Universität für Bodenkultur Wien (BOKU)
University of Natural Resources and Life Sciences, Vienna

Institut für Rechtswissenschaften
Institute of Law

Report:
Human Control of Machine Intelligence
Professor Joanna J. Bryson (University of Bath)
13 December 2018

Human Control of Machine Intelligence

"The greatest challenges of appropriately regulating Artificial Intelligence (AI) are social rather than technical." With this statement, Professor Joanna J. Bryson, University of Bath (UK), set the stage for her lecture on "Human Control of Machine Intelligence" on 13 December 2018 at the University of Natural Resources and Life Sciences, Vienna. In her lecture, Bryson explained why we need to hold humans accountable if we want to maintain control over AI.

Defining Artificial Intelligence

Because humans identify as intelligent, we often assume that intelligence means being "human-like". According to Bryson, this is not the case when we are talking about AI. Instead of imagining AI as an artificial human being, we should picture AI as an intelligent artefact, deliberately built to facilitate our intentions.

Bryson defined intelligence as the capacity to do the right thing at the right time, translating perception into action. This requires computation, which is the physical transformation of information using time, energy and space. In the course of evolution, humanity's winning strategy has been sharing knowledge and mining each other's prior experience. Recently, this successful concept was utilized by machine learning. Thus, AI's computation potential has been growing significantly, feeding on human knowledge through machine learning. However, this pace of growth will slow down once AI has reached the frontiers of human knowledge.

Maintaining control through transparency and accountability

Before AI becomes intellectually equal to humans, we should find a way to maintain control over it. Bryson argued that only those in control – human beings – should be held accountable.

Human accountability requires AI transparency. The goal of AI transparency is not about complete comprehension of AI, but providing sufficient information in order to hold human beings accountable. To understand human behaviour, for example, it is not necessary to understand how neurons in the brain are connected either. As long as humans can understand the reasons behind the AI's behaviour, there is no need to know every bit of computed data.

In this context, Bryson referred to the example of deaths caused by driverless cars. Accounts of the accident and the car's perception are available within a short period of time, since it is regulated as a part of the automobile industry. Because we can retrace how the car accident occurred, it is easier to solve the question of who is to be held accountable. In contrast, shell companies are an example of badly distributed accountability and control, thereby making it difficult to enforce legal penalties.

Through a transparent design of AI it is possible to say what went wrong if an error occurs. This requires documenting the software engineering process as well as logging the AI's training data and

performance. As an example of AI transparency, Bryson introduced a live visualization interface used for observing AI reasoning. This visualization helps naïve users to better comprehend actions performed by simple robots.

Regulating AI through legal personhood?

Bryson is of the opinion that humans who build and use AI should be held accountable rather than the machines themselves. Therefore, she opposes granting legal personhood to AI. She argues that our justice system works through dissuading people from illegal behaviour through the feeling of dysphoria caused by societal or physical isolation. Applying this system to machines will not work. The idea of punishing AI for any wrongdoing with human penalties like imprisonment relies on the belief that AI is or will be humanlike, which Bryson strongly rejects. If humanity wants to maintain control, responsibility must remain in human hands.

Bryson is convinced that more laws regarding AI are not necessarily needed, but rather tools to enforce them. Every aspect of an artefact follows human design decisions. Therefore, regulators should motivate developers to produce clear and safe code for AI. Bryson concludes that once we start legislating and adjudicating for human accountability, AI transparency will inevitably follow.

Group discussion

After Bryson's lecture, the audience discussed the role of accountability and transparency of AI in more depth with Bryson. As regards the audience's question concerning due diligence and standards of AI, Bryson advocated transnational regulation and named the EU as a particularly promising framework for tackling challenges such as the regulation of AI. Among other aspects, the roles of corporations and government in minimizing AI risks were debated. Using her wide knowledge of legal, technical and social aspects of AI, Bryson led through an inspiring discussion that gave way to many new thoughts.

Nino Gamsjäger, January, 2019