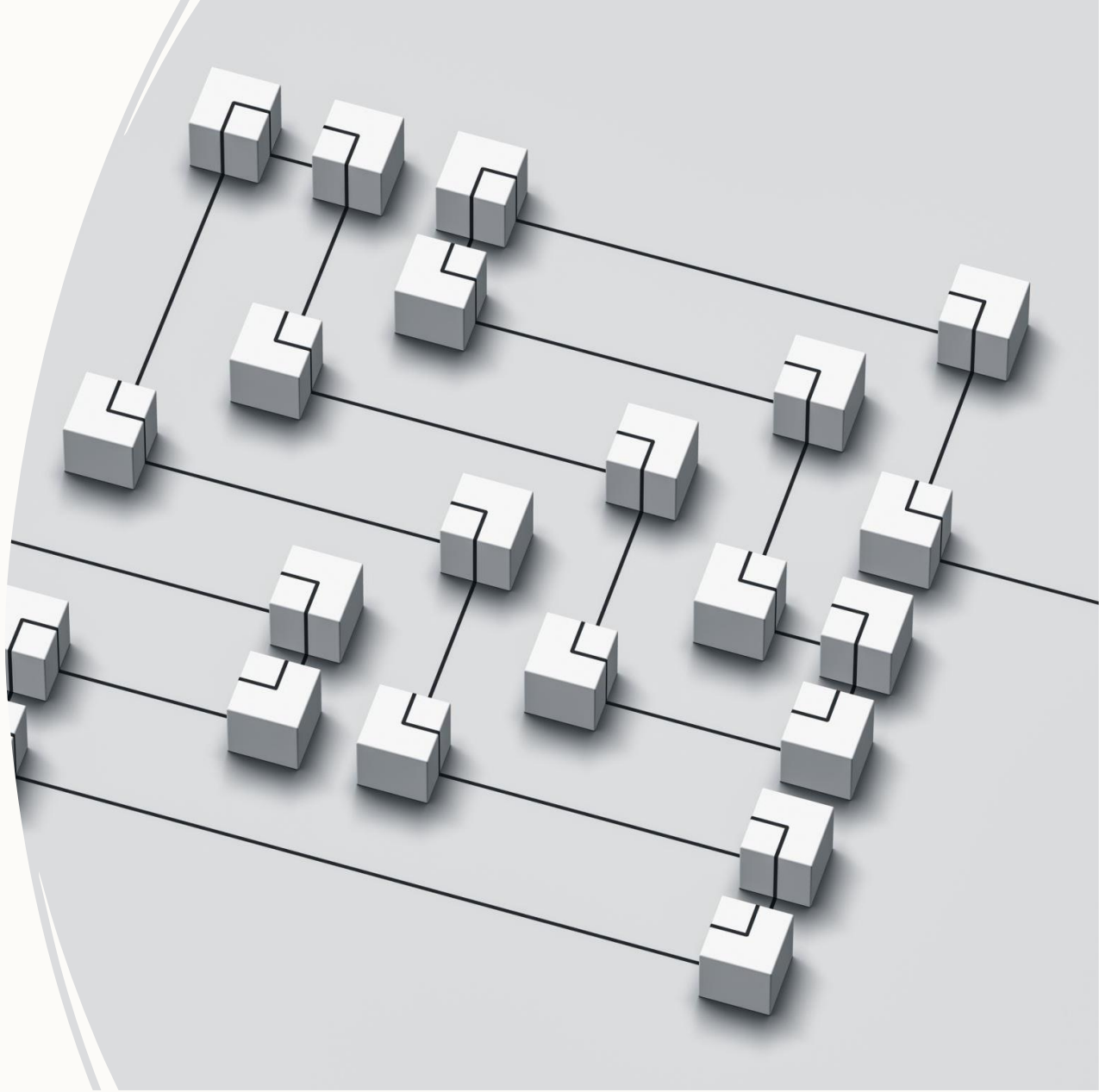
COMP 3056 Professional Ethics in Computing

Week 10 Trustworthy AI

Anthony Bellotti

Learning outcomes

- AI Risks
- Principles of AI Assurance
- Autonomous Systems and Encoded Ethics
- Fairness and Bias
- Brittleness and “Artificial Stupidity”
- Explainable AI



University of Nottingham
UK | CHINA | MALAYSIA

AI Risks

Artificial Intelligence (AI) Risks

- We can divide AI risks into 5 broad categories:

1. **Data related**: Data is an important resource for AI; poor quality or misuse can create problems.
2. **Performance**: Whether AI performs to benefit all its stakeholders, or not.
3. **Human/AI interaction**: How humans interact with the AI presents a risk.
4. **Objective**: How we code up the objective for the AI can be problematic, with unintended consequences.
5. **Social risks**: Broad social impact of AI, in general.

1. AI Data Related Risks

- **Poor Data Quality**

  Data source may be unknown or unreliable. Measured data may not be accurate or wrongly entered.

- **Data Bias**

  Collected data may present a bias related to how it is collected, and/or for protected classes.

- **Misuse of Data**

  Even if data is reliable, the data may be used without proper permission or in an inappropriate way, that differs for the intended reason for its collection.

- **Poor Security**

  Data may be used correctly but may not be stored, anonymized or communicated securely within the AI system.

2. AI Performance Risks

- **Poor Aggregate Performance**

  The AI may not perform well at the assigned task, in general (poor model fit). It may generally give biassed outcomes (poorly calibrated). It may be evident that more training data is required to improve performance.

- **Overfitting & Brittleness**

  Machine learning-based AI uses training data to learn. Often it may perform extremely well on training data, but poorly on new data. This is called overfitting. Similarly, the AI may adapt poorly to changes in data (brittleness), e.g. over time or place.

- **Sub-population Performance Bias**

  Even if aggregate performance is good, the AI may perform badly for a specific sub-population. This could be a protected class or a combination (intersectionalism).

- **Unfairness**

  The AI may make decisions that are unfair to a particular group, in some sense. It turns our there are several different definitions of fairness, so the AI stakeholders and developers need to determine which applies.

3. Human / AI Interaction Risks

- **Human-Computer Interface Design**

  Poor interface design may lead to improper use of the AI, or misunderstanding of its decision, or miscommunication of objectives.

- **Opaque System**

  If the AI cannot explain its decisions, it means we may not be able to understand how it works, and may mean the decision or the reason for the decision is wrong and we would never know. We may also need to explain decisions to customers or other stakeholders.

- **Over- or Under-reliance**

  End-users may over-rely on AI system, or have irrational mistrust of AI.

- **Misuse of AI**

  The AI may be designed in a way that makes it available for misuse, either by the organization that develops it, or by external parties that make use of the AI.

4. AI Objective Risks

For many AI systems, especially machine learning, the requirement for the AI is specified using an *objective function*, often as a mathematical formula, and the AI system is optimized around this.

This presents its own risks.

- **Ethical misalignment**

The AI may not correctly embody ethical principles and hence may behave in an unethical way.

- **Misaligned objective**

  Converting a complex requirement to formulae may lead to errors, or unintended consequences.

  (*e.g. King Midas story*).

- **Gaming the objectives**

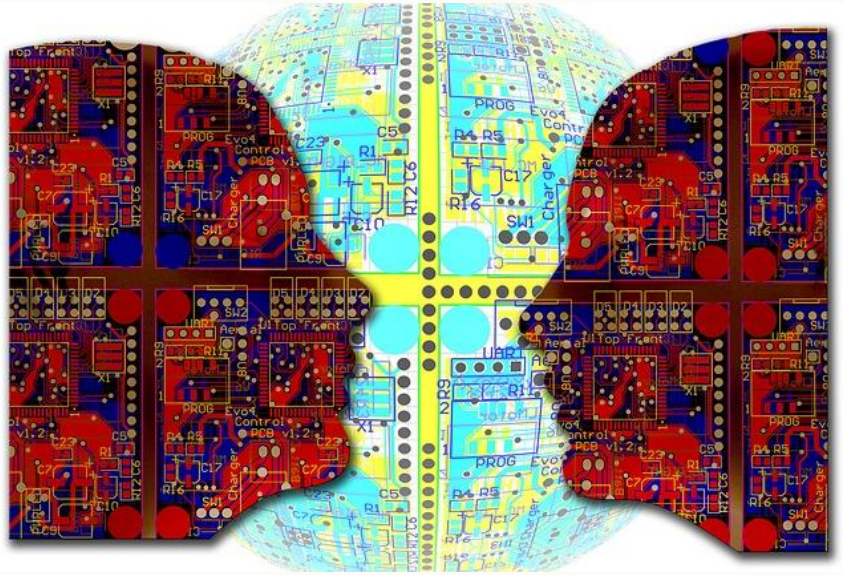
  The AI may find a way to meet an objective, but in a way that is inappropriate.

  (*e.g. easiest way to prevent a car crash is to ensure the car does not move*!).

5. AI Social Risks

Important political and social topics to think about, but we will not cover them further in this class:

- Lack of expertise in AI.
- Singularity Risk.
- Human behaviour modification
  (e.g. loss of autonomy; cyborgs).
- Reflexivity (AI's influence on itself).
- Risk to the economy and jobs.
- Exploitation of human workers in AI development.
- Climate Risk.



University of Nottingham
UK | CHINA | MALAYSIA

Principles of AI Assurance

AI Assurance

Broad principles to mitigate AI Risk:

- Society will benefit from regulation and governance of AI.
- Training and professional accreditation of AI professionals.
- Well-established methods for AI development, validation and deployment.
- Validation and monitoring of AI systems.
- Understanding the conditions when we should not use AI.
- Clear responsibility, accountability and governance of AI.

OECD AI Principles

- Organisation for Economic Co-Operation and Development (OECD).
- International framework.
- https://oecd.ai/en/ai-principles

University of Nottingham
UK | CHINA | MALAYSIA

Autonomous Agents & Encoded Ethics

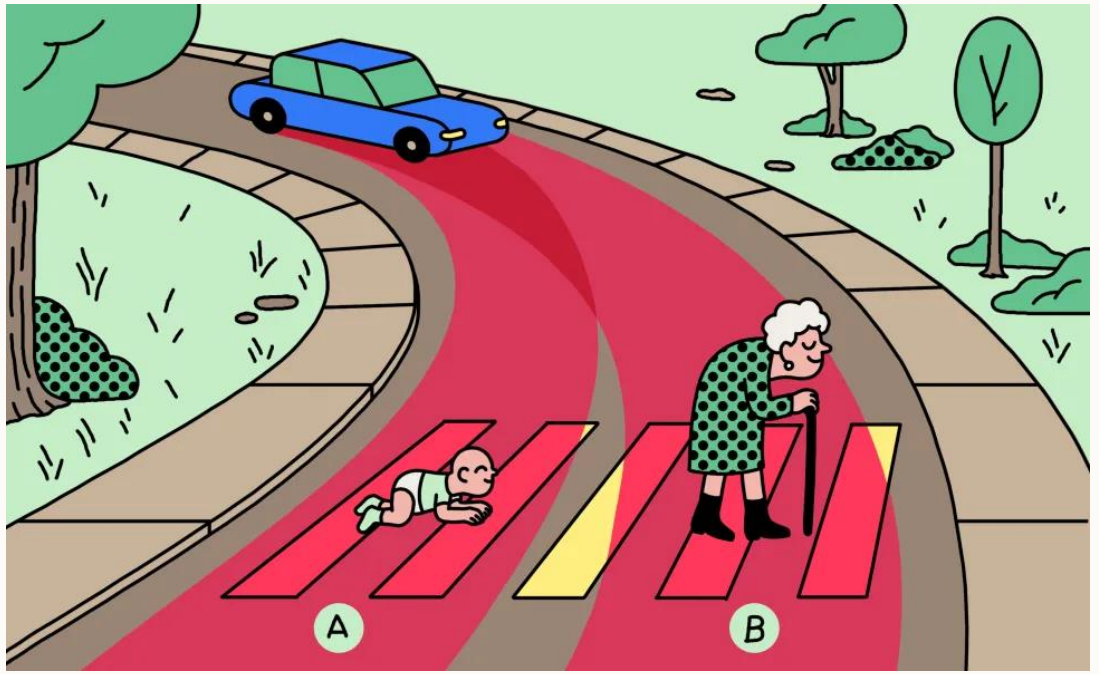
Autonomous Agents

AI is deployed as **decision support** or as **autonomous agent**:

- Example 1. A medical diagnosis is a decision support system since it provides recommendation for human medical doctors.
- Example 2. A self-driving car is an autonomous agent.
- Just like human decision making, AI decision making can have ethical consequences.
- There is human oversight over decision support AI, so ethics can be monitored and decisions over-ridden (*at least, in principle*).
- However, autonomous agents must make there own ethical decisions. So, how do we embed ethics? Who (what) takes responsibility?

Autonomous Vehicles & Ethics

- Consider a self-driving car that suddenly encounters a situation where a baby and an old lady are crossing a highway.
- An a split-second, the AI can compute that it will not have time to slow down. It computes it has three options:
  A. Hit the baby.
  B. Hit the old lady.
  C. Swerve off the road, hit a tree, and harm the passenger.
- How should we program the AI to act?



MIT Study (2018):

https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/

Autonomous Vehicles & Ethics

Continuation of Self-Driving Car example:

- What if the AI knew its driver was a 72 year old convicted violent criminal on probation from jail.
- Additionally the AI knows he has a cancer and doctors have diagnosed he has less than 6 months to live.
- Would that make a difference?
- **Utilitarian** “cost-based” approach: how much is each person “worth”?
- Is it ethical to code our judgements about people into AI.

Autonomous Agents & Embedded Ethics

- Even if we can program the preferred ethical response for specific scenarios, is it possible to provide code for all possible situations an autonomous agent can meet?
- A pattern of ethical scenarios, but each scenario is unique with it's own special conditions.
- Alternative is to provide a meta-program for the autonomous agents to spontaneously generate its own ethical decisions.
- … *at least, in the long run* (may be hard to do now).

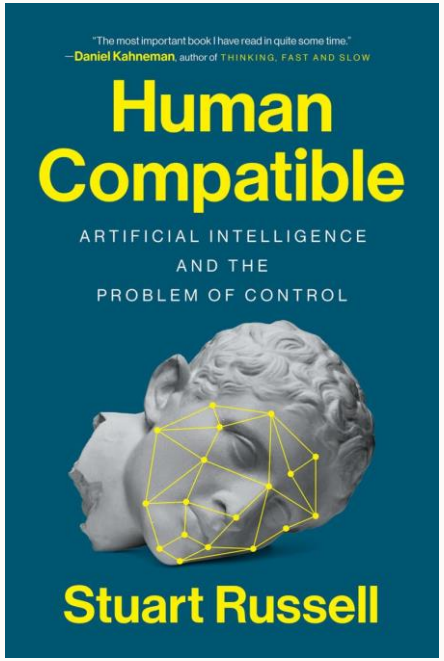
Autonomous Agents & Embedded Ethics

"We had better be quite sure that the purpose put into the machine is the purpose which we really desire."

Norbert Wiener in 1960 (cybernetics founder)

Prof. Stuart Russell's general ethical principles (2017):

1. Human-focused. The AI's only objective is to maximize the realization of human values.
2. Humility. The AI is initially uncertain about what those values are.
3. Learning. Human behaviour provides information about human values.



Stuart Russell is professor of computer science at the University of California, Berkeley

University of Nottingham
UK | CHINA | MALAYSIA

Fairness and Bias

We want AI to be fair.
But what do we mean by “fair”?

Fairness and AI

Consider these examples:

- A mother gives her two children some sweets: 5 for Timmy and 2 for Sammy. Sammy complains he gets less and this is *unfair*. But his mother explains that Timmy did his homework but Sammy didn't, so Timmy gets an extra reward and that is *fair*.

- A husband and wife are the same age, live in the same home, drive the same family car, have been driving for the same time, and neither have had a car accident. They both pay the same insurance premium. However, statistically men are more risky than women, in general, and if this were taken into account in the pricing, men should pay 10% more than women. The wife is unhappy with this and thinks it is *unfair* that women are "subsidizing" men's riskiness. However, her husband is happy with it the way it is.

Fairness and AI

- In computer science, we have some definitions of fairness:

  - **Independence**. The AI decision should be independent of any individual characteristic.
  - **Separation**. Same as Independence but the independence is within each outcome type being modelled.
  - **Sufficiency**. The AI decision should be sufficient to account for all features and characteristics,

- However, each of these definitions are different and generally to achieve one means discarding the other two.
- So which do we choose?
  - ✓ Depends on considerations of: business requirements, legislation, local ethical viewpoints (*cultural* relativism).

- Source: Fairness and Machine Learning: Limitations and Opportunities, Solon Barocas, Moritz Hardt, and Aroind Narayanan (online 2023), fairmlbook.org

Fairness definitions

Some notation.

- Let Y be the outcome the AI is modelling;
- Let A be the sensitive attribute we wish to check for fairness;
- Let R be a score or decision from the AI about an individual.
- $X \perp Z$ means X is independent of Z.

Example.

For credit scoring example, AI makes decisions about individuals applying for loans:

- Y = outcome: 1 if they default (do not repay the loan), 0 if they do repay.
- A = gender (Male/Female).
- R = credit score; higher value means the AI thinks they are more likely to default. A decision to give the loan is based on this; e.g. $R > t$ means the applicant gets the loan (for a fixed t).

Fairness definition 1: Independence

$$R \perp A$$

meaning R (score/decision) is independent of A.

- In terms of probabilities, for any two values a, b of A:

$$P(R = r \mid A = a) = P(R = r \mid A = b)$$

- The second (car insurance) example is of this type.

Fairness definition 2: Separation

$$R \perp A \mid Y$$

Similar to Independence, but within different outcome types.

- In terms of probabilities, for any two values a, b of A:

$$P(R = r \mid A = a, Y = 0) = P(R = r \mid A = b, Y = 0)$$
$$P(R = r \mid A = a, Y = 1) = P(R = r \mid A = b, Y = 1)$$

Fairness definition 3: Sufficiency

This definition states that the AI outcome should be sufficient to account for all variables, including the sensitive attribute.

For any r, a:

$$P(Y = 1 \mid R = r) = P(Y = 1 \mid R = r, A = a)$$

- The first example (children) is an example of Sufficiency, with doing homework as sensitive attribute.

Fairness definitions: Approximations

- In practice, there will not be an exact match between measured probabilities on data.
- Hence, approximations may be used.
- For example, for Sufficiency,

$$P(R = r \mid A = a) \approx P(R = r \mid A = b)$$

Fairness Example

- Consider this table of observations within each group of R, Y, A.

	R=0		R=1	
	Y=0	Y=1	Y=0	Y=1
A=0	2	8	15	5
A=1	10	5	15	15

- We calculate the following probabilities:
  - $P(R = 1 \mid A = 0) = \frac{2}{3}$
  - $P(R = 1 \mid A = 1) = \frac{2}{3}$
  - $P(Y = 1 \mid R = 1) = \frac{20}{50} = \frac{2}{5}$
  - $P(Y = 1 \mid R = 1, A = 1) = \frac{15}{30} = \frac{1}{2}$
- *So which fairness definition is being followed by this AI?*

University of Nottingham
UK | CHINA | MALAYSIA

Brittleness and "Artificial Stupidity"

Brittleness: Examples

- Tesla car smashes into emergency vehicles (Jan 2018).

- One of 11 accidents of this kind.
- The AI autopilot had problems recognizing stationary vehicle with flashing lights.
- (*Now fixed*!)

- ChatGPT prompt:

  suppose X is a column vector and I is the identity matrix. Let Y be the transpose of X. Then, is det(XY+I)=YX+1?

- ChatGPT:

  No, det(XY+I)=YX+1 is generally not true.

- Why?

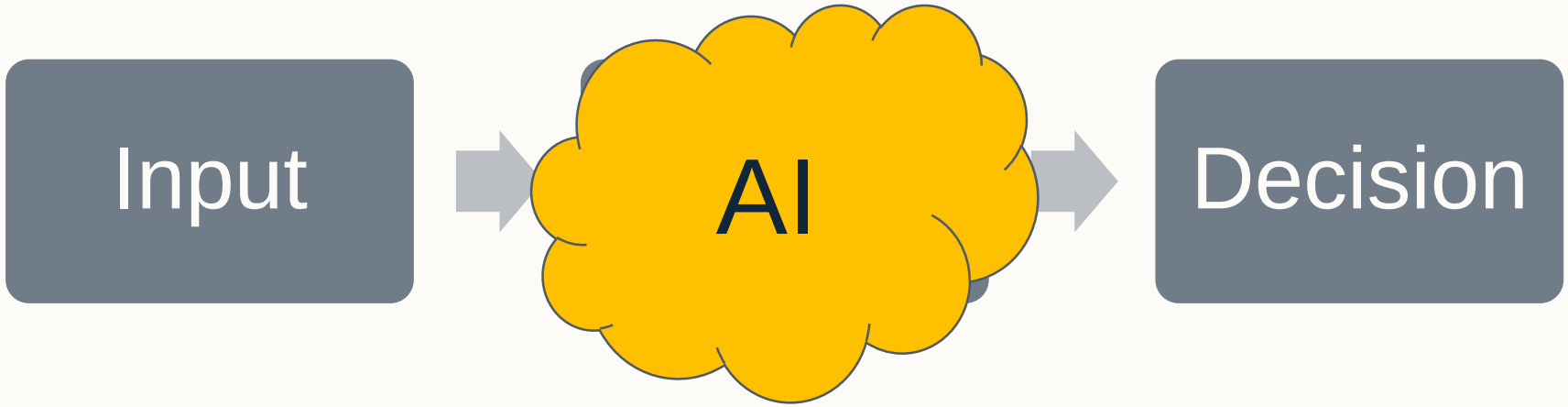
  … they're not generally equal because one is the determinant of an n×n matrix, and the other is a scalar …

University of Nottingham
UK | CHINA | MALAYSIA

Explainable AI

Black Box

- Often AI systems are too complex for us to understand how they are working and making decisions.
- Hence we call them Black Boxes.



- Examples include Deep Neural Networks, Random Forests, Gradient Boosted Trees, and Support Vector Machines.
- This becomes a problem if we need to explain decisions made about people.
- This is an ethical requirement, but may also be required by law (e.g. lending in USA and UK).

Explainable AI Metrics

- Several standard metrics have been developed to help to understand AI decisions.

- These typically work by measuring the impact of input features on the final decision.

  - **LIME**. Local Interpretable Model-agnostic Explanations
  - **SHAP**. SHapley Additive exPlanation
  - **PDP**. Partial Dependency Plots

- Reference: Christoph Molnar, **Interpretable Machine Learning** at https://christophm.github.io/interpretable-ml-book/.

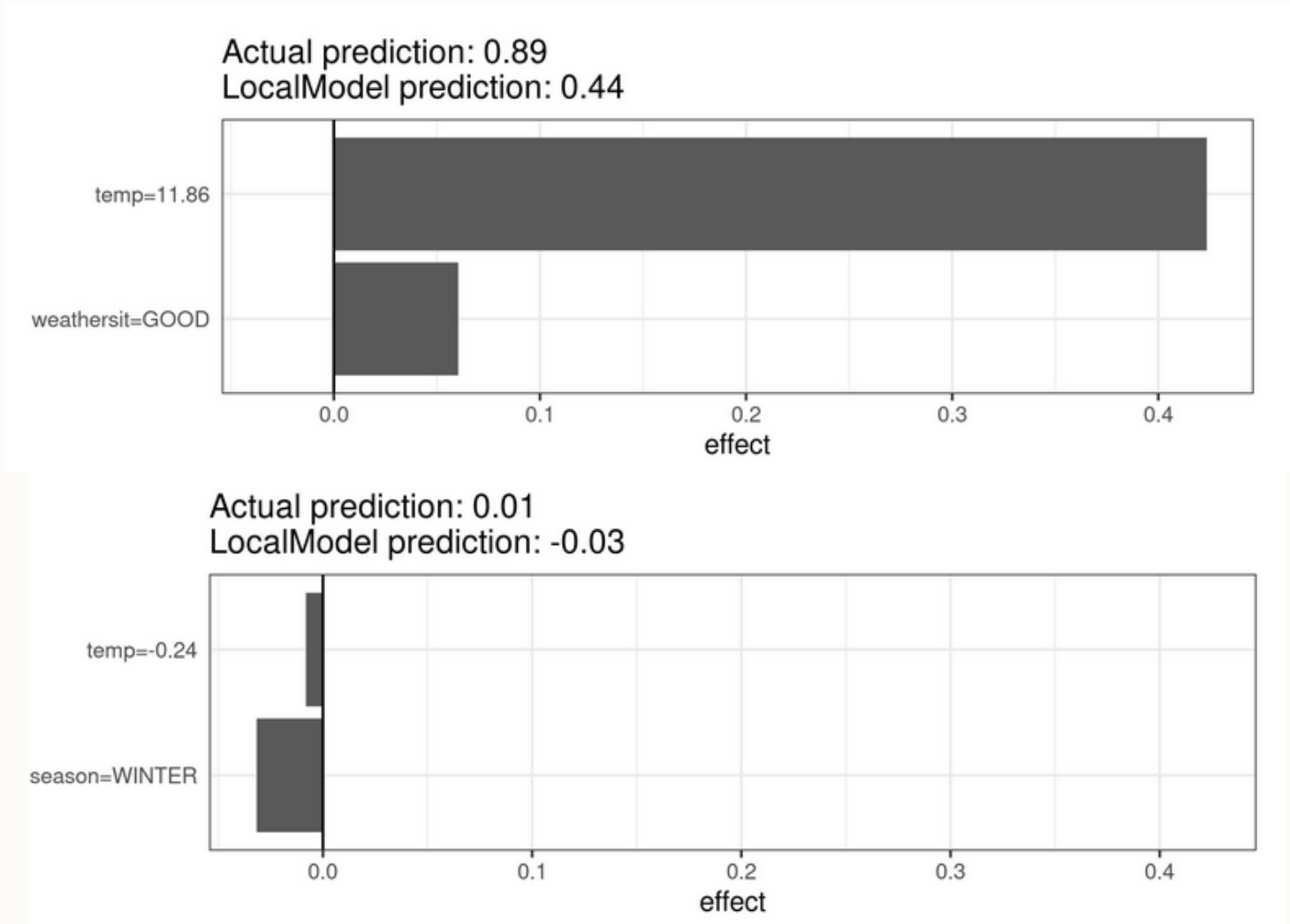
LIME

- Local Interpretable Model-agnostic Explanations.

- For a given input x and AI model f that outputs a decision $f(x) = \hat{y}$, build an ***approximate*** local linear model around $f(x)$.

- Report the coefficients of this local model.
  - Positive coefficients mean positive contribution to $\hat{y}$.
  - Negative coefficients mean negative contribution to $\hat{y}$.

LIME example



LIME explanations for two instances of the bike rental dataset.

Warmer temperature and good weather situation have a positive effect on the prediction.

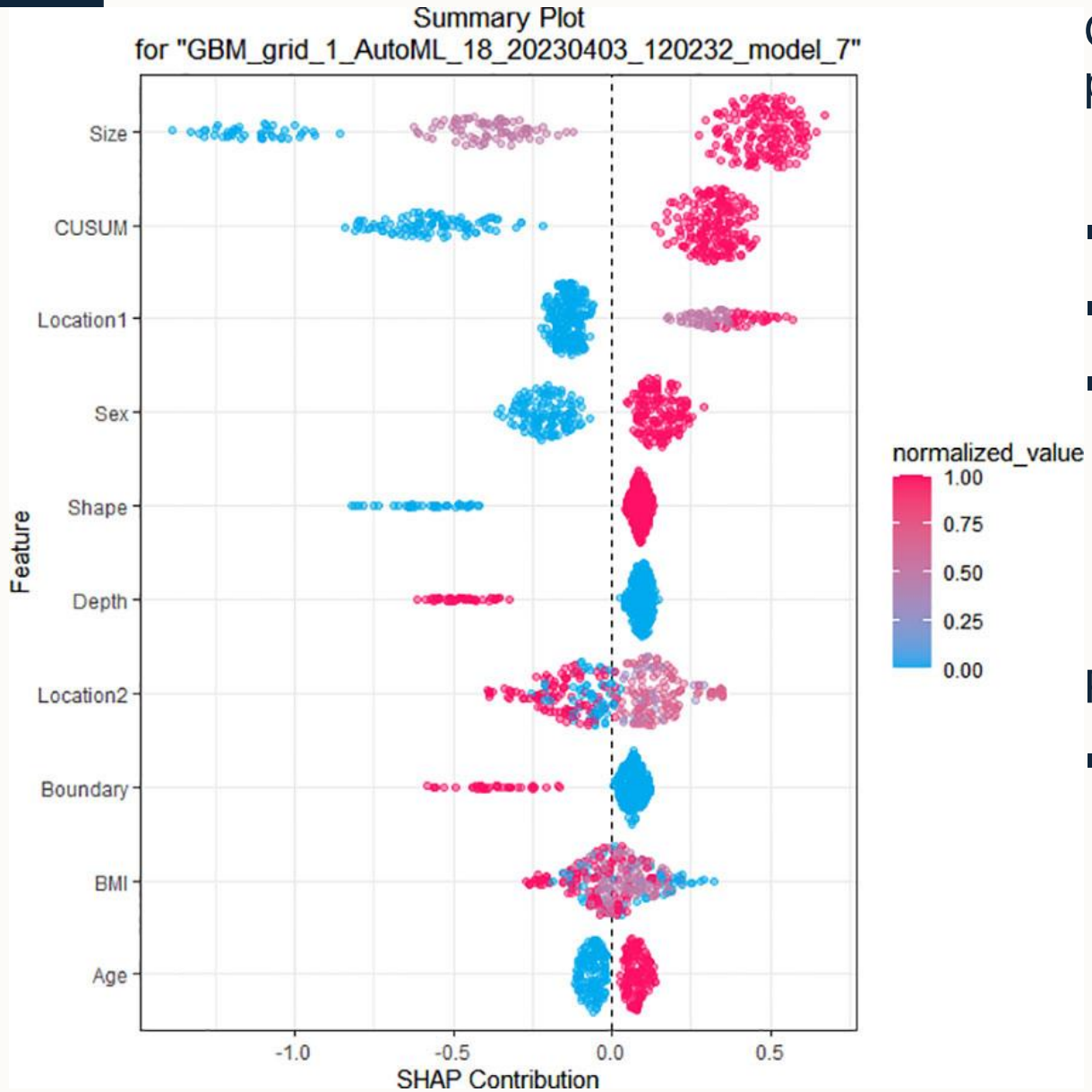
The x-axis shows the feature effect: The weight times the actual feature value.

From Molnar's book.

SHAP

- SHapley Additive exPlanation.
- Treat input features as “players in a game” to model an instance.
- Uses game theory to generate Shapley values to measure this.
  (details omitted)
- Thus feature contributions to models can be shown in SHAP summary plots.

SHAP Example



Contribution of features to difficult surgical procedure.

- Model: Gradient Boosting.
- Shape shows distribution of values in feature.
- Values close to 1 (red) mean higher contribution.

From:

- Liu L, Zhang R, Shi D, Li R, Wang Q, Feng Y, Lu F, Zong Y, Xu X, *Automated machine learning to predict the difficulty for endoscopic resection of gastric gastrointestinal stromal tumor*, Frontiers in Oncology 13, 2023

Workshop

- Fairness and Bias Analysis.