Was eine KI niemals können wird, wird auch der Mensch niemals können

Überlegungen zum Begriff der KI und deren Grenzen

Julian Braunwarth

Zusammenfassung (Deutsch)

In der interdisziplinären Debatte über die Grenzen und Möglichkeiten von KI gibt es ein Spektrum an Positionen, wie sie unterschiedlicher kaum sein könnten. Dabei scheinen die verschiedenen Lager in der Debatte häufig sehr unterschiedliche Vorstellungen davon zu haben, was der Begriff der KI überhaupt umfasst.

Um der Frage, was einer KI prinzipiell möglich oder unmöglich ist, sinnvoll nachzugehen, wollen wir deshalb zunächst klären, was man unter dem Begriff KI überhaupt verstehen kann bzw. sollte. Dabei soll dafür argumentiert werden, hier einen weiten und prinzipiell technologieoffenen Begriff von KI zu verwenden. Als „intuition pump“ soll uns hier ein Gedankenexperiment dienen, bei dem sich hypothetische Wissenschaftlerinnen in der Vergangenheit mit den Grenzen und Möglichkeiten von Technologie auseinandersetzen und ihre Argumente dabei auf die Dampfmaschinentechnologie beziehen. Selbst wenn deren Argumentation im Bezug auf diese Technologie tatsächlich zutreffen würde, wäre sie heute im Bezug auf die aktuelle Debatte um die Grenzen und Möglichkeiten von Technologie wohl vollkommen obsolet. In ähnlicher Weise könnte man älteren KI-kritischen Argumenten (wie bei Dreyfus, Searle oder Lucas) unterstellen, lediglich die Grenzen von symbolisch-logikbasierter KI zu behandeln – und analog dazu der aktuellen KI-Kritik, sich nur auf die aktuelle Technologie der künstlichen neuronalen Netze zu beziehen.

Mit einem technologieoffenen Begriff von KI geht einher, dass sich Argumente im Bezug auf die Grenzen von KI nicht zu sehr auf die Grenzen einer bestimmten Technologie stützen können. In einem Beispiel für eine solche Argumentation soll gezeigt werden, dass selbst die am denkbar weitesten entwickelte KI sich niemals vollständig selbst verstehen kann.

Im Anschluss daran wird dafür argumentiert, dass die Entwicklung einer KI – in einem weit gefassten Sinne – mit Fähigkeiten auf mindestens menschlichem Niveau zumindest nicht unmöglich ist. Wenn dies aber der Fall wäre, so müssten wir alle Fähigkeiten, die wir einer KI prinzipiell und grundsätzlich absprechen, auch dem Menschen absprechen. In anderen Worten hieße das: Was eine KI niemals können wird, wird auch der Mensch niemals können.

What AI Can Never Do, Humans Can't Do Either

Reflections on the concept and the limits of artificial intelligence

Julian Braunwarth

Abstract (English)

In the interdisciplinary debate about the limits and possibilities of AI, there is a spectrum of positions that could hardly be more different. The various camps in the debate often seem to have very different ideas of what the term AI actually encompasses.

In order to address the question of what is possible or impossible for AI in principle, I first want to clarify what can or should be understood by the term AI. I will argue in favour of using a broad and technology-open concept of AI. A thought experiment in which hypothetical scientists of the past discuss the limits and possibilities of technology and relate their arguments to steam engine technology will serve as an intuition pump. Even if their arguments were actually correct for that technology, they would probably be completely obsolete today in the current debate about the limits and possibilities of technology. Similarly, older AI-critical arguments (such as those of Dreyfus, Searle or Lucas) could be criticised for only addressing the limits of symbolic logic-based AI – and, analogously, current AI criticism for only addressing the current technology of artificial neural networks.

A technology-open concept of AI implies that arguments regarding the limits of AI cannot rely on the limits of any particular technology. As an example of such a technology-independent argument, I will show that even the most advanced AI can never fully understand itself.

It is then argued that the development of an AI – in a broad sense – with at least human-level capabilities is at least not impossible. In this case, however, we would also have to deny humans all the abilities that we deny an AI in principle. In other words, this would mean that what AI can never do, humans can't do either.