Søren From Soelberg, M.A.

Towards a philosophy of explainable artificial intelligence for its use in critical infrastructure systems

GRADUIERTENKOLLEG KRITIS

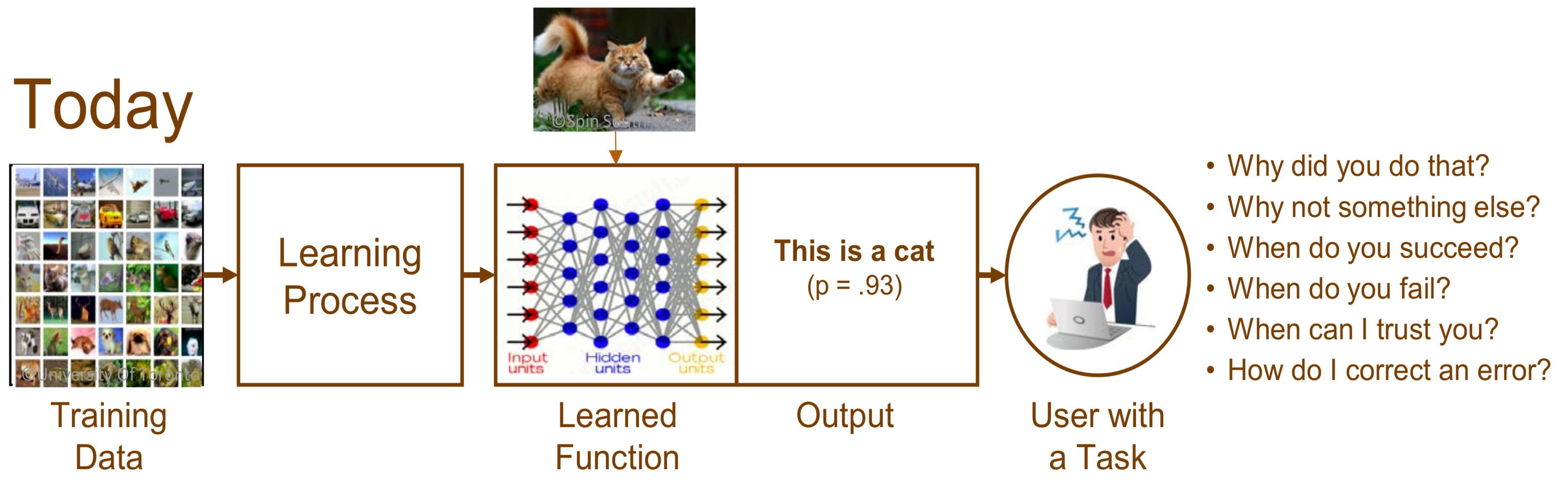
Philosophy of Technology / Prof. Nordmann und Prof. Rüppel

2. cohort

Subject & Research Question

How can we trust complex systems such as artificial intelligence (AI)?

- What kinds of explanation (or justification) are needed to satisfy our worries about the reliability of AI in critical infrastructure applications?
- To what extent does the special case of explaining AI generalise to the broader issues of understandability, explainability, trust and justification in human interaction with complex systems?



Current challenges with contemporary AI implementations (David Gunning & DARPA@IJCAI 2016, p. 3)

Interim Results

AI explainability remains under-defined

- The lack of a philosophically grounded notion of explainability impedes current efforts in AI research. Philosophy is still playing proverbial catch-up with the AI engineers.

Explainability engenders trust

- If critical infrastructure is to enjoy the benefits of contemporary and future AI technology, AI must inspire the trust and confidence of society at large, to which end explainability plays an important, but not all-encompassing part.

Are our expectations too high?

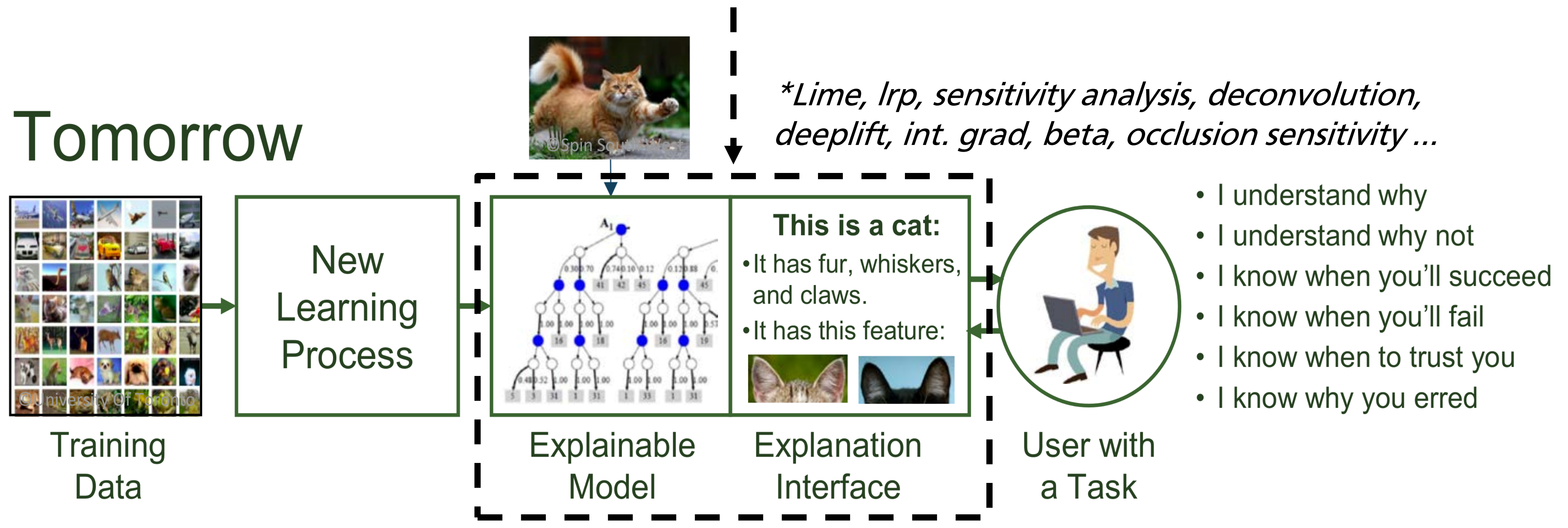
- In research literature and public policy, AI agents are subjected to higher expectations than to those human decision makers are generally held. A successful definition of AI explainability must consider the hitherto barely-charted territory of stakeholder requirements and expectations of AI with regard to advisory, semi- and fully autonomous AI decision making.

Methods

A philosophical framework adapted to the needs of AI explanation efforts is needed

- A philosophically grounded framework for the analysis of AI explainability is to be explicated based on cutting-edge literature in philosophy of explanation, philosophy of causation & philosophy of science.
- A representative spectrum of contemporary explainable AI models* is to be reviewed through said framework to verify to what extent they actually explain, are trustworthy, and appropriate for implementation in critical infrastructure domains.

**Lime, lrp, sensitivity analysis, deconvolution, deeplift, int. grad, beta, occlusion sensitivity …*



The asserted solutions of explainable AI (David Gunning & DARPA@IJCAI 2016, p. 3)

Cooperation & Highlight

Collaboration with colleague Tilman Beck M.Sc. on Explainable Natural Language Processing.

Interdisciplinary working groups Temporalities of (Transport) Infrastructure and Southern Theory Group.

Collaboration with the IANUS Peace Lab/Forum interdisziplinäre Forschung.

Highlight: The visiting Mercator Fellow Dr. Sabine Höhler from Kungliga Tekniska Högskolan (KTH) Stockholm.

Integration into the Research Programme

Construction of KRITIS/Function Failure/Protection

- Helping ensure the safe and trustworthy implementation of AI in critical infrastructure domains cuts across all three major research areas in KRITIS.

Criticality

- The notion of criticality as the criteria necessitating explainability is often invoked in contemporary explainable AI research without any theoretical grounding. The comprehensive research done by the 1. cohort on the concept of criticality is necessary to fill this gap.

Spatiotemporal transformation

- The development of AI systems is subjected to multiple push (technological, commercial) and pull (legal, ethical, societal) effects, which influences AI's layered mode of introduction, application, and adaptation into society.

Goal: Delivering foundational research that helps facilitate a safe and effective application of AI in the development, maintenance and protection of critical infrastructure.

TECHNISCHE UNIVERSITÄT DARMSTADT

DFG