Durán & Jongsma: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI (Durán and Jongsma (2021))

Robert Michels*

December 15, 2024

1 Summary

Some of the main arguments and claims from the paper:

- Examples of applications of AI in medicine:
  - detection of illnesses in image material such as X-rays
  - prioritise information or patient files
  - provide recommendations for medical decision-making
- Epistemic opacity of AIs is a problem in this context: making a medical decision or diagnosis involves moral responsibility, for a decision or a diagnosis to be morally responsible, it has to be possible to give reasons

*robert.michels@edu.ulisboa.pt

for it, but epistemic opacity makes it impossible to give suitable reasons for the output of the AI

- Some people argue that AI should not at all be applied in medicine for this reason, but this is problematic, since AI offers substantial benefits (e.g. ability to handle extremely large amounts of data in short time)
- The authors propose *computational reliabilism* (CR) to circumvent the problem of epistemic opacity: only the reliability of the AI in producing the right output for an input matters; what kind of internal process leads the AI to give the right output and whether we can know or understand this process does not matter (this makes CR an *externalist* position (recall the distinction between internalism and externalism in classical epistemology))
- This means that CR is not a variant of XAI (explainable AI) since it offers no way to solve the problem of epistemic opacity, but merely avoids it
- The authors' argument for this strategy: AI is like other complex medical instruments like MRI-machines – they can also inform the medical decisions of a doctor, even if that doctor does not know how the instrument works internally – if this lack of knowledge does not make the instruments ineligible for informing a medical decision, then lack of knowledge due to epistemic opacity also doesn't make AI ineligible to inform medical decision
- The authors mention four indicators for the reliability of a process: verification and validation of process, robustness, history of successful implementations, and the expert knowledge which went into developing/training the AI
- According to them, AIs should be seen as giving input on medical decisions, not as being autonomous decision makers – human should always be the decision-maker
- Three relevant issues regarding decision-making based on data:

  - Different interpretations of medical data possible even if the same normative leading principle (e.g. saving life of the patient) is followed
  - Different normative leading principles possible
  - Status of illnesses as such and treatment is often controversial in medicine

- To be useful, a medical AI should be adapted to these issues, e.g. providing different diagnoses relative based on different relevant normative leading principles for the same case

2 Discussion

Some points raised in our discussion:

- Is the 'revenge' argument against XAI/explainable AI given in the paper (XAI provides explanations which themselves may be opaque) a good argument? – seems problematic, for example Zerilli's intentional stance approach to XAI is not completely epistemically opaque (though it could be argued folk psychology is somewhat) and can be used in successful explanations (w.g. of what went wrong in the wolf/dog mis-classification example)

- Why do we trust human experts in medicine, whose expertise is also epistemically inaccessible to us, but not AIs? Maybe because of social, moral factors? The human expert are committed to certain values, the AI is not

- Does reinforced and supervised learning of an AI correspond to learning of humans?

- Maybe AI is analogous to a pet like a dog: dog is also trained to do things which align with our value (e.g. sheep dog learns to herd sheep), even though it does not have this values itself (dog follows its instinct and exercised behaviour)

- Is there a clear difference between regular (non-AI) algorithms and an AI used for the same purpose? Currently often only a gradual difference, plus a lot of hype and marketing

References

Durán, J. M. and Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, 47(5):2020–106820.