What if machines made fairer decisions than humans?

Automated decision-making by systems that use machine learning to dynamically improve performance are still seen as lacking the 'human perspective' and flexibility to adapt to the particular nuances of specific cases. But perhaps, as they lack the 'cunning' to hide their biases, automated systems actually make fairer decisions than do humans, when these decisions are based on data that have been properly curated.

Machine learning systems can perform tasks for which they were not initially developed. While it is usually possible – and very effective – to develop specific algorithms to solve well-defined problems, this is not always the case when confronting more complex situations. In these instances, it may be more efficient to explore ways in which the machine can develop or adjust its own decision-making algorithms, rather than for human programmers to try and specify every step, taking all possible nuances into account.

© pdusit / Adobe Stock

Machine learning (ML) is a form of artificial intelligence (AI) in which computers develop their own decision-making processes for situations that cannot be directly and satisfactorily addressed by available algorithms. The process is adjusted through the exploration of existing data on previous similar situations that include the solutions found at the time. The system trains itself to come to a decision-making process whose solutions would match those of the training examples, when confronted with the corresponding initial data. The assumption is that new problems arising in similar situations can interpolate between those examples and, therefore, the appropriate solutions will follow similar lines.

The task of programmers is no longer to identify the specific steps constituting the right algorithm to solve the problem, but to find the right data or set of examples that will lead the ML system to adjust its decision-making process properly. The risks of this method are obvious, as it cannot be guaranteed that the resulting system will extrapolate situations in a meaningful way when the problem deviates significantly from the original learning data. The advantages are also evident, as this method facilitates solutions for very difficult problems in a dynamic, self-adjusting and autonomous way.

Potential impacts and developments

Machine learning systems do not rely on a direct expression of knowledge, but on implicit information emanating from a wealth of data. The broader and more balanced the dataset is, the better the chances will be of obtaining a valid result; but there is no *a priori* way of knowing whether the available data will suffice to collect all aspects of the problem at hand. The outputs of systems based on AI can be biased owing to imbalances in the training data, or if the data source is itself biased with respect to ethnicity, sex or other factors. A typical example is the poor results some facial recognition systems present when identifying black women, because not enough images of that specific population were used in the learning process. This leads to biases related to the sampling of data and results in pernicious decisions and discrimination. Although these biases are bad enough, they are not the only ones possible; biases can already be present in the data that reflect previous decisions, which are not guaranteed to be correct. This potential discrimination against minorities and other population groups leads to major ethical concerns. Machine learning systems therefore act as a mirror of society and replicate previous biases that become assimilated as a result.

A further problem relates to complicated accountability in AI, because so many actors and different applications are involved. When AI misbehaves, or the output is incorrect, who is responsible for the error?

Author: Andrés García Higuera, Scientific Foresight Unit (STOA)

PE 729.501 – March 2022

The user that benefits from the system, its owner, the manufacturer or the developers? The situation is currently unclear and that is why transparency and traceability are so important in AI – to provide complete and continuous information on how the AI tool is designed, developed, validated and used in day-to-day practice. 'Blaming the machine' has become the new way to designate a scapegoat, but someone has to be clearly accountable if we want actions to be taken towards correcting malfunction and possible biases. This leads to a problem of acceptance and trust. Both the party affected by the decision and the one that will be accountable for it need to rely on the system and accept its outputs. Even if the performance of an AI system is high, reliable, secure and unbiased, it may still be rejected, because the parties affected do not understand or trust the technology. In this respect, improving education on AI, as well as involving different stakeholders throughout the whole development process might increase AI acceptability and applicability.

Finally, there are privacy and security concerns, both in normal circumstances and in case of cyberattacks that can affect results or compromise data protection. It is therefore important to build more robust and reliable systems, as well as to increase the layers of protection in AI tools.

Anticipatory policy-making

Researchers are currently working on solutions to detect and compensate for biases in the data used for training ML systems and to obtain AI tools capable of ensuring a fair and safe use independently of sex, gender, age or ethnicity. There is always a trade-off between limiting access to some information on grounds of confidentiality to mitigate bias and reducing the accuracy of the AI mechanism. Additionally, data can be anonymised and access to it can be allowed solely on grounds of legitimate interest. However, both these solutions present important safety risks and it is not always clear what can be defined as legitimate interest. Furthermore, the different availability of data depending on population can also affect the process.

Having access to all the information seems always to be the best way to ensure a good result. Although AI systems may appear to be black boxes, it is possible to introduce mitigating measures, such as the tracking and explainability of the decisions taken using methods like SHAP or LIME. These methods allow checks on the reasoning followed in a specific decision-making process, by highlighting the conditions and data used and their effect on making a final choice. The user or supervisor can thus decide whether the result is sufficiently justified depending on the context at a more personal level. This leads to the question of who that supervisor should be, as well as the need for auditing depending on the level of risk for different applications.

Artificial intelligence has become a sector with huge economic potential and Europe cannot lag behind on innovation as a result of over-regulation. Regulatory sandboxes set up temporary reprieves from regulation to allow technology and the related legislation to evolve together. Rather than increasing regulation, it is crucial to ensure that existing rules, such as the EU's General Data Protection Regulation (GDPR), cover all new aspects that may appear as the technology evolves. European legislation such as the proposed AI act (together with the data act proposal and the data governance act) may apply not only to algorithms but also to datasets, thereby enforcing the explainability of decisions obtained through systems based on ML.

The idea of setting up AI ethics committees to assess and provide certification for the systems or datasets used in ML is also proposed by organisations such as International Organization for Standardization (ISO) or European Committee for Electrotechnical Standardization (CEN). The Organisation for Economic Co-operation and Development (OECD) follows similar lines in its recommendations on AI. While setting up standards and certification procedures seems a good way to progress, it may also lead to a false impression of safety, as the ML systems and the datasets they use are dynamic and continue to learn from new data. A dynamic follow-up process would therefore also be required to guarantee that rules are respected following the FAIR principles of data management and stewardship (FAIR: Findability, Accessibility, Interoperability and Reusability). The European Parliament's Special Committee on Artificial Intelligence in a Digital Age (AIDA) presented a Working Paper on AI and Bias last November, paying special attention to data quality. It refers to the need to avoid 'training data that promotes discriminatory behaviour or results in underrepresentation of certain groups, and keeping a close eye on how feedback loops may promote bias'.