COMPUTATIONAL APPLICATIONS TO POLICY AND STRATEGY (CAPS)

Session 6 – Aspects of AI Policy, Strategy and Safety

Leo Klenner

Outline

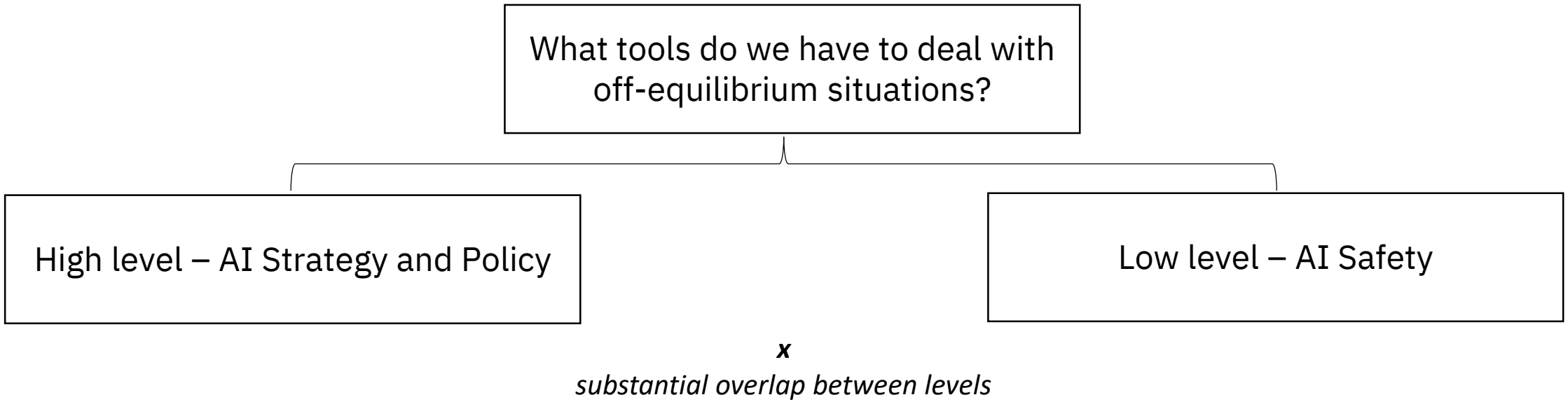
1. Recap
2. Overview of AI Strategy and Policy
3. AI Policy Case Study
4. Overview of AI Safety
5. Explanations of AI
6. Specifications for AI

Recap – 5/6 Completed

- We have come a long way:
  - Games as an environment to train and test AI
  - Python as a high-level language to write AI code
  - Implementing rule-based agents for StarCraft II
  - Running and evaluating reinforcement learning agents for StarCraft II
  - Primer on statistical learning and applications to game analysis

Today‘s Goal

> In previous sessions we tryed to get things to work
> Today we look at what happens when things do not work as planned



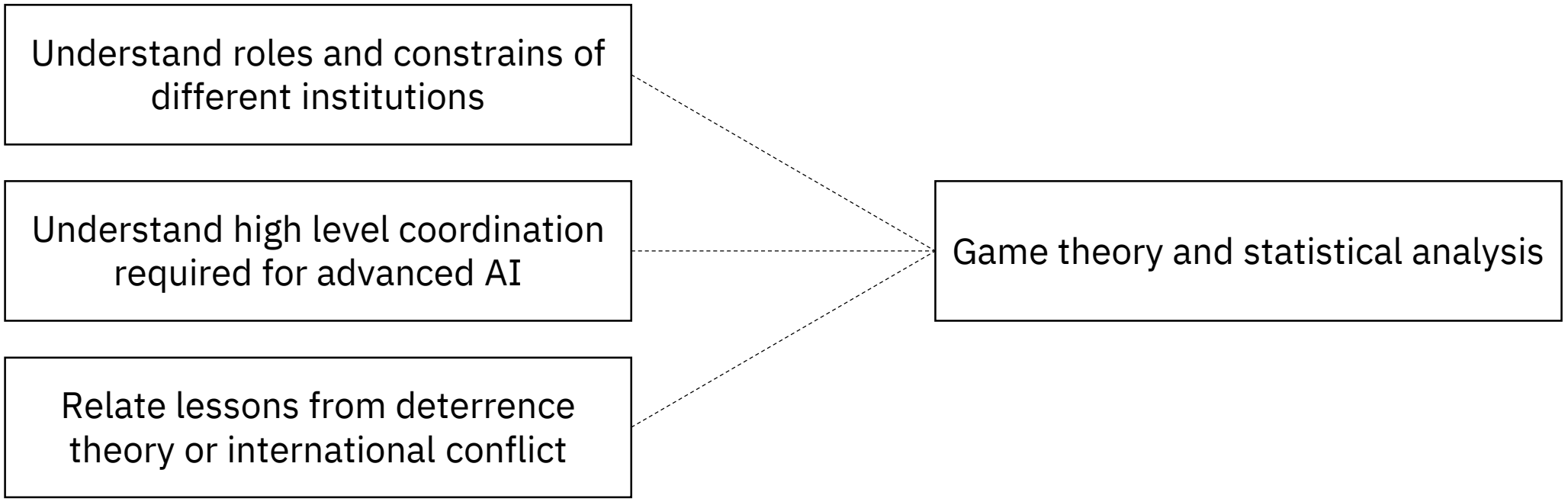
> AI Policy and Strategy

> AI Safety

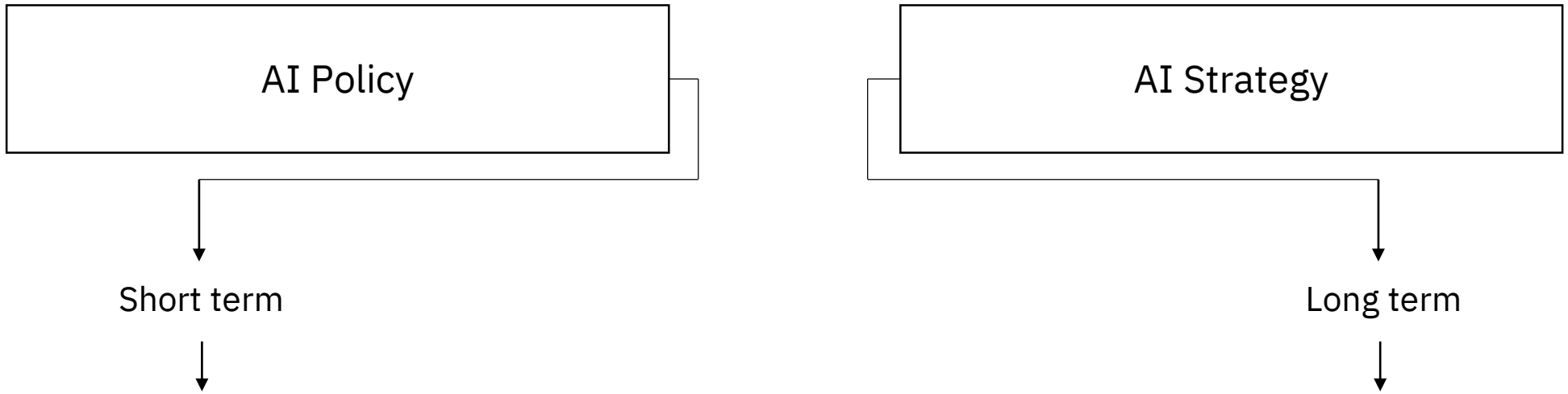
AI in an IR Context

"International relations is a valuable subdiscipline in dealing with AI."

~ Miles Brundage, Policy Research Scientist, OpenAI, *Guide to working in AI Policy and Strategy*



Analyzing Decisionmaking *on* AI



Understand and shape the context of AI implementations

Example of AI Policy Work

"As a Research Scientist at OpenAI, you'll be responsible for analyzing the AI policy landscape; developing and defining OpenAI's policy positions; representing those policy positions in public and private forums; and engaging with relevant stakeholders and information sources to further your own technical understanding of artificial intelligence. ... OpenAI has a range of policy interests relating to AI which include (but are not limited to): forecasting the rate of progress of AI technologies, analyzing how malicious actors may re-purpose AI, understanding how AI might change the geopolitical landscape, and exploring how AI will alter the makeup of the economies it is deployed into."

- OpenAI, Research Scientist, Policy, job description

Case Study of AI Policy – Wargaming

> As an AI Policy Researcher (contractor), you‘re tasked to identify the failure modes of the following scenario

> “Imagine logging onto a wargame named Athena to practice planning an air assault mission to seize blocking positions in support of an amphibious landing. The game forces you to complete the planning process while you talk to an Alexa-like application, who reminds you about forms of defense, the definitions of different tactical tasks, and relevant historical examples. As you play the game, an AI application captures the data and compares your use of cover and intersecting fields of fire, among other factors, to rate your performance while contributing to a larger database of how U.S. military professionals fight. At the end, Athena assesses the data and offers you constructive tips, comparing your efforts to those of top-performers.”
> - Benjamin Jensen et al., Wargaming with Athena: How to Make Militaries Smarter and Faster with Artificial Intelligence (War On the Rocks, 06/2018)

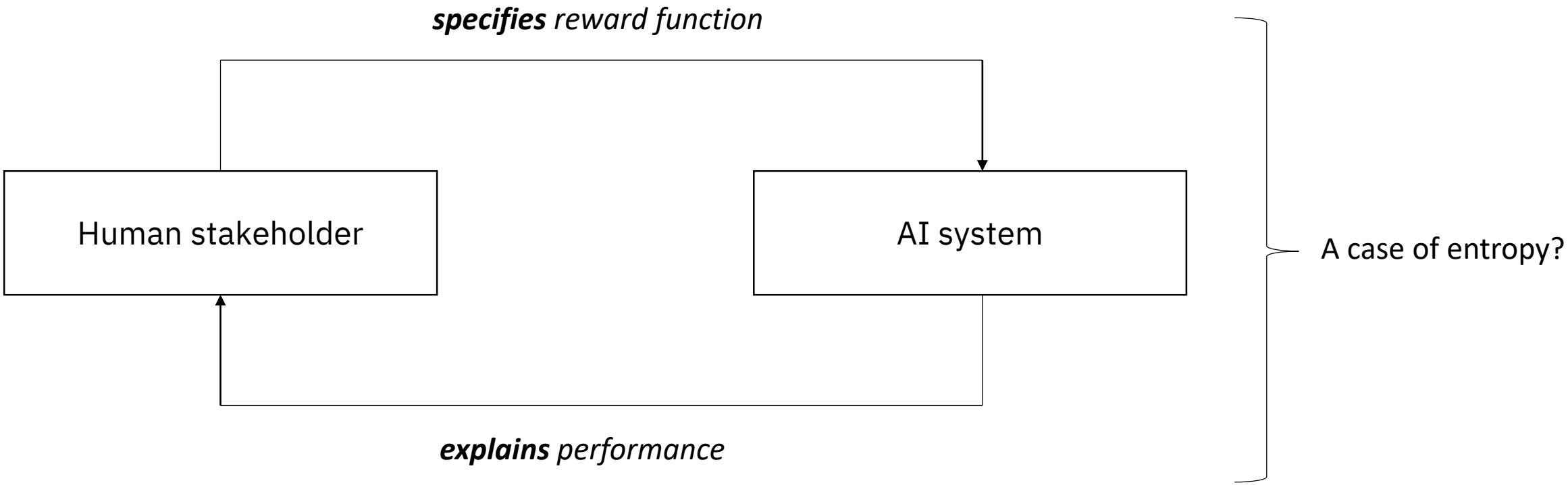
Evaluation of the Problem

- Do we understand the environment (deterministic v. stochastic)?
  - If the environment is deterministic, to what extent we extrapolate?
  - How much variance is in the training data?
  - What is the configuration of built-in AI opponents?
- How are recommendations deduced from performance data?
  - Does the system optimize for one optimal strategy?
  - How do we know this strategy is robust?
  - Does the system optimize within a human loop or across loops?
- Do we understand short-term v. long-term credit assignment of the task?
  - How to ensure balance across short-term and long-term planning?

> AI Policy and Strategy

> AI Safety

Human - AI Information Flows



Explainable AI

- AI systems are not self-explanatory
- Explainable AI deals with making these system understanble to humans
- Rapidly emerging field with wide spectrum of perspectives

- Derek Doran, et al. 2017. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.

Opaque Systems

Opaque systems. A system where the mechanisms mapping inputs to outputs are invisible to the user. It can be seen as an oracle that makes predictions over an input, without indicating how and why predictions are made. Opaque systems emerge, for instance, when closed-source AI is licensed by an organization, where the licensor does not want to reveal the workings of its proprietary AI. Similarly, systems relying on genuine "black box" approaches, for which inspection of the algorithm or implementation does not give insight into the system's actual reasoning from inputs to corresponding outputs, are classified as opaque.

Interpretable Systems

Interpretable systems. A system where a user cannot only see, but also study and understand how inputs are mathematically mapped to outputs. This implies model *transparency*, and requires a level of understanding of the technical details of the mapping. A regression model can be interpreted by comparing covariate weights to realize the relative importance of each feature to the mapping. SVMs and other linear classifiers are interpretable insofar as data classes are defined by their location relative to decision boundaries. But the action of deep neural networks, where input features may be automatically learned and transformed through non-linearities, is unlikely to be interpretable by most users.

Comprehensible Systems

Comprehensible systems. A comprehensible system emits symbols along with its output (echoing Michie's *strong* and *ultra-strong machine learning* [11]). These symbols (most often words, but also visualizations, etc.) allow the user to relate properties of the inputs to their output. The user is responsible for compiling and comprehending the symbols, relying on her own implicit form of knowledge and reasoning about them. This makes comprehensibility a graded notion, with the degree of a system's comprehensibility corresponding to the relative ease or difficulty of the compilation and comprehension. The required implicit form of knowledge on the side of the user is often an implicit cognitive "intuition" about how the input, the symbols, and the output relate to each other. Taking the image in Figure 3 as example, it is intuitive to think that users will comprehend the symbols by noting that they represent objects observed in the image, and that the objects may be related to each other as items often seen in a factory. Different users may have different tolerances in their comprehension: some may be willing to draw arbitrary relationships between objects while others would only be satisfied under a highly constrained set of assumptions.

Making AI Understandable I

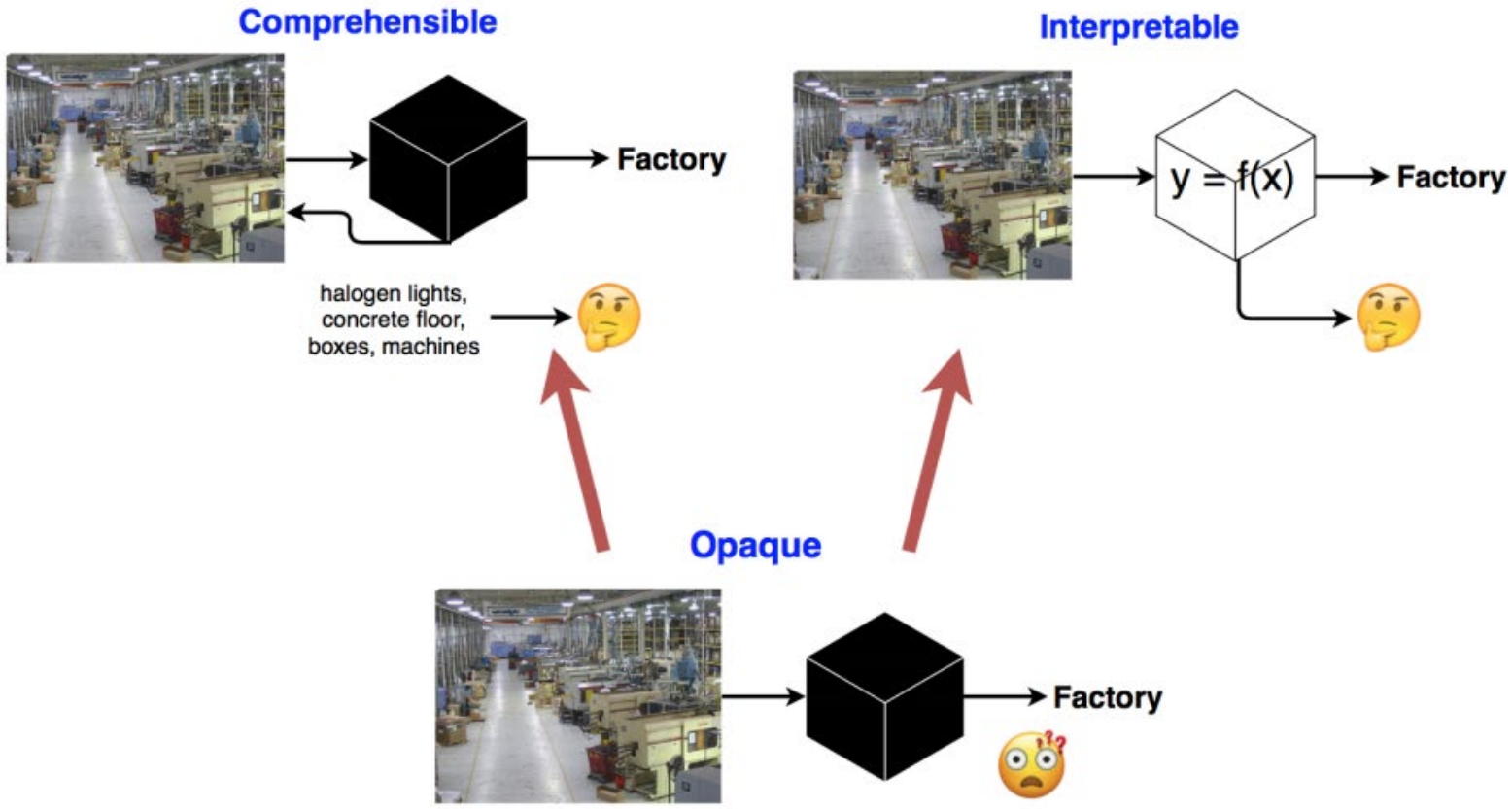


Fig. 3: Relation between opaque, comprehensible, and interpretable AI.

Making AI Understandable II

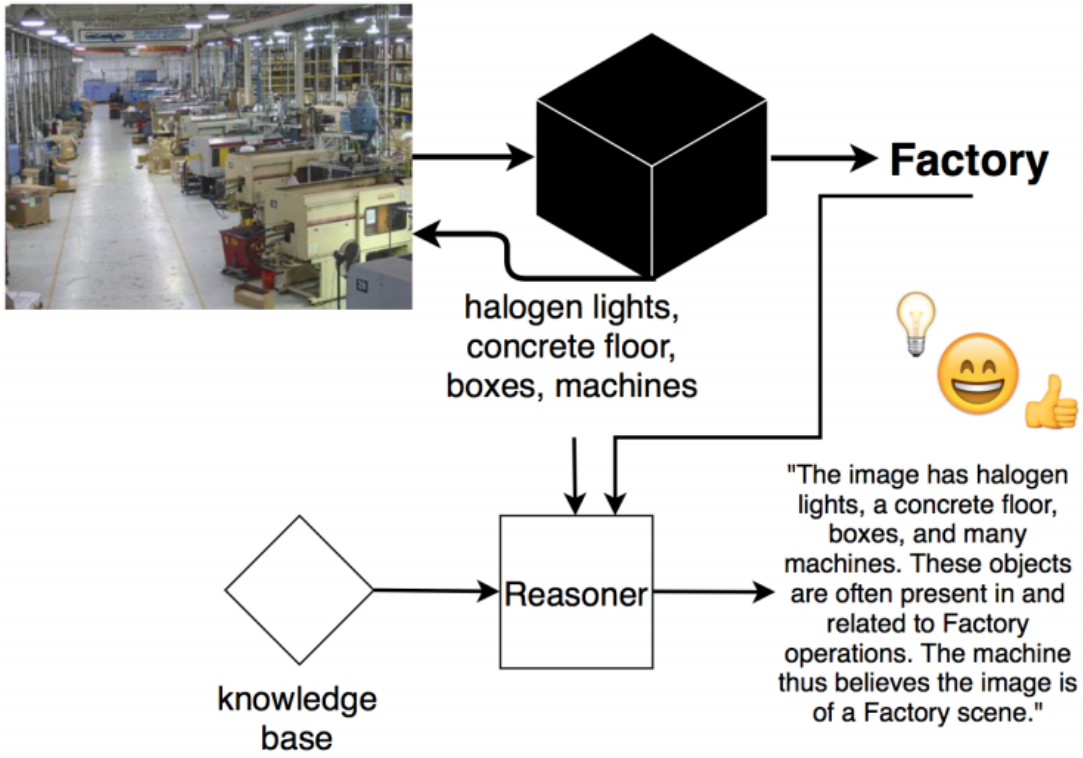
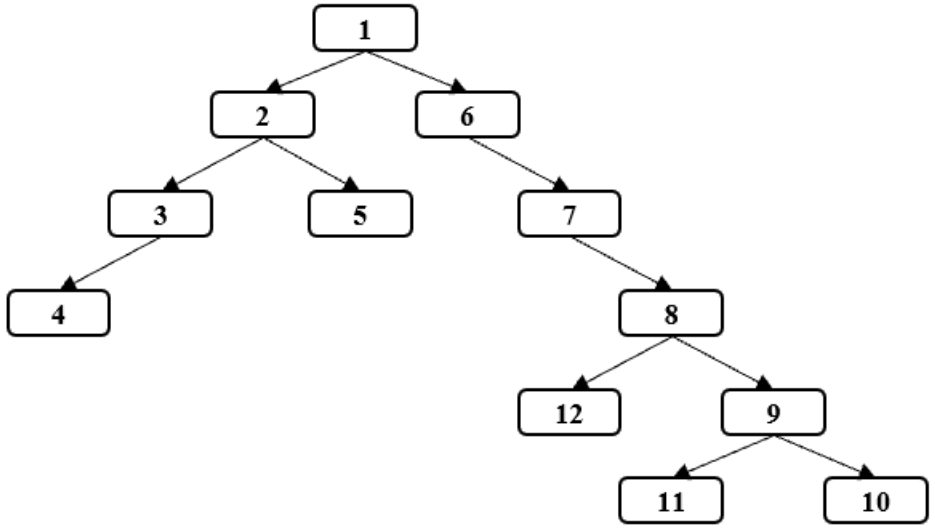


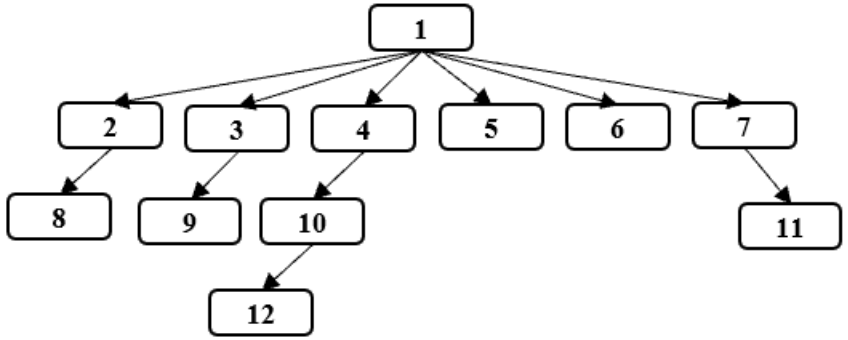
Fig. 5: Augmenting comprehensible models with a reasoning engine. This engine can combine symbols emitted by a comprehensible machine with a (domain specific) knowledge base encoding relationships between concepts represented by the symbols. The relationships between symbols in the knowledge based can yield a logical deduction about their relationship to the machine's decision.

Different Stakeholders, Different Explanations

> Optimal foraging for information in an environment changes based on stakeholders
> Hence, different stakeholders have different optimal explanations



Foraging path of human accountant of AWS



Foraging path of human team member of AWS

Specifications

> AI systems do not learn by themselves, they require specifications on how to learn.

> Specification problems arise when there is a mismatch between the **ideal specification** and the **revealed specification** => the AI system does not behave as we expected and want it to.

> Petro A. Ortega, et al. 2018. Building safe artificial intelligence: specification, robustness and assurance.

Different Levels of Specifications

- **ideal specification** (the "**wishes**"), corresponding to the hypothetical (but hard to articulate) description of an ideal AI system that is fully aligned to the desires of the human operator;
- **design specification** (the "**blueprint**"), corresponding to the specification that we *actually use* to build the AI system, e.g. the reward function that a reinforcement learning system maximises;
- and **revealed specification** (the "**behaviour**"), which is the specification that best describes what *actually happens*, e.g. the reward function we can reverse-engineer from observing the system's behaviour using, say, inverse reinforcement learning. This is typically different from the one provided by the human operator because AI systems are not perfect optimisers or because of other unforeseen consequences of the design specification.