Hard Choices and Hard Limits for Artificial Intelligence

Bryce Goodman, University of Oxford -- bwgoodman@gmail.com

The stakes and the question

AI predictions already outperform human judgement, and this trend is poised to continue. In the coming years, many of the decisions currently made by people will – and *should be* – delegated to AI.

The question: Are there *theoretical* limits on AI in decision making -- choices that AI cannot, and should not, resolve.

The argument

Hard choices occur because alternatives are “on a par”: one is neither better, worse nor equal to the other and yet the alternatives are comparable (Chang 2002).

Hard choices reveal limits on the utility model of rational choice, which says that agents’ choices are governed by a utility *function*.

Machine learning applications require a utility (aka reward or objective) function, and so will also be limited by hard choices.

The case of machine fairness shows that only human agency is capable of resolving hard choices in the design of AI systems.

Therefore, there are *theoretical* limits on AI in decision making.

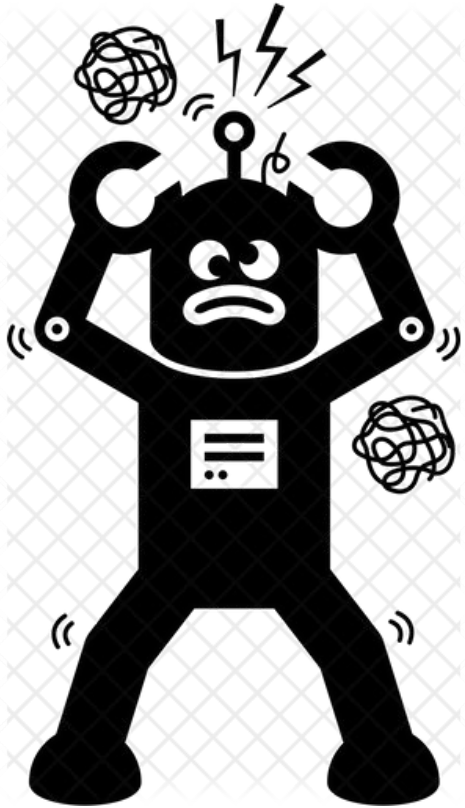

Image Credit: Iconscout.com