forum can pose questions.
- The actor may face consequences.



(Algorithmic) Accountability

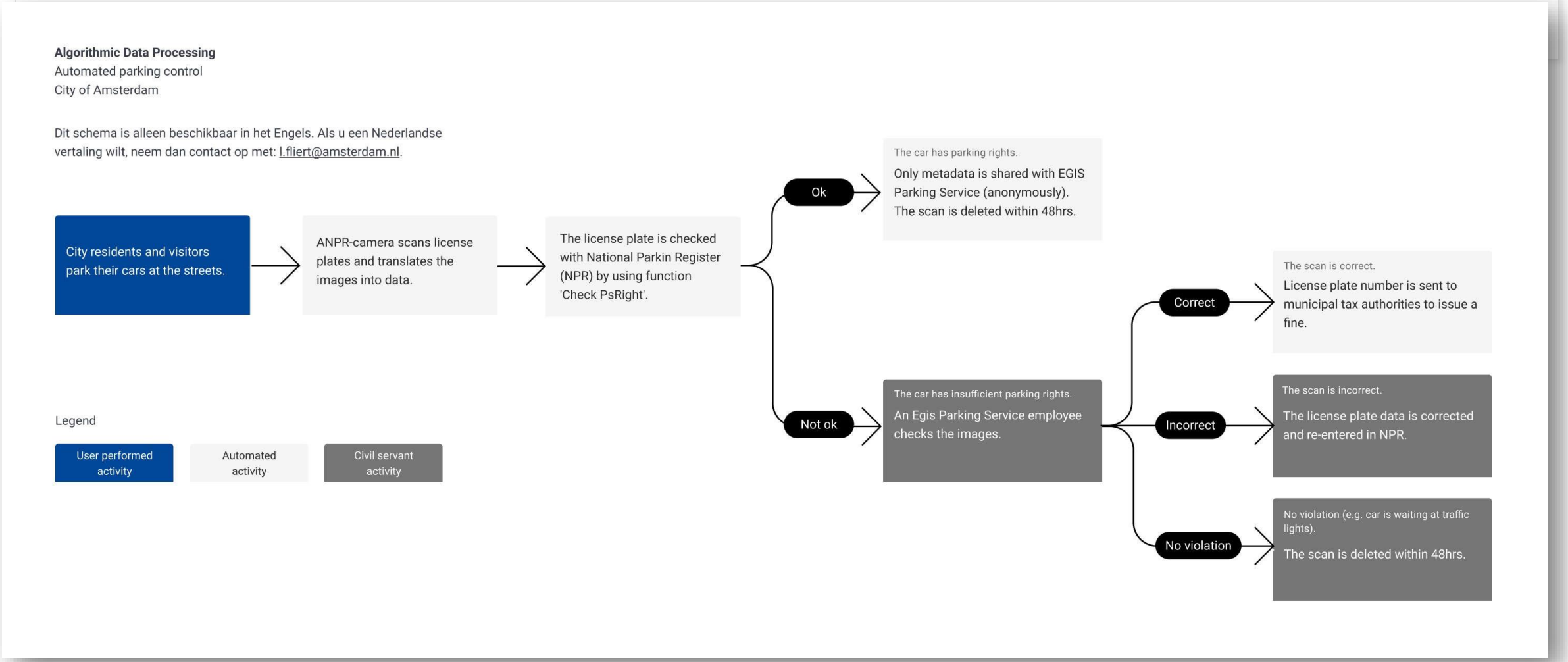
- Things may often go wrong...
- When it is the case, we want to assign blame... and start to look for accountable/responsible (human) agents [if we are lucky to find!]
- A new trend is blaming AI or the algorithms that make such decisions.

(Algorithmic) Accountability

- Accountability requires finding moral (ethical) or legal agents (e.g., people who are designing, deploying algorithms in organizations).
- Accountability is related to moral agency.
  - An agent should be able to act with reference to right and wrong.
- Under different ethical theories, the moral agent will be accountable accordingly (e.g., Rescue Robot).

Automated Parking Control (Amsterdam)

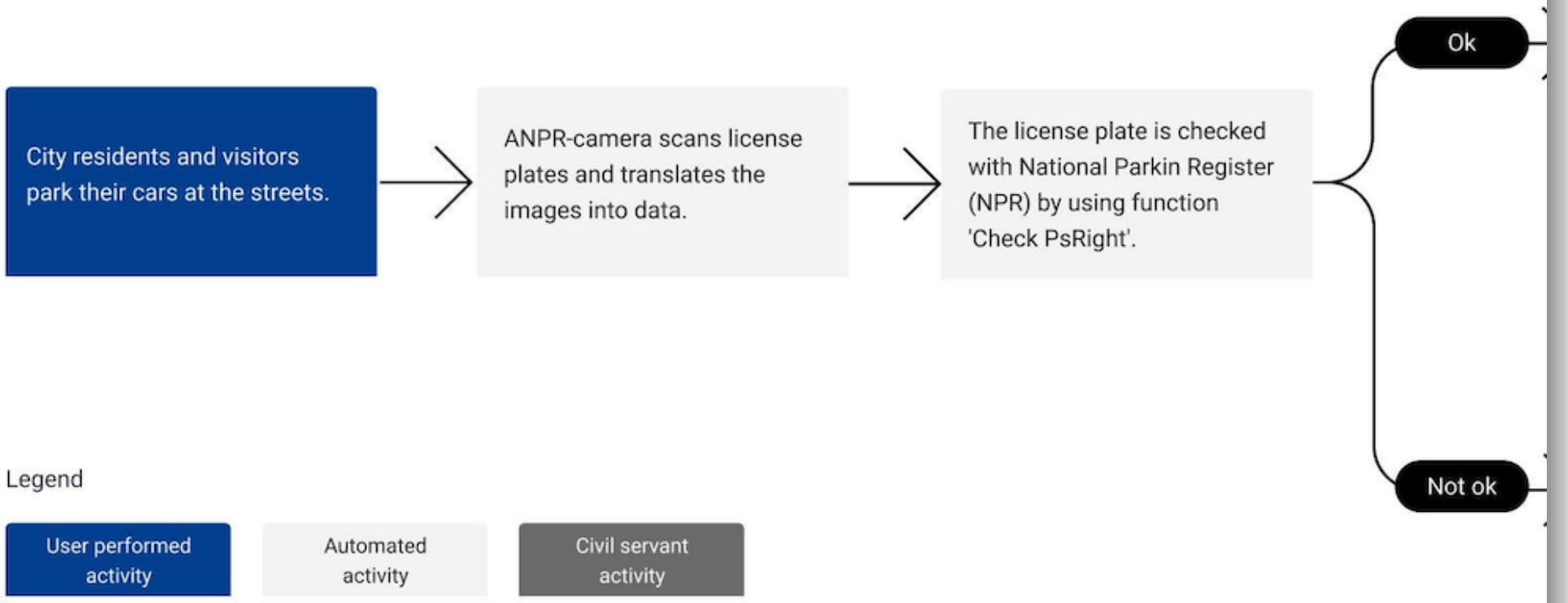
Automated Parking Control (Amsterdam)



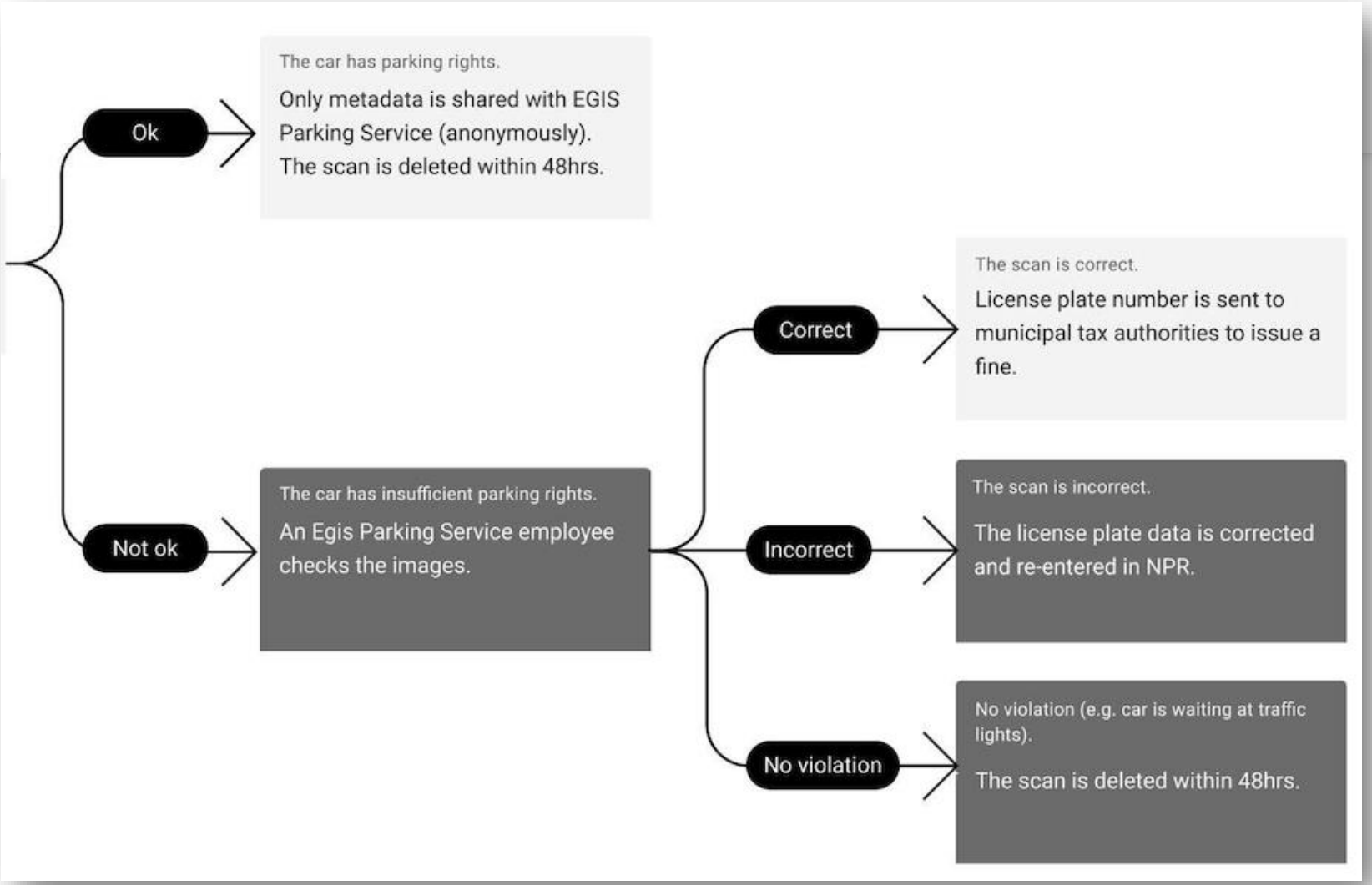
https://open.amsterdam/woo-zoeken/detail/993959fb-b6d0-4c46-aab1-617e69100e89 (last document update: 21 May 2025)

Algorithmic Data Processing
Automated parking control
City of Amsterdam

Dit schema is alleen beschikbaar in het Engels. Als u een Nederlandse vertaling wilt, neem dan contact op met: l.fliert@amsterdam.nl.



https://open.amsterdam/woo-zoeken/detail/993959fb-b6d0-4c46-aab1-617e69100e89 (last document update: 21 May 2025)



https://open.amsterdam/woo-zoeken/detail/993959fb-b6d0-4c46-aab1-617e69100e89 (last document update: 21 May 2025)

Reflection Time

1. What are the benefits?
2. What are the harms?

Responsibility

ART

Responsible AI

- Responsible AI provides directions for action, i.e., a code of behavior for AI systems and people.
- The consequences of decisions made can be ethically significant
  - Autonomous systems may still behave in a non-ethical manner.
- AI systems that put human well-being at the core of the development process are also likely to be adopted by humans.

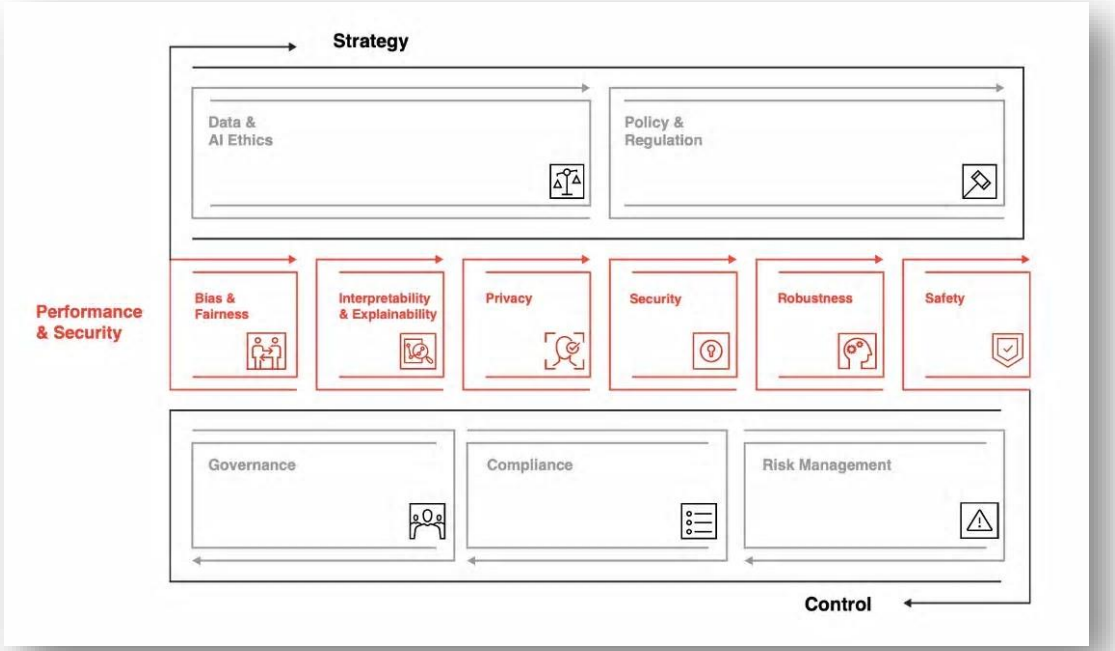
Responsible AI --- in practice

Google

Recommended practices

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment

pwc



Chatbots and Legal Responsibility

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

The judge wrote that Air Canada's customers had no way of knowing which part of its website – including its chatbot – relayed the correct information. Photograph: NurPhoto/Getty Images

- Air Canada tried to claim the bot was a *separate legal entity.* This didn't work!

"It makes no difference whether the information comes from a static page or a chatbot."

https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit (16 Feb 2024)

Transparency

ART

Transparency

- Many other terms: "explainability", "understandability", "interpretability"
- Transparency in AI:
  - supports access to justifications for decisions when needed. In public sector, people should also know how to contest and appeal.
  - addresses the right to know (e.g., GDPR).
  - helps in understanding and managing risks.

Why is transparency hard?

- Many stakeholders are involved…
- Contexts, user profiles, questions to be answered vary largely.
- How to explain the workings of a "black box" model?
  - Explanations could be added by design (e.g., interactive interfaces are great to explore models)
  - The use of simpler models works sometimes!
- How much transparency should we provide?

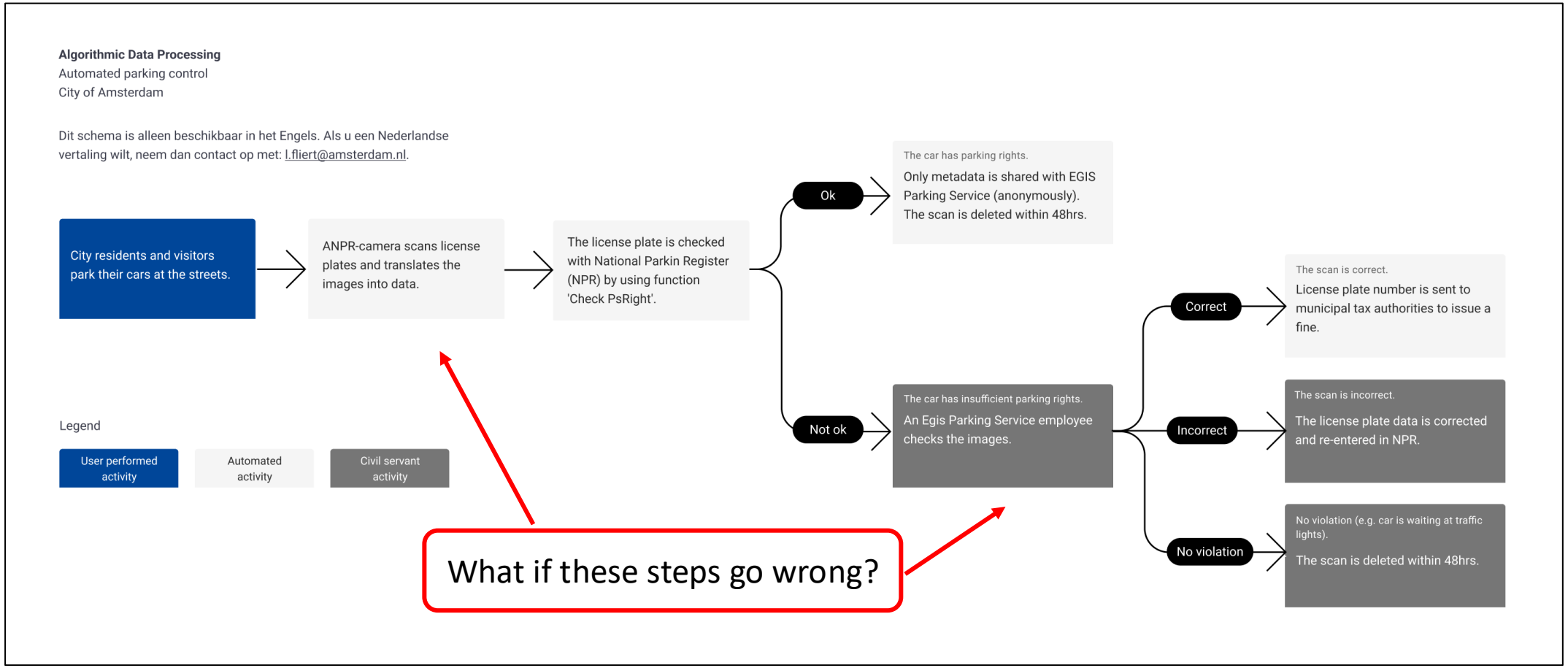
Major Findings from the literature on explanations

According to Miller, explanations are:

- Contrastive
  "Why event P happened instead of some event Q?"
- Selected (influenced by cognitive biases)
  (Partial) explanations are based on selected factors
- Not driven by probabilities
  Effective explanations are causal, not the most likely explanations
- Social/interactive
  Explanations for the user

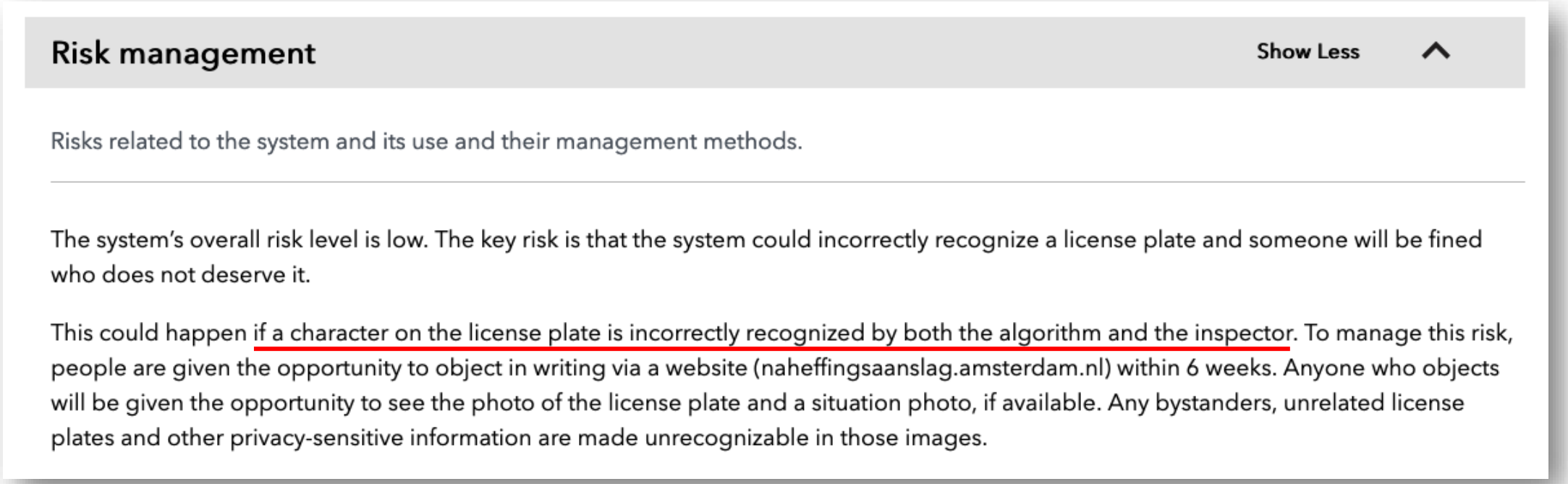
Miller, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267 (2019), 1-38.

Transparency: Automated Parking Control



https://open.amsterdam/woo-zoeken/detail/993959fb-b6d0-4c46-aab1-617e69100e89 (last document update: 21 May 2025)

Transparency: Automated Parking Control



Risk management

Show Less

Risks related to the system and its use and their management methods.

The system's overall risk level is low. The key risk is that the system could incorrectly recognize a license plate and someone will be fined who does not deserve it.

This could happen if a character on the license plate is incorrectly recognized by both the algorithm and the inspector. To manage this risk, people are given the opportunity to object in writing via a website (naheffingsaanslag.amsterdam.nl) within 6 weeks. Anyone who objects will be given the opportunity to see the photo of the license plate and a situation photo, if available. Any bystanders, unrelated license plates and other privacy-sensitive information are made unrecognizable in those images.

Transparency: Automated Parking Control

Data processing

Show Less

The operational logic of the automatic data processing and reasoning performed by the system and the models used.

Model architecture

The service uses license plate recognition algorithms to locate and process the license plate data from the camera data stream. Algorithms are used to locate the license plate from the image data, to adjust the images for identification, to identify the individual characters of the license plate, and to validate the plate contents against national license plate characteristics.

After a successful plate identification and processing, license plate data is sent to the National Parking Register for further processing. NPR's algorithm checks the validity of parking rights for the license plate in a given time and location (for technical information on the NPR algorithm, see the information on their website: https://nationaalparkeerregister.nl/fileadmin/files/Mobiel_parkeren /Interface_Description_v7.6.pdf). A positive response means the car has valid parking rights in place, and the license plate scan data can be removed in 48 hours. For license plates with invalid parking rights, the case is transferred to the cities tax department, which connects to the RDW database to link the license plate with the car ownership data, and to deliver a fine.

Content	Attachment
System architecture description	Automated parking control Attach architecture image

They provide 58 pages to explain the algorithm!

Summary

- Beneficial/Harmful AI Systems
- Characteristics of Trustworthy Autonomous Systems
  - Autonomy, Adaptability, Interaction
- The ART Principles
  - Accountability
  - Responsibility
  - Transparency