Explainability and Responsibility: Developing Trustworthy Artificial Intelligence Models

S. Baker

Lecturer, College of Science and Engineering
James Cook University, Cairns, Queensland, Australia
stephanie.baker@jcu.edu.au

Abstract:

In recent years, artificial intelligence (AI) has become pervasive for decision making in a wide range of fields, including healthcare and biomedicine. AI offers significant potential to the healthcare domain; however, it is not without risk. Without caution, AI models developed based on health data can learn to perpetuate biased, incomplete, outdated, or incorrect information. In the context of healthcare, poor decision making by AI models can have life-altering consequences. The need for trustworthy AI models has given rise to the field of responsible AI, which prioritizes characteristics including fairness, transparency, accountability, and safety. The question of how to achieve responsibility remains an active topic in the literature. Explainable AI techniques, which focus on revealing the inner workings of AI models, have been repeatedly shown to improve transparency of AI models. However, recent research shows that explainability can address much more than transparency – in fact, it can address all characteristics required to achieve responsibility.

In this talk, we will first discuss responsible AI and the range of ways that responsibility has been defined. Then, we will introduce methods of achieving explainability in AI systems – both through explainability by design, and through post-hoc explanation techniques that can be implemented for any AI model. We then draw these two fields together, illustrating how explainability can be leveraged to meet the key characteristics required of responsible AI models in healthcare and related domains. This talk will illustrate that explainability is the first stepping stone to achieving truly responsible AI systems.

Biography:

Dr. Stephanie Baker is a Lecturer in the Electronic Systems & Internet of Things Engineering Department at James Cook University. She is a computer systems engineer with expertise in artificial intelligence for healthcare applications. Her current interests include non-invasive health monitoring using wearable and non-contact sensor data, and the use of explainable artificial intelligence (XAI) techniques to improve clinician trust in artificial intelligence tools.