University of Bern
Discussion in the seminar “CAS AML M5 Philosophy and Ethics of Extended Cognition and Artificial Intelligence”
Nov, 15th, 2024

Paper: James H. Moor, Are There Decisions Computers Should Never Make? Nature and System 1 (1979), 217-229

1. context:

We now turn to ethical issues concerning AI.

2. Summary of the paper

See also slides by Timothée

a. Is it correct to say that machines can make decisions?
   - Distinction between decisions in the narrow and the broad sense: narrow sense: choose this or that, might be random picking; broad sense: including deliberation.
   - Interesting question: Can computers decide in the broad sense?
   - Possible answer: No, they only calculate.
   - But we can redescribe what they do, and this redescription is very common.
   - Better answer: Yes, they can decide, e.g. when playing checker (using an old version of machine learning they “reason” what the next move should be).
   - Additional argument: Humans could reason in precisely the same way, and if they did, we would say that they deliberate and decide.
   - Objection: consciousness needed for decisions.
   - Counter: many human decisions are unconscious.
   - Objection: humans, and not computers, decide.
   - Counter: computers can decide, even if humans decide to use them; cf. delegation of power. Power is not ability.
   - If we don’t say that computers decide, we don’t appreciate their importance.

b. When can computers decide?
   - There are limitations to the power to decide, e.g. halting problem.
   - But humans are not better.
   - Empiricist approach (taken to be a commonsensical position):
     i. How well computers decide is an empirical matter.
     ii. There is evidence about the answer given particular tasks.
     iii. So far, we don’t know how well computers do with decisions.
   - Computer pioneers exaggerate the deciding power of computers.
   - Other have denied it; e.g. Hubert L. Dreyfus. He is correct to point out that human behavior cannot be programmed at the level of everyday language, but there may be a deeper level, such that computers can imitate humans.
   - Example of selecting hypotheses: Computers are good at this.

c. How can we assess the decision-making power of computers?
   - In principle: track record and justification
   - Two types of decisions: clear standards and fuzzy standards (these are two extreme cases; there can be cases in between).
   - For clear standards, the track record matters; if good track record, no interest in justification; if not good, justification not of interest, either.
   - Fuzzy standards: justification more important; e.g. professor grading essays; maybe, there is not the correct grade, but the grades can be better or worse justified.

    - Justification can indeed be given by computers: Mycin: recommender system; contains rules, which can be used for justification.
    - Further basis for decisions not interesting (e.g. physical level); we don't understand human decisions from this perspective either.
    - It does not matter whether the decisions have a similar base as those of humans (after all, they are supposed to be better).

d. Are there decisions that computers should never make?
    - Weizenbaum: computers do not share our concerns, are outsiders.
    - But outsiders may be better at deciding certain things (e.g. physician for an alien society).
    - Three possible principles:
        i. Computers should never decide if people want to decide, cf. pleasure of deciding. Problem: There may be other reasons to let computers decide. If computers make decisions better, then they may decide, even if humans want to decide.
        ii. Computers should only decide in cases where they are better than humans. Problem: sometimes other considerations, e.g. deciding boring or time-consuming.
        iii. Computers should never override human decisions. Counterexample: traffic: it may empirically be true that traffic goes overall better if humans cannot override computer decisions.
    - Instrumental approach: Let computers decide if this serves our main goals better.
    - Implication: Computers shouldn't decide about our basic values.
    - Still, many decisions not about basic values, so computers could in principle take many decisions.
    - Still, there is fear about dehumanization. The real problem is responsibility. Even if computers decide, they are not full human persons, cannot be sued; so humans should continue to take responsibility.

3. Some discussion points

How machine decisions differ from human decisions

- Today, humans frame most computer decisions, e.g. it is specified by humans when a computer should decide and what the options are. Humans make these kinds of framed decisions too, but they can also make spontaneous decisions (e.g., I suddenly decide to stand up).
    - It might be that we humans don't want machines to make these kinds of spontaneous decisions.
        - Some people had the intuition that this kind of capacity might lead to risks.
        - There is also the issue that we want machine decisions to have a justification that spontaneous decisions might not have.
    - In certain cases, however, e.g. for care of elderly people, we might actually want spontaneous machine decisions (e.g. make a joke or bring someone a cup of coffee).
- Computers can't feel pain or emotions. However, pain and emotions are important for human decisions as they often ground them.
    - Again, the lack of this capacity might avoid the risk of having biased machine decisions.
- Humans have the ability to factor in the context
- Maybe, only humans can read intentions. This is important for the interpretation of laws (what did the lawgiver intend?). The reading of intentions is important for decisions that involve the law (recall the example of the poem of a merciful king that we discussed).

    - Maybe this applies more to GOFAI (Good Old-Fashioned AI) as modern connectionist AI seems better at factoring in context (see, for example, chatGPT).
        - Similarly, Dreyfus's critique of AI models might not hold anymore. (Note that this is not a weakness of the paper as Moor doesn't rely on Dreyfus's arguments but rather goes against them).

- Only humans may be able to apply mercy (same poem).

Machine decisions: responsibility and justification

- Who "takes the decision" in the case of a CV pre-selection for a job application?
    - Computer, executor, software engineer?
        - The latter option doesn't seem plausible.
    - Linked to the question of responsibility.
- Does the computer need to mimic human decisions?
    - It helps, as it makes it easier to understand and trust them.
    - Symbolic rule-based AI has the advantage of (typically) being more interpretable than modern ML or DL approaches.
        - This opacity of connection AI also makes it harder to avoid biases in machine decisions.
- There is a case of double standards here. After all, humans themselves make a lot of bad decisions.
    - Good to raise the standards for machines (where we can do so).
    - Still, one can ask: Are these double standards of society justified?
        - Yes, because humans, unlike machines, can bear responsibility.
        - No, at some point, the benefit of responsibility assignment (which applies to human decisions) should be outweighed by better performance.
            - For autonomous driving, we wondered at what point this cutoff is. When should we allow it?
            - We mentioned that autonomous driving is a case where performance can be measured, i.e., as number of deaths per mile.
- Is the justification really important? (as important as Moor thinks?)
    - One worry was that justifications are quite subjective.
    - However, to some extent, a justification can be judged objectively (for example comparing two different justifications of grades evaluators gave to a thesis).
    - We mentioned that LLMs can produce talk that looks like a justification.
        - But maybe the justification doesn't reflect the mechanism they actually rely on (see the paper *Language Models Don't Always Say What They Think*).
    - It seems like everybody agreed that justification is indeed important.
- Which decisions should machines not make?
    - Some people thought that machines should not make decisions in life-threatening situations.
        - Basically, you can use them in low-stakes situations but not high-stake ones.
    - Moor argues that machines shouldn't make decisions about our life's goals and ultimate values (e.g., decide if I should prioritize my career or my role as a parent). Decisions that involve only subordinate goals may be handed over to machines.

cb/pb