Position: **In Defence of Post-hoc Explainability**

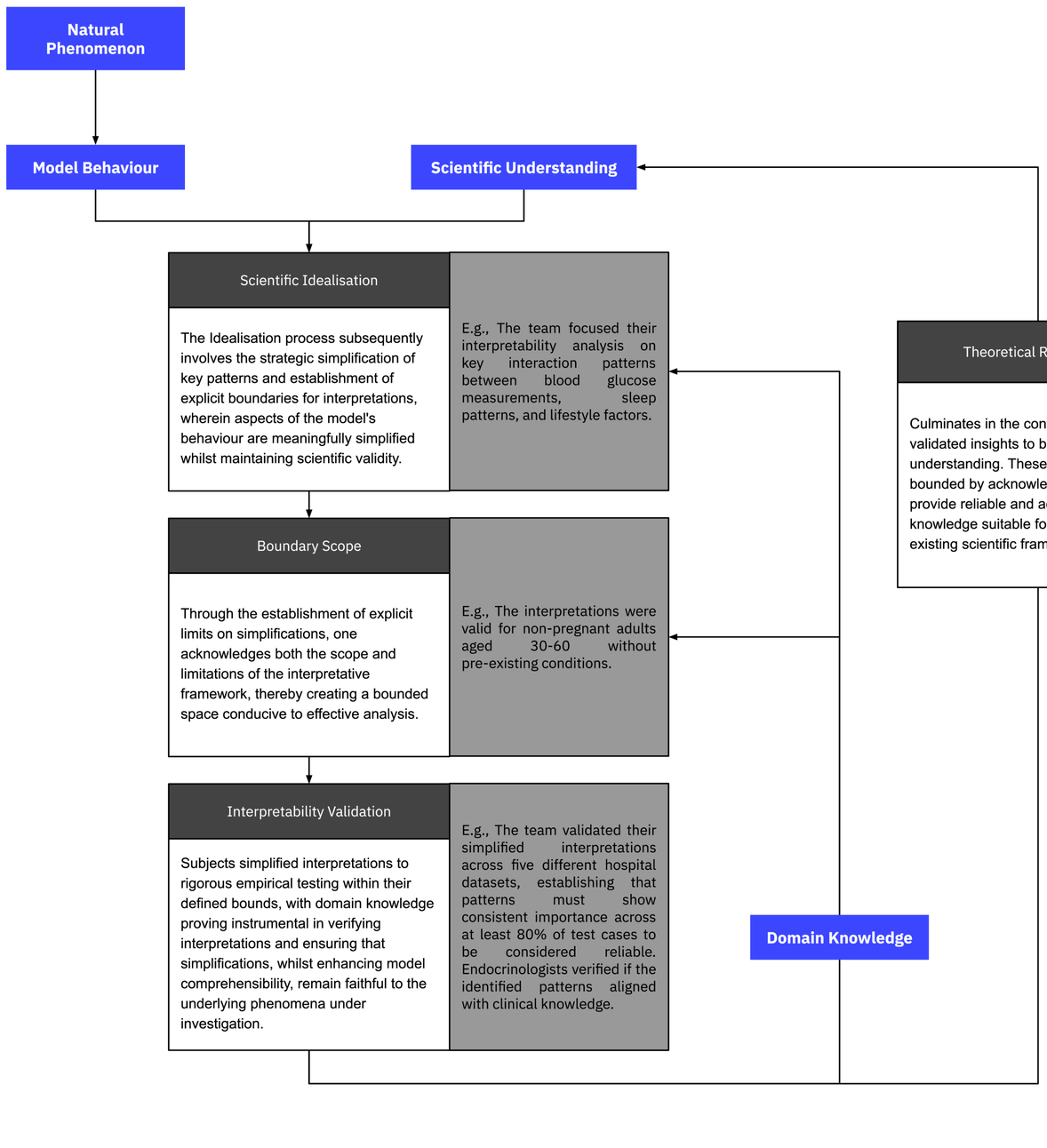
XAI # Philosophy of Science # AI for Science # Epistemic Justification

Author: Nick Oh (socius)

TL;DR

Whilst complex AI models are increasingly used in scientific research, their "black box" nature creates tension with scientific understanding. Some argue we should only utilise simpler, inherently interpretable models, especially for high-stakes decisions. This paper introduces Computational Interpretabilism - a framework that justifies using complex models with post-hoc explanations. Drawing parallels with human expertise, the paper shows how post-hoc explanations can generate valid scientific insights when properly bounded and validated. The key insight is that complete mechanical transparency is not necessary for legitimate scientific knowledge.

Bounded Factivity



We establish two key principles that together justify post-hoc interpretability methods in scientific ML. Mediated Understanding reveals *how* scientific knowledge emerges through structured interactions between models, methods, and domain knowledge, while Bounded Factivity demonstrates *why* such mediated processes can be epistemically valid despite their approximative nature.

Mediated Understanding

Direct access to model mechanics is not necessary for scientific insight [Sullivan, 2022, Beisbart and Räz, 2022]. Instead, understanding emerges through structured interpretation of model behaviour, and the relationship between model understanding and phenomenon understanding is reciprocal.

Bounded Factivity

Rather than demanding complete factivity – perfect correspondence between interpretation and model mechanics – we advocate for truth within explicitly acknowledged limits and simplifications; just as traditional scientific models advance understanding despite their simplifications.

Mediated Understanding