Regulation of AI and Corresponding Explainability Practices

Keri Grieman*, Joseph Early**

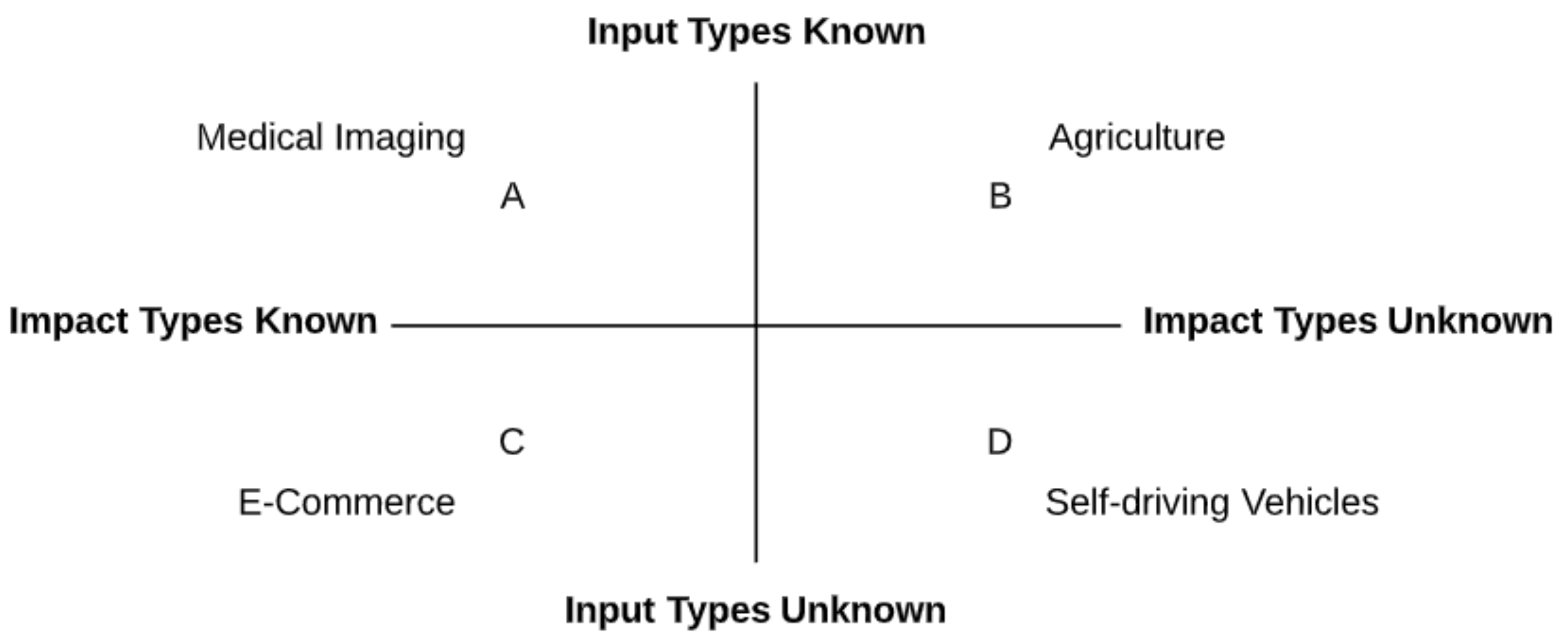
*Queen Mary University of London

**University of Southampton

Regulation of AI

Given the law's focus on regulating humans, regulation of AI may require a re-imagining of some core legal concepts. Civil wrongs (as opposed to criminal wrongs) focus on a duty of care, breach of duty, causation, and harm. There are questions such as foreseeability and reasonableness: was the harm reasonably foreseeable? Did a person act reasonably to prevent a harm occurring? These questions are more difficult where an AI is involved: a human is not making 'decisions' at all levels, and interpreting why decisions were made becomes increasingly difficult the more complex the AI.

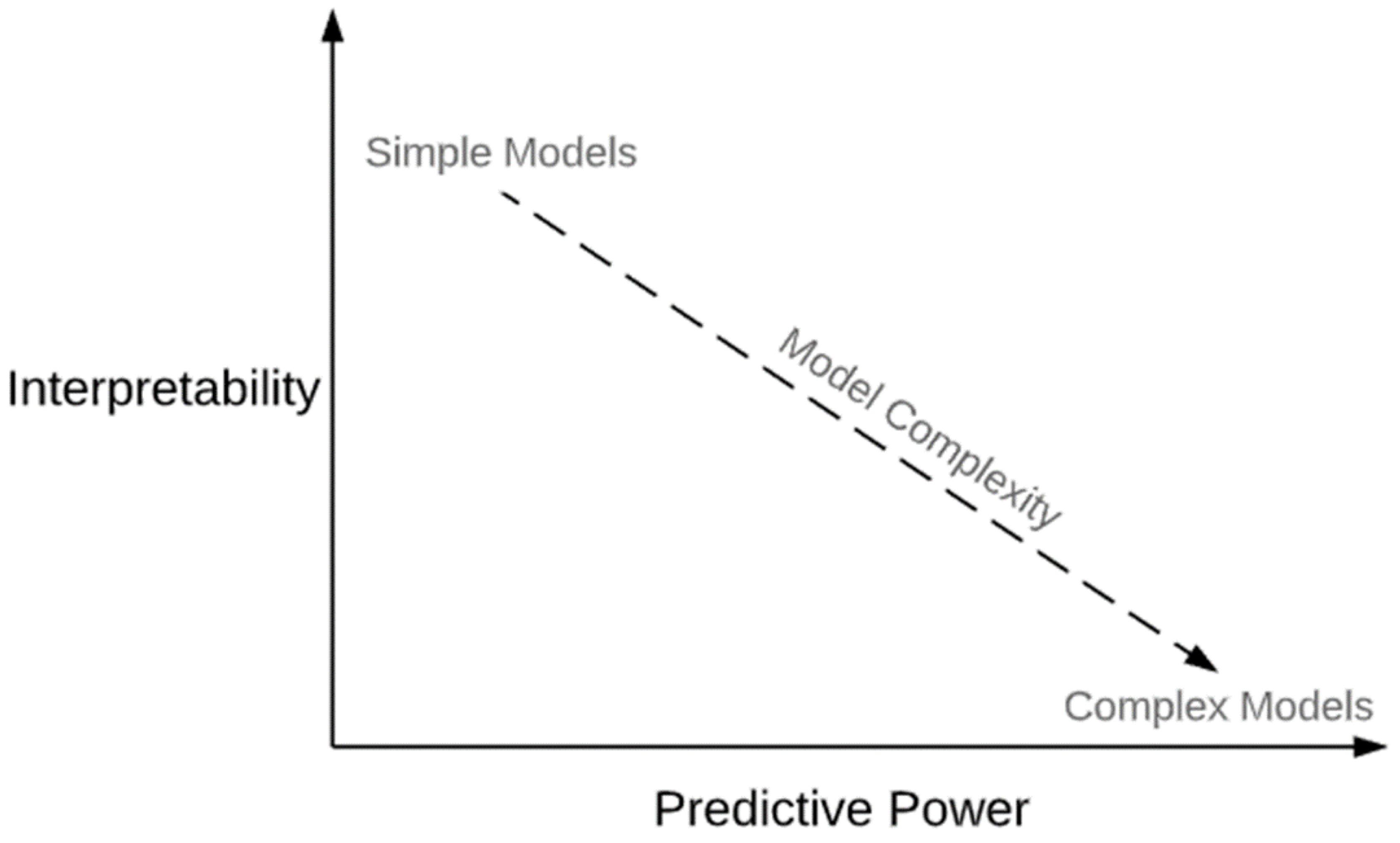
Potential Regulatory Frameworks



A potential framework for regulation is to consider the input to and impact of an AI system. To a point, these follow parallels with other regulation: the more foreseeable a harm, the more care must be taken to prevent it. However, where certain elements are unknown, other tools must be used.

Explainability for Regulation

The recent trend in machine learning and AI research has been towards more complex models that are able to achieve high performance on difficult tasks. Unfortunately, this transition from simple models to complex models comes with a loss of model interpretability, raising concerns about trust and guarantees of safe performance.



Model interpretability can be regained by using explainability tools. Explanations provide human understanding of autonomous decision making, helping to bridge the gap between the technical side of AI development and its regulation. Explainability requirements could be used as regulatory tools where input and/or impact are unknown.

Contact

Keri Grieman - kgrieman@turing.ac.uk

Joseph Early - jearly@turing.ac.uk