Abstract: Dual Virtues for AI Systems?

Contemporary AI-models are a far cry from epistemically competent: while the achievements of a programme like AlphaGo are very impressive, it cannot hold a conversation, LLMs like ChatGPT can hold something that resembles a conversation but lack all other epistemic competences. At the same time it is not very clear what epistemic competence even consists of: What does it take to be a good epistemic agent?

Virtue epistemology is the subdiscipline in epistemology that has investigated what it takes for an epistemic agent to be excellent. Clearly, the approaches of virtue epistemology cannot be translated 1:1 onto research on AI-models because human cognisers possess different cognitive resources than AI-models. However, the normative demands on human cognisers may also be applied to computer models if we aim for them to acquire human-like epistemic competences.

Now, virtue epistemology consist of two camps that are often taken to be incompatible: virtue reliabilists and virtue responsibilists. The former define virtues as capacities that reliably produce true beliefs, and auxiliary virtues as capacities that augment the reliability of other virtues. Reliabilist virtues are capacities like facial recognition or perfect pitch. The latter camp defines virtues as excellent habits that help us to manage our epistemic life well and be responsible for our beliefs and convictions. They are modelled on Aristotelian moral virtues. Examples of responsibilist virtues are curiosity, judiciousness, or epistemic humility.

A recent trend in virtue epistemology has been to propose ways how these two camps can be harmonised. Elsewhere, I have argued that reliabilism and responsibilism are excellences of the different types of cognitive processing which humans use. Reliabilist virtues are the disposition of automatic, fast, and heuristic *Type 1* cognition modules to function excellently, i.e. reliably. Responsibilist virtues, are the dispositions of slow, controlled, and rule-based Type 2 cognition to function excellently. Excellent Type 2 cognition goes beyond mere reliability, it includes problem-solving competence, investigatory excellence, as well as the monitoring of other cognitive processing. Consequently, virtue reliabilism and virtue responsibilism do not compete; instead, they complement each other.

AI models do *not* possess such a dual process structure, so why should we care about dual virtues? Human epistemic excellence cannot be achieved with only one of the virtue types. Just being reliabilistically virtuous will limit cognisers to the contexts in which they are reliable, and it makes them passive recipients of information. Just being responsibilistically virtuous on the other hand will be highly inefficient, because a cogniser will first have to sort through raw unprocessed perceptual data. From this, we may infer that a competent AI-model will also needs both types of competences. Interestingly, the two paradigmatic types of contemporary AI-models appear to function either like Type 1 or like Type 2 cognition. There are the statistical models, trained on data to extract patterns – this is part of the function of Type 1 cognition. Second there are the rule based AI models which partially implement features of Type 2 cognition.