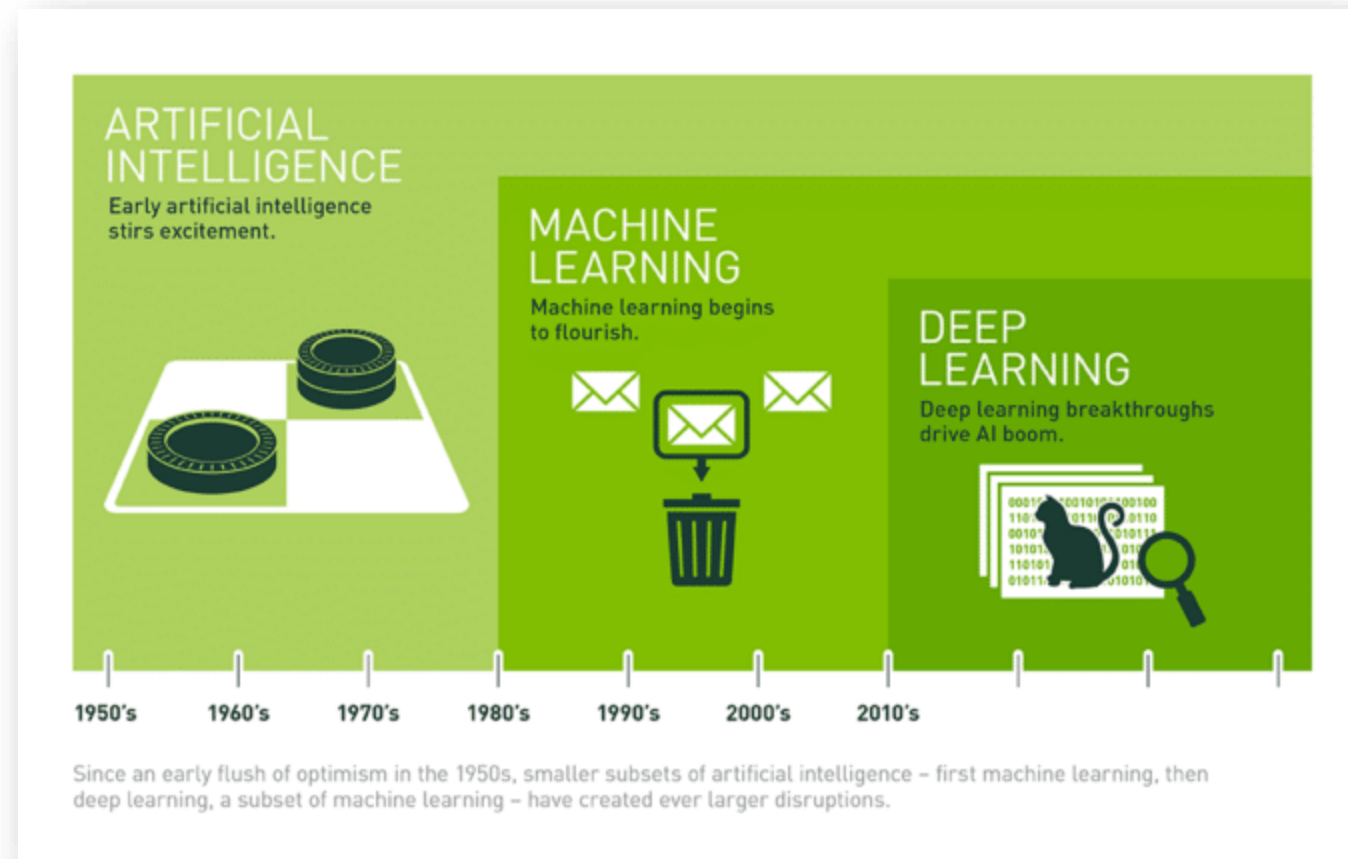


The ART Principles

Accountability, Responsibility, Transparency

AI is not ML, DL ...






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](#)

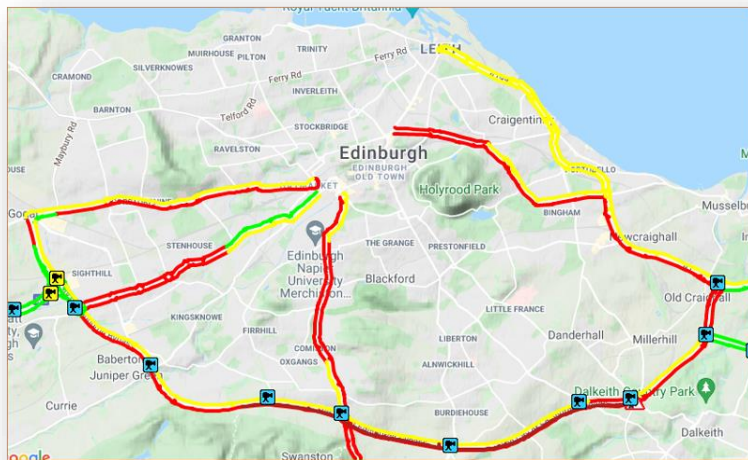
71k Accesses | **538** Citations | **3622** Altmetric | [Metrics](#)

 [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 successful¹. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives². 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



Start learning Scottish Gaelic!



balach



cú



cat



muc



AI has great potential (if controlled)

- AI can bring significant benefits to society.
 - e.g., climate change, cure to diseases ...
- As we mentioned so far in the lectures, AI can produce undesirable impacts.
 - e.g., amplifying biases, discrimination, misinformation, manipulation ...

Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case



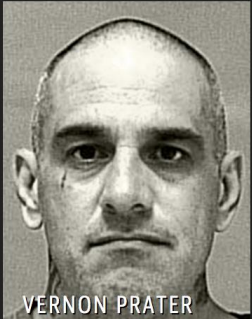

By [Jay Stanley](#), Senior Policy Analyst, ACLU Speech, Privacy, and Technology Project

JUNE 2, 2017 | 1:30 PM

TAGS: [Privacy & Technology](#)



Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Self-driving Uber car involved in fatal accident in Arizona

It's believed to be the first pedestrian fatality attributed to a self-driving vehicle.



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**

AI has great potential (if controlled)

- AI can bring significant benefits to society.
 - e.g., climate change, cure to diseases ...
- As we mentioned so far in the lectures, AI can produce undesirable impacts.
 - e.g., amplifying biases, discrimination, misinformation, manipulation ...
- We need to find an ethically acceptable way of designing technology that can benefit the society.

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

Fix Technology by more Technology!

DO NOT TRY THIS AT HOME... OR ANYWHERE, FOR THAT MATTER —

Alexa suggests 10-year-old put a penny on partially exposed plug

"Alexa: Stop recommending stupid and dangerous things."

ERIC BANGEMAN - DEC 28, 2021 7:10 PM UTC



"Customer trust is at the center of everything we do and Alexa is designed to provide accurate, relevant, and helpful information to customers," an Amazon spokesperson said in a statement. "As soon as we became aware of this error, we took swift action to fix it."

The Landscape of AI Ethics Principles

- A Google Scholar search reveals 82100 results for "AI Ethics Principles" query.
- Reaching a unique set of AI ethics principles is almost impossible.
- We will focus on some common themes.
- 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 Lenca](#) & [Effy Vayena](#) 

Nature Machine Intelligence 1, 389–399 (2019) | [Cite this article](#)

37k Accesses | 413 Citations | 734 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-

Findings from Jobin *et al.*'s paper

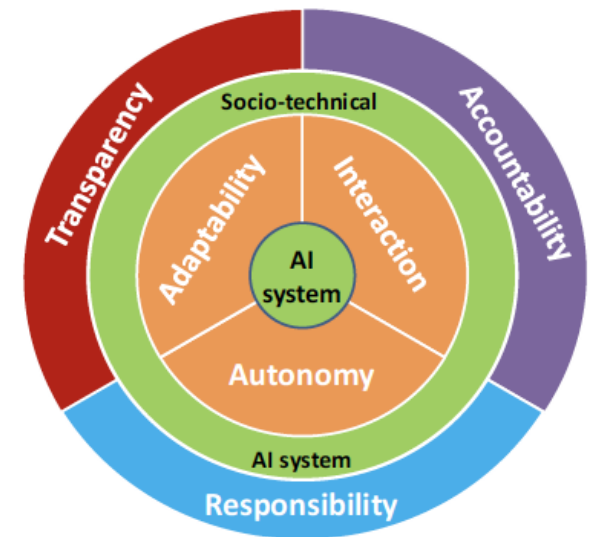
- **Transparency** (appeared in 87% of the documents),
- **Justice and Fairness** (81%),
- **Non-maleficence** (71%),
- **Accountability/Responsibility** (71%),
- **Privacy** (56%),
- **Beneficence** (49%),
- **Freedom and Autonomy** (40%),
- **Trust** (33%),
- **Sustainability** (17%), **Dignity** (15%),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

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**, which does not necessarily mean that the autonomous systems behave in an ethical manner.
- AI systems that put **human well-being at the core** of the development process are also likely to be adopted by humans, who have a say more than before.

The ART Principles for Trustworthy Autonomous Systems

- **A**ccountability
 - The system explains and justifies its decision to users and relevant parties.
- **R**esponsibility
 - The focus is on how the socio-technical systems operate.
- **T**ransparency
 - It is about the data being used, methods being applied, openness about choices and decisions.



The ART Principles for Trustworthy Autonomous Systems

ART is essential to build social trust in Autonomous Systems

- Accountability
 - The...
- Responsibility
 - The...
 - It is...
- Transparency
 - It is about...



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.
- Remember the **Rescue-Robot** example. Under different ethical theories, the moral agent will be accountable accordingly.

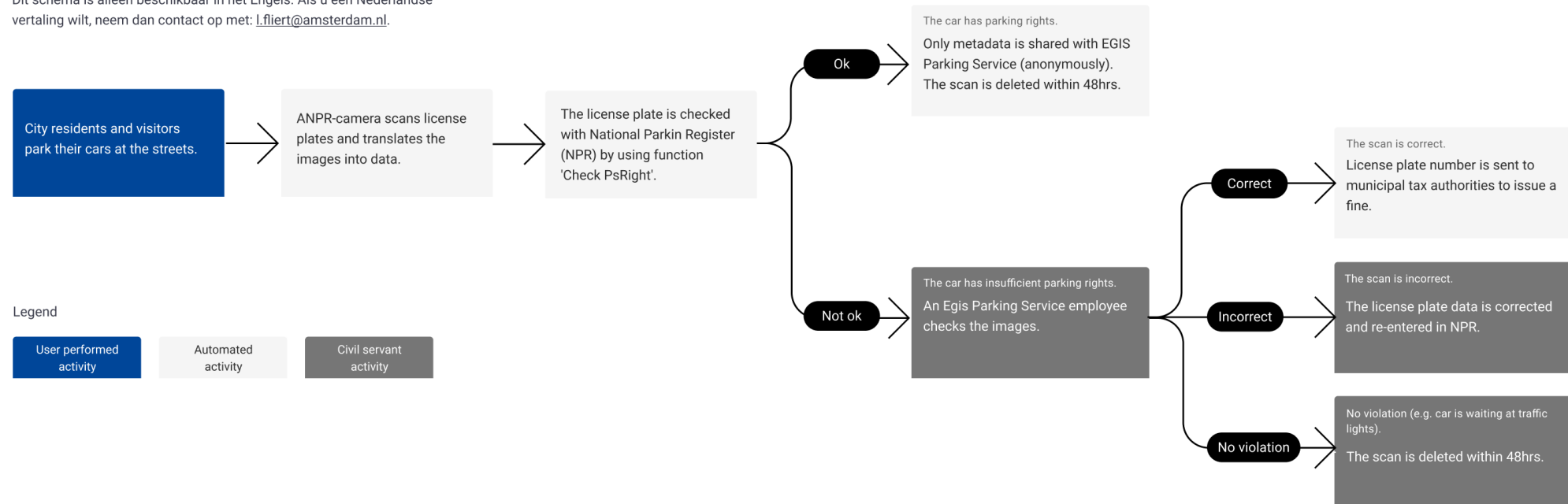


Automated Parking Control (Amsterdam)

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: I.fliert@amsterdam.nl.



Reflection Time

- What did you like/dislike about this architecture?
- What are the benefits/**harms** of such a system?



Responsibility

ART

Responsibility

- The forum should have **power** over actors.
- We are talking about **liability** when the AI system acts in an unexpected way.
- In such cases, the **immediate** responsible actors are the developers and the manufacturers.



Responsible AI --- in practice



Recommended practices

Use a human-centered design approach



Identify multiple metrics to assess training and monitoring



When possible, directly examine your raw data



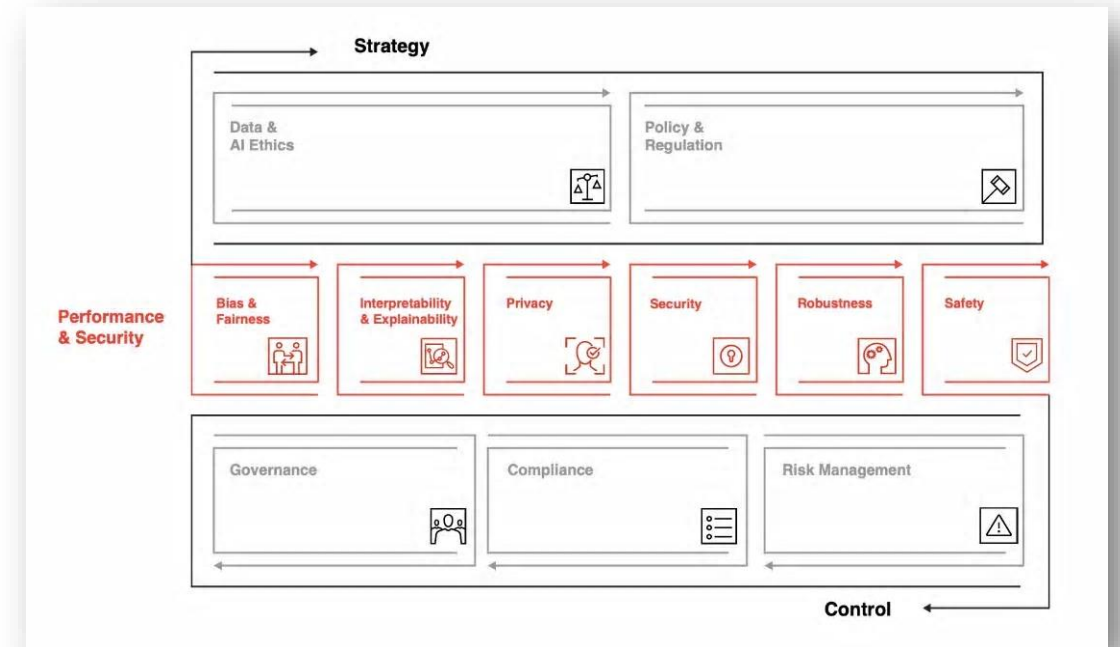
Understand the limitations of your dataset and model



Test, Test, Test



Continue to monitor and update the system after deployment



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."

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). For example, a participation information sheet should include all details about data lifecycle.
 - helps in **understanding and managing risks**. For example, an organization can be responsible and accountable if it knows the inner workings of their offered solutions.

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

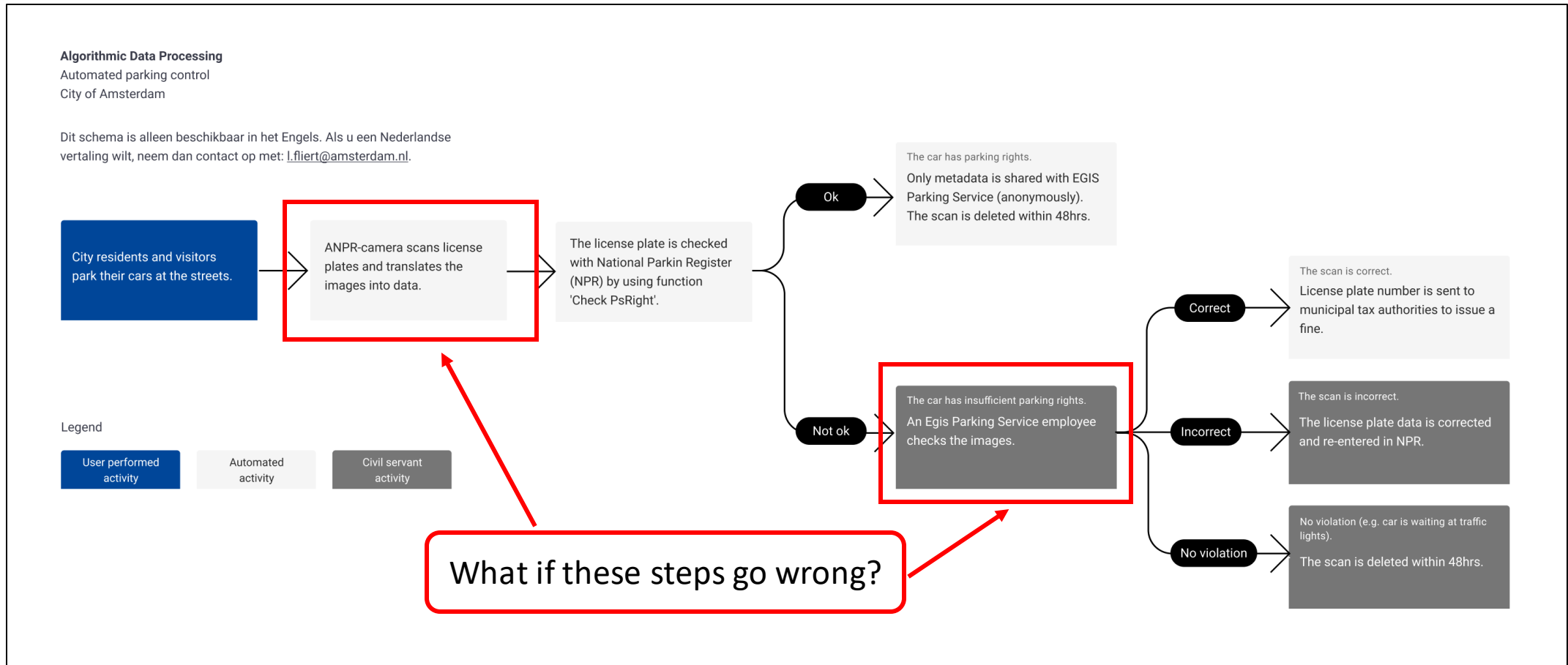
Explanation in Artificial Intelligence:
Insights from the Social Sciences

Tim Miller

*School of Computing and Information Systems
University of Melbourne, Melbourne, Australia*

tmiller@unimelb.edu.au

Transparency: Automated Parking Control



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!

Why is transparency hard?

- We are talking about sociotechnical systems; hence we are dealing with **many stakeholders**.
- Contexts, user profiles, questions to be answered **vary** largely.
- A data scientist may need to learn more about unjust biases in their data, whereas a user may be interested in something different.
- **How to explain** the workings of a "black box" model?
 - Explanations could be added by design, but this requires careful engineering to have a usable solution (e.g., interactive interfaces are great to explore models)
 - The use of simpler models works sometimes!
- **How much transparency** should we provide? We do not want to make our systems vulnerable to attacks at the same time.

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](#)

i [Matters Arising](#) to this article was published on 14 October 2020

i 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 successful¹. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives². 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

Transparency and reproducibility in artificial intelligence

[Benjamin Haibe-Kains](#) ✉, [George Alexandru Adam](#), [Ahmed Hosny](#), [Farnoosh Khodakarami](#), [Massive Analysis Quality Control \(MAQC\) Society Board of Directors](#), [Levi Waldron](#), [Bo Wang](#), [Chris McIntosh](#), [Anna Goldenberg](#), [Anshul Kundaje](#), [Casey S. Greene](#), [Tamara Broderick](#), [Michael M. Hoffman](#), [Jeffrey T. Leek](#), [Keegan Korthauer](#), [Wolfgang Huber](#), [Alvis Brazma](#), [Joelle Pineau](#), [Robert Tibshirani](#), [Trevor Hastie](#), [John P. A. Ioannidis](#), [John Quackenbush](#) & [Hugo J. W. L. Aerts](#)

Nature **586**, E14–E16 (2020) | [Cite this article](#)

15k Accesses | **53** Citations | **507** Altmetric | [Metrics](#)

Addendum: International evaluation of an AI system for breast cancer screening

[Scott Mayer McKinney](#) ✉, [Marcin Sieniek](#), ... [Shravya Shetty](#) ✉ [+ Show authors](#)

Nature **586**, E19 (2020) | [Cite this article](#)

3694 Accesses | **5** Citations | **2** Altmetric | [Metrics](#)

i The [Original Article](#) was published on 01 January 2020

Addendum to: *Nature* <https://doi.org/10.1038/s41586-019-1799-6> Published online 01 January 2020

To assist with the replication of our results, we have expanded the [Supplementary Methods](#) of our Article to provide more detail on how our deep learning system was trained. This includes additional optimization hyperparameters, as well as a more exhaustive description of the data augmentation strategy. Revised [Supplementary Methods](#) are presented in

Summary

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

