Algorithms also make mistakes (in: *Trouw*, 26 June 2018, page 27)

Self-learning algorithms do not have the wisdom in lease, argues Marc Steen. Knowledge of underlying processes is crucial to understanding where things go wrong.

Thanks to the debacle with Cambridge Analytica and Facebook, we have become more aware of the importance of online privacy and data protection. To ensure this, the General Data Protection Regulation came into force in Europe in May. But there is a much bigger problem: the problem of self-learning algorithms that draw all sorts of conclusions based on stacks of data. Conclusions that we believe indiscriminately, without explanation. Professor of psychology Mark van Vugt calls for a debate on this in his scientific section (Trouw, 2 June). That debate is necessary and useful. Sometimes you have to dig a spade deeper.

Van Vugt mentions an algorithm that would distinguish criminals from non-criminals on the basis of photos of faces. Two professors from the University of Washington looked at this critically and explained that a bias had crept into the algorithm through the photos with which the algorithm was trained. The algorithm is trained with two sets of photos: police photographs of convicted criminals, who look sullen, and random photographs of professional websites, on which the men smile.

The algorithm has thus learned to distinguish between looking and smiling. Not to recognize criminals. For example, there are numerous examples of algorithms that draw conclusions. In case law, in finding a job, in obtaining a loan or mortgage, in choosing medical treatments. And that without explanation about the underlying process. That way we cannot criticize that process properly. That should be better.

Man and machine

To improve our ability to ask critical questions, we can use three concepts: agency, transparency and accountability of algorithms. Agency (the ability to act autonomously) is about the distribution of power between man and machine. This is a sliding scale. On one side is the man who uses the machine as a tool. It is then crucial that he understands how the machine works. On the other side of the scale the machine is central and man has become a gear. He reads what he has to do on his screen and does not understand much of it. “Computer says no”. It seems clear to me that we do not want to become gears.

Understand and explain

Transparency is about the extent to which an algorithm is transparent. Can you understand and explain it? But companies or governments often do not want to or cannot tell how their algorithms work. This may be due to intellectual property or secrecy, but it may also be because they themselves do not understand how their (self-learning) algorithm works exactly. Accountability is about being accountable. For example about what is fair or fair. Algorithms use data from society. There is discrimination and that is how (inadvertent) discrimination ends up in the algorithms. There are examples of discrimination in the judiciary in the US. Or about the degree of accuracy. Algorithms do work with large numbers, but make mistakes on an individual level. Like when a policeman upheld rapper Typhoon unjustly (he turned out not to be criminal). In such cases you also need transparency.

I hope that we, as citizens and as consumers, can ask more critical questions about algorithms, and thus force governments and companies to use algorithms that fit into a just society.

Marc Steen, senior researcher at TNO

Ook algoritmen maken fouten (in: *Trouw*, 23 juni 2018, pagina 27)

Zelflerende algoritmen hebben niet de wijsheid in pacht, betoogt Marc Steen. Kennis van onderliggende processen is cruciaal om te begrijpen waar het misgaat.

Dankzij het debacle met Cambridge Analytica en Facebook zijn we ons meer bewust geworden van het belang van online privacy en dataprotectie. Om die te waarborgen, is in mei de AVG (Algemene Verordening Gegevensbescherming) van kracht geworden in Europa. Maar er is een veel groter probleem: het probleem van zelflerende algoritmen die op basis van stapels data allerlei conclusies trekken. Conclusies die we klakkeloos geloven, zonder uitleg. Hoogleraar psychologie Mark van Vugt roept in zijn wetenschapsrubriek op tot een debat daarover (Trouw, 2 juni). Dat debat is nodig en nuttig. Soms moet je een spade dieper graven.

Van Vugt noemt een algoritme dat op basis van foto's van gezichten criminelen van niet-criminelen zou onderscheiden. Twee hoogleraren van de Universiteit van Washington keken hier kritisch naar en leggen uit dat er een bias (vertekening) in het algoritme is geslopen door de foto's waarmee het algoritme is getraind. Het algoritme is getraind met twee sets foto's: politiefoto's van veroordeelde criminelen, die nors kijken, en willekeurige foto's van professionele websites, waarop de mannen glimlachen.

Het algoritme heeft dus geleerd om onderscheid te maken tussen nors kijken en glimlachen. Niet om criminelen te herkennen. Zo zijn er talloze voorbeelden van algoritmen die conclusies trekken. In de rechtspraak, in het vinden van werk, in het verkrijgen van een lening of hypotheek, in het kiezen van medische behandelingen. En dat zonder uitleg over het onderliggende proces. Zo kunnen we dat proces niet goed bekritiseren. Dat moet beter.

Mens en machine

Om beter te worden in het stellen van kritische vragen kunnen we drie begrippen gebruiken: agency, transparantie en accountability van algoritmen. Agency (het vermogen om autonoom te handelen) gaat over de verdeling van macht tussen mens en machine. Dit is een glijdende schaal. Aan de ene kant staat de mens die de machine gebruikt als gereedschap. Het is dan cruciaal dat hij begrijpt hoe de machine werkt. Aan de andere kant van de schaal staat de machine centraal en is de mens verworden tot een tandwiel. Hij leest op zijn scherm wat hij moet doen en begrijpt er weinig van. Computer says no. Het lijkt mij duidelijk dat we geen tandwielen willen worden.

Begrijpen en uitleggen

Transparantie gaat over de mate waarin een algoritme inzichtelijk is. Kun je het begrijpen en uitleggen? Maar bedrijven of overheden willen of kunnen vaak niet vertellen hoe hun algoritmen werken. Dat kan zijn vanwege intellectueel eigendom of geheimhouding, maar het kan ook zijn omdat ze zelf evenmin begrijpen hoe hun (zelflerende) algoritme precies werkt. Accountability gaat over het kunnen afleggen van verantwoording. Bijvoorbeeld over wat rechtvaardig of eerlijk is. Algoritmen gebruiken data uit de maatschappij. Daar komt nu eenmaal discriminatie voor en zo komt er (onbedoeld) discriminatie in de algoritmen terecht. Er zijn voorbeelden van discriminatie in de rechtspraak in de VS. Of over de mate van accuratesse. Algoritmen werken wel met grote getallen, maar maken fouten op individueel niveau. Zoals toen een politieagent rapper Typhoon onterecht staande hield (hij bleek niet crimineel). In zulke gevallen heb je ook transparantie nodig.

Ik hoop dat we, als burgers en als consumenten, meer kritische vragen kunnen gaan stellen over algoritmen, en daarmee overheden en bedrijven dwingen om algoritmen te gebruiken die passen in een rechtvaardige samenleving.

Marc Steen, senior onderzoeker bij TNO.