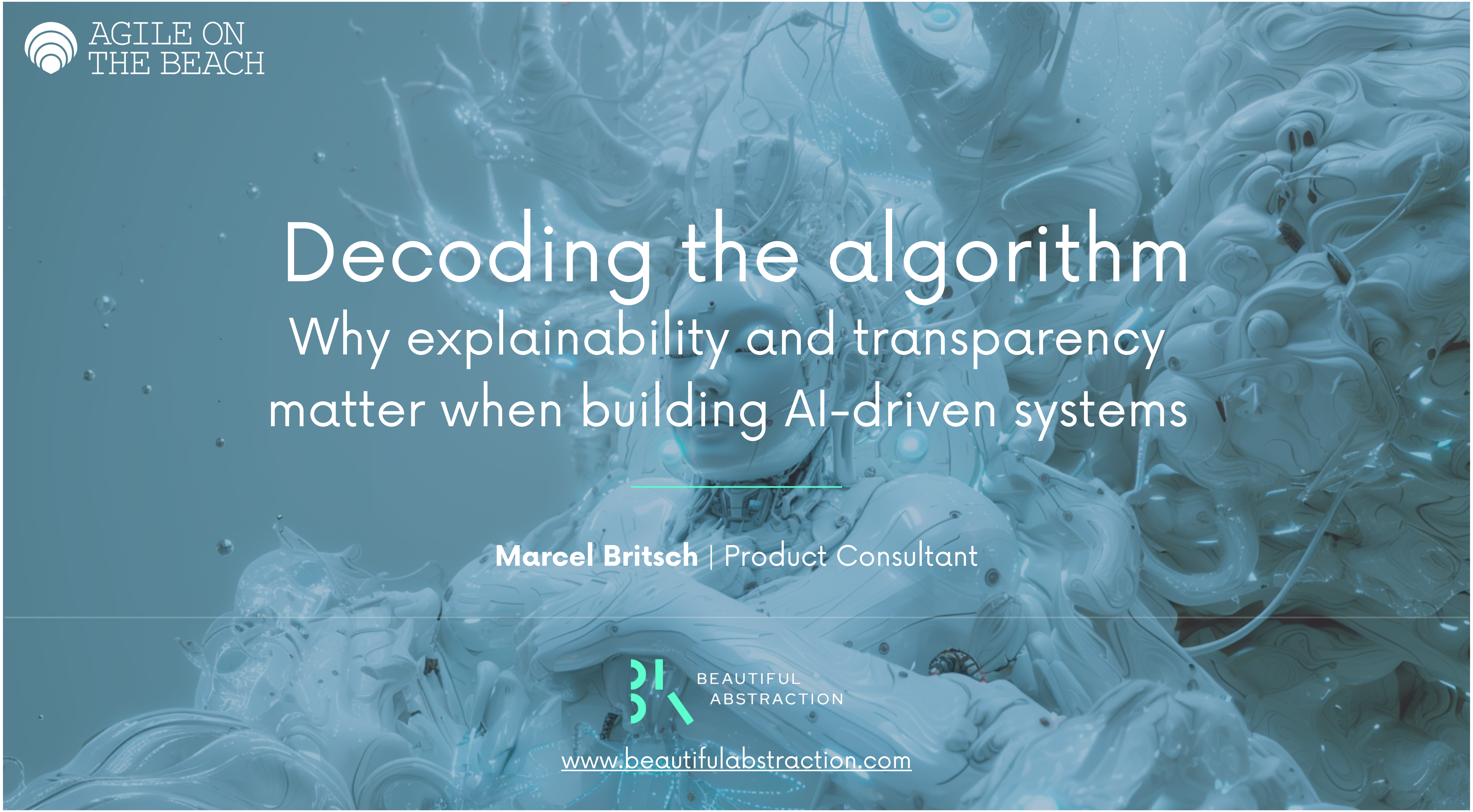


AGILE ON THE BEACH

Decoding the algorithm

Why explainability and transparency matter when building AI-driven systems

Marcel Britsch | Product Consultant

BEAUTIFUL ABSTRACTION

www.beautifulabstraction.com

AGILE ON
THE BEACH

How civilisation~~s~~ ends

Musings on ~~existential~~ risk

image: Midjourney

image: Midjourney

What Existential Risks?

image: Midjourney

Natural Disasters

image: Midjourney

Biotechnology

image: Midjourney

Nano technology

image: Midjourney

Physics Experiments

image: Midjourney

Nuclear annihilation

image: Midjourney

Unaligned AI

image: Midjourney

Where is everyone?

Fermi Paradox - Enrico Fermi

image: Midjourney

image: Midjourney

Are they hiding?

Dark Forest Hypothesis, Cixin Liu

Maybe there is no-one else?

Berserker hypothesis, Fred Saberhagen

image: Midjourney

The Great Filter

Robin Hanson

image: Midjourney

Are we the only (intelligent) life that ever evolved?

image: Midjourney

Or

Did everyone else perish at it this very point?

image: Midjourney

We may haver just entered the Great Filter.

image: Midjourney

Why does it matter?

Existential risks are bad, but so are minor fuck-ups.

image: Midjourney

Why transparency and explainability matter when building AI-driven systems

Today

- What is algorithmic decision making?
- Why can it be problematic?
- How does AI feature in this?
- What is explainability and why do we want it?
- What does the 'the law' say?
- How does explainably done well look like?
- How do you rationalise outputs?
- What are the benefits of explainability?

Always consult with your trusted data scientist and legal counsel

What is algorithmic decision making?

Classification:

Ceci n'est pas un chat.

Prediction:

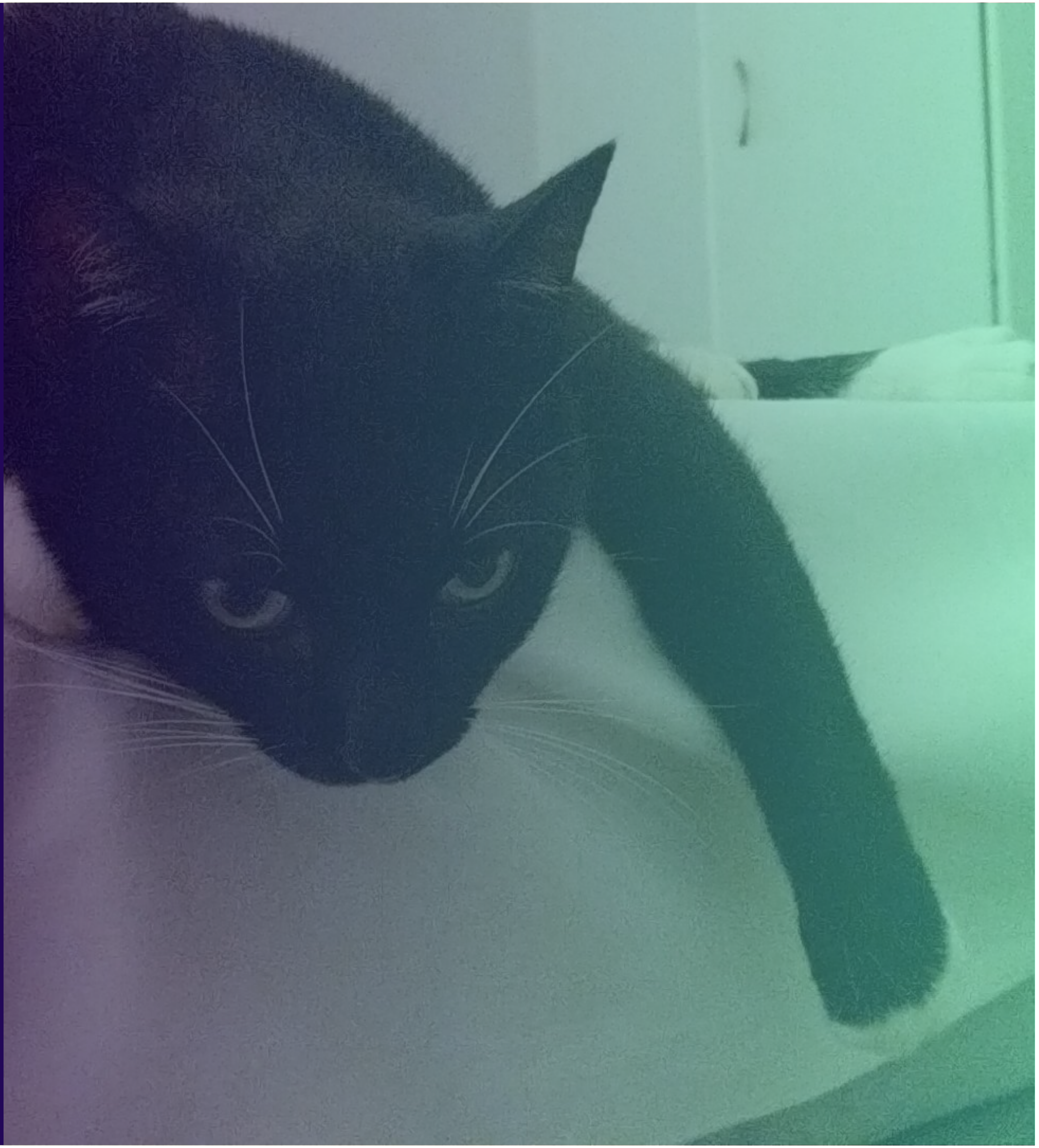
This cat will bite!

Recommendation:

Buy this cat toy!

Generation:

Have a cat picture



Algorithmic decision making: the process of making decisions based on or assisted by outputs from algorithms, with or without human involvement.

5 examples...

Where we might use algorithmic decision making

AIRLINE SEAT ALLOCATION

How do we assign seats to maximise cost, cater for customer wishes and balance the plane?

Where will they sit?

INSURANCE OR LOAN UNDERWRITING

What is the risk this customer brings?

Should they get a policy and what should they pay?

911 POLICE INCIDENT CALL TRIAGE

What is the incident profile?

Which incidents do we prioritise, who do we send there, how fast do they need to be there and what should they expect?

CANCER DIAGNOSIS

Is this cancer, and if so, what type is it?

What treatment options will I recommend?

The problem with algorithmic decision making

5 examples

Risks

What could possibly go wrong (for the end-user)?

AIRLINE SEAT ALLOCATION

You might not sit next to your partner and you might be sulking for a while.

INSURANCE OR LOAN UNDERWRITING

The algorithm might unfairly discriminate and you might not get a loan, or have to pay a higher premium and you may end up in even more debt, eventually lose your house, your job, go to prison.

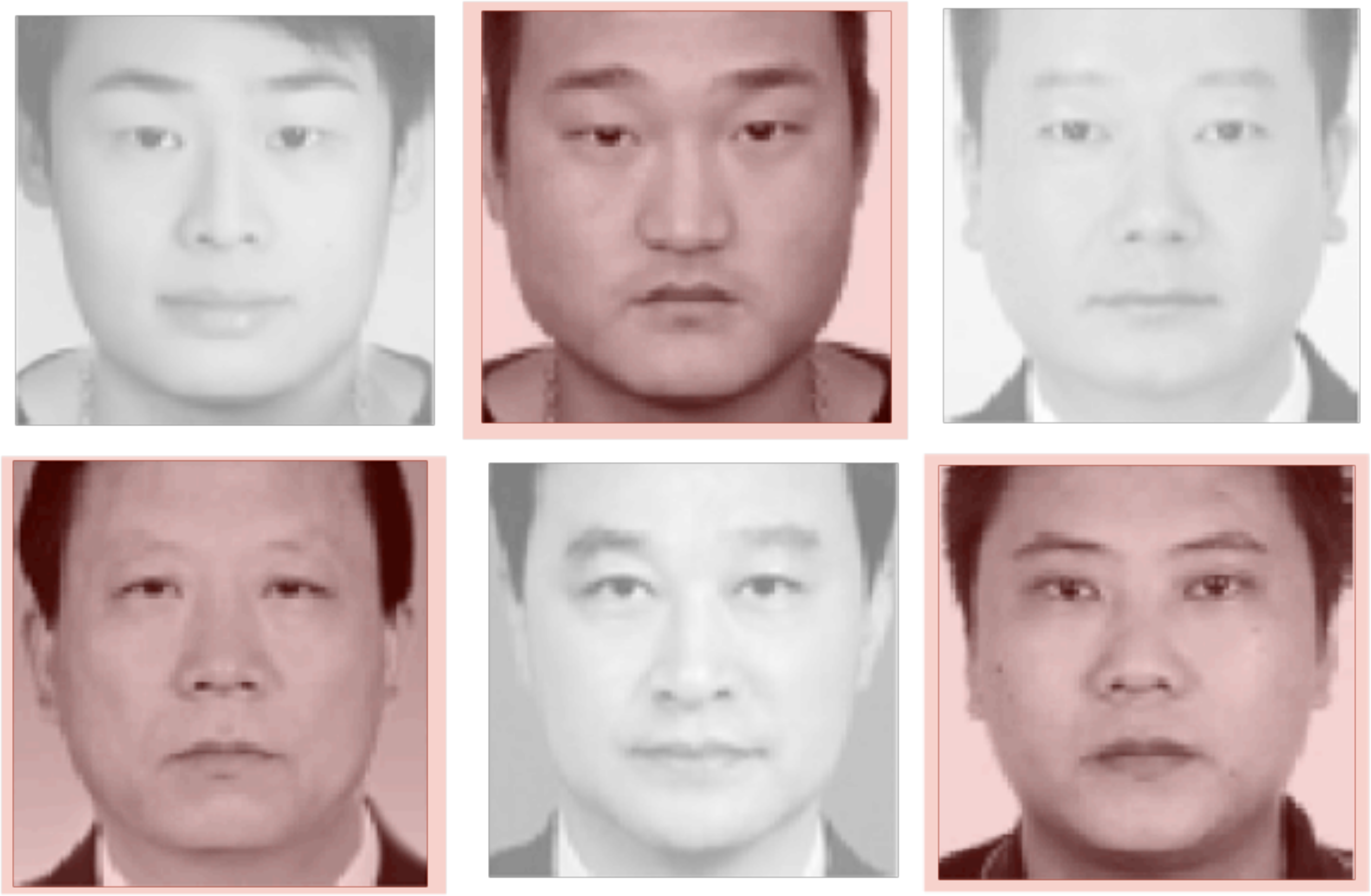
911 POLICE INCIDENT CALL TRIAGE

The algorithm might 'misrepresent' the situation and convince the operator to send a SWAT team to your house, and rather than being Doxed you get shot.

CANCER DIAGNOSIS

The algorithm might not clearly articulate the confidence level and you may get the wrong cancer treatment.

Who's a criminal?



visual from **Automated Inference on Criminality using Face Images by Wu, Zhang** , 2016

Automation is great – until it isn’t

And AI?

Intelligence is

reasoning, understanding, problem solving...

“skill-acquisition efficiency.”

Francois Chollet

Artificial Intelligence is

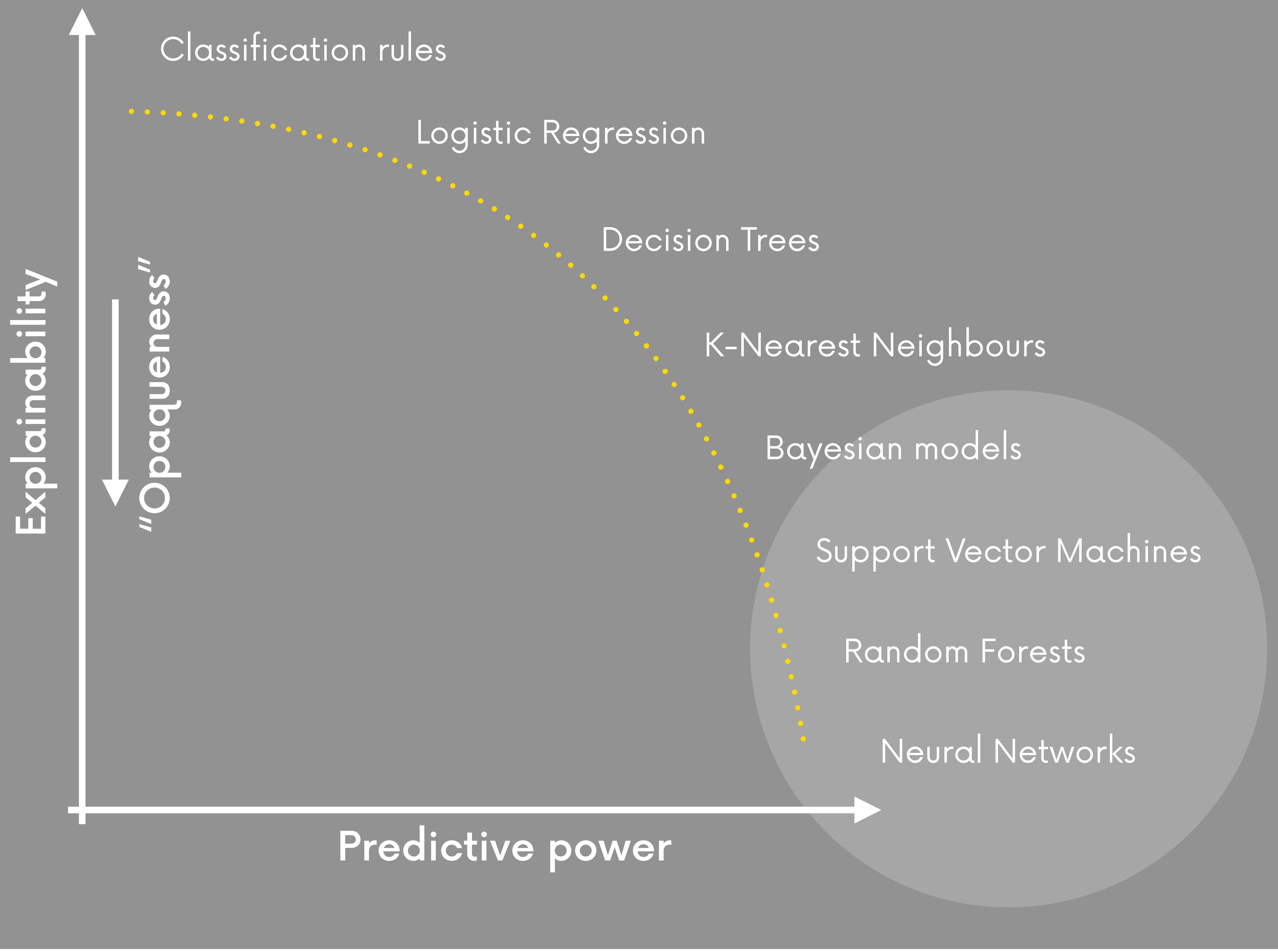
“systems capable of performing complex tasks that historically only a human could do, such as reasoning, making decisions, or solving problems.”

Coursera

“to generate outputs such as content, predictions, recommendations, or decisions which influence the environment.”

EU AI Act

Power vs Explainability



Explainability

Explainability helps us rationalise (and make understandable) the outputs of a system in relation to the inputs we provided

Interpretability & Explainability

What's the **relationship** between inputs and outputs?

How do we make the system behaviour **understood** by its users?

How do we rationalise the outputs of a system as we use them to take action?

Explainability allows us to **assure** that the decision made by or following algorithmic outputs are ‘good’ ones.

Explainability is **important** now, as our algorithms' power and reach create extensive risk.

Explainability can be **hard** (especially if and because we don't always know what's going on inside a model)

5 examples

Explanations

What would our end-users like explained to them?

AIRLINE SEAT ALLOCATION

Not sure customers really care about explanation (you might for your news recommendation algorithm).

INSURANCE OR LOAN UNDERWRITING

Customers will want to understand why their loan or policy was rejected.

911 POLICE INCIDENT CALL TRIAGE

An operator will need to understand why the system triaged the call as high risk house invasion vs Doxing and which factors were significant (e.g. message content, stress levels of caller, background noise).

CANCER DIAGNOSIS

A radiologist will want to understand which factors made the system diagnose cancer, with what level of confidence?

What the law says

A broad regulatory landscape:

- Umbrella 'acts' for general concerns
- AI (algorithm) specific acts
- Domain specific acts

Common principles:

> “safety, security and robustness; appropriate transparency and explainability;
> fairness, accountability and governance;
> and contestability and redress.”

“MHRA’s AI regulatory strategy ensures patient safety and industry innovation into 2030”

Regulations

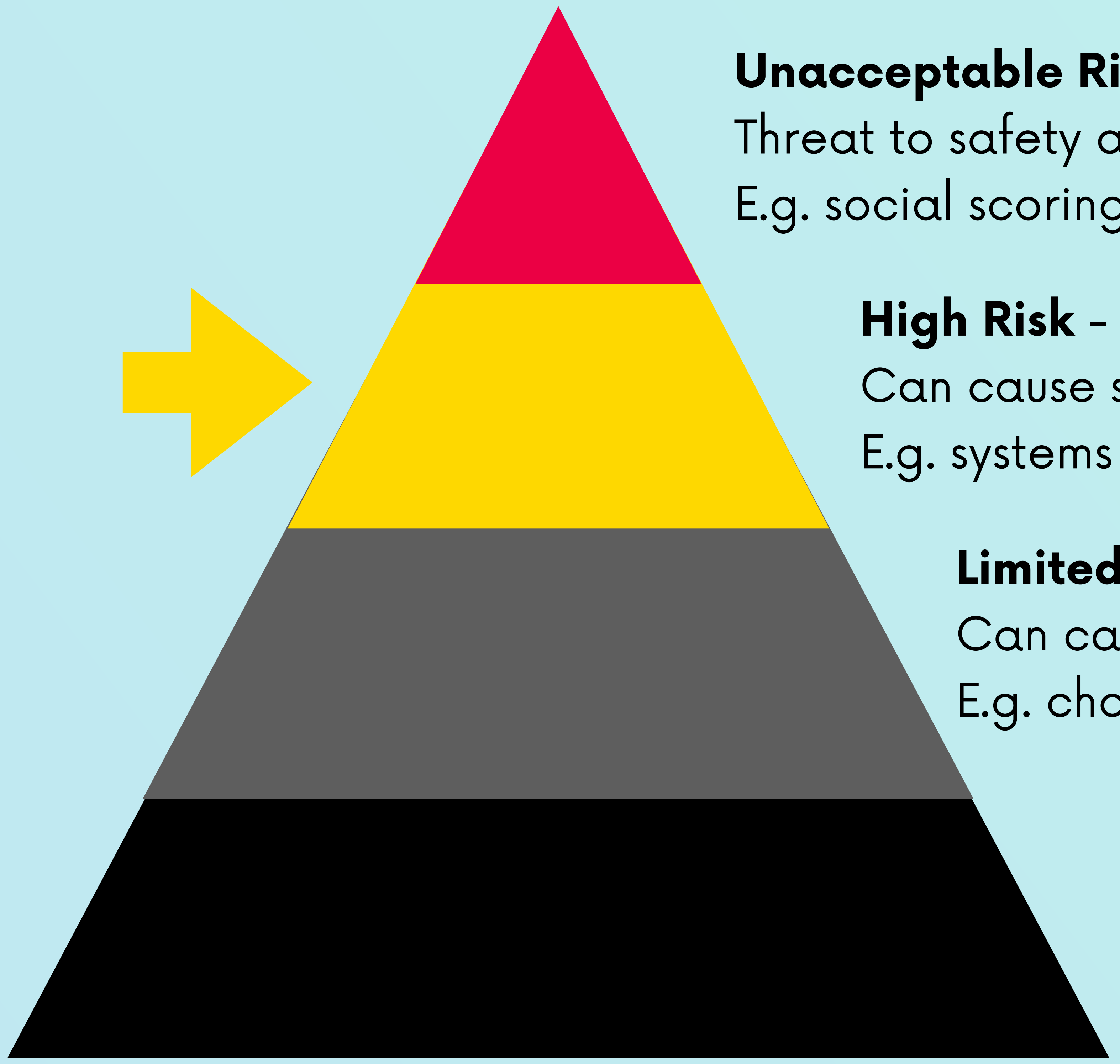
What 'the law' says...

AI ACT

(EU AI REGULATION)

Proportionality, transparency, traceability, human oversight.

AI SYSTEM CLASSIFICATION



Unacceptable Risk - prohibited
Threat to safety and livelihood.
E.g. social scoring

High Risk - risk mitigation, transparency, human oversight
Can cause significant harm.
E.g. systems in education, legal, law enforcement, medicine

Limited Risk - indication of AI use, transparency
Can cause transparency issues
E.g. chatbot, AI generated content

Minimal Risk - no obligations
E.g. spam filters, AI in video games,

Military applications - out of scope

Based on EU AI Act

Regulations

What 'the law' says...

AI ACT

(EU AI REGULATION)

Proportionality, transparency, traceability, human oversight.

PSD2

(EU PAYMENT SERVICES REGULATIONS)

Clear and understandable information about credit-worthiness and fraud.

Regulations

What 'the law' says...

AI ACT

(EU AI REGULATION)

Proportionality, transparency, traceability, human oversight.

PSD2

(EU PAYMENT SERVICES REGULATIONS)

Clear and understandable information about credit-worthiness and fraud.

GDPR

(UK/EU PRIVACY REGULATIONS)

- the right to be **informed** about the use of automated decision making, the logic involved and the data used as well as the envisaged consequences / impact
- the right to **access** to the data
- The right to **intervention** (correction etc)
- the right to **object** to the use of personal data (in certain circumstances)

‘**AI**’ is a red herring: it’s all about ‘**algorithms**’ and their impact

Careful where you make
life-changing decisions

“[the sole use of AI] does not meet requirements for a **human based judgement** to be used in marking decisions.

But it is also our view – by virtue of taking a **precautionary principle** – that the potential for **bias**, **inaccuracies** and a lack of **transparency** in how marks are awarded could introduce unfairness into the system.”

Ofqual policy paper, 2024

Useful resources

- Explaining decisions made with AI
  ICO / Alan Turing Institute
  https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

- Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World
  Stanford University
  https://hai.stanford.edu/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world

What is a good ‘explanation’?

A **good explanation** allows users and recipients of algorithmic decision-making to understand outcomes and optimise for positive outcomes.

Good explanations reflect information and presentation needs in terms of usecase, domain, expectations and capabilities

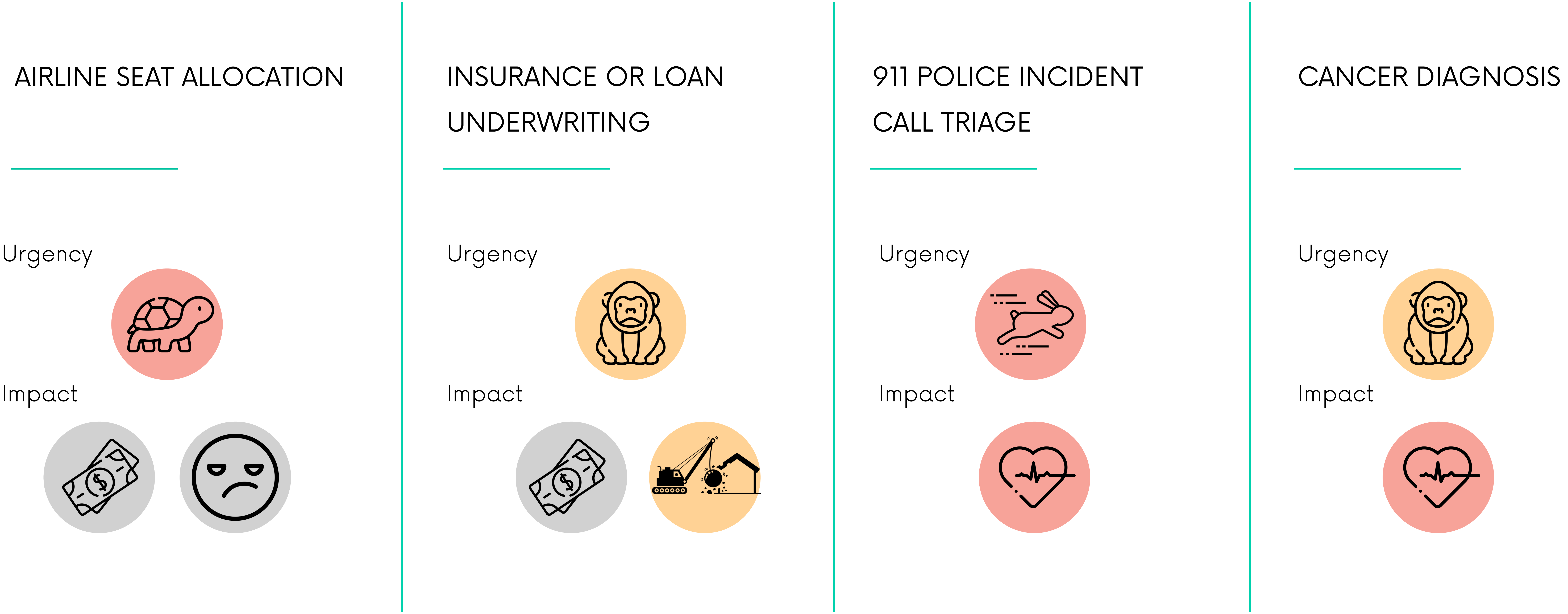
They are

- user centric
- contextual
- meaningful & understandable

5 examples

Urgency & impact

How important is it I act now, and how important is it that I get it right?



AIRLINE SEAT ALLOCATION

Urgency

Impact

INSURANCE OR LOAN UNDERWRITING

Urgency

Impact

911 POLICE INCIDENT CALL TRIAGE

Urgency

Impact

CANCER DIAGNOSIS

Urgency

Impact

5 examples

Level & type of explainability

What information do I need (as end-user) and how do I need it?

AIRLINE SEAT ALLOCATION

Explainability level

- None

INSURANCE OR LOAN UNDERWRITING

Explainability level

- Medium level of detail to support users 'understanding' and challenging'
- Focus on justification

911 POLICE INCIDENT CALL TRIAGE

Explainability level

- Minimal information for fast decision-making with easy to absorb qualifiers
- Focus on impact / adverse impact

CANCER DIAGNOSIS

Explainability level

- In-depth rationale, full qualifiers for detailed analysis, examples and comparisons
- Focus on supporting detailed analysis

How does an explanation ‘look’ like?

- Cover process and outputs
- Rationale
- Responsibilities
- Data

- Fairness
- Safety & Performance
- Impact
- Touchpoints
- Think information design!

Explaining decisions made with AI
ICO & The Alan Turing Institute

https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/?q=r

Develop systems in an **explainability-aware** fashion, across the entire SDLC

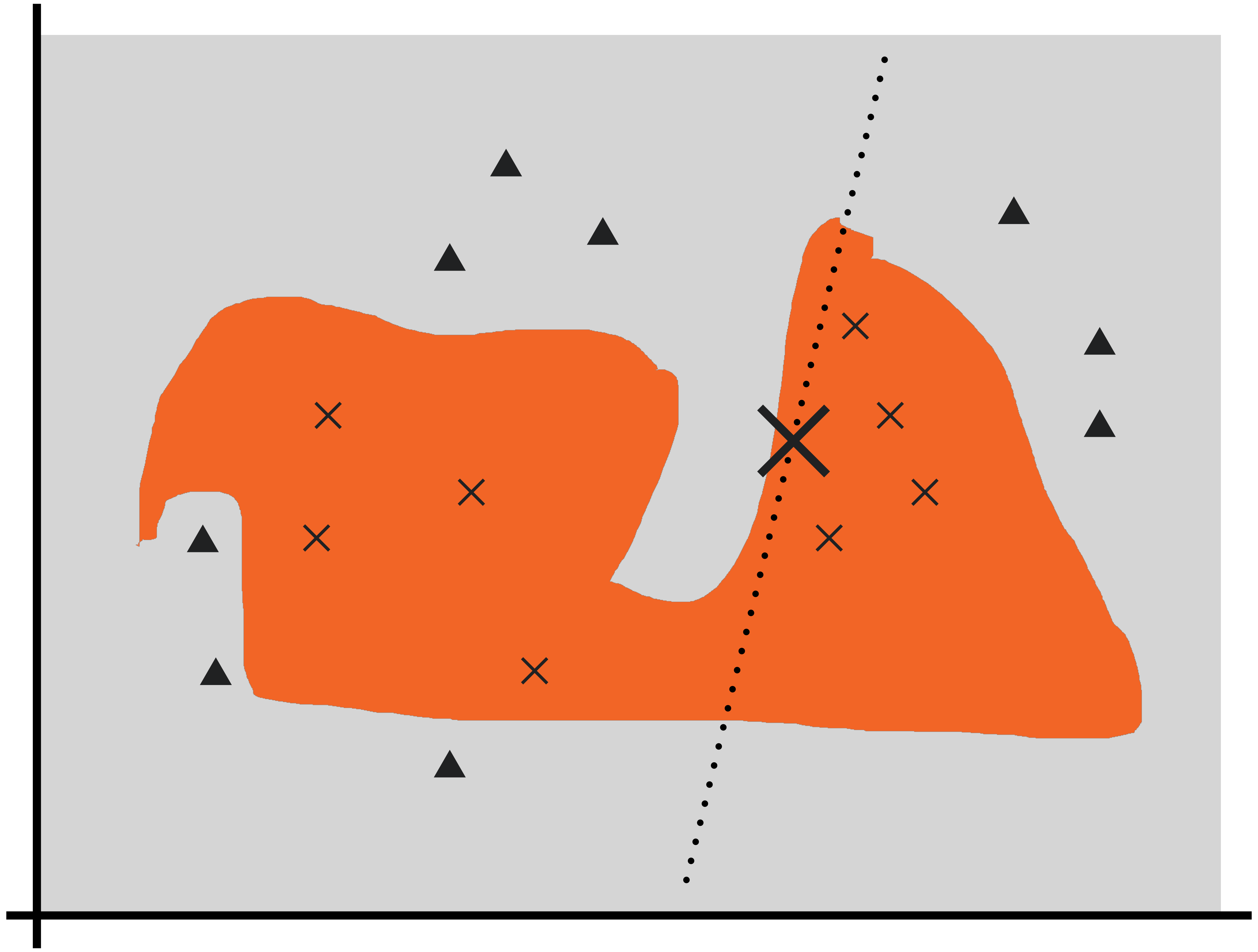
How to rationalise outputs

Strategies to achieve explainability

- Rationalise inherently explainable algorithms
- Use ensembles
- Use post-rationalising tools

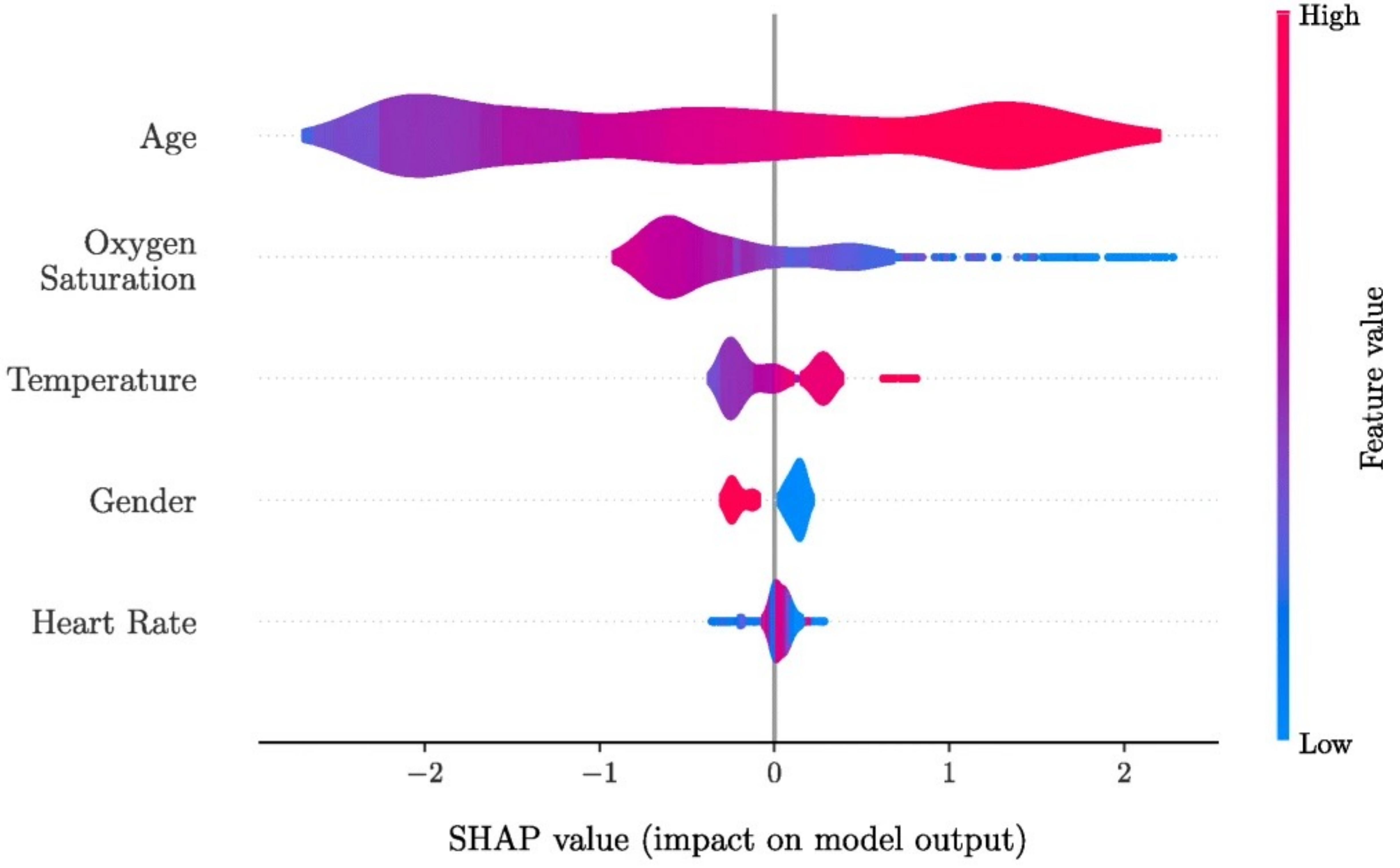
LIME

Local Interpretable Model-Agnostic Explanation



SHAP

SHapley Additive exPlanations

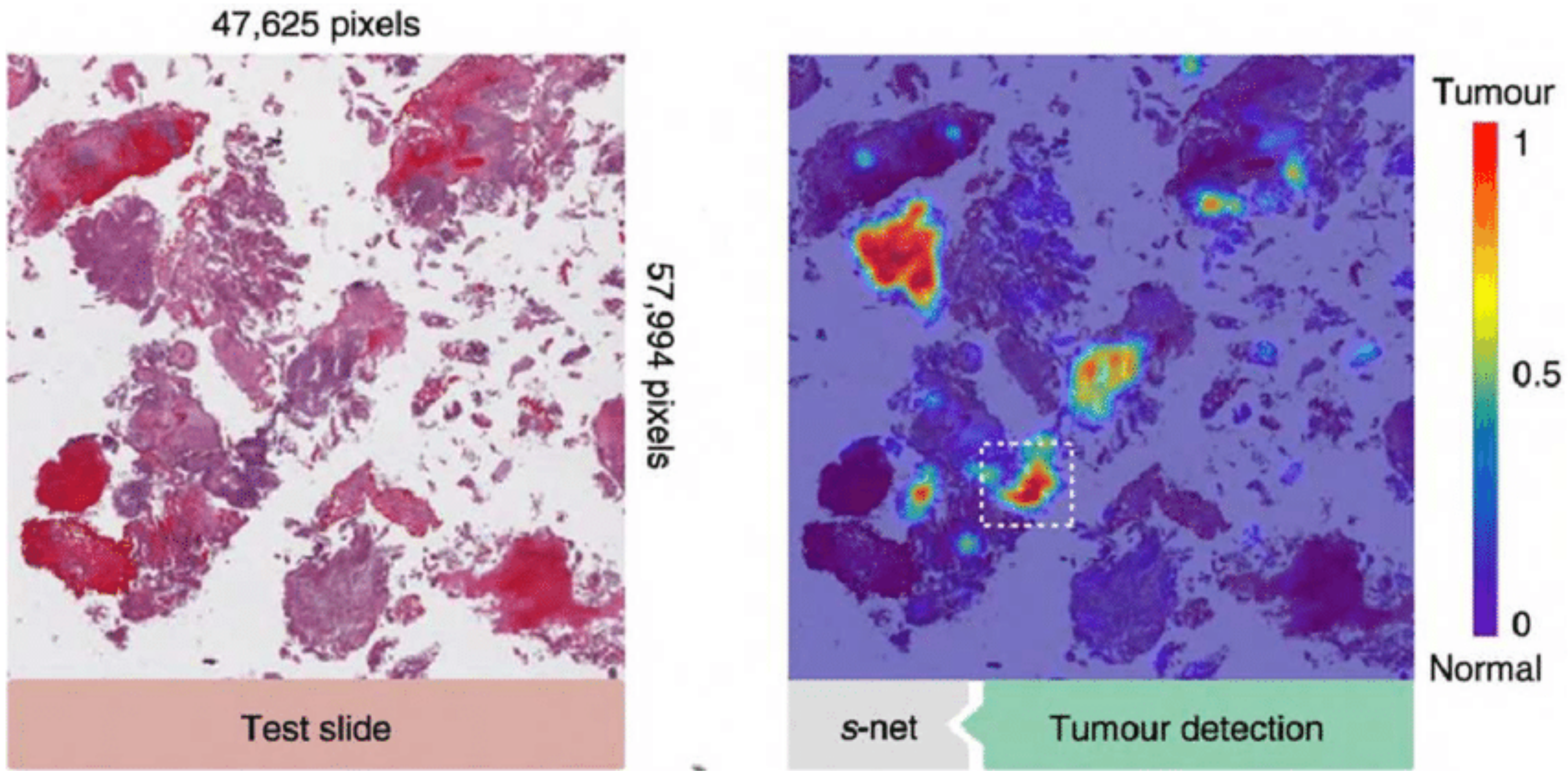


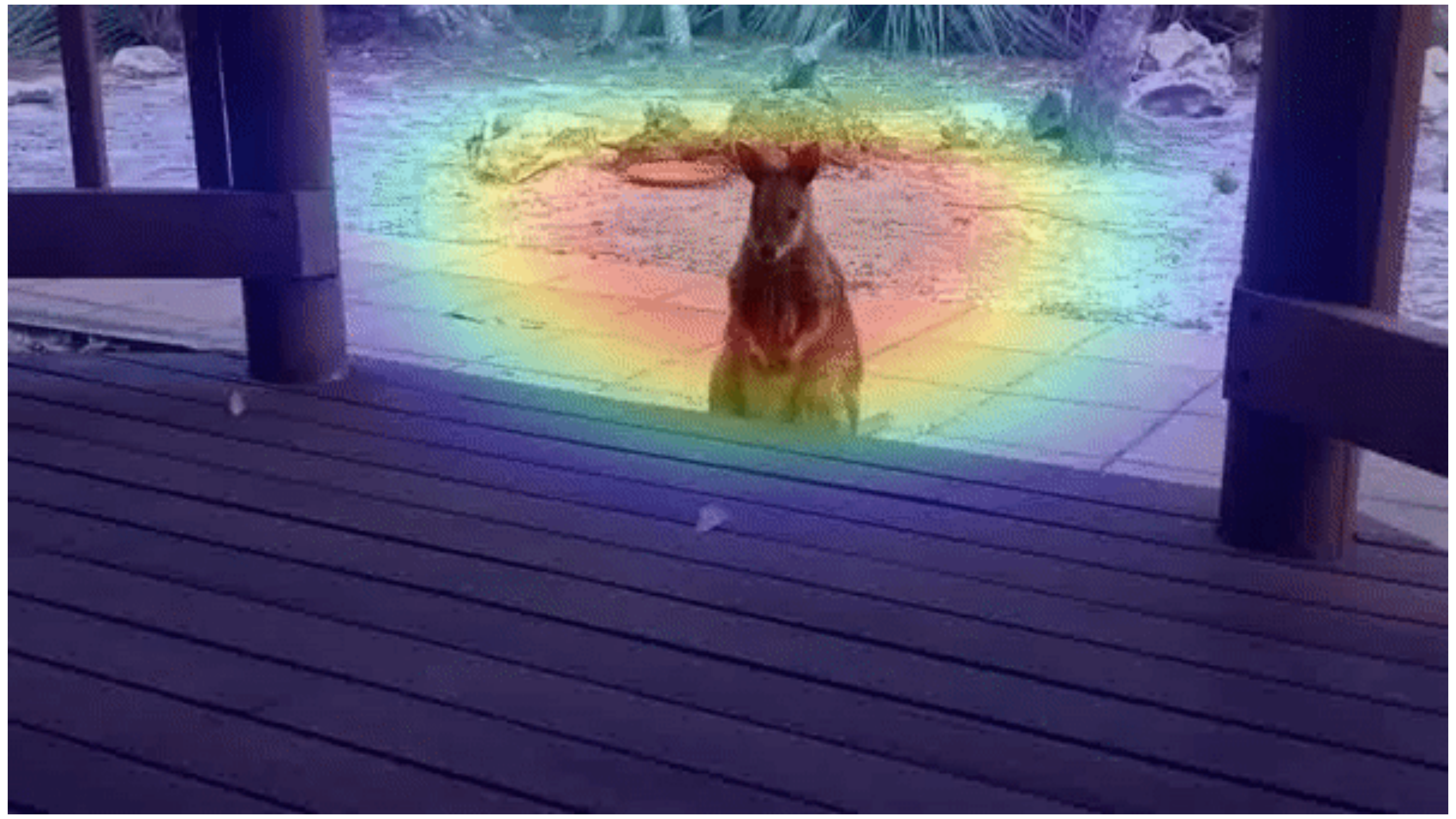
(b) Mortality Model without Lab Values

From predictions to prescriptions: A data-driven response to COVID-19
Https://link.springer.com/article/10.1007/s10729-020-09542-0

GradCAM

Salience mapping / Gradient Class Activation Mapping





Activation mapping: https://www.sciencedirect.com/topics/computer-science/class-activation-mapping

Salience Map: https://github.com/frgfm/torch-cam

More…

Transparency by algorithm type and discussion of supplementary models

- https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/annexe-2-algorithmic-techniques/
- https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/annexe-3-supplementary-models/

Benefits

Benefits of explainability

- Compliance
- Governance
- Trust / Reassurance
- Better outcomes
- Human flourishing

Compliance by design

https://bit.ly/compliancebydesign

Ethical product management

https://bit.ly/ethicalpm

My current project

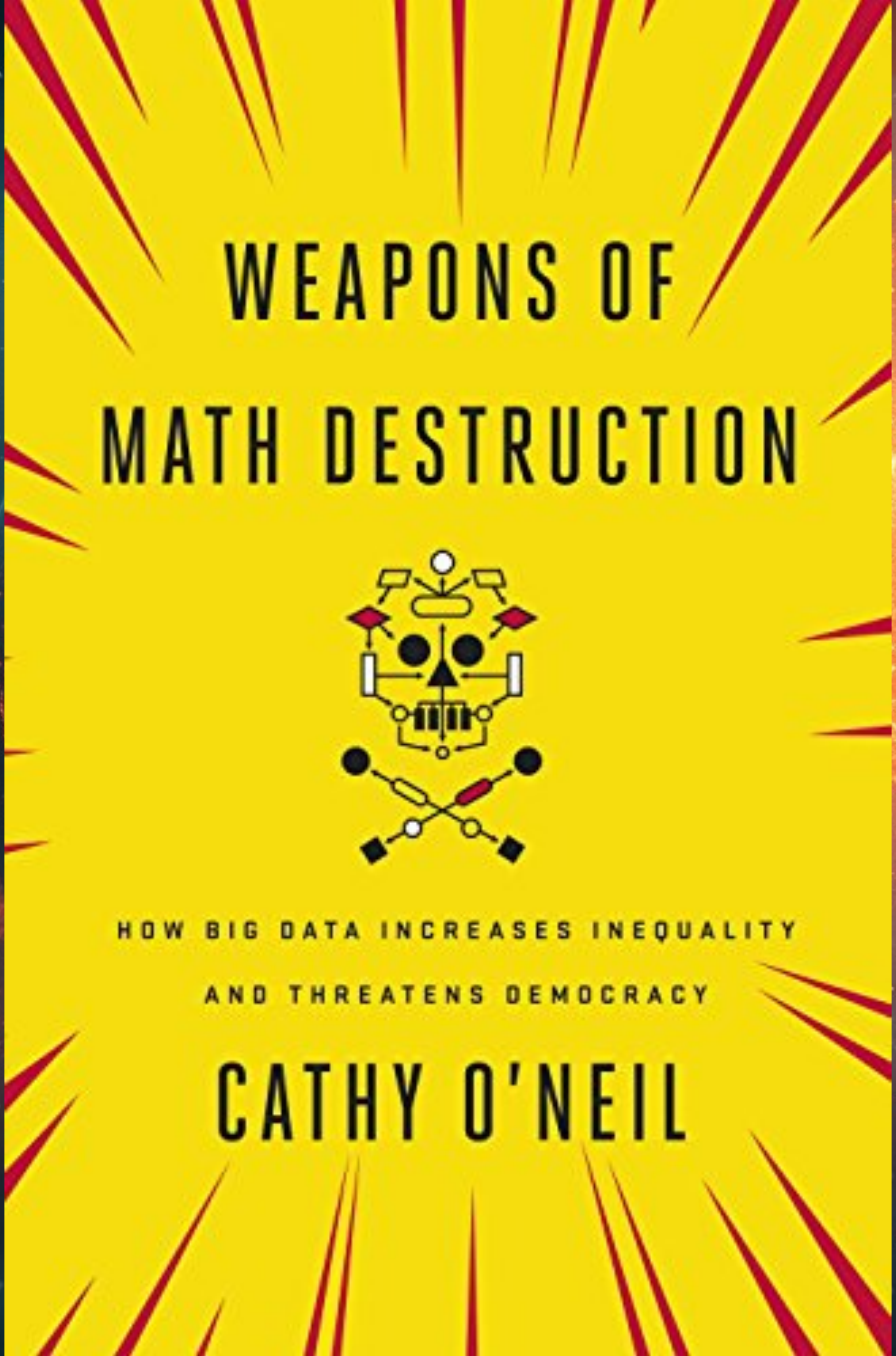
Conclusion

- **Ensure outcomes that are non-discriminatory, safe, and supportive of individual and societal wellbeing**
- Tread with caution
- Explainability is rationalising how inputs lead to outputs
- Explainability is important and beneficial
- Explainability needs to be carefully designed to cater for user, usecase and context
- There are transparent and black box algorithms (explainability for the latter is harder)
- Avoid the temptation of AI
- Consider cost / benefit / environmental and societal impact
- **Choose the right model and explainability approach**

Predatory algorithms create nothing less than a death spiral of modelling.

Dr. Cathy O'Neil, Mathematician
Author of 'Weapons of Math Destruction'

image: Midjourney



WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

image: Midjourney

Let’s not loose the future!

image: Midjourney

We need

- **awareness** of our actions
- **mitigation strategies** to handle technologies that do not exist yet
- **ethics frameworks**

and do this at **individual, local and global level.**

image: Midjourney

Questions?

Presentation deck...

bit.ly/explai

I'd love to hear from you...

BEAUTIFUL ABSTRACTION

Web

www.beautifulabstraction.com

marcel.britsch@beautifulabstraction.com

THE DIGITAL BUSINESS ANALYST
Agile musings and ramblings

Blog

www.thedigitalbusinessanalyst.com

THE BURN UP
ALL THINGS AGILE

Podcast

www.theburnup.com