Making AI less risky business

Business Applications of AI Risk Management

Alex Shepherd - AI Client Manager @ BSI - 25th Oct 2024

Who am I?

Studied Cognitive Science and Artificial Intelligence at the University of Edinburgh

Worked as a Data Scientist for over 3 years in the Pharma and Life Sciences Sectors

Now work as an AI auditor at the British Standards Institute (BSI), additionally work as an AI Governance advisor to startups.

My expertise currently lies in the field of Responsible AI

The national standards body for the UK Government

- Contributes to standards development
- Audits clients across the globe
- Impartially advises on compliance to standards.

Goals

- Understand the current AI risk landscape
- How to classify risks against the common AI ethical principles
- Apply the Responsible AI framework to manage for AI risks and produce more trustworthy technologies
- Understand tradeoffs between each of the AI ethics principles and the relationships between each principle when controlling for AI risks

Contents

- What is AI risk?
- Why should we bother about it?
  - Introduction to Responsible AI (RAI) Framework
- Where does AI Risk come from?
- Types of AI Risk
- How can we treat AI risks?
  - Example dilemmas and applications of RAI Framework
- Tradeoffs

What does AI Risk look like?

AI Risks

Wild-west tech gone wrong

A Chevrolet dealership in the US utilised Generative AI, as part of a customer-facing chatbot solution.

A customer was able to alter the behaviour of the chatbot via prompt injection, to accept customer's offer to purchase a 2024 Chevrolet Tahoe for $1 as "legally binding" with no "takesies backsies".

https://colombiaone.com/2024/01/08/gm-dealership-chatbot-ai/

AI Risks

Chatbots giving unavailable discounts

In 2022, Air Canada's chatbot promised a discount that was, according to their policies, not available to a customer.

This led to a reactive legal challenge, which ended up with Air Canada paying the customer in damages and tribunal fees.

"It should be obvious to Air Canada that it is responsible for all the information on its website, … It makes no difference whether the information comes from a static page or a chatbot."

British Columbia Civil Resolution Tribunal member Christopher Rivers

https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know

AI Risks

Dataset bias -> Model bias

Amazon's deployed a CV screening model to facilitate the screening of top candidates for technical roles.

They later identified that, given the dataset of successful employees were (white) males, the model behaviour was able to focus on male-like features in CV to more likely select male candidates. Female candidates were thus being automatically excluded from these job opportunities.

https://www.reuters.com/article/idUSKCN1MK0AG/

AI Risks

Lack of independent checks

Michael Cohen, former lawyer for Donald Trump, used Google Bard to generate non-existent legal case citations. These false citations were unknowingly included in a court motion by Cohen's attorney, David M. Schwartz. (1)

Additionally there is a report of lawyers being fined for misleading the court based on a legal brief containing multiple non-existent case citations. (2)

(1) https://www.reuters.com/legal/ex-trump-fixer-michael-cohen-says-ai-created-fake-cases-court-filing-2023-12-29/
(2) https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/

AI Risks

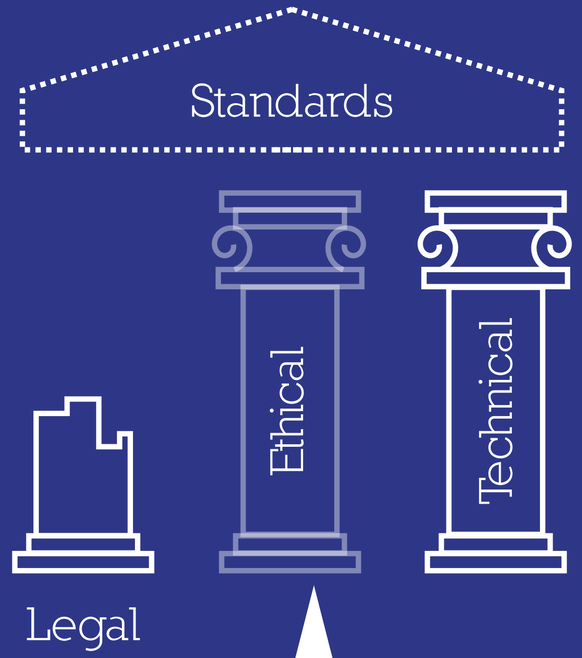
Ethics washing

Reports of Google Researchers having to vet their research publications, whenever they mentions concerns or something worrying. (1)

Google's former Co-lead of Ethical AI Research Team, Timnit Gebru was allegedly fired over a publication discussing the financial and environment costs of LLMs, as well as LLM's impact on marginalised populations (2)

Further examples provided in (3)

Article on counteracting ethics-washing



Let's just say our systems are ethical, when they really aren't.

(1) https://www.theguardian.com/technology/2021/feb/26/google-timnit-gebru-margaret-mitchell-ai-research

(2) https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases

(3) https://medium.com/@aphiegover/the-ethics-washing-of-ai-952321d8a70d#:~:text=practice%20of%20feigning%20ethical%20consideration,while%20not%20genuinely%20acting%20on

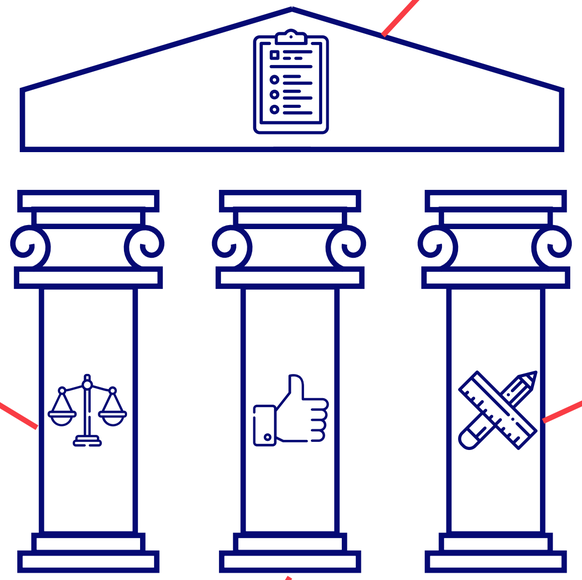
Responsible AI Framework

Tool to diagnose and treat AI risks

Responsible AI Framework

Standards

Independent verification of Legal, Ethical and Technical aspects, building trust with consumers, regulators, suppliers and national bodies



Legal

Proactive compliance to AI, data and fundamental rights-related matters

Technical

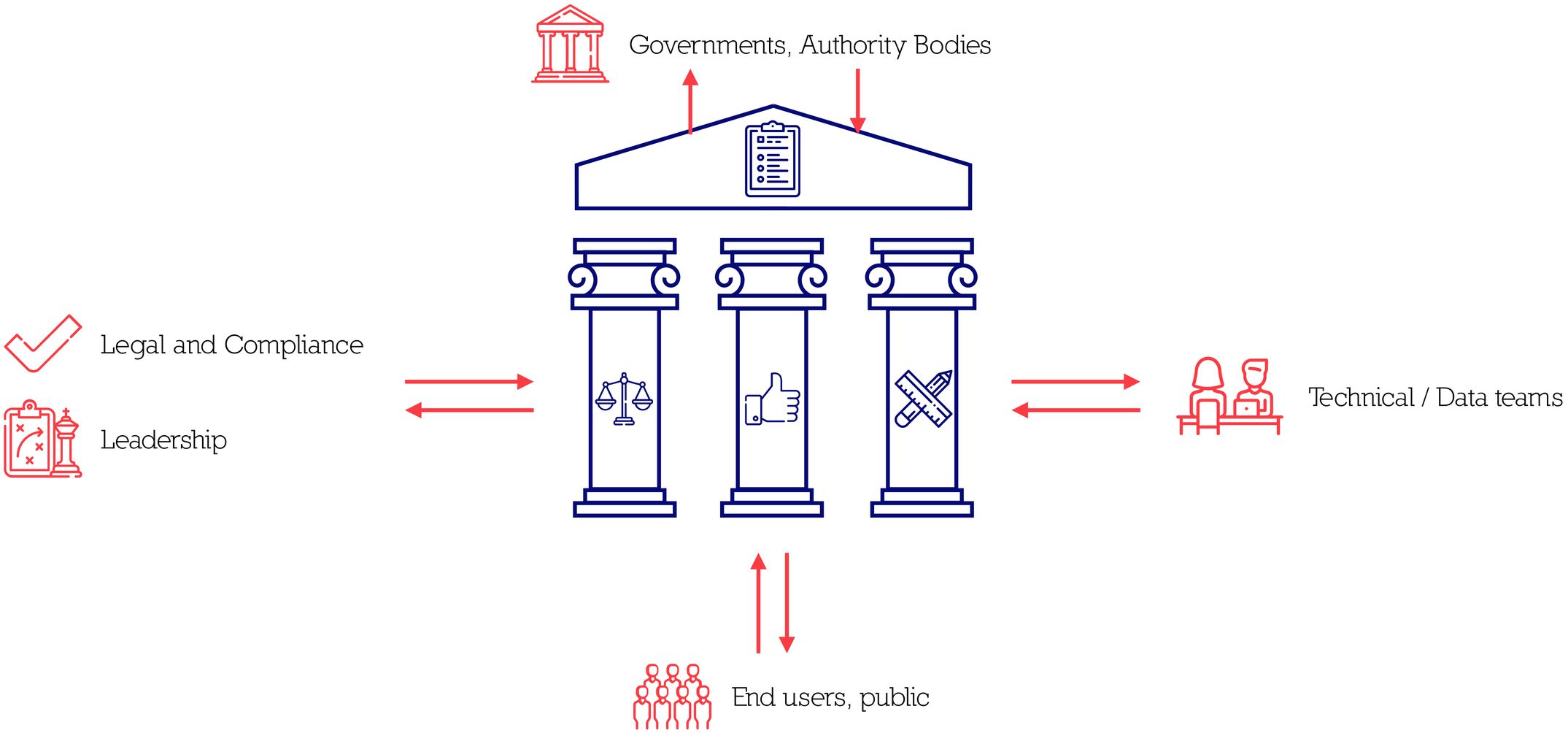
Technical implementation of design

Ethical

Incorporate AI ethics principles by design

Responsible AI Framework

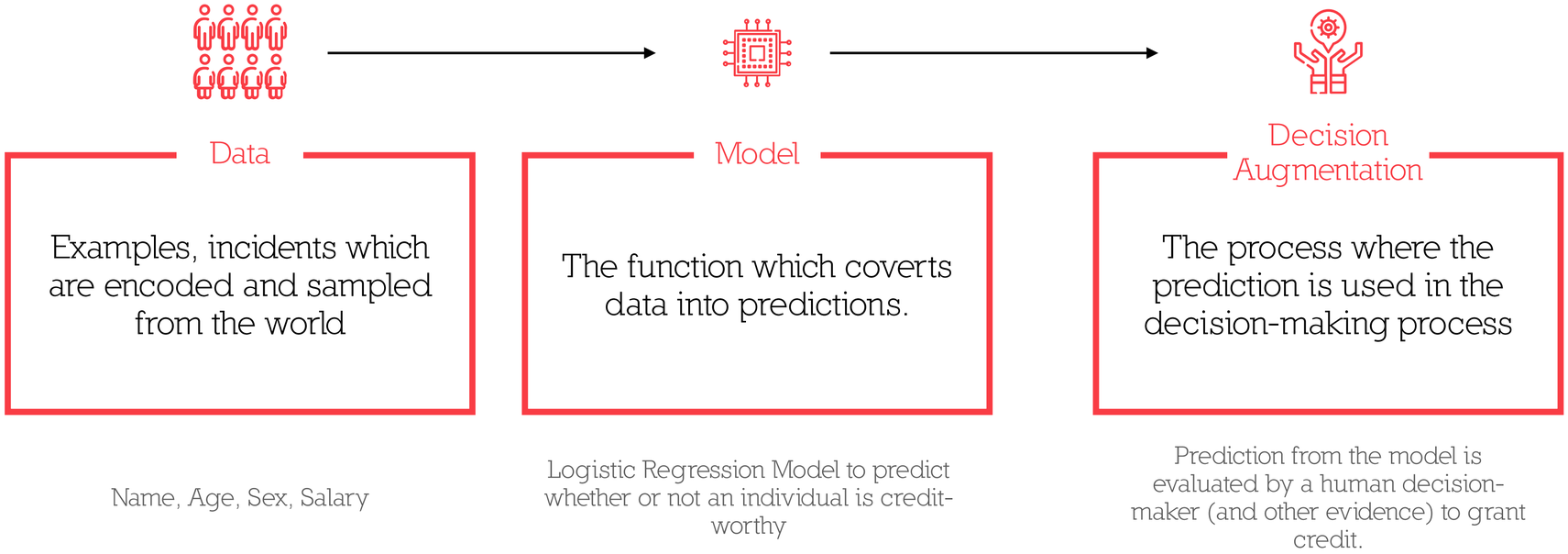
Building an ecosystem of collaboration and innovation



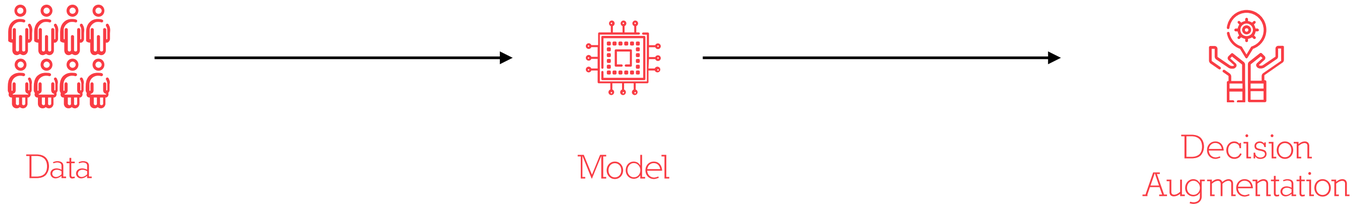
Types of AI Risk

AI Value Chain + AI Ethics Principle

AI Value Chain



AI Value Chain



Radiologist further examines sections that model pointed out as being highly probable as cancerous

Whole-slide images of breast tissue with labelled cancerous regions

Convolutional Neural Network to produce probability that area is cancerous

AI Value Chain

Data → Model → Decision Augmentation

Data

Whole-slide images of breast tissue with labelled cancerous regions

Model

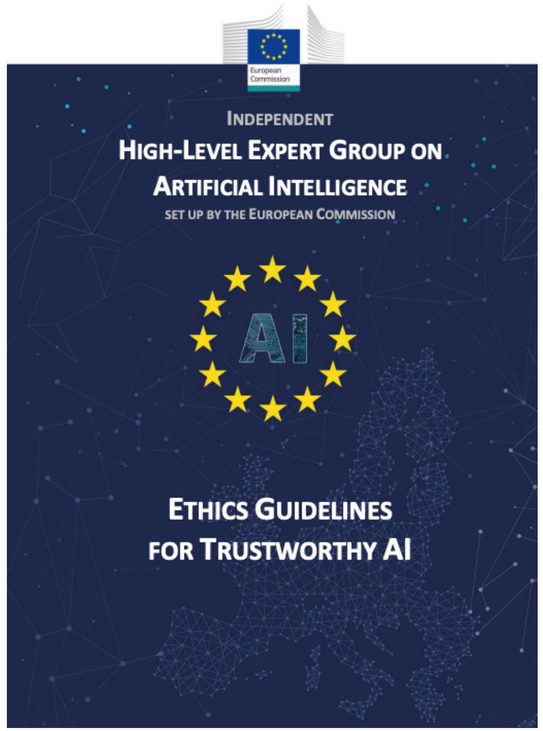
Convolutional Neural Network to produce probability that area is cancerous

Decision Augmentation

Radiologist further examines sections that model pointed out as being highly probable as cancerous

AI Ethics Principles

- Privacy
- Transparency
- Fairness
- Technical Robustness + Safety
- Human Agency + Oversight
- Societal + Environmental well-being
- Accountability



For more info on all of the Experts involved: https://ec.europa.eu/futurium/en/european-ai-alliance/ai-hleg-steering-group-european-ai-alliance.html
Ethics guidelines for trustworthy AI report: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Classifying AI risk

"No takesies backsies"

A Chevrolet dealership in the US utilised Generative AI, as part of a customer-facing chatbot solution.

A customer was able to alter the behaviour of the chatbot via prompt injection, to accept customer's offer to purchase a 2024 Chevrolet Tahoe for $1 as "legally binding" with no "takesies backsies".

AI Value Chain

+

AI Ethics Principle

Classifying AI risk

"No takesies backsies"

A Chevrolet dealership in the US utilised Generative AI, as part of a customer-facing chatbot solution.

A customer was able to alter the behaviour of the chatbot via prompt injection, to accept customer's offer to purchase a 2024 Chevrolet Tahoe for $1 as "legally binding" with no "takesies backsies".

AI Value Chain

Model

Model

+

AI Ethics Principle

Technical Robustness + Safety

Human Oversight

Classifying AI risk

Predicting your sexual orientation

Netflix's recommendation engine's ability to predict "propensity to like queer content" discovered sexual orientation of user from recommendation engine.

AI Value Chain

+

AI Ethics Principle

https://www.bbc.com/news/technology-66472938

Classifying AI risk

Predicting your sexual orientation

Netflix's recommendation engine's ability to predict "propensity to like queer content" discovered sexual orientation of user from recommendation engine.

AI Value Chain

Model

+

AI Ethics Principle

Privacy

https://www.bbc.com/news/technology-66472938

AI Risk Treatment

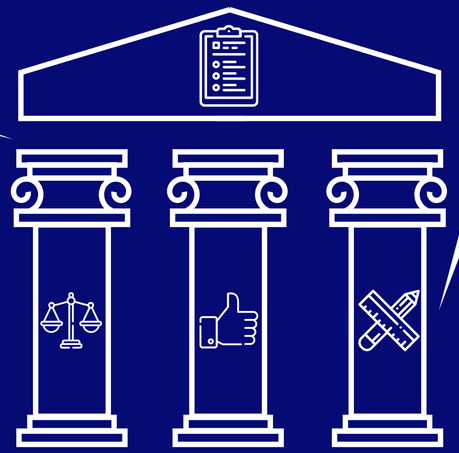
AI Risk Treatment

Dilemma: Model Attacks

A bank developed technology to enable users to access their bank account through facial recognition.

You discover that your friend can gain access by simply holding up a picture of you in front of the phone camera.

Legal?
Standards?



Technical?

Dilemma = Actual İncident

Prevention of Model Attacks

Facial-recognition locks on delivery lockers were easily opened by a group of fourth-graders in a science-club demo using only a printed photo of the intended recipient's face, leaving contents vulnerable to theft.

Legal?
Standards?

- EU Aİ Act - Art 15
- (EU Aİ Act - GPAİ + Systemic Risk)
- İSO 42001

Technical?

- Define potential data and model attacks
- Attack evaluation
- Training with adversarial examples

https://incidentdatabase.ai/cite/223

AI Risk Treatment: Model Attacks

Legal + Standards

- EU AI Act - Robustness + Safety for High-Risk AI Systems
- EU AI Act - GPAI + Systemic risk
- ISO 42001

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act Risk Hierarchy

EU AI Act for
AI developers

Increasing risk to
Health, safety and/or Fundamental Rights

Unacceptable
e.g. Social scoring

High Systems affecting access to education, employment, healthcare & justice.

Limited e.g. Chatbots, deep-fakes etc.

Minimal e.g. Spam filters, Video games

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act Risk Hierarchy

High	Systems affecting access to education, employment, healthcare & justice.

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act - High-Risk Obligations

- Conformity assessment (audit) BEFORE the AI offering is placed on the market. (Article 6.1b)
- Human Oversight to monitor and address anomalies. (Art 14)
- AI system will need to be registered in central EU database (Article 51)

Art 15

"designed and developed [to] achieve an appropriate level of accuracy, robustness, and cybersecurity … throughout their lifecycle"

"resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates."

"solutions to address AI specific vulnerabilities shall include … measures to prevent, detect, respond to […] attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks"

https://artificialintelligenceact.eu/article/15/

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act - GPAI

GPAI definition

"an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks"

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act - GPAI

GPAI + Systemic Risk

- Has high impact capabilities (proposed as having cumulative amount of computation used for training is above 10^{25} floating point operations (FLOPs))
- Evaluated by the European Commission and/or scientific panel, based on criteria, including, but not limited to:
  - Number of model parameters
  - Dataset size and/or quality (e.g. measured in tokens)
  - Estimated energy consumption for training
  - Model's "reach" (made available to > 10000 European businesses, number of registered end-users)

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act - GPAI

GPAI + Systemic Risk

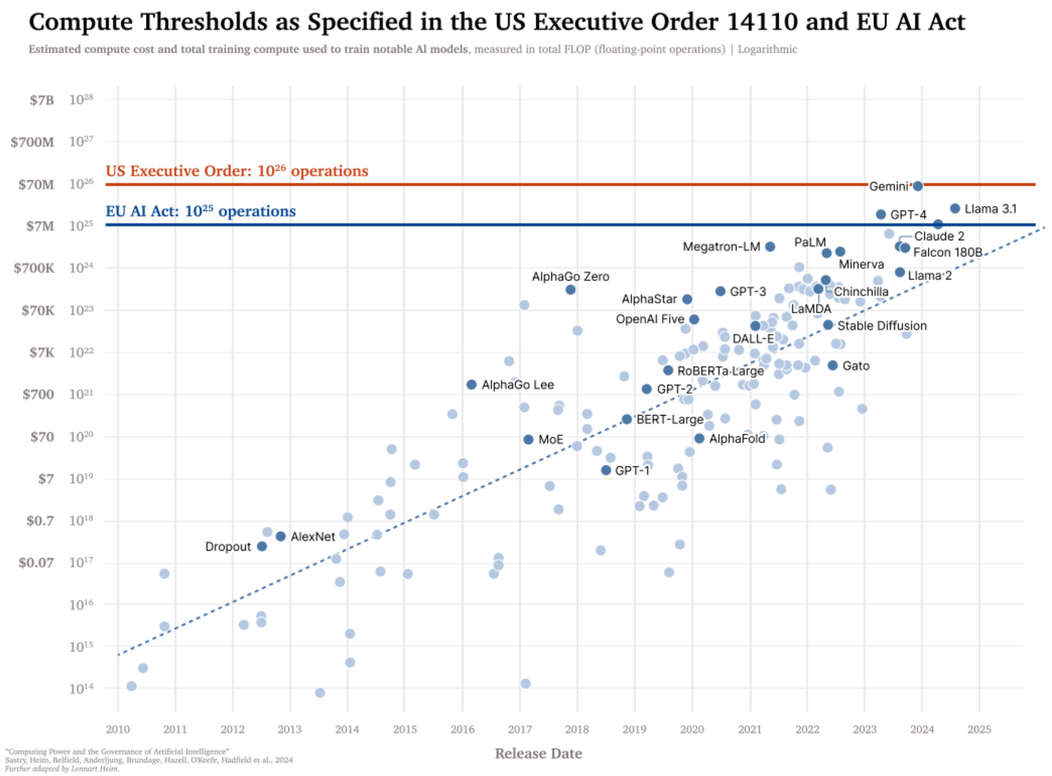


Figure 1: Training compute has been increasing at a fast rate, doubling roughly every 6 months (4× per year). The US AI EO introduces reporting requirements for models trained with more than 10^{26} operations. The EU AI Act presumes a GPAI model poses systemic risk and imposes a variety of requirements for models trained with more than 10^{25} operations.

Heim, L. (2024). Training Compute Thresholds: Features and Functions in AI Governance. *arXiv preprint arXiv:2405.10799*.

AI Risk Treatment: Model Attacks

Legal + Standards: EU AI Act - GPAI + Systemic Risk Obligations

- Cybersecurity protection is enabled
- Adversarial testing (stress-testing model)
- Risk assessment and mitigation
- Document and report serious incidents to the AI Office

AI Risk Treatment: Model Attacks

Legal + Standards: ISO 42001

The ISO 42001 standard was published in Dec 2023 to empower organisations to develop an AI Management system, which is highly aligned with the EU AI Act.

The standard focusses on the concept of an AI management system, which consists of:

- Defining organisational objectives
- Evaluating and managing risks and opportunities
- Build trust into AI systems

Webinar series on ISO 42001

Treating against Model Attacks

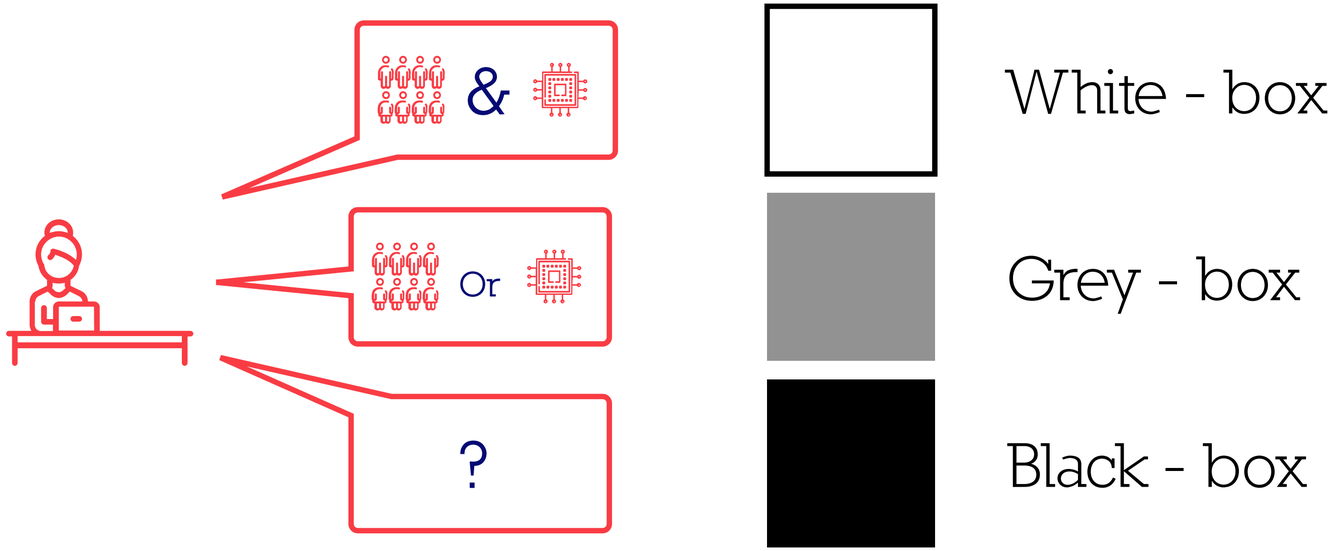
Technical: Model Attacks Terminology

- Adversary Knowledge
- Data Poisoning
- Model Evasion

Treating against Model Attacks

Technical: Model Attacks Terminology

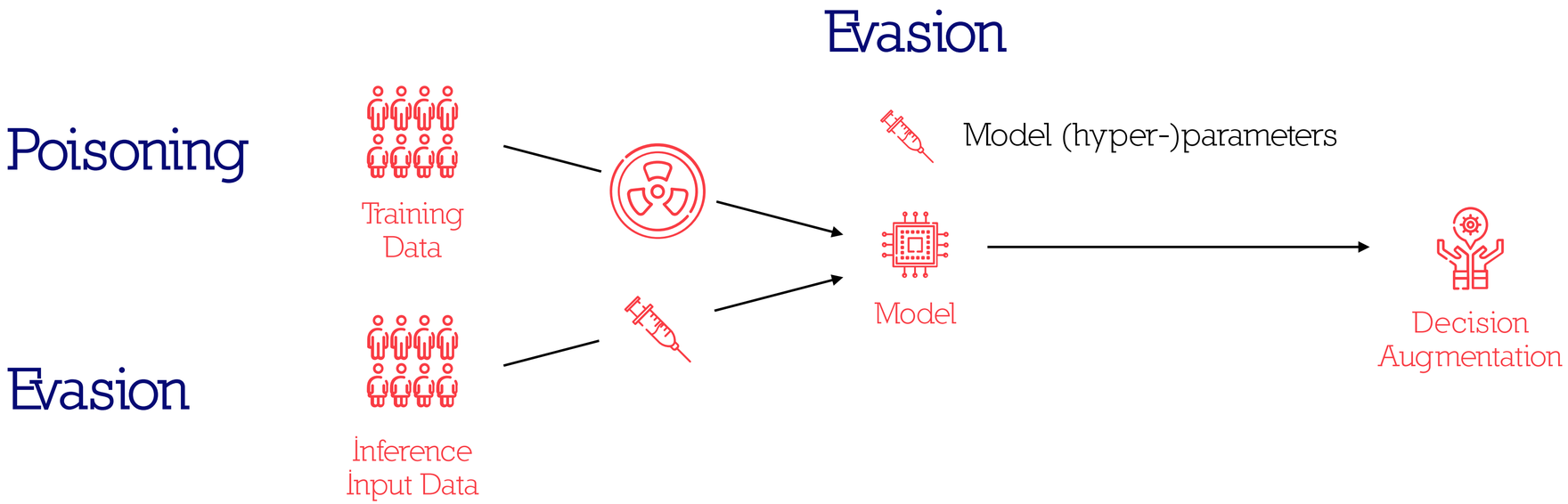
Adversary Knowledge



Bountakas, P., Zarras, A., Lekidis, A., & Xenakis, C. (2023). Defense strategies for adversarial machine learning: A survey. *Computer Science Review*, *49*, 100573.

Treating against Model Attacks

Technical: Model Attacks Terminology



Bountakas, P., Zarras, A., Lekidis, A., & Xenakis, C. (2023). Defense strategies for adversarial machine learning: A survey. *Computer Science Review*, *49*, 100573.

Treating against Model Attacks

Technical: Model Attacks

Poisoning

- Input Feature Manipulation
- Label Manipulation
- "Backdoor" attacks

Evasion

- Inference input data manipulation

"Two McAfee researchers demonstrated how using only black electrical tape could trick a 2016 Tesla into a dangerous burst of acceleration by changing a speed limit sign from 35 mph to 85 mph."

https://www.ibm.com/docs/en/watsonx/saas?topic=atlas-evasion-attack

Bountakas, P., Zarras, A., Lekidis, A., & Xenakis, C. (2023). Defense strategies for adversarial machine learning: A survey. *Computer Science Review*, *49*, 100573.

Treating against Model Attacks

Action points

Attack Evaluation

Measuring attack specificity (does the attack focus on modifying specific examples or a general class of examples).

Evaluating attack performance in classification models via changes to false positive and false negative rates.

Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., & Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. *Computer Science Review*, *34*, 100199.

Model training with adversarial examples

Including perturbed training examples (examples with noise) can build resilience into the final model against attacks.

Bountakas, P., Zarras, A., Lekidis, A., & Xenakis, C. (2023). Defense strategies for adversarial machine learning: A survey. *Computer Science Review*, *49*, 100573.

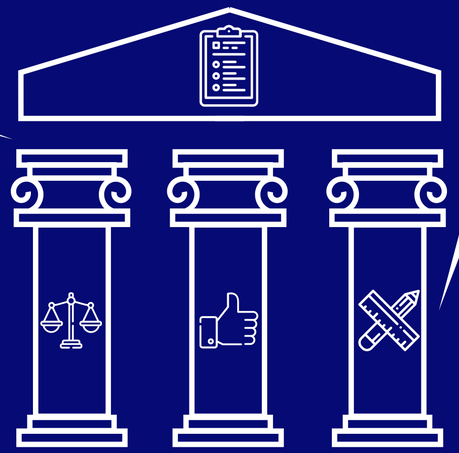
Principles

AI Risk Treatment

Dilemma: Social Impact

A doctor chatbot advises patients to overdose on pain-killers to reduce pain.
Overdosing can result in death.

Legal?
Standards?



Technical?

Dilemma = Actual İncident

Social impact

Belgian man provoked by Chat-GPT to commit suicide to ‘‘sacrifice himself to stop climate change’’

Legal?
Standards?

- EU Aİ Act - GPAİ + Systemic Risk
- İSO 42001 - Aİ İmpact Assessment

Technical?

- Logging of user activity˝
- Human vs Aİ Control debate˝
- Aİ Auditing˝
- Post-marketing surveillance strategy

https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-

AI Risk Treatment: Social Impact

Legal + Standards

- EU AI Act - Human Oversight Obligations for High-Risk AI Systems
- EU AI Act: GPAI + Systemic Risk
- ISO 42001 - AI Impact Assessment

AI Risk Treatment

Legal + Standards: EU AI Act - GPAI

GPAI + Systemic Risk

- Has high impact capabilities (proposed as having cumulative amount of computation used for training is above 10^{25} floating point operations (FLOPs))
- Evaluated by the European Commission and/or scientific panel, based on criteria, including, but not limited to:

Number of model parameters

Dataset size and/or quality (e.g. measured in tokens)

Estimated energy consumption for training

Model's "reach" (made available to > 10000 European businesses, number of registered end-users)

AI Risk Treatment

Legal + Standards: EU AI Act - High-Risk Obligations

High-Risk AI

Art 14

"appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use"

"to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance"

"to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system"

"to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state"

https://artificialintelligenceact.eu/article/14/

AI Risk Treatment

Legal + Standards: ISO 42001

A unique part of the ISO 42001 standard is the documented assessment of an AI system's impact on society

Will certain individuals/ groups affected?

Can it affect social structures/politics?

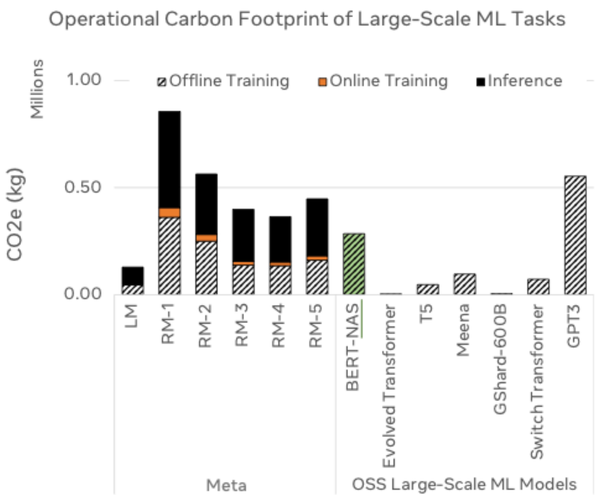
Communicate to all interested parties

Webinar series on ISO 42001

AI Risk Treatment

Academia: The Environmental Costs of AI

1



*Training footprint **only**

3

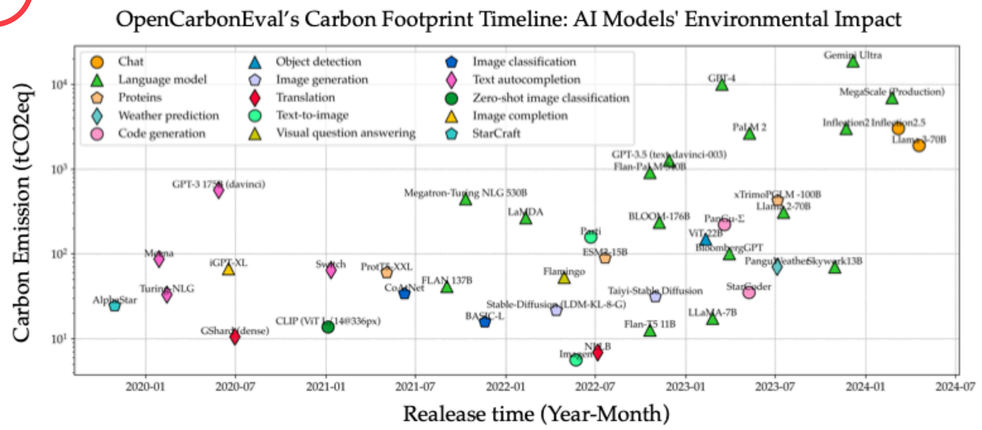


Figure 1: Large-scale models' environmental impact covering 42 large-scale AI models across 15 tasks. OpenCarbonEval enables the estimation of carbon emissions for various models, facilitating a more sustainable AI development process.

2

"A ChatGPT-like application with estimated use of 11 million requests/hour produces emissions of **12.8k metric ton CO2/ year**, 25 times the emissions for training GPT-3. Inference is critical to environmental and power cost."

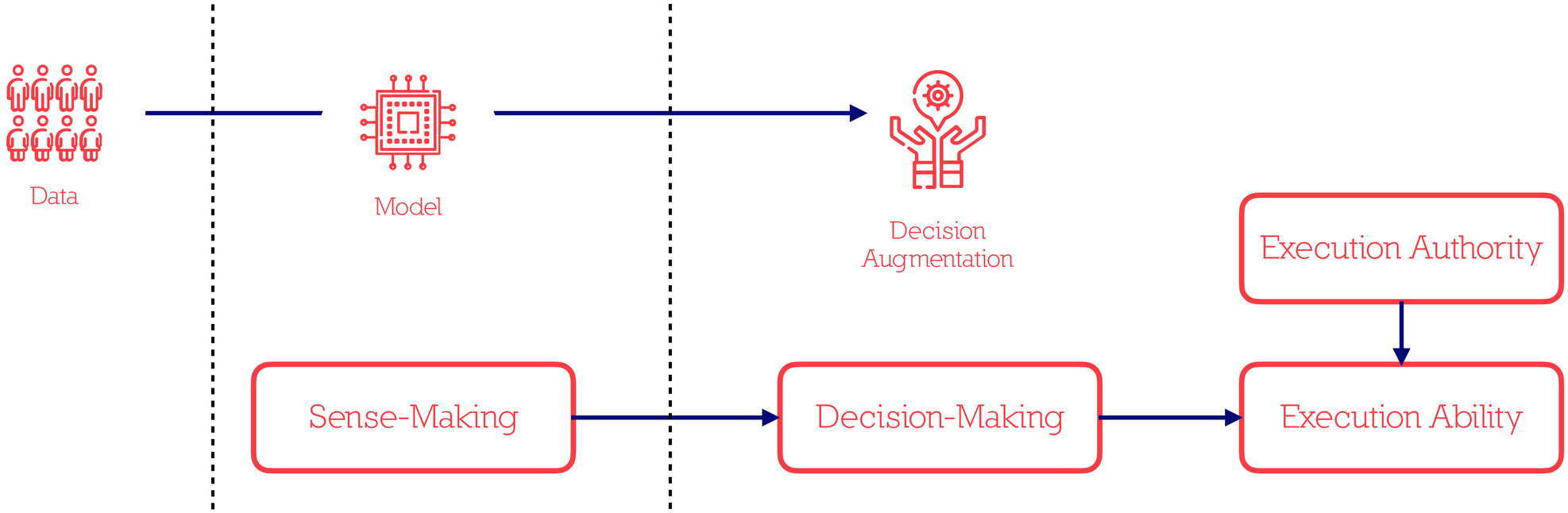
4

"For example, the IEA estimates that datacenters' energy consumption will double from 2022 to 2026, equalling Japan's energy consumption, while approximate 660 million people still lack access to electricity based on UN reports."

(1) Wu, C. J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., ... & Hazelwood, K. (2022). Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, *4*, 795-813.
(2) Chien, A. A., Lin, L., Nguyen, H., Rao, V., Sharma, T., & Wijayawardana, R. (2023, July). Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd workshop on sustainable computer systems* (pp. 1-7).
(3) Yu, Z., Wu, Y., Deng, Z., Tang, Y., & Zhang, X. P. (2024). OpenCarbonEval: A Unified Carbon Emission Estimation Framework in Large-Scale AI Models. *arXiv preprint arXiv:2405.12843*.
(4) Yu, Z., Wu, Y., Deng, Z., Tang, Y., & Zhang, X. P. (2024). OpenCarbonEval: A Unified Carbon Emission Estimation Framework in Large-Scale AI Models. *arXiv preprint arXiv:2405.12843*.

AI Risk Treatment

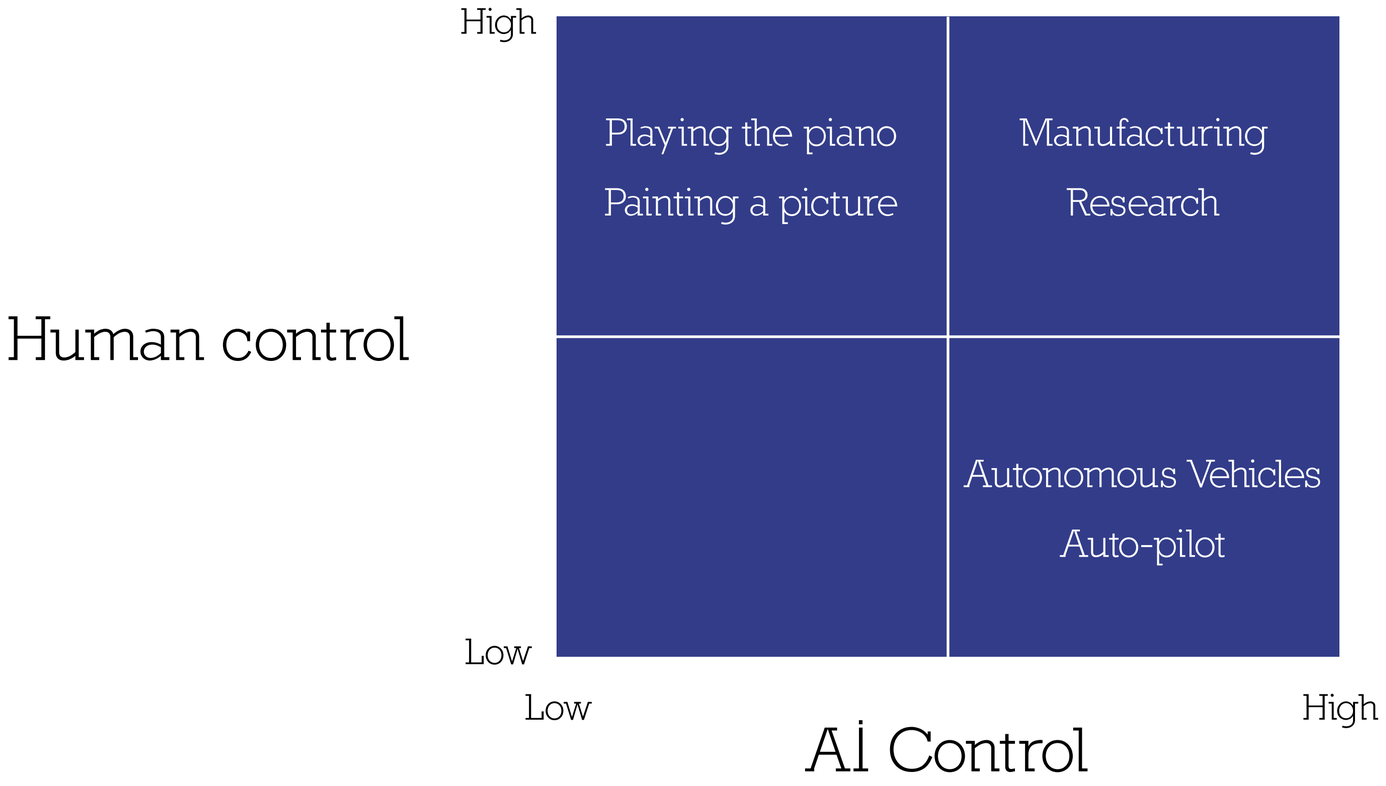
Academia: Human vs AI Control



Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., ... & Xu, W. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391-437.

AI Risk Treatment

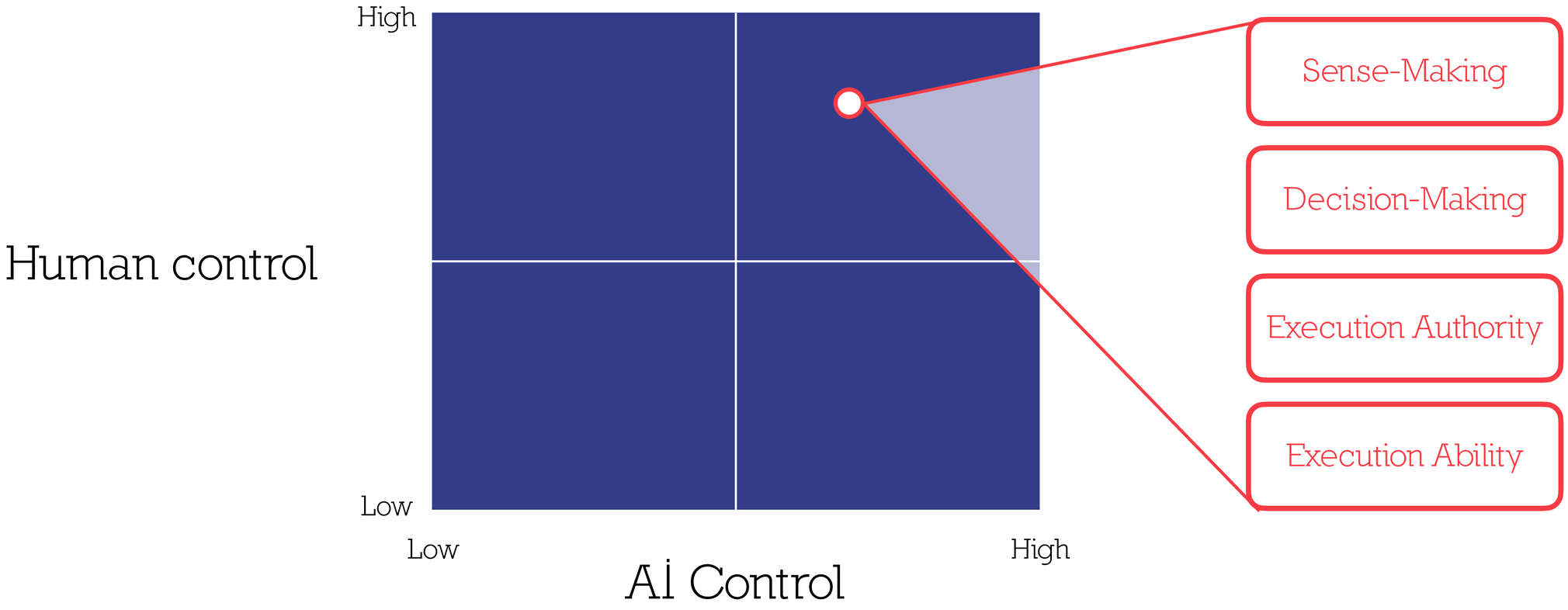
Academia: Human vs AI Control



Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.

AI Risk Treatment

Academia: Human vs AI Control



Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.
Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., ... & Xu, W. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391-437.

AI Risk Treatment

Academia: Limitations to Human Oversight

- Degree of training required to undertake the role of human overseer.
  - Under qualified will simply accept model performance, through automation bias.
  - (Over-)qualified will tend to over-rely on their own judgement rather than model predictions.
- Lack of incentive structures around human oversight
- Human oversight can create false sense of security and trust in an AI system

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681.

AI Risk Treatment

Action points

AI Impact Assessment	To document particular societal and environmental risks a particular AI system can create through it's lifecycle (from inception to decommissioning).
	ISO 42001
AI Auditing	Impartial review of AI systems from a credible person can ensure compliant and competent oversight.
	Ojewale, V., Steed, R., Vecchione, B., Birhane, A., & Raji, I. D. (2024). Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *arXiv preprint arXiv:2402.17861*.
Post-marketing surveillance strategy	To have a plan to provide resource (people, tooling) and mechanism to feedback model usage to management and developers to assess and mitigate evolving AI risks.

Trade-offs

Tradeoffs

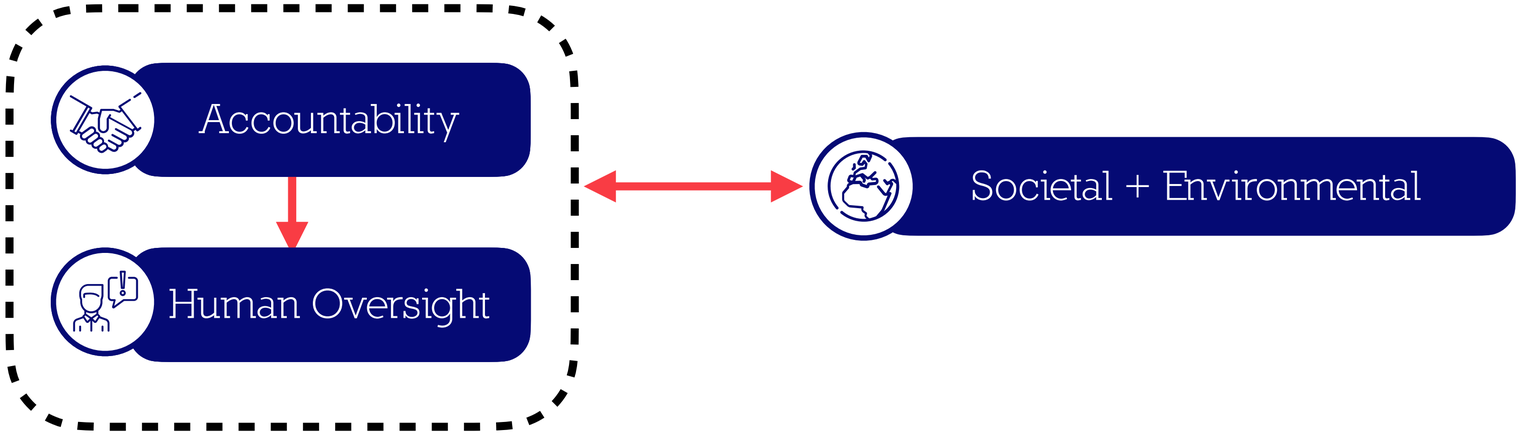
Exercises

Draw a diagram to connect all AI Ethics Principles together

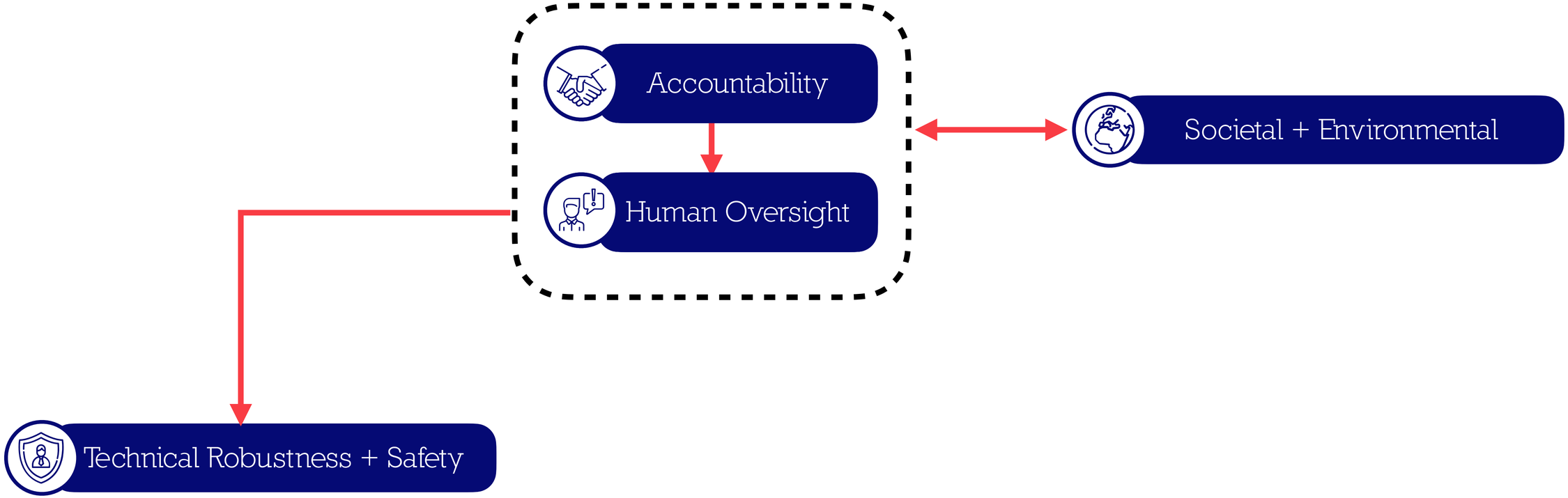
	Privacy	Transparency	Fairness	Technical Robustness + Safety	Human Agency + Oversight	Societal + Environmental well-being	Accountability
Privacy		Y					
Transparency							
Fairness							
Technical Robustness + Safety							
Human Agency + Oversight							
Societal + Environmental well-being							
Accountability					Y		

Tradeoff Overview

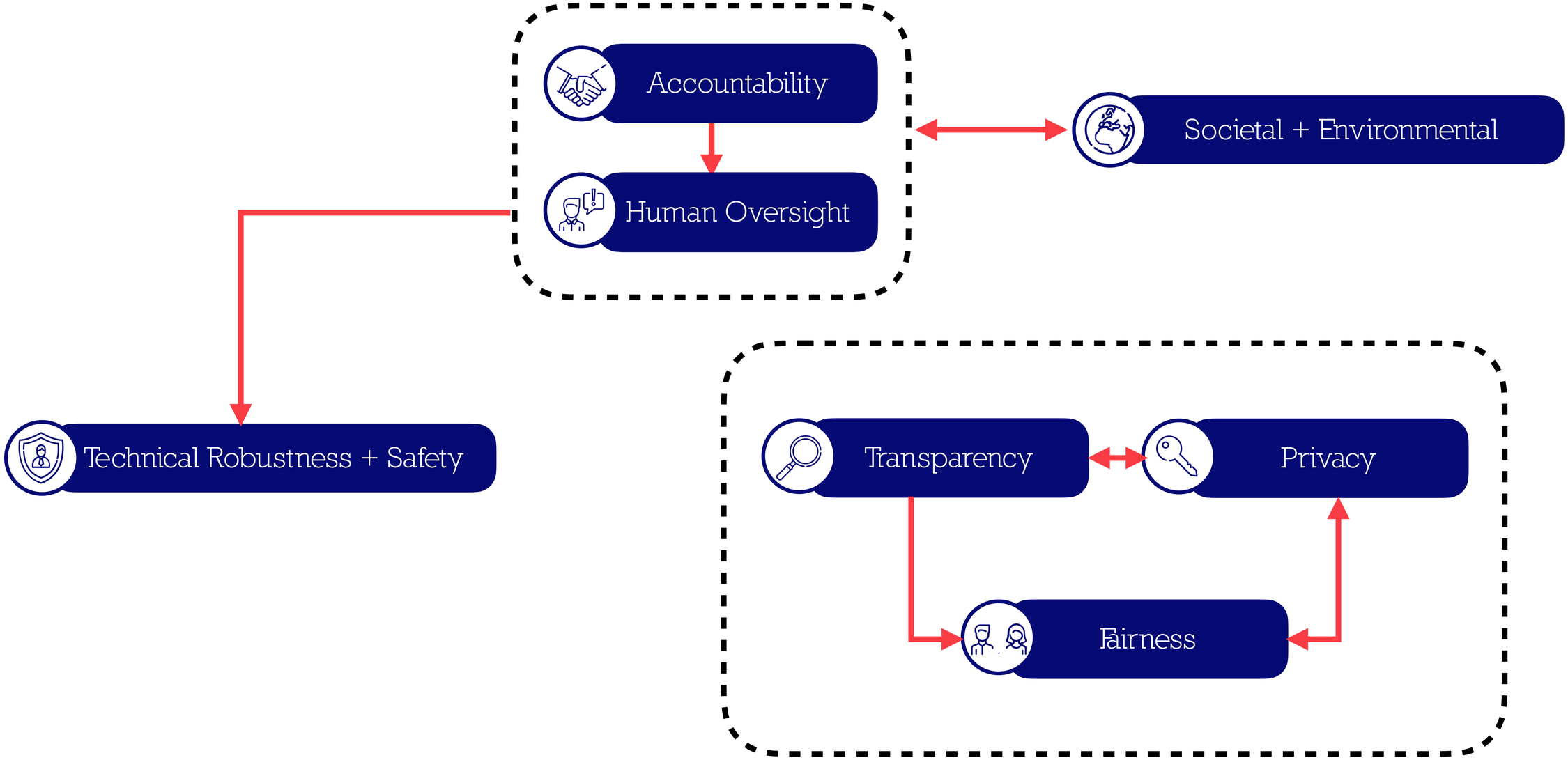
Tradeoff Overview



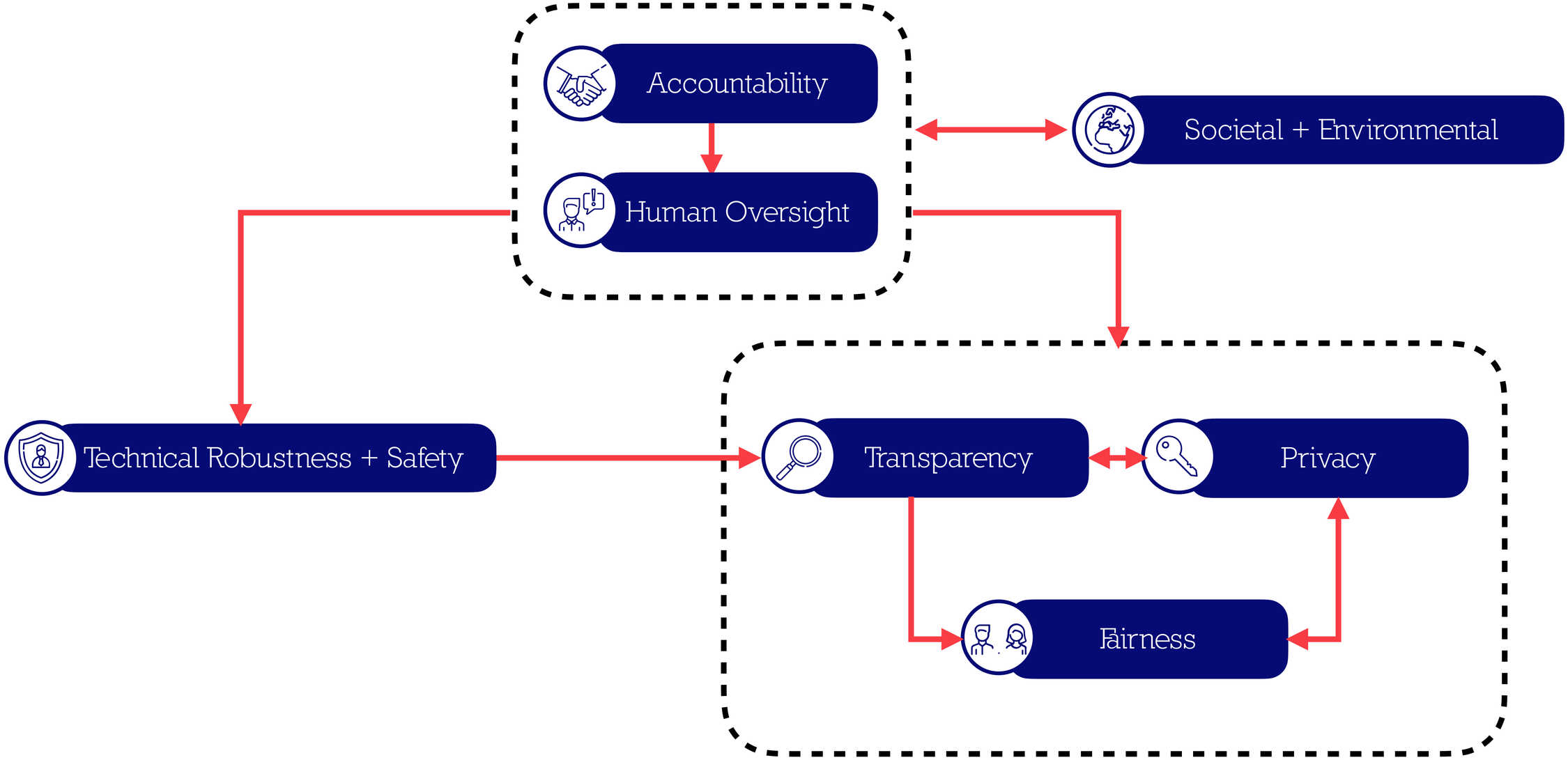
Tradeoff Overview



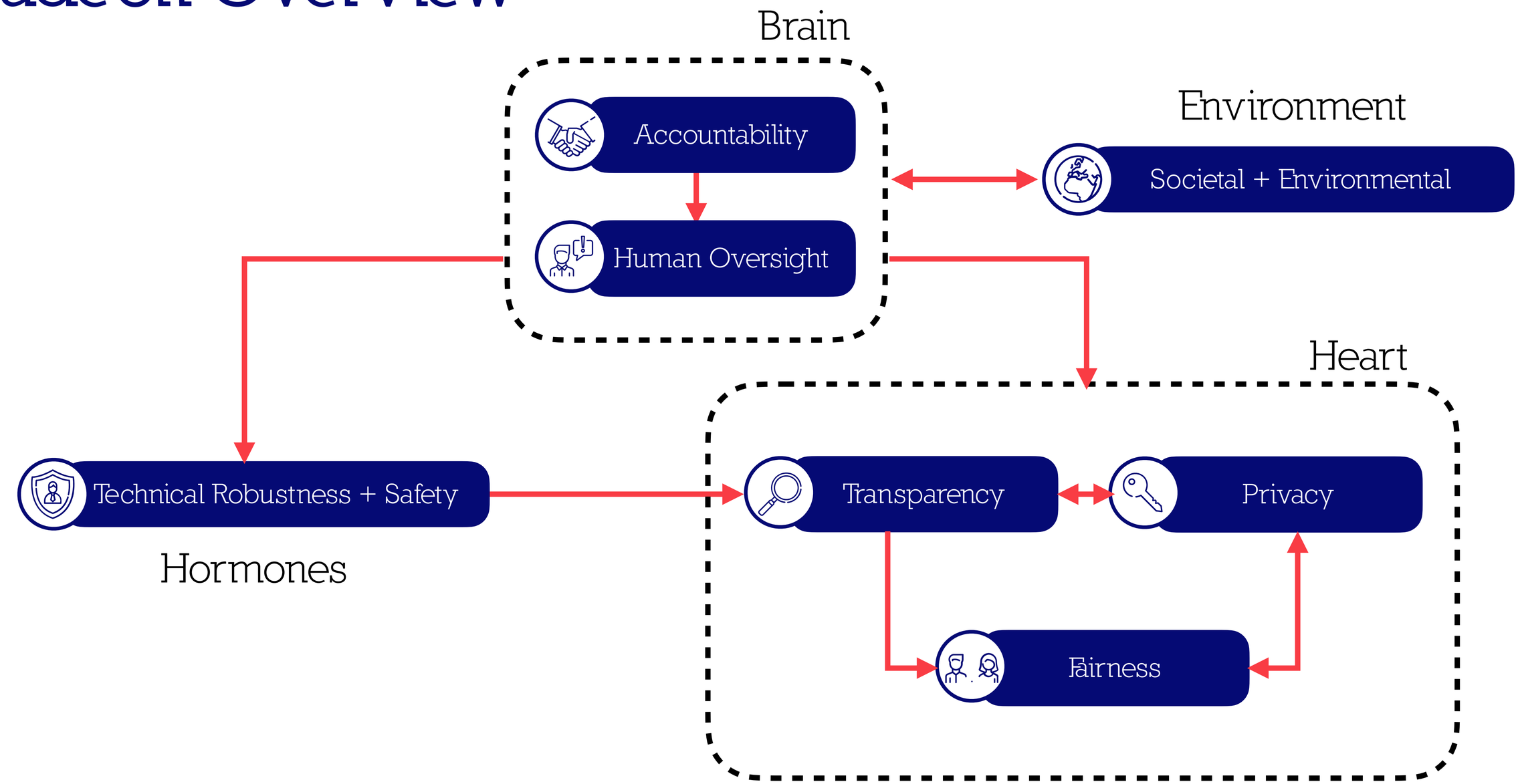
Tradeoff Overview



Tradeoff Overview

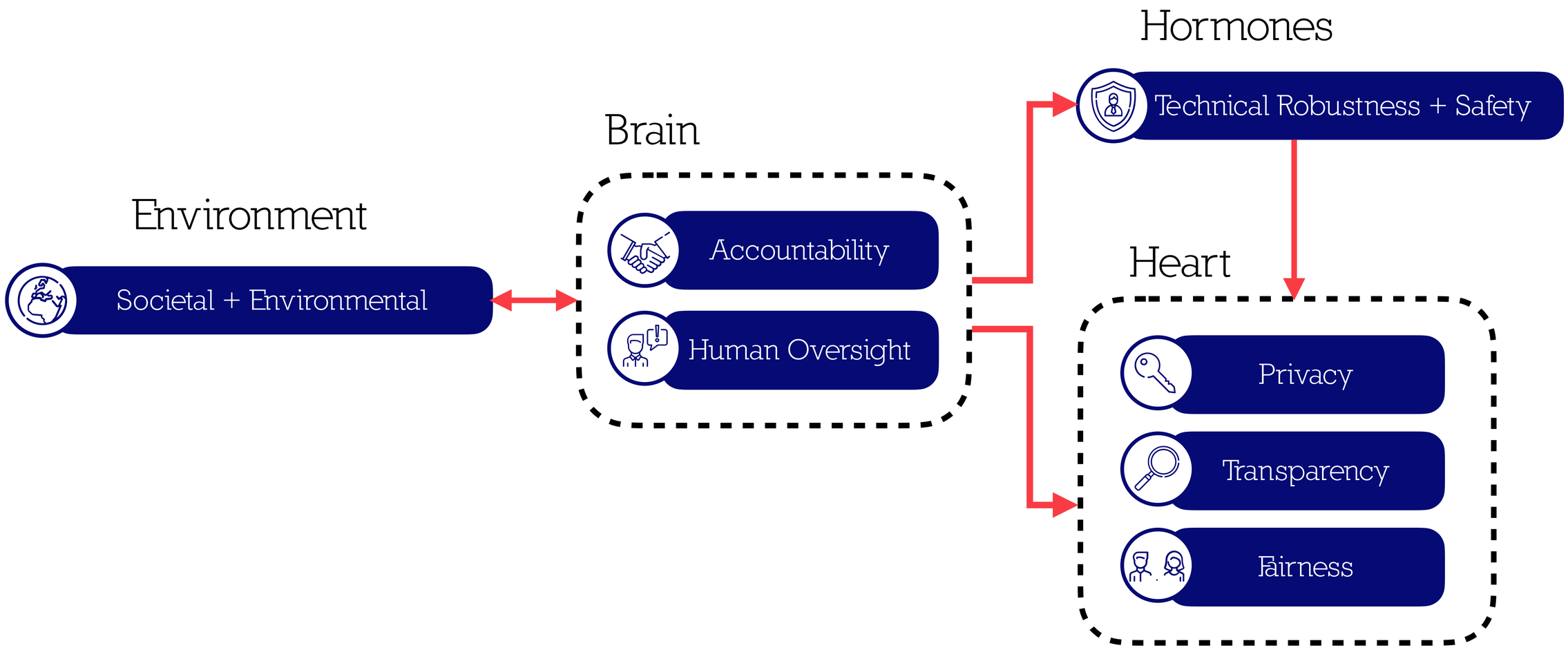


Tradeoff Overview



Tradeoffs

Putting the R into AI Projects



Tradeoffs

Putting the R into AI Projects

Environment

- What does our market need?
- What are our climate/environment obligations?
- How does this align with our mission?

Brain

- What are our (internal) auditing procedures?
- Who will evaluate our systems? What intervals?
- How will we feed this back to suppliers/clients?

Hormones

- What is our product success metrics?
- What cybersecurity considerations have we made?

Heart

- How we ensure privacy when using personal data?
- How can we understand model predictions?
- How can we make sure that the model produces equitable outcomes?

To recap…

Goals

- Understand the current AI risk landscape˝
- How to classify risks against the common AI ethical principles
- Apply the Responsible AI framework to manage for AI risks and produce more trustworthy technologies
- Understand tradeoffs between each of the AI ethics principles and the relationships between each principle when controlling for AI risks

Classifying AI risk

AI Value Chain

Data

↓

Model

↓

Decision Augmentation

+

AI Ethical Principle

- Privacy
- Transparency
- Fairness
- Technical Robustness + Safety
- Human Agency + Oversight
- Societal + Environmental well-being
- Accountability

Classifying AI risk

A Chevrolet dealership in the US utilised Generative AI, as part of a customer-facing chatbot solution.

A customer was able to alter the behaviour of the chatbot via prompt injection, to accept customer's offer to purchase a 2024 Chevrolet Tahoe for $1 as "legally binding" with no "takesies backsies".

AI Value Chain

Model

Model

+

AI Ethics Principle

Technical Robustness + Safety

Human Oversight

Tradeoffs

Putting the R into AI Projects

Environment

- What does our market need?
- What are our climate/environment obligations?
- How does this align with our mission?

Brain

- What are our (internal) auditing procedures?
- Who will evaluate our systems? What intervals?
- How will we feed this back to suppliers/clients?

Hormones

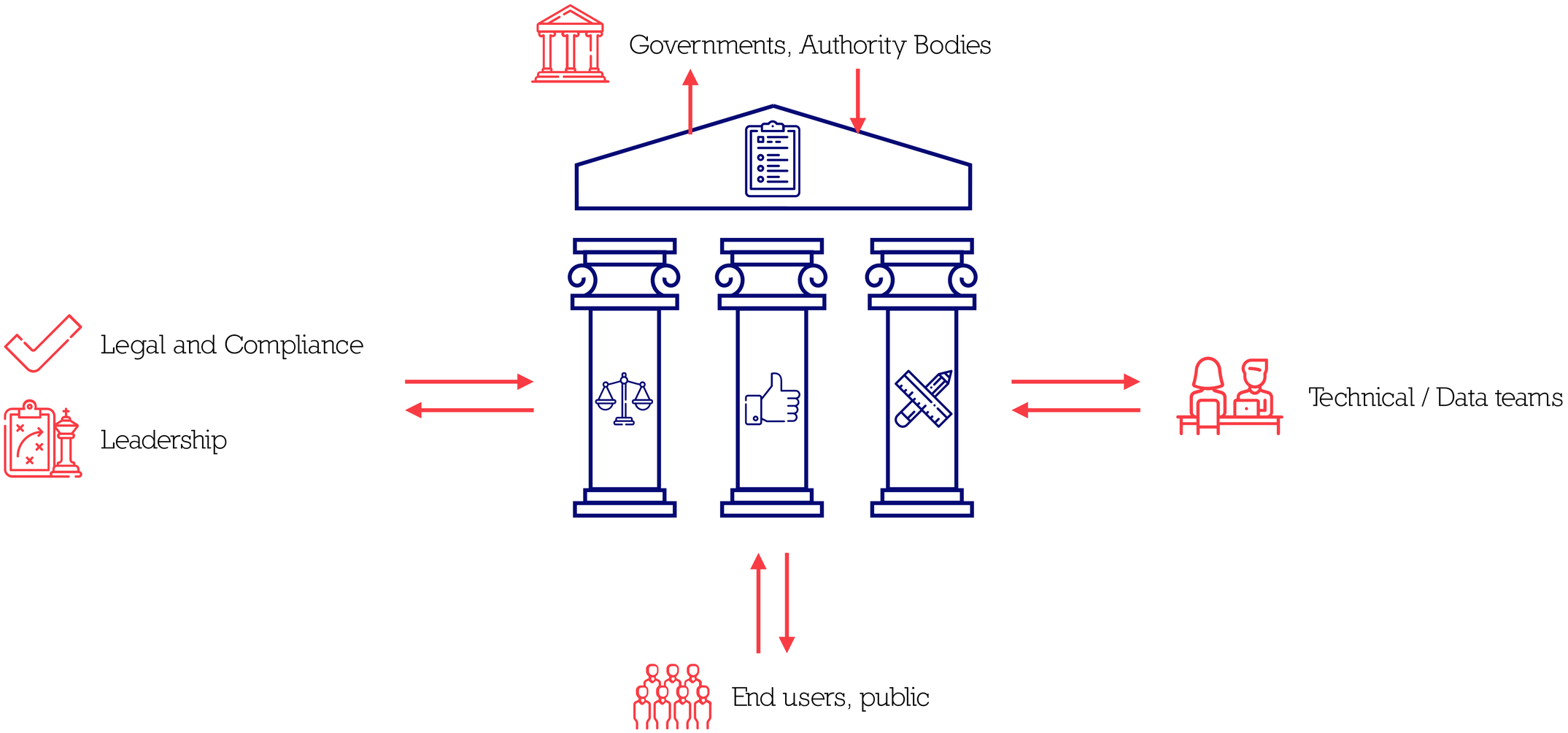
- What is our product success metrics?
- What cybersecurity considerations have we made?

Heart

- How we ensure privacy when using personal data?
- How can we understand model predictions?
- How can we make sure that the model produces equitable outcomes?

Responsible AI Framework

Building an ecosystem of collaboration to identify and treat AI risk



Questions and Discussions