AI OPACITY VS PATIENT'S AUTONOMY IN DECISION-MAKING

Jose Luis Guerrero Quinones, Independent Researcher (Spain)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: bioethics, artificial intelligence, medical decision-making, black box medicine, opacity

1. Artificial Intelligence in healthcare settings

1.1. Machine learning and Artificial Intelligence

- Functioning of Machine Learning (ML):
  - Two algorithms working on parallel: learner and classifier.
  - Neuronal networks: simple trainable mathematical units (neurons) organised in layers learning from the layer below. Increase of 'semantic density'.
  - Decision-making is the probabilistic way to transforming data into "knowledge" in order to make a prediction.

1.2. The problem of opacity in AI

- There is a break with the formalisation of procedural algorithms and semantic knowledge in Deep Learning Networks (DLN) that implies that their outputs are not based on well-defined or explicit criteria.
  - Thus, a new type of lack of transparency: in-principle-impossible to know/discern.
  - The functionality of DLN depends on the parameters used by the algorithm, the weights linked to the different inputs and the internal connection strengths of the different layers of neurons to provide the wished output, and not the algorithm itself. The weight of the links is hardly intelligible.
- Determining the trust in a model is especially important when an individual has to decide based on the AI prediction.
  - Doctors are in a better position to decide based on an AI system if an intelligible explanation is provided.
  - Humans usually have previous knowledge of the field where the AI is applied, which can be used to accept or reject a prediction if the understand the reasoning behind it
- No right to transparency and explainability of AI. There is only the need to offer *ex ante* information and access rights about basic aspects of the system's functionality.
  - Full transparency is thought to be a 'panacea', but both its accessibility and its comprehensibility present problems.
  - Transparency could be required if problems arise, similar to what happens in human mistakes in healthcare.
  - Making AI more explainable makes it less complex, which limits its

performance.

1.3. The impact on patient's autonomy

- It means to find the balance between decision-making power humans retain and the one that is delegated to AI.
- Impossibility to maintain the current notion of autonomy in our current digital era.
  - o Black boxes preclude the workability of an autonomy notion that includes independence and intentionality, for the lack of transparency, and thus of information, hinders the patient from obtaining relevant information to aid for her decision-making.

2. Preventing a negative impact of AI opacity

- Healthcare and medicine can highly benefit from improvements in AI and Deep Learning (DL) due to the enormous amount of data generated.
  - o DL can accept multi-type data, which of much relevance for healthcare applications (e.g., computer vision, language processors, robotic-associated surgery).
  - o Great impact to improve diagnosis and treatment: towards personalised medicine. ▪ Improvement of diagnostic accuracy.

    ▪ ML will become indispensable for clinicians.
- What is a satisfactory explanation? Two alternatives:
  - o Model-centric: understanding how the machine works.
  - o Subject-centric: local relation between one input and one output, relevant for patient prognosis.
- Main thesis: diagnostic skill is not essential for medical practitioners + relying on doctors' diagnosis where there are better computer-based procedures would be detrimental for healthcare.
  - o Ethical implications of using less accurate diagnostic systems (humans) that are proved to perform worse than others (AI).
  - o The relevant concern is how to improve the quality of health care by revising our current model of medical decision-making, using the best of both humans and computer-based systems
- Human control must be always maintained; thus, humans remain the ultimate controllers of the decision-making process.
- Meaningful human control: AI can and should be controlled by humans.
  - o Humans retain decisional authority: AI decision support systems as auxiliary tools for decision-making.
- Right to withdraw from AI diagnosis and treatment + right to be offered an alternative type of intervention.

3. Enhancing patient's autonomy by using AI

- In medical care settings the patient always delegates in others the obtainment of knowledge about her illness/condition. Autonomy remains 100% on the patient, but what matters is that she gets the information from a doctor who used an AI as a tool. Thus, it would be a problem of trust, and not decision-making; the latter concerns the physician exclusively (that is, how much they delegate of AI), and it may also convey responsibility issues.

- Respect for autonomy and intelligibility of AI systems.
  - ○ Disclosure of information to patients is only relevant in cases where risk management must be considered for decision-making.
  - ○ There are areas where the use of AI will not change at all the patient's autonomy to make decisions through informed consent.
  - ○ The key feature is to disclose clear, meaningful, and comprehensible information, so a general explanation of the AI system functionality will suffice to protect the patient's autonomy through informed consent.
  - ○ In legal terms, the obligation to inform the patient about the use of AI is not always mandatory; only when the use of AI is front-end, patients' autonomy requires clear notification to be dealing with a machine.
- Big Data provides patients with an unsurmountable amount of complex health data that can effectively improve their autonomy. Patients have more control over health data that they can themselves gather through wearable devices, for example, which might support personal autonomy. There is much more information available that patients could use to conform and consent to AI diagnostic and decision-making.

  ○ AI systems must be designed to include and consider patient-centred data. ○ Inherent
  sensitivity of health-related data and potential vulnerability of patients. ○ Is it at all possible
  to retain control on how the patient's data will be used in opaque diagnosis
  systems? What would, then, imply that a person has control over her data?
- Defence of a system where the implementation of AI is only partial, where is very unlikely that ML ever surpasses a level of conditional automation, where humans will remain in control of oversighting algorithm interpretation of images and data.