We need virtues for ethical algorithms (translated from a Dutch article in *iBestuur*, februari 2019; https://ibestuur.nl/podium/we-hebben-deugden-nodig-voor-ethische-algoritmes)

Algorithms and artificial intelligence (AI) are increasingly used in government and public administration. For example, there was the "Unsolicited advice on the effects of digitization on constitutional relationships" of the Council of State, on automated decision-making, self-learning systems and chain decision-making. An interesting topic. Controversial too. In the context of governance, you want to be able to intervene in automatic decision-making and trace back decisions.

All kinds of technical, economic, legal and ethical issues are involved. We can summarize the technical and economic issues as: more and more can be done, and it can always be cheaper and easier. Marlies van Eck investigated the legal issues of "automated chain decisions" ("Citizens insufficiently legally protected"). And Marrietje Schaake argued for better legislation ("Artificial intelligence belongs in the law"). All very useful and desperately needed. Regarding ethical issues, however, there is bad news: there are no "ready-made methods" to resolve ethical dilemmas ("How do you keep the balance between AI, big data and ethics?"). But fortunately there is also good news: professionals can learn to deal ethically with algorithms and AI by cultivating virtues such as self-control, modesty, justice, courage and perspective.

Virtue ethics

I advocate the use of virtue ethics. This ethical tradition is aimed at creating a just society in which people can flourish and cultivate virtues that contribute to other people's and their own flourishing. Cultivation refers to the development of feelings, thoughts and actions in the direction of justice and of flourishing. It refers to o bringing feelings, thinking and acting into alignment. You feel that something is unjust, you figure out how you can rectify this, and you take action.

Virtue ethics is also a way to promote the 'human dimension' ("The computer can also be wrong"). It is not about rules, but about improvising; always from a specific context. Take "courage", a classic virtue. If I am an old man, not very agile, and I see a fight in the street, then courage is: bring myself to safety and call 112. I would act recklessly if I intervened. But if I am a strong guy and trained in martial arts, for me courage means: quietly intervening and asking a bystander to call 112. Standing aside would be cowardly. It's always about finding the appropriate middle (or mean).

In her book "Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting," Shannon Vallor, professor at Santa Clara University in Silicon Valley, proposes to use virtue ethics in developing and applying new technologies. She explains that the other known ethical traditions (consequentialist ethics and duty ethics) do not work for new technologies. For consequentialist ethics you would need to know exactly all the pros and cons and weigh them against each other. For duty ethics you would need to find universal rules that apply in all circumstances. But that is not possible with new technologies, precisely because they have not yet crystallized; you do not know the pros and cons in detail and you cannot anticipate all circumstances.

Let us take a brief look at some relevant "techno-moral" virtues (from Technology and the Virtues).

1. Self-control. People who work on algorithms need to cultivate self-control. The temptation can be great to collect all kinds of data from all kinds of organizations, to link them to each other, and to use all kinds of analyses. However, such gluttony can turn into a mess and illegal practices. Self-control can help to strive for minimal use of data and understandable algorithms. For example an algorithm that only uses data from its own organization (not from other organizations) and an algorithm such as "if A is larger than X, than B" (no algorithms that you cannot understand and thus cannot explain).

2. Modesty. Another virtue that can help is modesty. This is about knowing what we do not know; about the shortcomings of algorithms. You know, e.g., that algorithms can produce false positives and false negatives, and you take that into account. You can cultivate modesty by visualizing levels of (un)certainty in the algorithm's outcomes, and by enabling people to check and correct errors. You take the so-called 'truth' of an algorithm with a grain of salt.

3. Justice. We strive to have a fair distribution of advantages and disadvantages, and to safeguard fundamental rights, such as dignity and equality. Unintentional injustices can creep into an algorithm just like that. A notorious example played in 2015: Google then launched image recognition based on machine learning, and labelled two dark teenagers as "gorillas." This was because only photos of light-coloured people were used in the training data. They could have prevented this by making the training data a better reflection of the population.

4. Courage. The virtue of courage refers to a balancing between hope and fear, about finding the right balance between blind faith in technology and unfounded hope on the one hand, and indiscriminate rejection of technology and unfounded fear on the other. Courage deals with guiding the usage of technology, not stopping technology. People can cultivate the virtue of courage by carefully trying things out, e.g., in controlled experiments, and critically monitoring the consequences of this experiment, both positive and negative, and by continuously adjusting things.

5. Perspective. The virtue of perspective is also needed, and in particular the combination of different perspectives. One moment you zoom-in on the level of the algorithm, e.g., for predictive policing, and try to solve biases in the algorithm. At another moment you zoom-out to the level of society, and you look at inequal opportunities for different groups in society and their effects on different levels of crime. You can also change perspective from different fields: how people look at this algorithm from data science, from administrative law, and from economics (cost-benefits).

Professional ethics

Virtue ethics is pre-eminently a *professional* ethic: an ethics that offers inspiration and direction and handles to professionals and the ways they feel, think and act. Sometimes people ask me: *How do you operationalize virtue ethics?* You cannot capture virtue ethics in numbers that you can add, or in rules that you can write down and that apply universally. You can, however, effectively put virtue ethics into practice by actually working with it.

Get started practically

Suppose you are working on the design or implementation or deployment of algorithms and AI in government and public administration. Then you can start with a short self-examination: Which *of these virtues do I need?* Self-control? Modesty? Justice? Courage? Perspective? Which of those virtues do I want to cultivate in order to do my work better? Then you can find ways to cultivate these virtues. You can say "no" to a new feature (self-control). You can give more attention to the shortcomings of algorithms (modesty). You can ring the bell if you suspect that the application of an algorithm will undermine justice (justice). You can come up with an experiment and set it up. You can switch between zooming in on the algorithm and zooming out to society (perspective).

And if that delivers good results, then you do it again. This is how you cultivate virtues. Another way to cultivate virtues is to learn from others. That is why it is always necessary to celebrate successes and to share learning experiences.

Marc Steen works as a senior researcher at TNO

We hebben deugden nodig voor ethische algoritmes, iBestuur, februari 2019

https://ibestuur.nl/podium/we-hebben-deugden-nodig-voor-ethische-algoritmes

Het gaat steeds vaker over algoritmes en artificiële intelligentie (AI) in overheid en openbaar bestuur. Zo was daar het 'Ongevraagd advies over de effecten van de digitalisering voor de rechtsstatelijke verhoudingen' van de Raad van State, over geautomatiseerde besluitvorming, zelflerende systemen en ketenbesluitvorming. Een mooi onderwerp. Controversieel ook. In de context van bestuur wil je immers ook in kunnen grijpen in automatische besluitvorming en beslissingen kunnen traceren.

Er spelen allerlei technische, economische, juridische en ethische issues mee. De technische en economische issues kunnen we samenvatten als: er kan steeds meer, en het kan steeds goedkoper en makkelijker. Marlies van Eck onderzocht de juridische issues van 'geautomatiseerde ketenbesluiten' ('Burger onvoldoende juridisch beschermd'). En Marrietje Schaake pleitte voor betere wetgeving ('Kunstmatige intelligentie hoort in de wet'). Allemaal zeer nuttig en hard nodig. Wat betreft de ethische issues echter, is er slecht nieuws: er zijn geen 'kant-en-klare methoden' om ethische dilemma's op te lossen ('Hoe houd je de balans tussen AI, big data en ethiek?'). Maar gelukkig is er ook goed nieuws: professionals kunnen leren om ethisch om te gaan met algoritmes en AI door het cultiveren van deugden zoals zelfbeheersing, bescheidenheid, rechtvaardigheid, moed en perspectief.

Deugdethiek

Ik pleit voor de inzet van deugdethiek. Deze ethische traditie is gericht op het creëren van een rechtvaardige samenleving waarin mensen kunnen floreren, en die mensen aanzet om deugden te cultiveren die daaraan bijdragen. Het cultiveren verwijst naar het ontwikkelen van gevoelens, gedachten en handelingen in de richting van rechtvaardigheid en van floreren. Het gaat om het op één lijn brengen van voelen, denken en handelen. Je voelt dat iets onrechtvaardig is, je zoekt uit hoe je dit kunt rechtzetten, en je onderneemt actie.

Deugdethiek is ook een manier om de menselijke maat terug te brengen ('De computer kan het ook fout hebben'). Het gaat namelijk niet over regeltjes, maar over improviseren; steeds vanuit een specifieke context. Neem 'moed', een klassieke deugd. Als ik een oude man ben, niet zo goed ter been, en ik zie een vechtpartij op straat, dan is moed voor mij: mezelf in veiligheid brengen en 112 bellen; ik zou overmoedig handelen als ik tussenbeide zou komen. Maar als ik een sterke kerel ben en getraind in vechtsport, dan betekent moed voor mij: rustig ingrijpen en een omstander vragen om 112 te bellen; aan de kant blijven staan zou laf zijn. Het gaat steeds om het vinden van het juiste midden.

In haar boek 'Technology and the virtues: A philosophical guide to a future worth wanting', stelt Shannon Vallor, hoogleraar aan Santa Clara University in Silicon Valley, voor om deugdethiek te gebruiken bij het ontwikkelen en toepassen van nieuwe technologie. Ze legt uit dat de andere bekende ethische tradities (gevolgenethiek en plichtethiek) niet werken voor nieuwe technologie. Voor gevolgenethiek moet je precies alle voor- en nadelen kennen en tegen elkaar afwegen. Voor plicht-ethiek moet je zoeken naar universele regels die in alle omstandigheden zouden moeten gelden. Maar dat is niet mogelijk bij nieuwe technologieën, precies omdat die nog niet uitgekristalliseerd zijn; je kent de voor- en nadelen niet in detail en je kunt niet anticiperen op alle omstandigheden.

Laten we enkele relevante 'technomorele' deugden kort onder de loep nemen (uit Technology and the Virtues).

1. Zelfbeheersing

Mensen die aan algoritmes werken kunnen zelfbeheersing cultiveren. De verleiding kan groot zijn om allerlei data van allerlei instanties te verzamelen, aan elkaar te koppelen, en daar allerlei analyses op los te laten. Zo'n gulzigheid kan echter uitdraaien op een puinhoop en op illegale praktijken. Zelfbeheersing kan helpen om te streven naar minimaal gebruik van data en naar begrijpelijke algoritmes. Bijvoorbeeld een algoritme dat alleen data vanuit de eigen organisatie gebruikt (niet van andere instanties) en een algoritme zoals 'als groter dan X, dan A' (geen algoritmes die je niet kunt begrijpen en dus niet kunt uitleggen).

2. Bescheidenheid

Een andere deugd die kan helpen is bescheidenheid. Dat gaat over het weten wat we niet weten; over de tekortkomingen van algoritmes. Je weet bijvoorbeeld dat er altijd false positives en false negatives uit een algoritme kunnen komen, en je houdt daar rekening mee. Je kunt bescheidenheid cultiveren door een proces zo te ontwerpen dat mensen in control zijn voor het interpreteren van twijfelgevallen, of door het algoritme uitleg te laten geven bij de uitkomsten, bijvoorbeeld door het visualiseren van onzekerheid. Je neemt de zogenaamde waarheid van een algoritme met een korreltje zout.

3. Rechtvaardigheid

Rechtvaardigheid gaat over het streven naar een eerlijke verdeling van voordelen en nadelen, en over het borgen van rechten, zoals waardigheid. Er kunnen zomaar, onbedoeld onrechtvaardigheden binnensluipen in een algoritme. Een berucht voorbeeld speelde in 2015: Google lanceerde toen image recognition gebaseerd op machine learning, en labelde twee donkere tieners als 'gorilla's'. Dat kwam doordat er alleen foto's van lichtgekleurde mensen waren gebruikt in de trainingsdata. Dat hadden ze kunnen voorkomen door de trainingsdata een betere afspiegeling te maken van de bevolking.

4. Moed

Dit betoog gaat over het sturen van technologie, niet het tegenhouden van technologie. De deugd van moed gaat over het balanceren tussen hoop en angst, over het vinden van het juiste midden tussen blind vertrouwen in technologie en ongegronde hoop aan de ene kant, en klakkeloos afwijzen van technologie en ongegronde angst aan de andere kant. Concreet kunnen mensen de deugd van moed cultiveren door dingen zorgvuldig uit te proberen, bijvoorbeeld in gecontroleerde experimenten, en kritisch te volgen wat de gevolgen van zo'n experiment zijn, zowel de positieve als de negatieve, en om dat continu bij te sturen.

5. Perspectief

Ook de deugd van perspectief is nodig, en dan met name het combineren van verschillende perspectieven. Het ene moment zoom je in op het algoritme, bijvoorbeeld voor predictive policing, en probeer je biases in het algoritme op te lossen. Het andere moment zoom je uit naar een niveau van de maatschappij, en kijk je naar verschillende kansen voor verschillende groepen in de samenleving en de effecten daarvan op criminaliteit. Ook kun je perspectief wisselen vanuit verschillende vakgebieden: hoe kijken mensen naar dit algoritme vanuit data science, vanuit bestuursrecht, en vanuit kosten-baten.

Professionele ethiek

Deugdethiek is bij uitstek een professionele ethiek: een ethiek die inspiratie en richting en handvatten kan bieden aan het voelen, denken en handelen van professionals. Soms vragen mensen mij: hoe operationaliseer je deugd-ethiek? Je kunt deugdethiek niet vangen in getallen die je kunt

optellen, of in regels die je kunt opschrijven en die altijd gelden. Je kunt echter deugdethiek prima operationaliseren door er concreet mee aan de slag te gaan.

Praktisch aan de slag

Stel dat je bezig bent met het ontwerpen of implementeren of gebruiken van algoritmes en AI in overheid en openbaar bestuur. Dan kun je starten met een kort zelfonderzoek: welke van deze deugden heb ik nodig? Zelfbeheersing? Bescheidenheid? Rechtvaardigheid? Moed? Perspectief? Welke van die deugden wil ik cultiveren om mijn werk beter te kunnen doen? Daarna kun je manieren vinden om één tandje beter te worden in één van die deugden. Je kunt eens 'nee' zeggen tegen een nieuwe feature. Je kunt de tekortkomingen van algoritmes meer aandacht geven. Je kunt aan de bel trekken als je vermoed dat de toepassing van een algoritme rechtvaardigheid onderuit haalt. Je kunt een experiment bedenken en opzetten. Je kunt schakelen tussen inzoomen op het algoritme en uitzoomen naar de maatschappij.

En als dat goede resultaten heeft, dan doe je het nog een keer. Zo cultiveer je deugden. Een andere manier om deugden te cultiveren is van anderen leren. Daarom is het ook altijd nodig om successen te vieren en leerervaringen te delen.

Marc Steen werkt als senior onderzoeker bij TNO