On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of 'I don't know'

Anders Hansen, University of Cambridge

We present the consistent reasoning paradox (CRP). Consistent reasoning -- at the core of human intelligence -- is the ability to handle tasks that are equivalent, yet described by non-equivalent sentences (tell me the time!/could you please tell me the time?). CRP: There are classes of problems for which there exists an AI that does not hallucinate and is correct on all the problems, given that they are described by specific sentences. However, if such an AI would emulate human intelligence and attempt to be consistently reasoning, it would hallucinate (wrong answers) infinitely often on the same problems. Moreover, it is strictly harder to detect if the AI was wrong than solving the original problem. In fact, one cannot detect the hallucinations (with probability $p > 1/2$) even with access to an oracle providing a correct solution to the problem. This implies that any trustworthy and consistently reasoning AI must be able to say 'I don't know', and that this is the optimal form of trust possible for any consistently reasoning AI. In summary: A super AI may avoid hallucinations, however, if such a super AI will emulate human intelligence, it becomes fallible and to maintain trustworthiness must be able to say 'I don't know'.