The Ethical Digital Architect

Eltjo Poort

ITARC 2025

Introducing

Eltjo Poort

Architect in the digital world

- **Vice President Consulting, CGI**
- **Architecture Practice Lead**
- **Owner of CGI's Risk and Cost Driven Architecture approach**
- **Consulting architect, troubled project assessments, second opinions**
- **Interested in ethical and responsible digital architecture, contributor to CGI's Responsible AI Framework**
- **Lector TU Eindhoven, VU Amsterdam**
- **Violinist, music composer and arranger**

Tweet

Grady Booch
@Grady_Booch

Every line of code represents an ethical and moral decision.

Empathy

The ability to empathize with people is an important basis of ethics.

Empathy just can't be captured in logical laws.

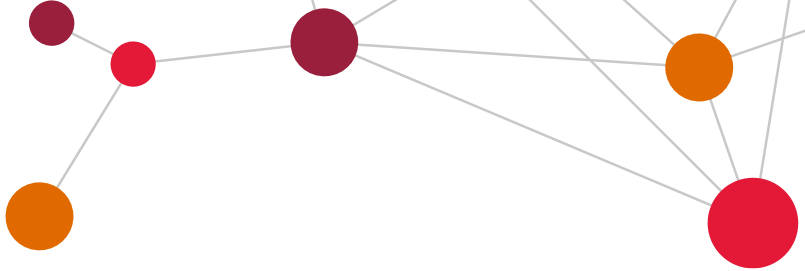
CNN Politics Trump Facts First CNN Polls 2025 Elections Watc

Elon Musk wants to save Western civilization from empathy

Analysis by Zachary B. Wolf, CNN

3 minute read · Published 5:38 PM EST, Wed March 5, 2025

The Ethical Digital Architect



- Why IT staff should care
- Examples of ethical impact
- Group discussions
- Plenary feedback
- Ethical practices and tools
- Conclusions

Why you should care

You can go to prison for making the wrong architectural decision!

Helped design software that conceals high pollutant levels if car is in regulator's lab.
Sentence: 40 months (plea deal)

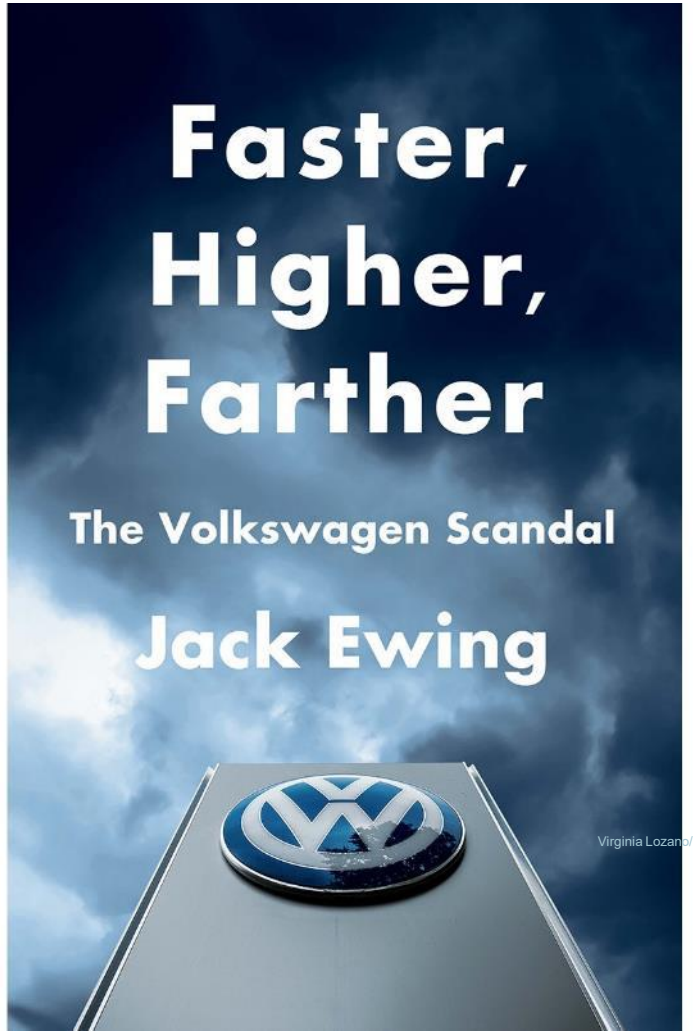
James Liang

Why you should care: the “acoustic function”

“It was a technical problem. An ethical problem? I do not understand why you say that… We had some targets for our technical engineers, and they solved it with some software solutions.” – Matthias Mülller, CEO VW

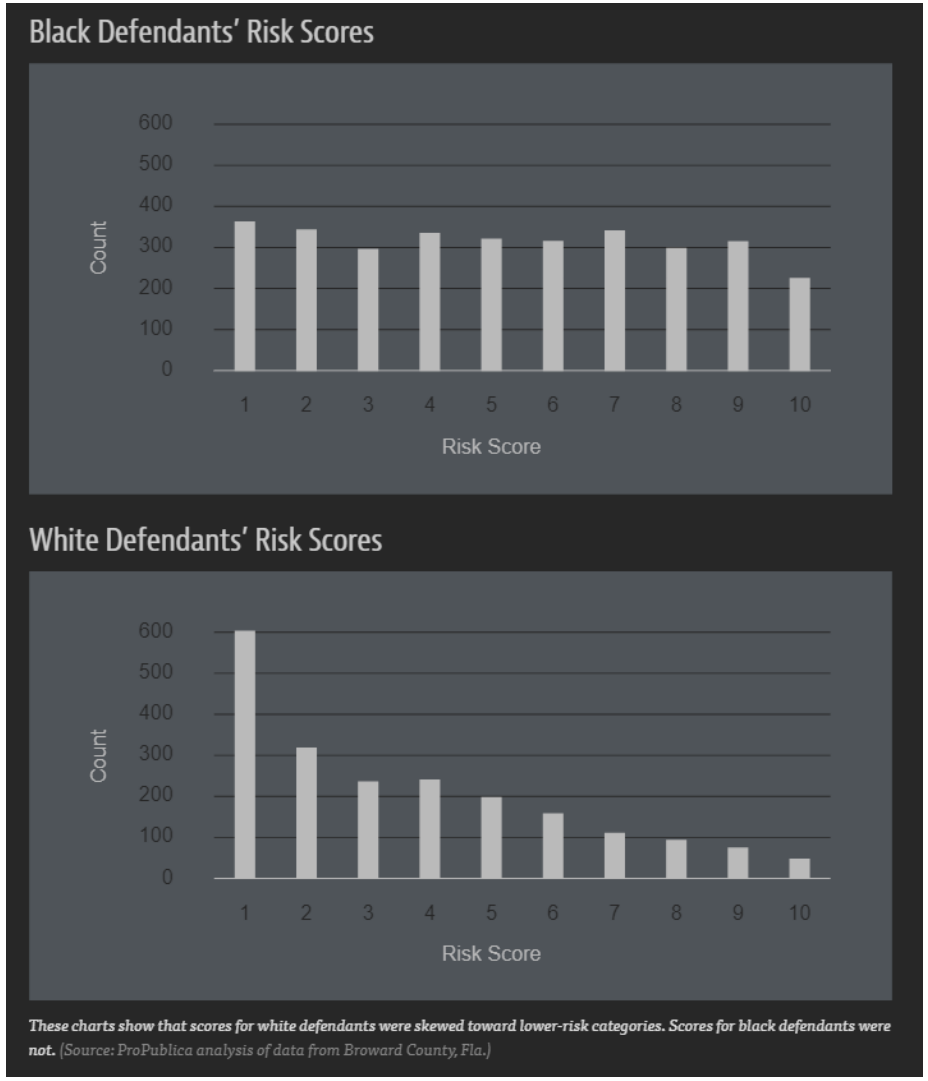
“Combination of fear and unrealistic expectations…”

“Many of those involved did not perceive that what they were doing was morally wrong.”



Virginia Lozano/Detroit News, via Associated Press

Machine bias



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

COMPAS is an algorithm that predicts recidivism based on 138 parameters and historic data (race is not a parameter).

- The system is used extensively by judges to set bail or parole conditions.
- Because it is based on historic data, it is inherently racist.
- The algorithm is owned by a for-profit company that refuses to reveal details.
- Supreme Court threw out a case that this violates defendants' due process rights.

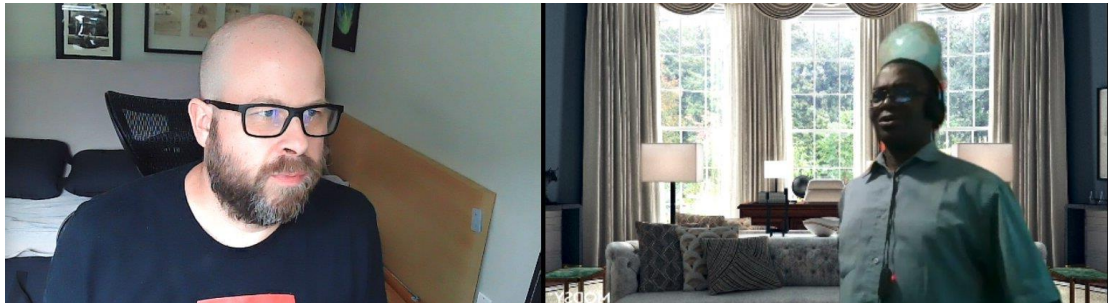
Racial bias in Zoom algorithm erases black faces

It started when Ph.D. student Colin Madland tweeted about a Black faculty member's issues with Zoom. According to Madland, whenever said faculty member would use a virtual background, Zoom would remove his head.

"We have reached out directly to the user to investigate this issue," a Zoom spokesperson told TechCrunch. "We're committed to providing a platform that is inclusive for all."

Twitter apologises for 'racist' image-cropping algorithm

The Guardian, September 2020



defaulted to show only the right side of the picture on mobile?

83 2.099 19,3K

Colin, but at home. · 19/09/2020 ···

Flipped the image....@Twitter is trash.

More examples?

Ethics reasoning tools

Ethics reasoning tools: Dialectic

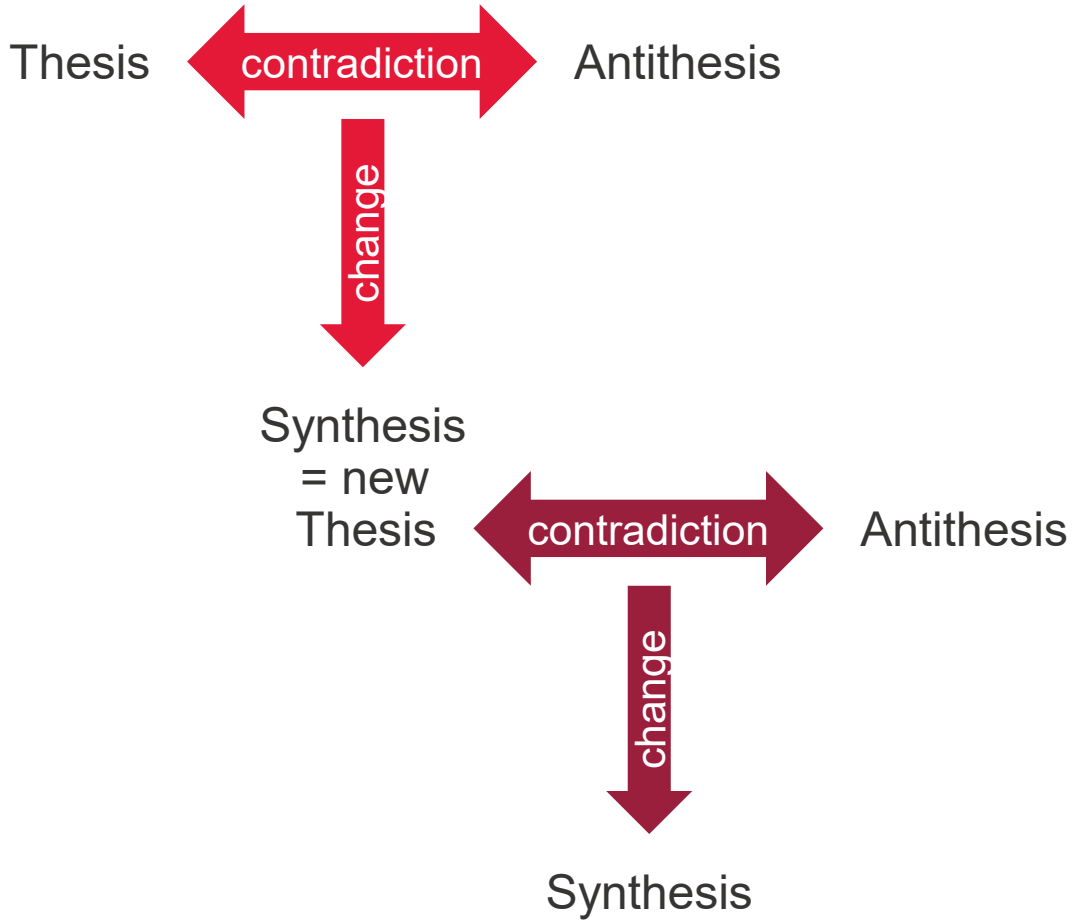
A discourse between two or more people holding different points of view about a subject but wishing to establish the truth through reasoned arguments.

Example: finding cheating software

Thesis: I need to tell the world

Antithesis: I will get fired

Synthesis: If others have found the cheating software too, we may tell the world together at less personal risk



Ethical movements

Often discussed when it comes to ethics in AI

Utilitarianism

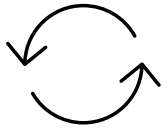
Ethics as maximizing well-being.

(Formulate a kind of KPIs for well-being and optimize accordingly)

Deontology

Ethics as a set of rules

(Thou shalt not kill etc)

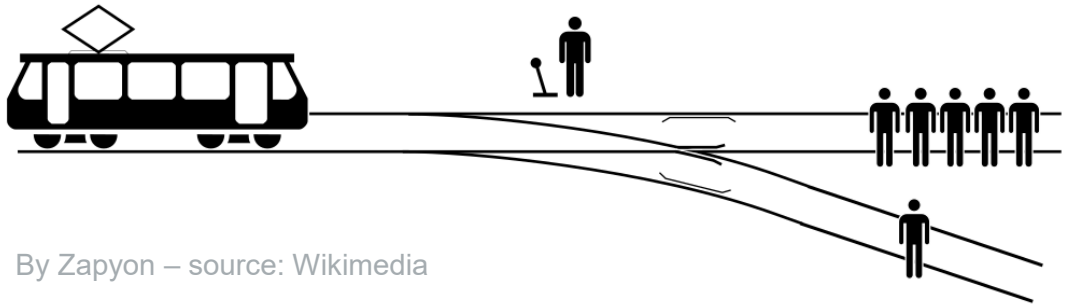


Rule utilitarianism

Combination of the two: set of rules based on maximizing well-being.

(For example, using an algorithm to determine which rules work best)

Ethics reasoning tools: Utilitarian trade-off



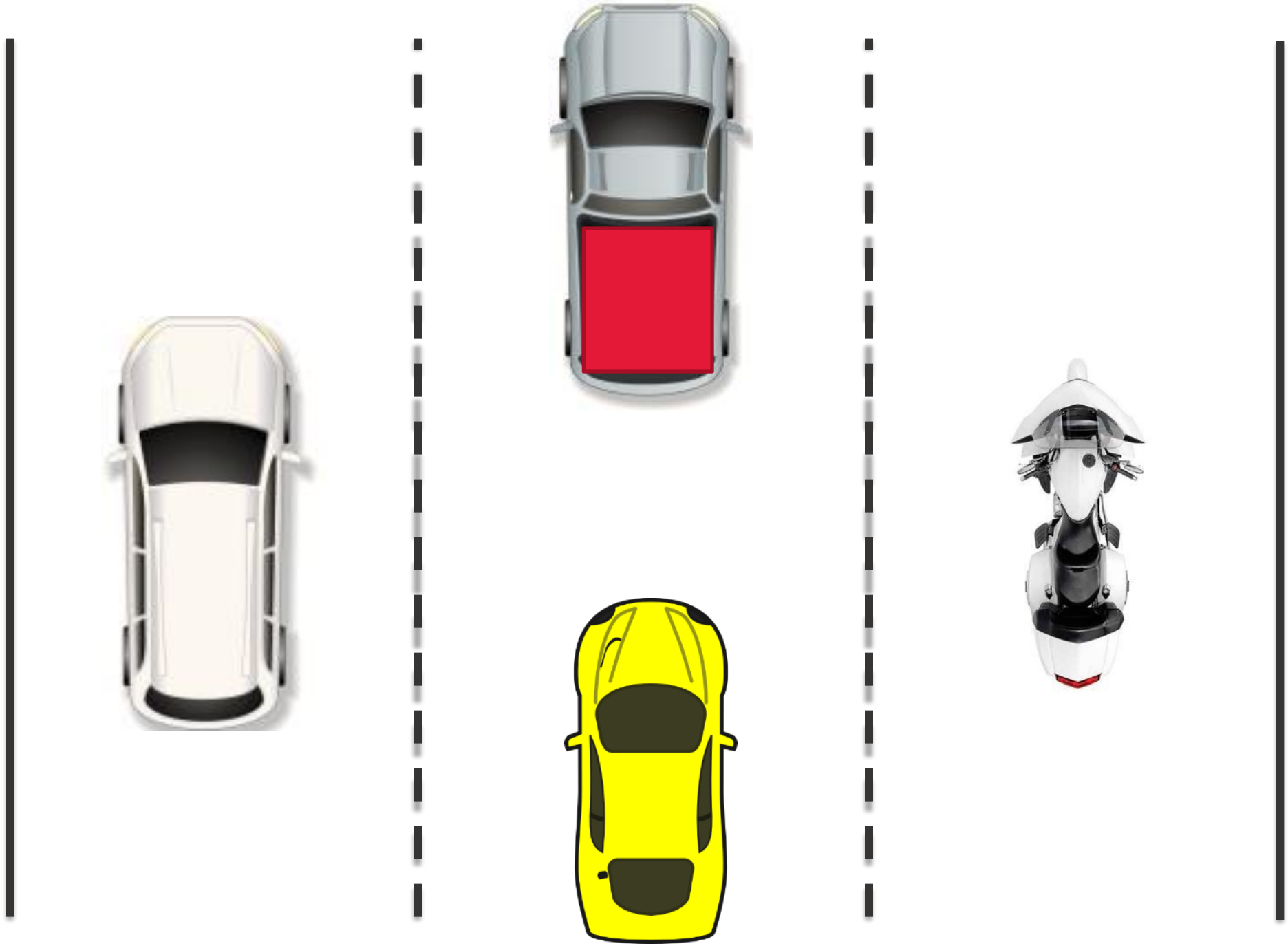
By Zapyon – source: Wikimedia

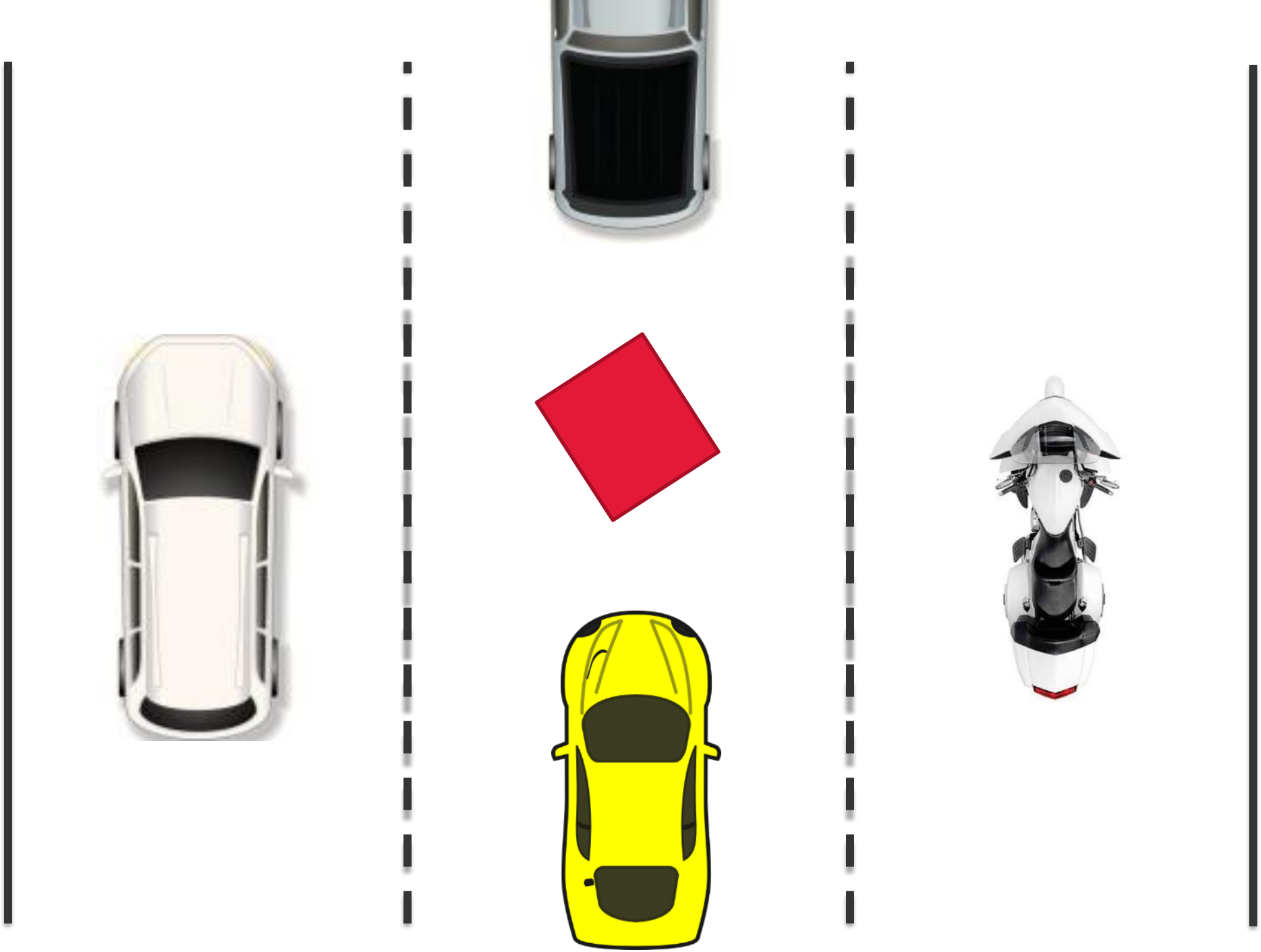
	Pull lever	Do nothing
Save lives	+++++	+
Avoid responsibility for deaths	-	+

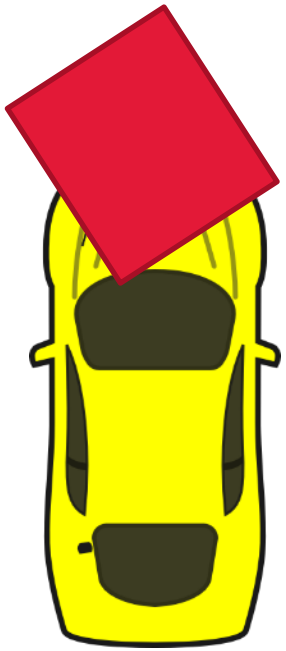
It’s not always a matter of life and death...

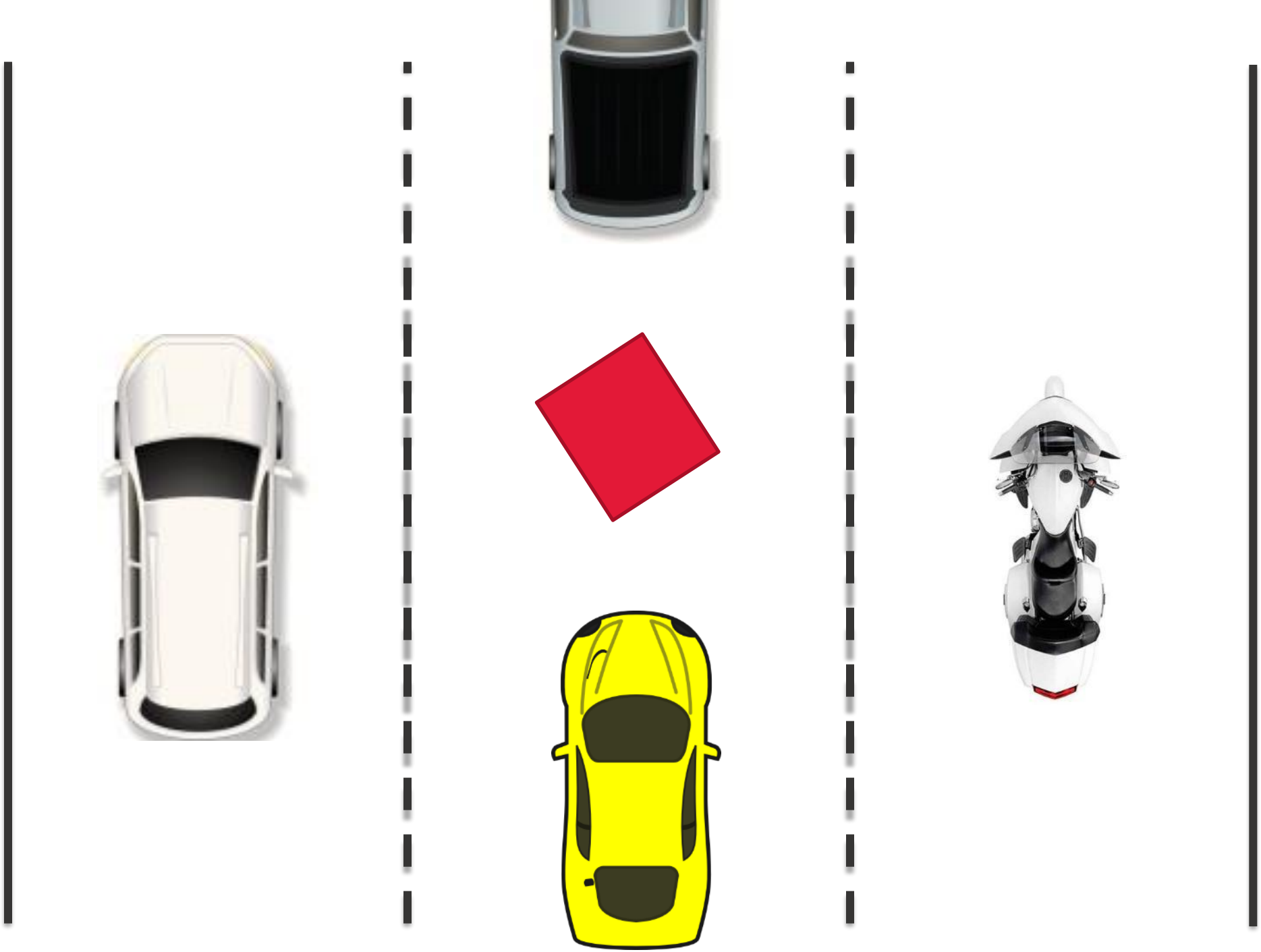
Who will be taken out of the queue?

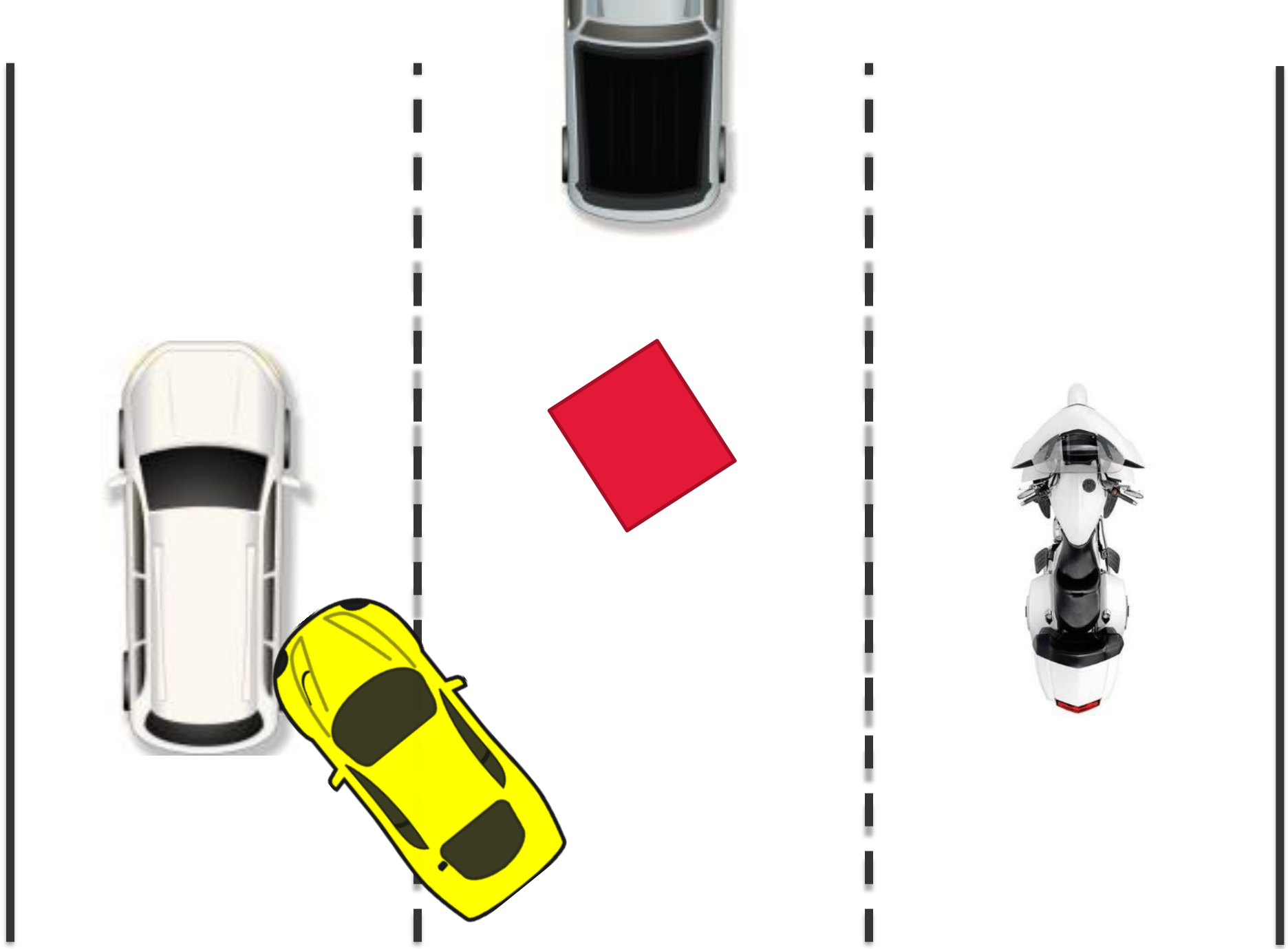
	Algorithm based on characteristics and historic data	Agent’s professional gut feeling	Random checks
Human autonomy	---	+++	-
Public safety	+++	++	---
Fairness	-	---	+++
Transparancy	?	---	+++

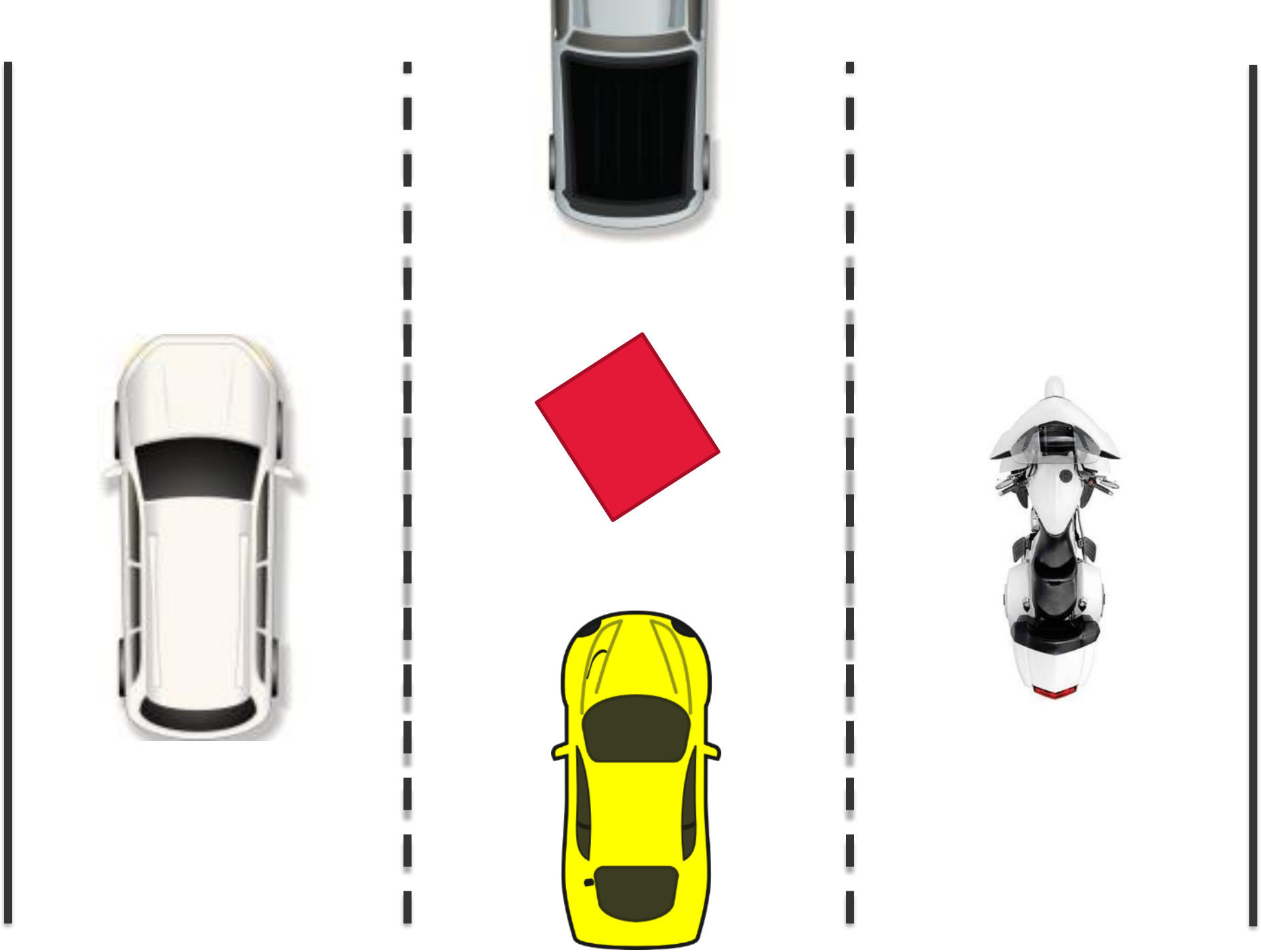


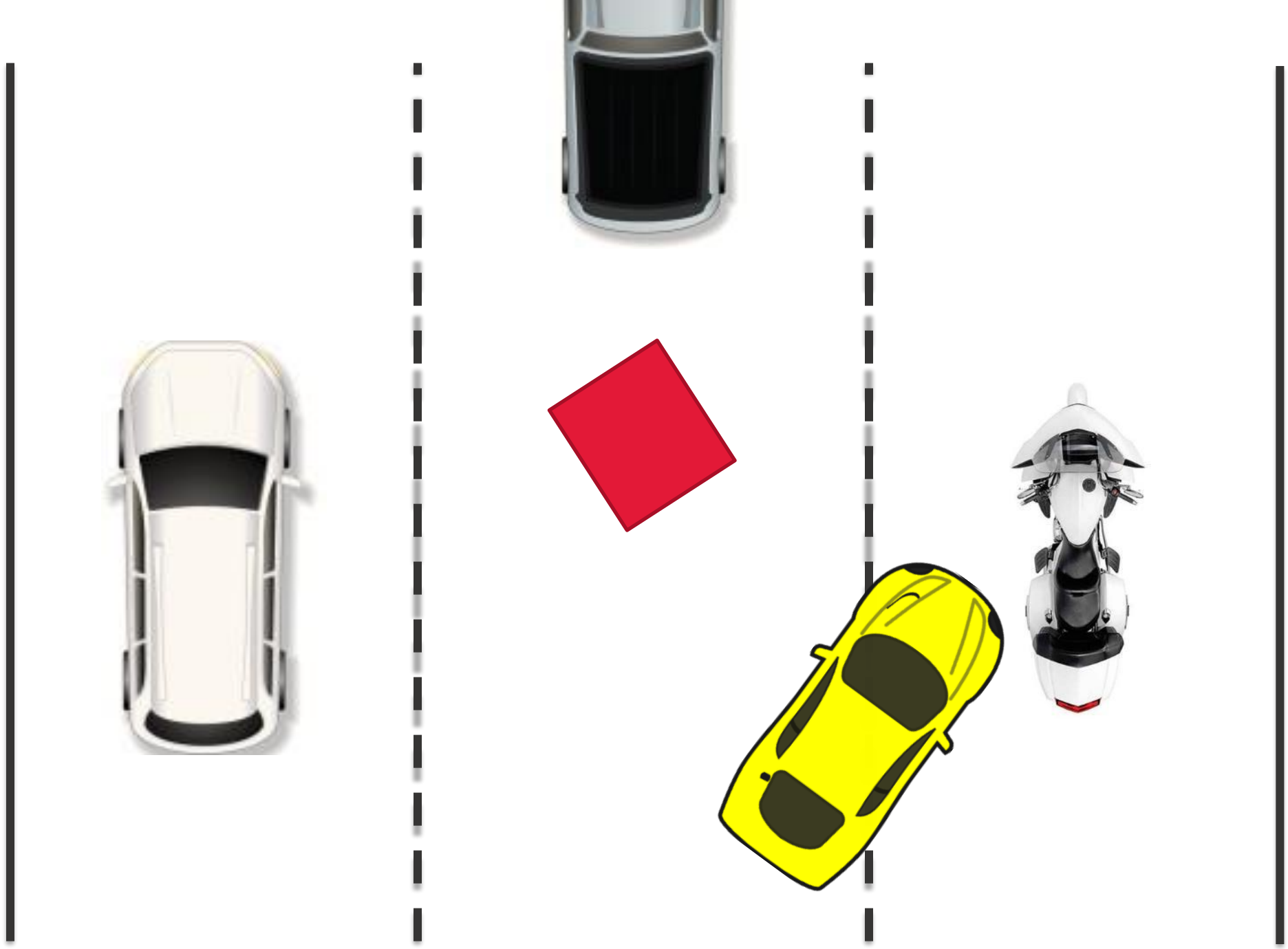


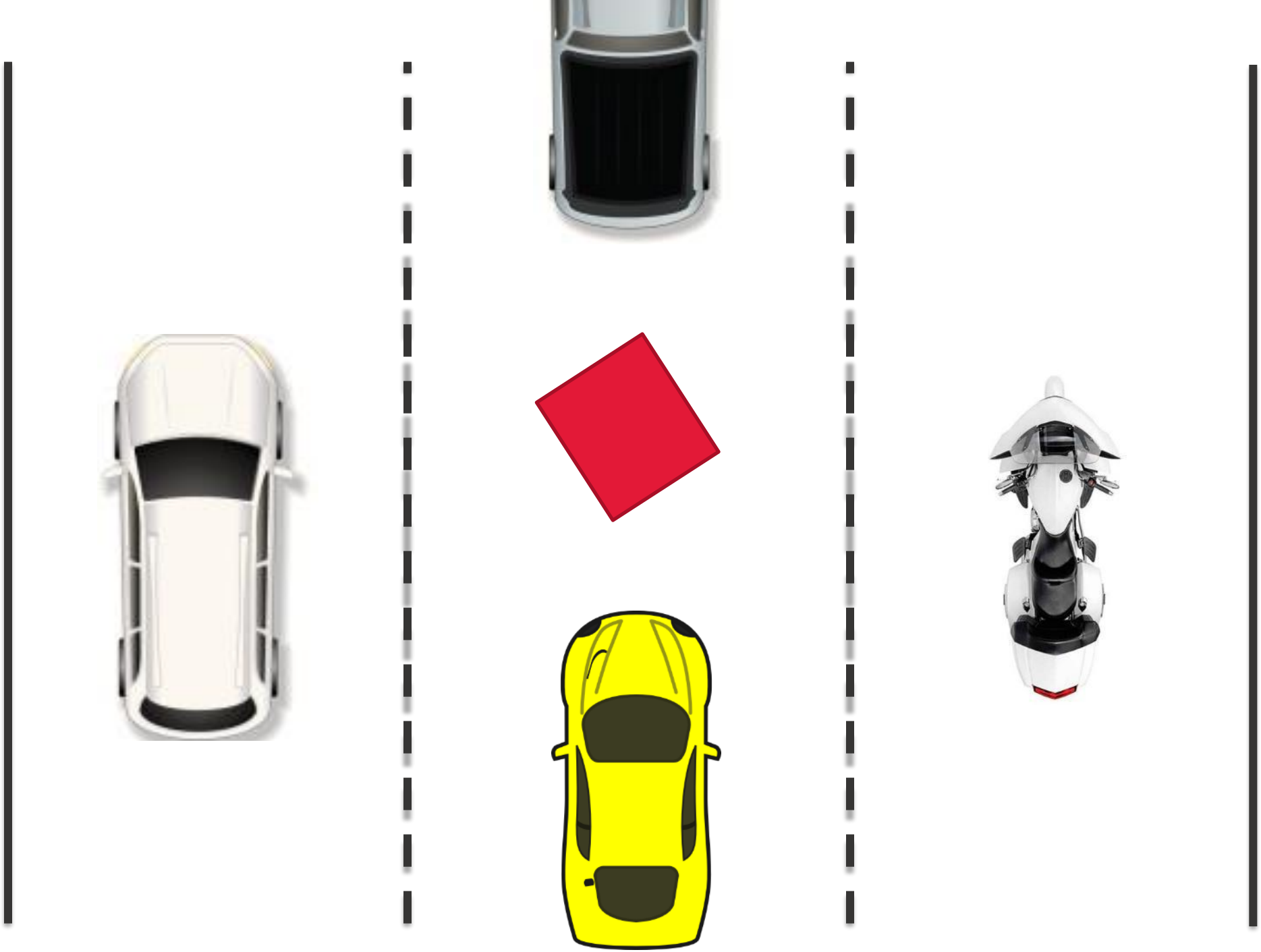








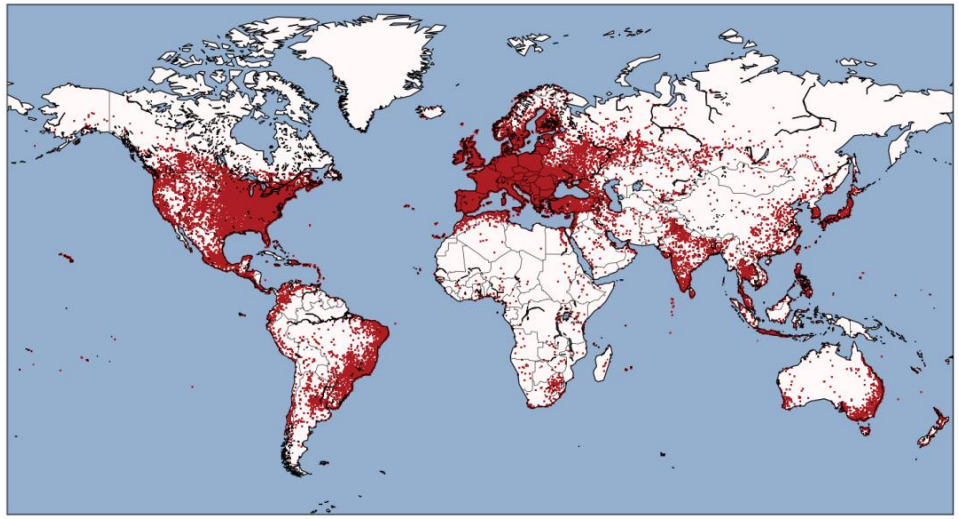


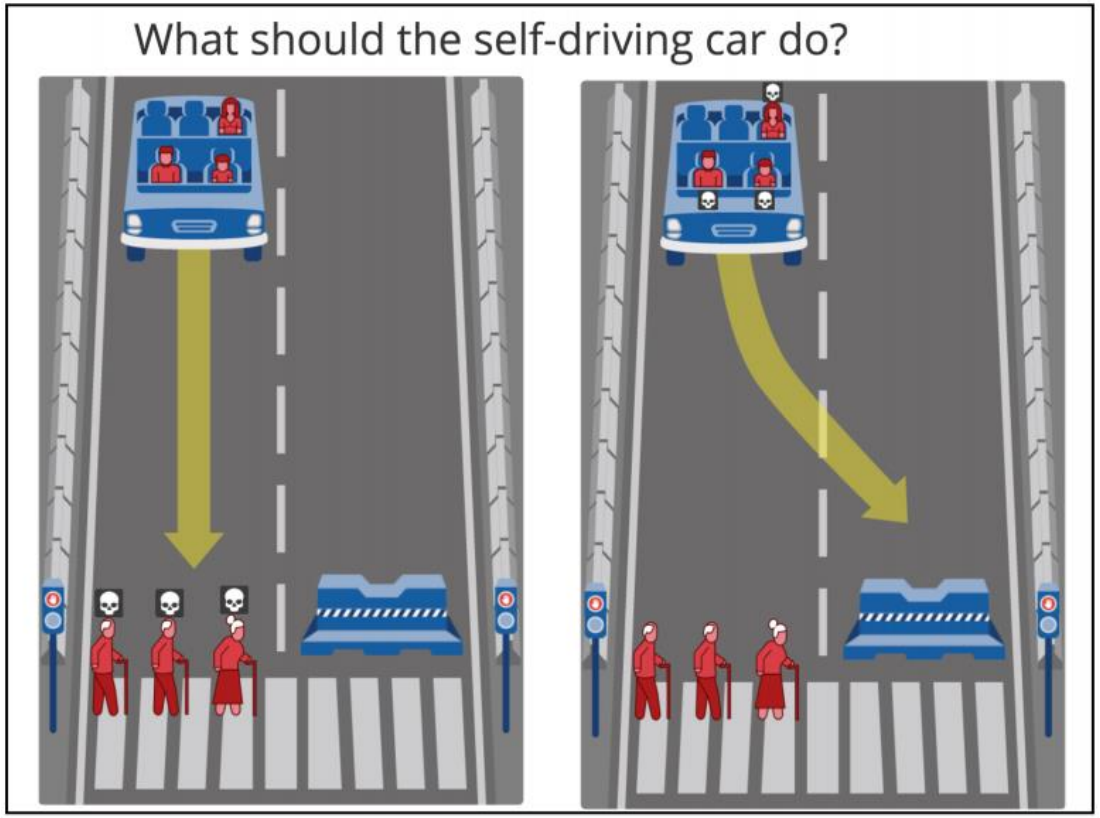


The Moral Machine Experiment

Online experimental platform designed to explore moral dilemmas faced by autonomous vehicles gathered:

- 40 million decisions
- From millions of people
- In 233 countries

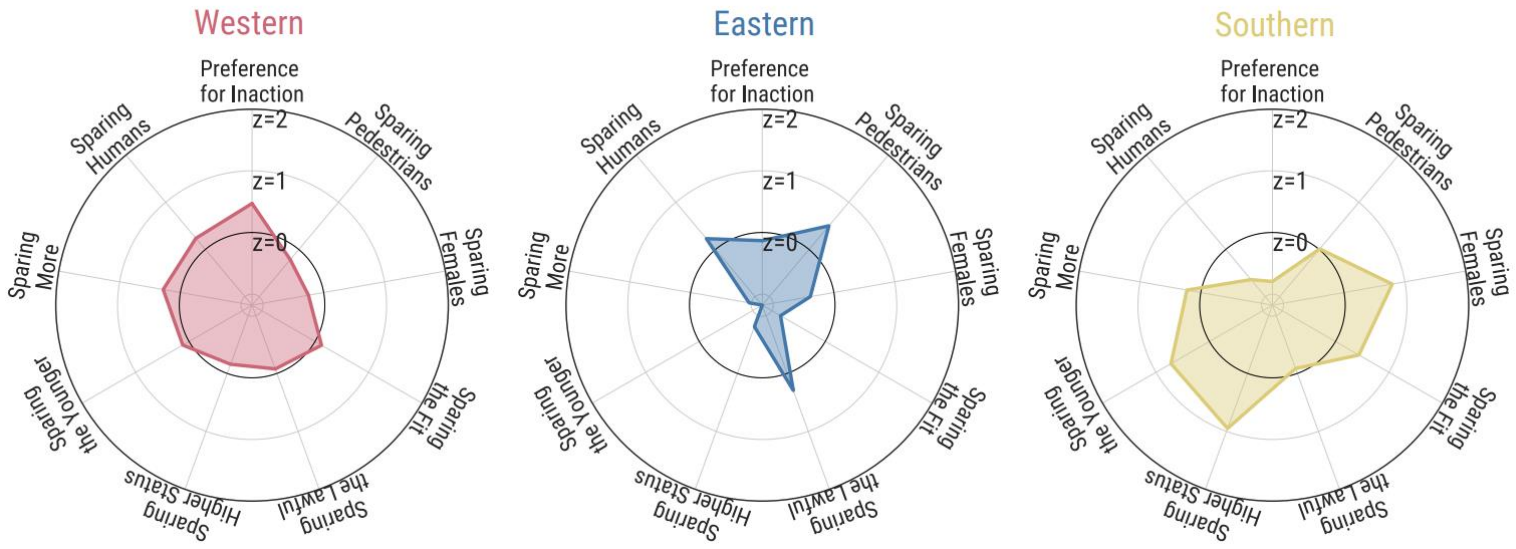


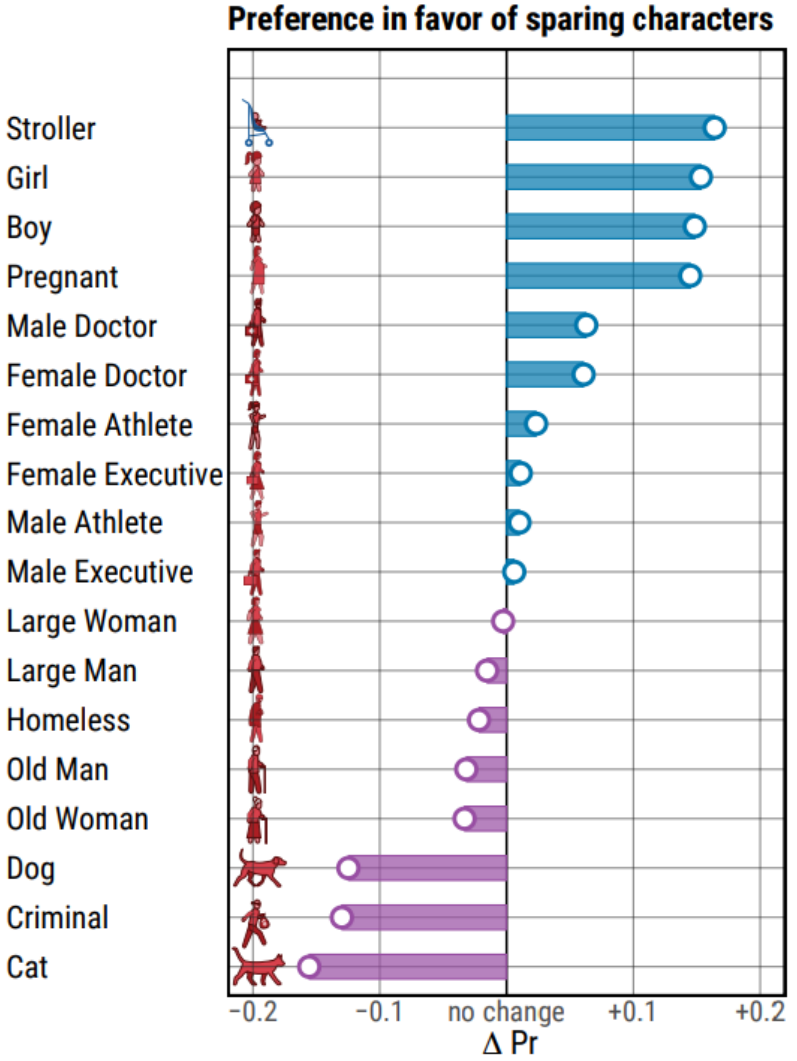


Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563,** 59–64 (2018).

The Moral Machine Experiment

- All cultures on average agreed on who should be spared over whom
- But significant cultural variation in the amount of preference





Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563,** 59–64 (2018).

Group discussions

Group discussions

Divide yourselves in groups of 4

Pick one of the scenarios presented

Discuss the scenario's ethical implications, using e.g. the reasoning techniques presented

- Try to find novel ways of responding to the situation
- What ethical/moral values are at stake?

Scenarios

- Difficult to appeal decisions
- Black box decision making
- More business at expense of privacy
- Discriminate by design
- Location guessing
- Big data

Reasoning tools

- Dialectic
- Utilitarianism (incl. trade-off)
- Deontology

Scenario 1: Difficult to appeal decisions

You are the lead architect for a system that handles requests for disability benefits. Currently, 40% of decisions to deny a citizen such benefits is appealed. Only 10% of appeals is successful. You are requested to design a change in the system that makes it less obvious to users how they can file an appeal after a negative decision, and that introduces an extra – apparently unnecessary – verification step that is known to be error-prone. You suspect that the department manager wants this change because of a target to lower the number of appeals to 30%.

Scenario 2: Black box decision making

You are an HR business partner in an organization that uses a system that helps rate resumes for job applications using a machine learning algorithm. Although ethnical origin is not part of the data fed to the algorithm, you start noticing a bias for people with English-sounding names. The bias may be caused by a similar bias in the historical training data, which was supplied by predominantly conservative firms. Your attempts to bring the bias to the attention of product management have been met with evasive responses.

Scenario 3: More business at the expense of privacy

You are the lead architect for a set of APIs (Application Programmer Interface) of a social networking system at a start-up that is quickly becoming successful. Product management asks you to design a change in an API that would give advertising partners a way to retrieve demographic data of all contacts of members that have agreed to the terms of use. Although strictly speaking legal, you strongly suspect that this is way more intrusive than members intended. When bringing this to the attention of management, their response is that the change is required to sustain the growth of the business.

Scenario 4: Discrimination by design

You are the product manager for a book lending registration product used by libraries throughout the world. Your sales team asks you to design a new configuration parameter named “Modesty”: if this parameter is set, the system will refuse loans of many categories of books if the member is female. Any attempts by female members to check out these categories of books are to be registered in the database. When asked, management tells you that the company is opening up new strategic markets, and that the survival of the company depends on the success in these new markets.

Scenario 5: Location guessing

You are a developer for a free messaging app that makes money by running ads. The advertisements are more profitable if the targeting algorithm includes the user's location, but many users do not give your app permission to use their device's location data.

Product management has recently asked you to see if you can deduce a user's rough location by looking at the geo data in photos taken by the user (your app has permission to open images in order to be able to attach them to messages).

Scenario 6: Big data

You are a data scientist working for a public order management organization. One of your systems uses a mood sensing algorithm based on open social media sources.

A senior public official has asked your team to extend the algorithm with a trigger that sends an alert if the mood threatens to escalate to violent levels. The alert should include information identifying the key instigators of the violence.

Experience the commitment®

Ethical practices and tools

Trickle-up: Ethical architecture is everyone's business

Even if you are not in charge of the most impactful ethical decisions, you can still make an impact by making sure the decisions you *do* have the mandate for are the right ones.

Examples

- Sensible defaults
- Avoiding roach motels
- …?

Trine Falbe – whitehatux.com

Edge cases

"We'll go live when the new system can handle 95% of present cases…"

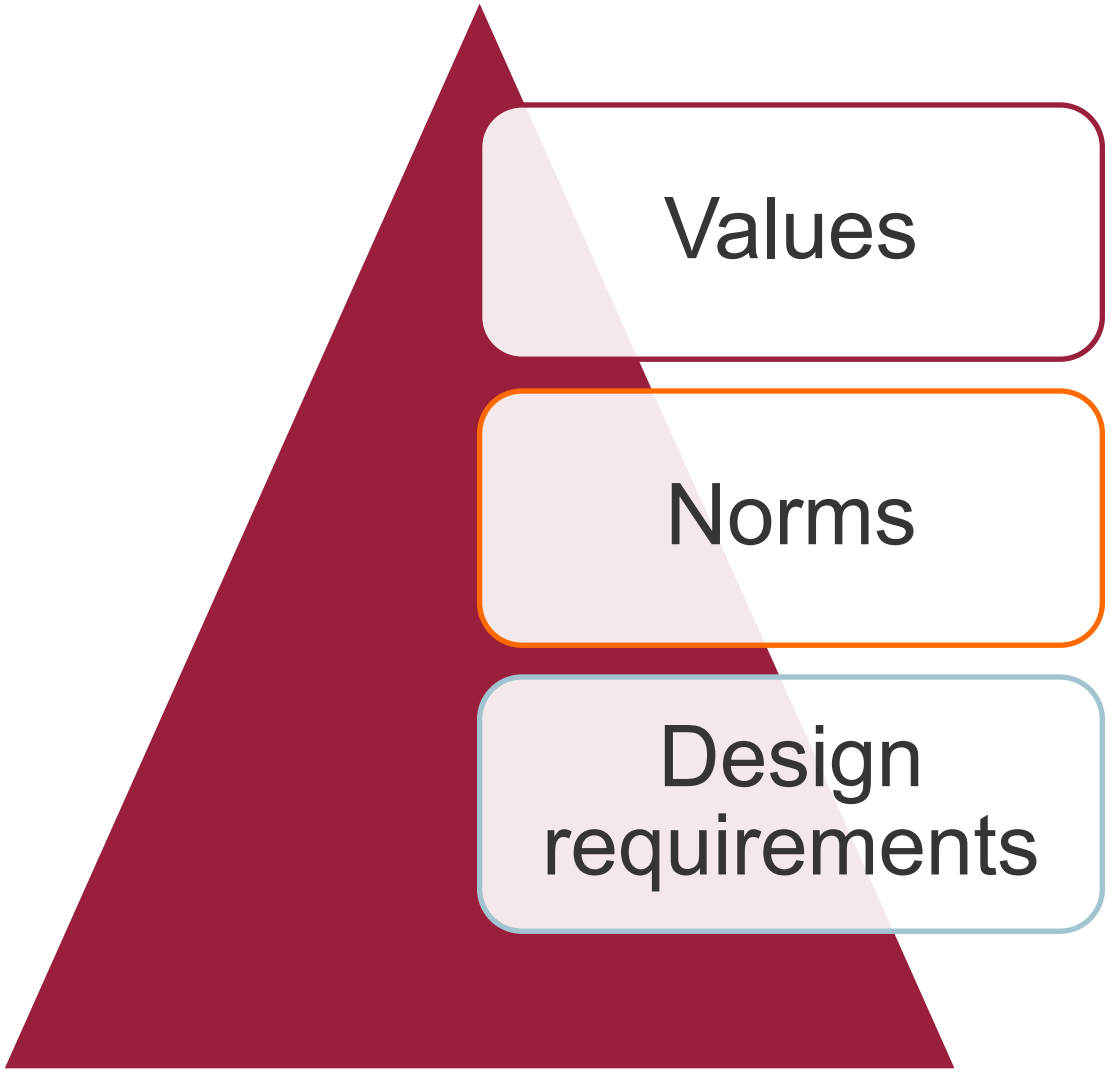
- Minimum Viable Product?
- Can make life very difficult for 5% of customers/citizens

Architect's responsibility

- When will architecture runway be ready for final 0.5%?
- Architect with Just Enough Anticipation, but anticipate edge cases
- Examples…?

Ethical Architecture: Design for Values

Delft University

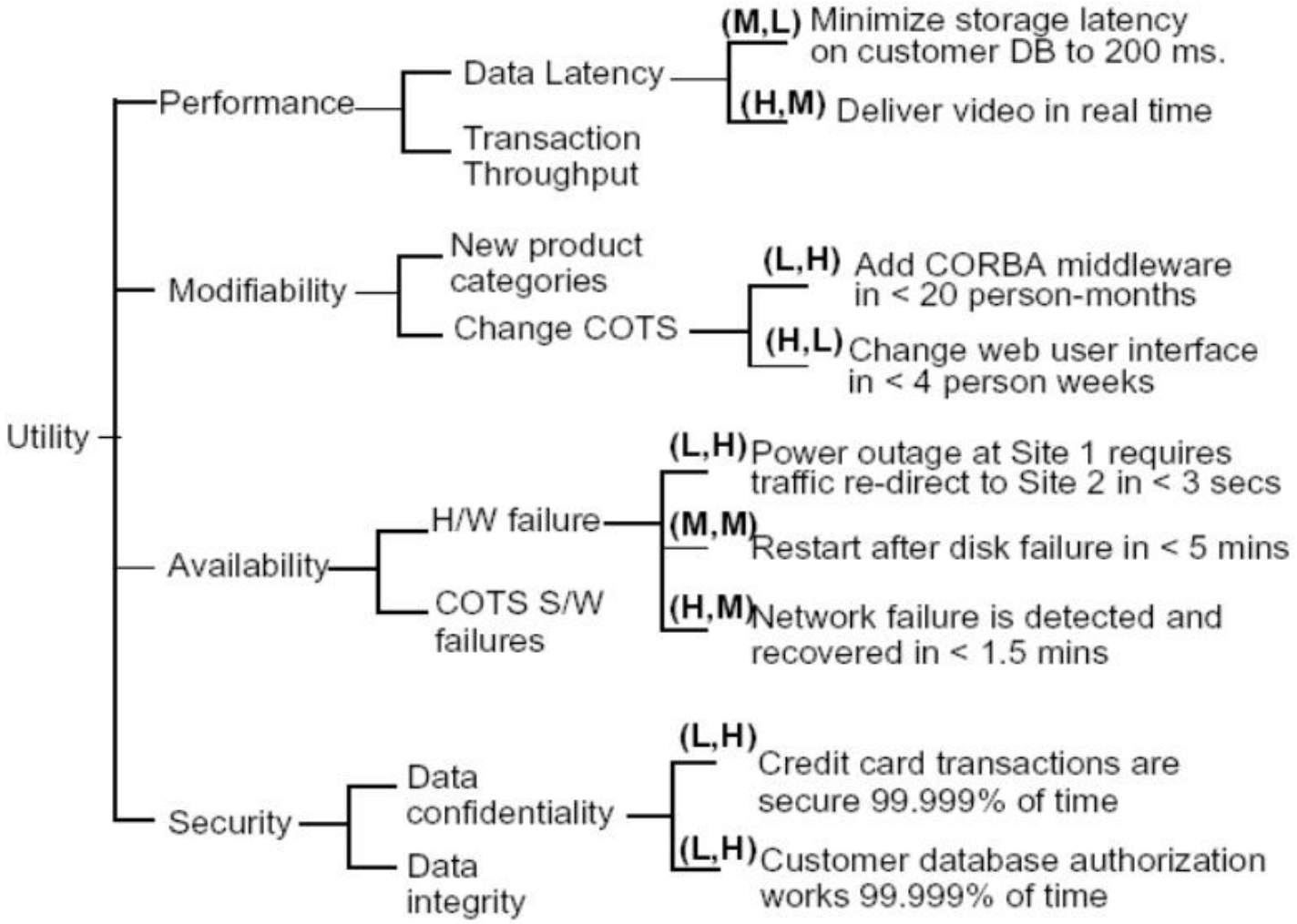


Conceptual tool for designers

From vague social/ethical values to concrete design requirements

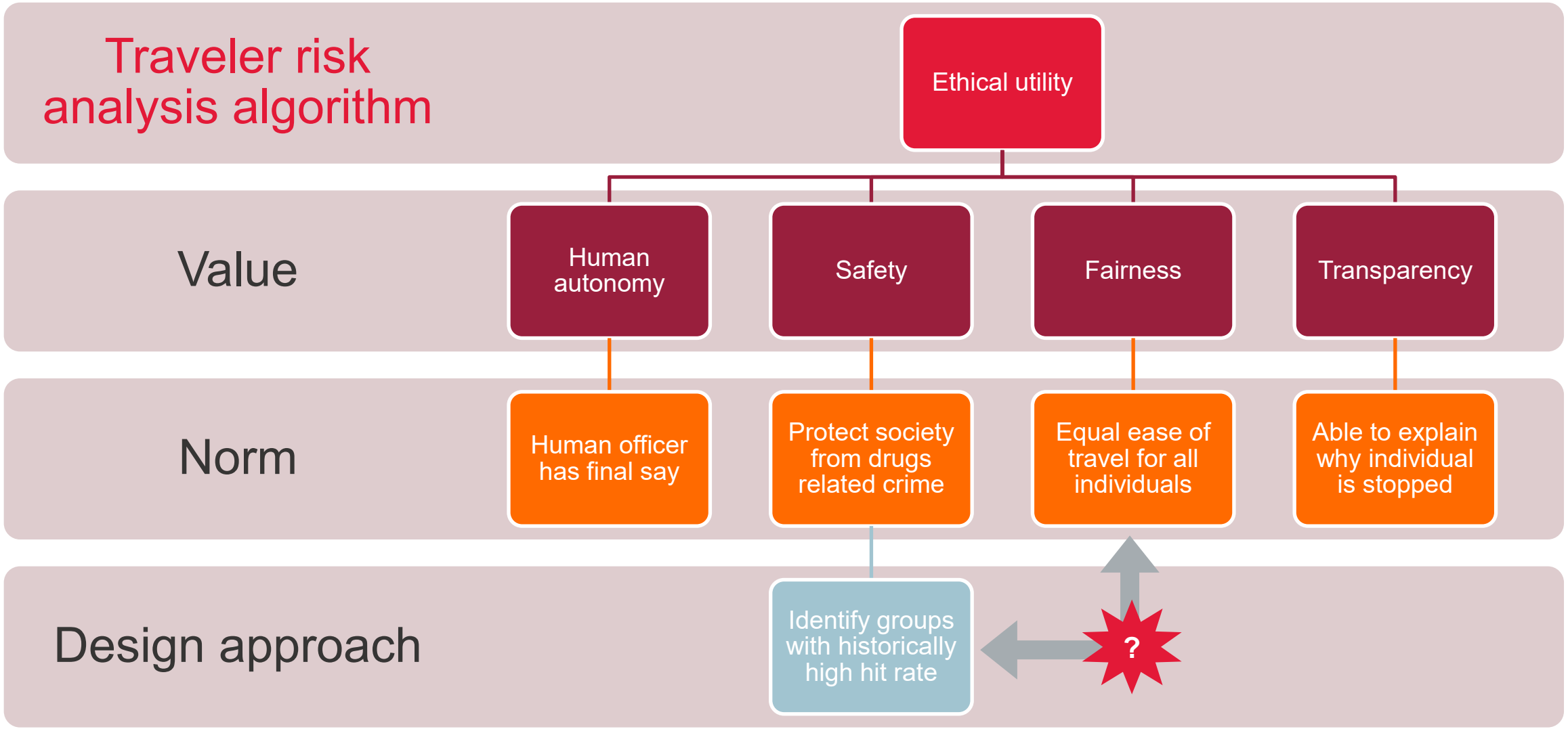
The ATAM Utility Tree

Software Engineering Institute, Carnegie Mellon

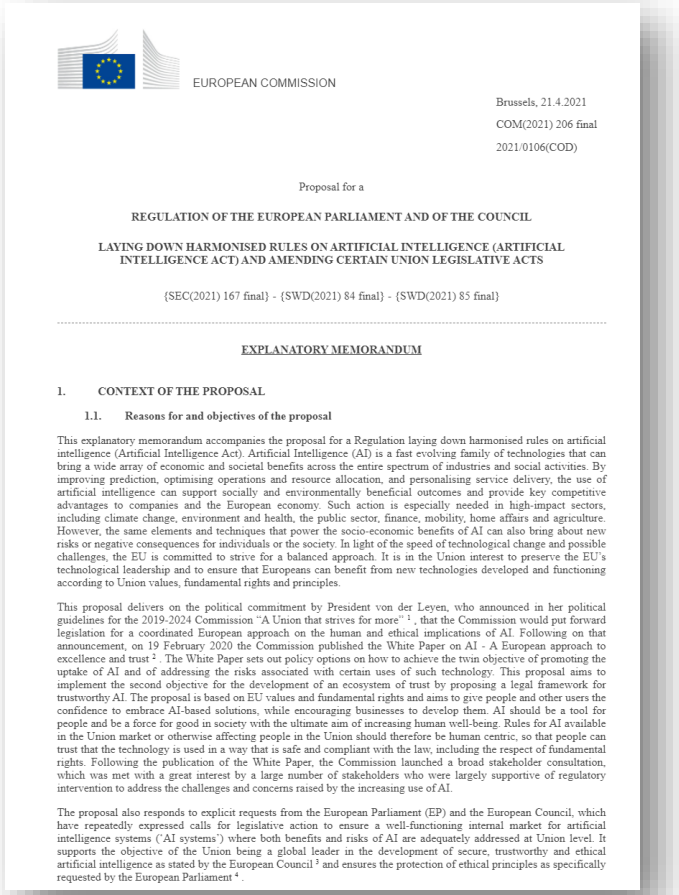


(Importance,Risk)
L=low
M=medium
H=high

Ethical Architecture: utility tree analysis
(Design for Values x ATAM)



EU is regulating application of AI



EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final
2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

EXPLANATORY MEMORANDUM

1. CONTEXT OF THE PROPOSAL

1.1. Reasons for and objectives of the proposal

This explanatory memorandum accompanies the proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Artificial Intelligence (AI) is a fast evolving family of technologies that can bring a wide array of economic and societal benefits across the entire spectrum of industries and social activities. By improving prediction, optimising operations and resource allocation, and personalising service delivery, the use of artificial intelligence can support socially and environmentally beneficial outcomes and provide key competitive advantages to companies and the European economy. Such action is especially needed in high-impact sectors, including climate change, environment and health, the public sector, finance, mobility, home affairs and agriculture. However, the same elements and techniques that power the socio-economic benefits of AI can also bring about new risks or negative consequences for individuals or the society. In light of the speed of technological change and possible challenges, the EU is committed to strive for a balanced approach. It is in the Union interest to preserve the EU's technological leadership and to ensure that Europeans can benefit from new technologies developed and functioning according to Union values, fundamental rights and principles.

This proposal delivers on the political commitment by President von der Leyen, who announced in her political guidelines for the 2019-2024 Commission "A Union that strives for more" [1], that the Commission would put forward legislation for a coordinated European approach on the human and ethical implications of AI. Following on that announcement, on 19 February 2020 the Commission published the White Paper on AI - A European approach to excellence and trust [2]. The White Paper sets out policy options on how to achieve the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of such technology. This proposal aims to implement the second objective for the development of an ecosystem of trust by proposing a legal framework for trustworthy AI. The proposal is based on EU values and fundamental rights and aims to give people and other users the confidence to embrace AI-based solutions, while encouraging businesses to develop them. AI should be a tool for people and be a force for good in society with the ultimate aim of increasing human well-being. Rules for AI available in the Union market or otherwise affecting people in the Union should therefore be human centric, so that people can trust that the technology is used in a way that is safe and compliant with the law, including the respect of fundamental rights. Following the publication of the White Paper, the Commission launched a broad stakeholder consultation, which was met with a great interest by a large number of stakeholders who were largely supportive of regulatory intervention to address the challenges and concerns raised by the increasing use of AI.

The proposal also responds to explicit requests from the European Parliament (EP) and the European Council, which have repeatedly expressed calls for legislative action to ensure a well-functioning internal market for artificial intelligence systems ('AI systems') where both benefits and risks of AI are adequately addressed at Union level. It supports the objective of the Union being a global leader in the development of secure, trustworthy and ethical artificial intelligence as stated by the European Council [3] and ensures the protection of ethical principles as specifically requested by the European Parliament [4].

Definition

- Software that uses techniques such as machine learning; logic- and knowledge-based; statistical approaches…

Prohibition

- Against AI systems that
  - cause harm or discriminate
  - use social scoring
  - real-time biometric indentification for law enforcement

Regulate

- High-risk AI systems subject to conformity assessment

EU Trustworthy AI Assessment Checklist

Human agency and oversight

- Fundamental rights
- Human agency
- Human oversight

Technical robustness and safety

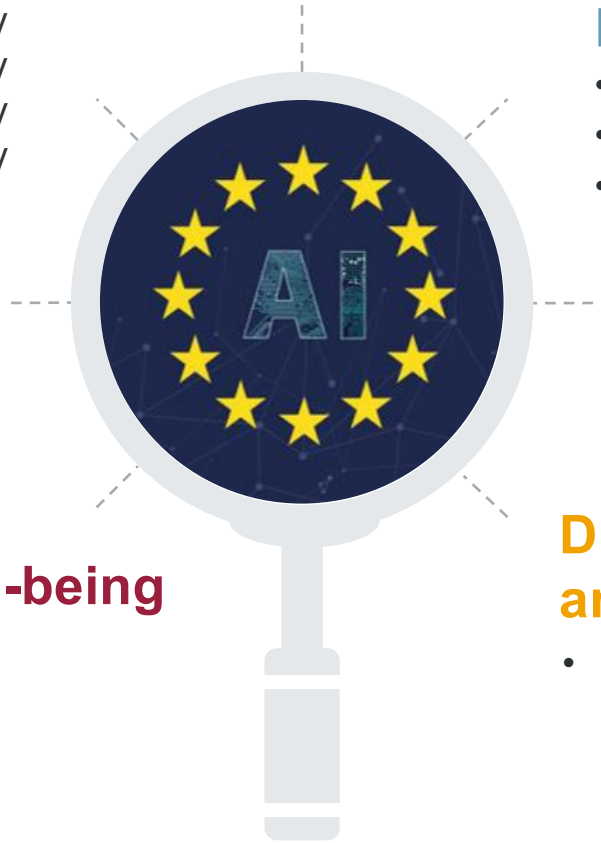
- Resilience to attack and security
- Fallback plan and general safety
- Accuracy
- Reliability and reproducibility

Privacy and data governance

- Respect for privacy and data protection
- Quality and integrity of data
- Access to data

Transparency

- Traceability
- Explainability
- Communication



Accountability

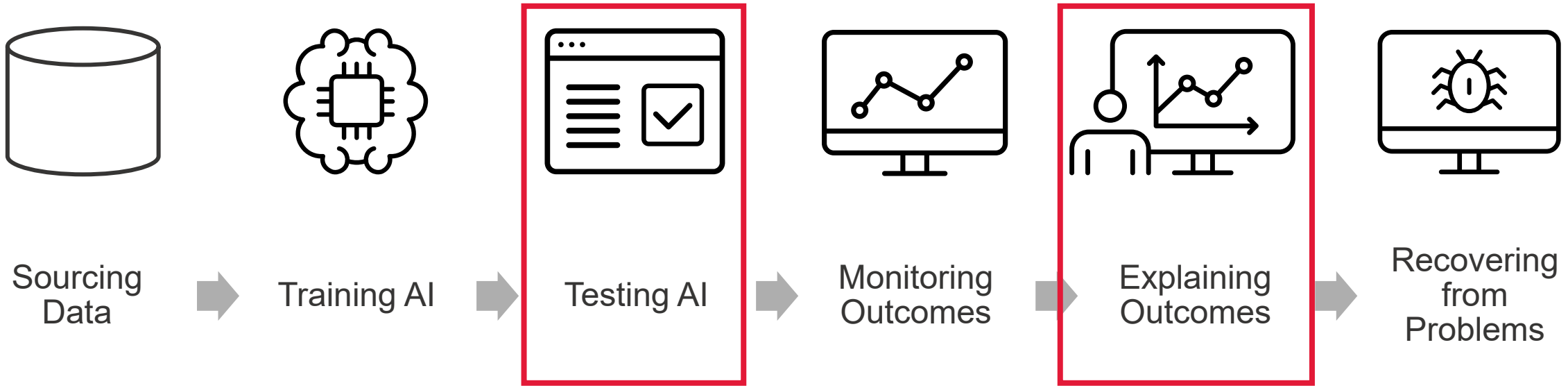
- Auditability

Societal and environmental well-being

Diversity, non-discrimination and fairness

- Unfair bias avoidance

Architectural strategies and tactics for ethical AI



Design for Testability

Ethical aspects to design tests for:

- **bias** (fairness & diversity)
- **accuracy** (trustworthyness)
- **resilience** to attacks (security)

BUT: You won't be able to test everything

Machine Learning models require additional testing to validate that the model is optimized for the right behavior/parameters.

Examples: finding tanks on photos, drone SAM simulation*

EU Countries will provide Regulatory Sandboxes to foster innovation**

*https://www.theguardian.com/us-news/2023/jun/01/us-military-drone-ai-killed-operator-simulated-test
**EU AI Act section VI Measures in Support of Innovation

Design for Explainability

How do you get an explanation out of a neural network?

- Vary inputs to determine which inputs have the greatest influence
- Input / feature weights (SHAP, LIME and integrated gradients)*

Beware: Neural Networks (including LLMs) have no reasoning capabilities

- Try making them solve a Cabbage/Goat/Wolf problem
- But they are great at bluffing

“Symbolic AI” (e.g. decision trees, expert systems) provide reasoning and explanations

- May be combined with LLMs through Retrieval Augmentation Generation (RAG)

* What is explainable AI, and why it is so important? | CGI US

Conclusion