Department of Philosophy

Lingnan University 嶺南大學 香港 Hong Kong

Dr. David Jacob Harrison

Hong Kong Catastrophic Risk Centre, Lingnan University

The Message Hidden Within the Pattern: A Reverse Alignment Problem for Debates in Artificial Intelligence

Date : 24 March 2024 (Monday)
Time : 16:45 - 18:00
Venue : LKK103 & Zoom Seminar

Here is the Zoom link to the seminar –

Abstract: This paper introduces what I call the reverse alignment problem [RAP] into the literature on AI alignment and the philosophy of technology. Much of the literature on the alignment problem depicts the issue largely as a technical or engineering problem concerning the right specification of the goals and objectives to be pursued to ensure the machine pursuing the objective approximates (is consistent with) the intentions of the designer. This paper builds on this literature by introducing a bidirectional interaction between human values and the machines we use to effectuate our intentions.

I argue that alignment works in both ways: we could align machines with the complex tapestry of human values (notoriously difficult); or we can reduce and simplify human values, preferences, and goals so as to be easier to satisfy—a process that is, or so I argue, already underway. Thus, the RAP identifies the way in which we are already paving the way for forms of alignment that may not be desirable, indeed, may be misaligned with the social values for which they were initially intended. Contextual and internal factors that cannot be directly observed present a challenge to machine learning systems, which rely both on reducing uncertainty and on formalizable, mathematical constructs on which it can be trained and from which it can draw inferences and generalisations. Pulling from a rich (though rarely cited in the alignment literature) body of work in the sociology and philosophy of science, I indicate how contemporary machine-learning techniques inherit 20th century behavioural science theories (Skinnerian behaviourism; Ekman's theory of emotions) that sought to reduce the complexity of human interiority into scientifically amenable behavioural markers—a trend amplified and magnified with modern theories of behaviour in machine learning. As philosopher of technology Karen Crawford has observed on emotion recognition in machine learning, "Techniques have been developed to reduce the messiness of feelings, interior states, preferences, and identifications into something quantitative, detectable, and trackable" (Crawford 2021). Building on this insight, I argue that the RAP is the process in which the human dimension of the alignment problem is surreptitiously modified, modulated, and manipulated to align more easily with machines: alignment achieved in the reverse direction—not by making machines more like humans, and therefore sensitive to contextual features of the value landscape, but by making humans more 'machine-like'.

All are welcome

For enquiries, please call 2616 7445 / 2616 7488 or e-mail to dphilo@ln.edu.hk