When artificial intelligence is discriminatory

May 16, 2017

Author

Eric Lavallée

Partner, Lawyer Partner, and Trademark Agent

Artificial intelligence has undergone significant developments in the last few years, particularly in respect of what is now known as deep learning.[1] This method is the extension of the neural networks which have been used for a few years for machine learning. Deep learning, as any other form of machine learning, requires that the artificial intelligence system be placed before various situations in order to react to situations which are similar to previous experiences.

In the context of business, artificial intelligence systems are used, among other things, to serve the needs of customers, either directly or by supporting employees interventions. The quality of the services that the business provides is therefore increasingly dependent on the quality of these artificial intelligence systems.

However, one must not make the mistake of assuming that such a computer system will automatically perform its tasks flawlessly and in compliance with the values of the business or its customers.

For instance, researchers at the Carnegie Mellon University recently demonstrated that a system for presenting targeted advertising to Internet users systematically offered less well-paid positions to women than to men.[2]In other words, this system behaved in what could be called a sexist way. Although the researchers could not pinpoint the origin of the problem, they were of the view that it was probably a case of loss of control by the advertising placement services supplier over its automated system and they noted the inherent risks of large-scale artificial intelligence systems.

Various artificial intelligence systems have had similar failures in the past, demonstrating racist behaviour, even to the point of forcing an operator to suspend access to its system.[3]

In this respect, the European Union passed in April 2016 a regulation pertaining to the processing of personal information which, except in some specific cases, prohibits automated decisions based on some personal data, including the “racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation […]”.[4] Some researchers wonder about the application of this regulation, particularly as discrimination appears in an incidental manner, without the operator of the artificial intelligence system intending it.[5]

In Québec, it is reasonable to believe that a business which would use an artificial intelligence system that would act in a discriminatory manner within the meaning of the Charter of Human Rights and Freedoms would be exposed to legal action even in the absence of a specific regulation such as that of the European Union. Indeed, the person responsible for an item of property such as an artificial intelligence system could incur liability in respect of the harm or damage caused by the autonomous action of such item of property. Furthermore, the failure to having put in place reasonable measures to avoid discrimination would most probably be taken into account in the legal analysis of such a situation.

Accordingly, special vigilance is required when the operation of an artificial intelligence system relies on data already accumulated within the business, data from third parties (particularly what is often referred to as big data), or when the data will be fed to the artificial intelligence system by employees of the business or its users during the course of a “learning” period. All these data sources, which incidentally are subject to obligations under privacy laws, may be biased at various degrees.

The effects of biased sampling are neither new nor are they restricted to the respect of human rights. It is a phenomenon which is well-known by statisticians. During the WW II, the U.S. Navy asked a mathematician named Abraham Wald to provide them with statistics on the parts of bomber planes which had been most hit for the purpose of determining what areas of these planes should be reinforced. Wald demonstrated that the data on the planes returning from missions was biased, as it did not take into account the planes that were taken down during these missions. The areas damaged on the returning planes did not need to be reinforced, rather the places which were not hit were the one that had to be.

In the context of the operation of a business, an artificial intelligence system to which biased data is fed may thus make erroneous decisions – with disastrous consequences for the business on a human, economic and operation point of view.

For instance, if an artificial intelligence system undergoes learning sessions conducted by employees of the business, their behaviour will undoubtedly be reflected in the system’s own subsequent behaviour. This may be apparent in the judgments made by the artificial intelligence system in respect of customer requests, but also directly in its capacity to adequately solve the technical problems submitted to it. Therefore, there is the risk of perpetuating the problematic behaviour of some employees.

Researchers of the Machine Intelligence Research Institute have proposed various approaches to minimize the risks and make the machine learning of artificial intelligence systems consistent with its operator’s interests.[6] According to these researchers, it would certainly be appropriate to adopt a prudent approach as to the objectives imposed on such systems in order to avoid them providing extreme or undesirable solutions. Moreover, it would be important to establish informed supervision procedures, through which the operator may ascertain that the artificial intelligence system performs, as a whole, in a manner consistent with expectations.

From the foregoing, it must be noted that a business wishing to integrate an artificial intelligence

system in its operations must take very seriously the implementation phase, during which the system will “learn” what is expected of it. It will be important to have in-depth discussions with the supplier on the operation and performance of his technology and to express as clearly as possible in a contract the expectations of the business as to the system to be implemented. The implementation of the artificial intelligence system in the business must be carefully planned and such implementation must be assigned to trustworthy employees and consultants who possess a high level of competence with respect to the relevant tasks.

As to the supplier of the artificial intelligence system, it must be ensured that the data provided to him is not biased, inaccurate or otherwise defective, in such a way that the objectives set out in the contract as to the expected performance of the system may reasonably be reached, thus minimizing the risk of litigation arising from discriminatory or otherwise objectionable behaviour of the artificial intelligence system. Not only such litigation can be expensive, it could also harm the reputation of both the supplier and its customer.

---

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
2. Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *In Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 598-617). IEEE; Datta, A., Tschantz, M. C., & Datta, A. (2015). Also see: Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92-112.
3. Reese, H. (2016). Top 10 AI failures of 2016. The case of *Tay*, Microsoft’s system, has been much discussed in the media.
4. Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
5. Goodman, B., & Flaxman, S. (2016, June). EU regulations on algorithmic decision-making and a “right to explanation”. *In ICML Workshop on Human Interpretability in Machine Learning* (WHI 2016).
6. Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). *Alignment for advanced machine learning systems* . Technical Report 20161, MIRI.