The Problem of Responsibility Assignment in AI

Among the vast issues emanating from the ethical and policy challenges of AI,a significant proportion deals with assigning blame or accountability. If a software does something that is considered wrong, or guides a hardware to do something wrong, who/ what is responsible? There are many examples of such instances nowadays, as AI is being used widely across domains and applications.

A hospital found that its AI software was mis-diagnosing patients – on test results it was grading the patients as healthy, whereas they were not healthy. Some patients suffered severe consequences as a result of this false negative result. In another instance, the AI software being used by a firm for recruiting candidates for job interviews, was found to be systematically discriminating against certain profiles. The firm was later sued for this. Another example is that of a facial recognition AI software refusing to enable a person of a certain ethnic background to use it, as it could not classify the individual as a legitimate user.

In all the cases mentioned above, the challenge for either courts, or management, or product makers was to identify who or what was to blame for the mistakes. A software for health diagnosis is created by a chain of contributors. They determine how its parts work and how it computes its results. When there is failure, the problem is to isolate and identify the specific link in the product chain that failed.

In operations and computer science research, there is a well-known credit assignment problem. This problem addresses the question of how credit or recognition can be assigned to the many steps that are taken to finally solve a problem. For example, if a problem is solved by ten steps, can it be the case that all the ten steps contributed equally to solve the problem, or were some steps more important than others in leading to the solution? This problem is solved in many ways, and forms the basis of many AI algorithms.

The responsibility or blame assignment problem is analogous to the credit assignment problem, with the difference that in this case the challenge is to find out who or what is responsible for a negative outcome rather than finding the contributions to the solution. As important and critical decisions move to algorithms, the responsibility assignment problem becomes important, as has been seen for many cases that have gone to the court.

Are the credit and responsibility assignment problems the same? In a neural network, for example, the design of the node is such that the output of any node, for given inputs, can be exactly computed. The challenge is to find which of the nodes eventually determined the result. However, the overall neural network system learns the weights that connects the nodes, and to that extent the system has agency. The credit assignment problem is of finding which of the nodes, after the weights are

assigned, are contributing most to the output. The responsibility assignment problem is that of finding the role of the neural network system, in a large collection of agents that take actions to achieve the final outcome. Some agents have control over their choices and actions, some don't.

In the domain of AI and its applications, the problem of agency has to take prominence. The outcome of an application, say a medical diagnosis, has many actors whose agency has to be understood and delineated. The agency of all actors is not the same, and some precede others in the decision chain, and hence their causal role has to be identified. We illustrate with one example.

Deep learning neural networks are often built with open source software tools, like Tensorflow and Keras. These tools were made by a community of developers, some supported by organisations like Google and Facebook, while others work independently. To build these tools they would have relied on other utility tools, like editors, compilers, and operating systems, that were built by other firms or groups. The agency of developers who built them is embedded in these utility tools. Those who build the AI tools inherit some constraints and affordances from using these utility tools, which affects their agency.

Organisations that build AI applications, like medical diagnosis software, select these AI tools, collect data for training, and apply the methods of deep learning to hone the software for use in a live context. This requires many decisions regarding design, selection of data for training and testing, tuning the models, validating against hold-out cases, and testing in a live environment.

Once a medical diagnosis application is deployed in a hospital, it may be used by technicians, nurses, or doctors. The specific arrangement and process in which decisions regarding diagnoses are made will be settled according to the routines and priorities of the doctors and the hospital staff. Here, the agency of the software and its developers is buried in the tools being used, whether software or hardware, and its presence is subtle and largely invisible.

In case of a failure, say a faulty diagnosis in which a patient suffers or loses theirlife, the responsibility assignment has to be called. Was it the fault of the software, or how it was trained, or how it was used by the technicians, or how the deep learner functioned? From aaccountability perspective, all the actors and agents have a role to play, some more than others.

As we increase our use of AI in our everyday lives and in business functions, the idea of agency and responsibility assignment will have to be brought into focus. As more decisions shift to algorithms and automation, it is imperative we understand how and why certain decisions were made, and which actor was involved in them. An analysis of 'how' and 'why' certain decisions are made needs to be taken into cognizance to capture the subtle differences in the agency of human and non-human

actors. This would be an important factor in assigning responsibility. The problem of responsibility assignment will be important for justice, as well as for policy.

Rahul De' is a professor of Information Systems, and Sai Dattathrani is a doctoral student, both at Indian Institute of Management Bangalore.

Rahul De': Professor Rahul De' teaches Information Systems and Artificial Intelligence at IIM Bangalore. His research interests are in ICT for development, open source, e-Government systems, and AI ethics. He has published four books and over 75 peer-reviewed academic articles.

Webpage: https://www.iimb.ac.in/user/67/rahul-de

Sai Dattathrani: Sai Dattathrani is a doctoral student in the Information Systems area at Indian Institute of Management Bangalore. AI and ethics are her areas of interest. She worked in the IT industry before pursuing her doctoral degree. She has a bachelor's degree in Engineering and master's in education technology.

Linkedin handle — https://www.linkedin.com/in/sai-dattathrani-91224436/

Rahul De

Visit: https://www.iimb.ac.in/user/67/rahul-de

Professor Rahul De' teaches Information Systems and Artificial Intelligence at IIM Bangalore. His research interests are in ICT for development, open source, e-Government systems, and AI ethics. He has published four books and over 75 peer-reviewed academic articles. Webpage: https://www.iimb.ac.in/user/67/rahul-de