Aristotle in times of AI: Virtue ethics as a guideline for digitization (translated from a Dutch article in *De Ingenieur*, May 2019; https://www.deingenieur.nl/tijdschrift)

In Europe we do not want to use artificial intelligence to sell stuff or to keep an eye on citizens, but to solve social problems. In doing so, we need to resort to virtue ethics, argues TNO researcher Marc Steen.

You find machine learning, algorithms and artificial intelligence (AI) increasingly often in devices, services and processes. This raises all kinds of social and ethical questions: about the addictive effects of social media, about the impact of bots on elections, about the profits of tech companies, about surveillance. In Europe we want to deal with this differently than in the US, where companies use AI to sell stuff, and unlike in China, where the government uses AI to monitor citizens. In Europe we want to use AI to solve social problems. And we also want to uphold values such as freedom, equality, solidarity, privacy and democracy.

To steer AI in the right direction, I advocate the use of virtue ethics. I follow Shannon Vallor, professor at Santa Clara University in Silicon Valley. In her book *'Technology and the virtues'*, she argues that we need virtue ethics for questions about emerging technologies such as AI.

Virtue ethics is an alternative to duty ethics and consequentialist ethics. With duty ethics you look for duties and rules that apply to everyone, always and everywhere. With consequentialist ethics you try to maximize positive effects and minimize negative consequences. But because AI is 'under development', those duties and rules are not yet clear. Moreover, it is difficult to estimate the scope of the consequences. A self-driving car has consequences not only for the driver and passengers, but also for pedestrians and cyclists, and for your family and your social life. For example, what are the effects if you are going to live in the countryside and commute four hours back and four hours back to your office in Brussels—while sleeping behind the wheel?

This raises questions like: *What kind of society do we want? What does the good life look like?* These are the kinds of questions addressed in virtue ethics. Aristotle, one of its founders, put it like this: Virtue ethics presupposes that there are ultimate goals for everything. A beech nut aims to grow into a beech tree. We, as people, aim to develop our human capabilities, e.g., our capabilities for reflection, communication, collaboration or creation—so that we can live meaningful and fulfilling lives. Virtue ethics is about creating a society in which people can develop their capabilities.

Unjust situations

Furthermore, virtue ethics is about cultivating virtues and finding the appropriate middle (or mean) in each situation. Suppose I see a fight on the street. If I am not agile, I act courageously if I stand aside and call 112. Intervention would be reckless. But if I am strong and trained in martial arts, I act courageously if I intervene. Doing nothing would be cowardly. Obviously, we need different virtues than the people in Aristotle's Athens. That is why Vallor describes twelve techno-moral virtues that are important to us. Here are four examples:

Self-control. If you are working on an algorithm, the temptation can be to extract data from all kinds of sources, link them to each other and apply all kinds of analyses. However, such gluttony can turn into a mess and illegal practices. If you want to cultivate self-control, you aim for minimal use of data and for algorithms that you can understand and explain, such as "if A is larger than X, then B".

Courage. This virtue is about finding the right balance between too much faith in technology and too much fear of technology. You can do that, for example, by carefully testing an algorithm, in small steps, in experiments, and critically monitoring and adjusting the algorithm and the processes around it. In addition, you need courage to pull the plug if the advantages do not outweigh the disadvantages.

Justice. Cathy O'Neill's book *'Weapons of math destruction'* is full of examples of how algorithms sustain or even exacerbate all kinds of unjust situations. You can cultivate justice, e.g., by avoiding bias in training data. Google did not do that when the company tagged a photo of two dark-coloured teenagers as gorillas in 2015. The responsible algorithm was trained on photos of light-coloured people.

Modesty. Suppose you are working on an algorithm that detects fraudsters. That algorithm also produces false predictions: sometimes it designates someone as a fraud who is not a fraud (false positive), sometimes it designates someone as a non-fraud who does fraud (false negative). You can cultivate modesty by designing a process that enables people to detect and correct such errors. Or you can visualize the degree of (un)certainty of the algorithm's outcome, so people can better interpret it.

How can you get started with virtue ethics? You can start with the question: what virtues do I need in this project? Then you will need to practice these virtues. You can say "no" to a feature (self-control). Or ring the bell if you see an algorithm producing unjust results (courage). You can also exchange experiences. And you can learn from people who embody one or more virtues in their work; examples are available on the DearEngineer.eu site. Virtue ethics is pre-eminently an ethics for engineers; they want to help build a fair society in which people can flourish.

Dr.ir. Marc Steen works as a senior researcher at TNO, including in the Responsible Value Creation with Big Data program.

Aristoteles in tijden van AI: Deugdethiek als richtsnoer voor digitalisering (De Ingenieur, mei 2019; https://www.deingenieur.nl/tijdschrift)

In Europa willen we kunstmatige intelligentie niet gebruiken om spullen te verkopen of burgers in de gaten te houden, maar om maatschappelijke problemen op te lossen. Daarbij moeten we onze toevlucht nemen tot de deugdethiek, betoogt TNO-onderzoeker Marc Steen.

Je komt machine learning, algoritmes en artificiële intelligentie (AI) steeds vaker tegen, in apparaten, diensten en processen. Dat roept allerlei maatschappelijke en ethische vragen op: over de verslavende werking van social media, over de invloed van bots op verkiezingen, over de winsten van techbedrijven, over surveillance. In Europa willen we daar anders mee omgaan dan in de VS, waar bedrijven AI inzetten om spullen te verkopen, en anders dan in China, waar de overheid AI inzet om burgers in de gaten te houden. Hier willen we AI gebruiken om maatschappelijke problemen op te lossen. En daarbij willen we waarden zoals vrijheid, gelijkwaardigheid, solidariteit, privacy en democratie overeind houden.

Om AI in goede banen te leiden, pleit ik voor de inzet van deugdethiek. Daarin volg ik Shannon Vallor, hoogleraar aan de Santa Clara University in Silicon Valley. In haar boek Technology and the virtues betoogt ze dat we deugdethiek nodig hebben bij vragen over opkomende technologieën zoals AI.

Deugdethiek is een alternatief voor plichtethiek en gevolgenethiek. Bij plichtethiek ga je op zoek naar plichten en regels die voor iedereen, altijd en overal gelden. Bij gevolgenethiek probeer je positieve gevolgen te maximaliseren en negatieve gevolgen te minimaliseren. Maar doordat AI volop in ontwikkeling is, zijn die plichten en regels nog niet helder. Bovendien is het lastig om de reikwijdte van de gevolgen in te schatten. Een zelfrijdende auto heeft niet alleen gevolgen voor de bestuurder en inzittenden, maar ook voor voetgangers en fietsers, en voor je gezin en je sociale leven. Je gaat bijvoorbeeld op het platteland wonen en vier uur heen en vier uur terug forensen naar je kantoor in Brussel. Je kunt immers slapen achter het stuur.

Wat willen we voor maatschappij? Hoe ziet het goede leven eruit? Dat soort vragen stel je vanuit de deugdethiek. Aristoteles, een van de grondleggers, schreef daar het volgende over. Deugdethiek veronderstelt dat er doelen bestaan. Een beukennootje heeft als doel om uit te groeien tot een beukenboom. Wij mensen hebben als doel om onze vermogens te ontwikkelen, zoals reflecteren, communiceren, samenwerken en creëren, zodat we betekenisvol kunnen leven. Daarom gaat deugdethiek over het inrichten van een maatschappij waarbinnen mensen hun vermogens kunnen ontwikkelen.

Onrechtvaardige situaties

Verder gaat deugdethiek over het cultiveren van deugden en over het vinden van het juiste midden in iedere situatie. Stel dat ik een vechtpartij zie op straat. Als ik niet goed ter been ben, handel ik moedig als ik aan de kant blijf staan en 112 bel. Ingrijpen zou overmoedig zijn. Maar als ik sterk ben en een vechtsport beoefen, handel ik moedig als ik ingrijp. Niets doen zou laf zijn. Nu hebben wij andere deugden nodig dan de mensen in Aristoteles' Athene. Daarom beschrijft Vallor twaalf techno-morele deugden die voor ons van belang zijn. Vier voorbeelden.

Ten eerste zelfbeheersing. Als je werkt aan een algoritme, kan de verleiding groot zijn om data uit allerlei bronnen te halen, ze 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. Als je zelfbeheersing wilt

cultiveren, streef je naar minimaal gebruik van data en naar algoritmes die je kunt begrijpen en uitleggen, zoals ‘als A, dan X’.

Ten tweede moed. Deze deugd gaat over het vinden van het juiste midden tussen te veel vertrouwen in technologie en te veel angst voor technologie. Je kunt dat bijvoorbeeld doen door een algoritme zorgvuldig te beproeven, in kleine stappen, in experimenten, en alles kritisch te monitoren en bij te sturen. Daarbij moet je dan ook de moed hebben om de stekker eruit te trekken als de voordelen niet opwegen tegen de nadelen.

Ten derde rechtvaardigheid. Het boek Weapons of math destruction van Cathy O’Neill staat vol voorbeelden van hoe algoritmes allerlei onrechtvaardige situaties in stand houden of zelfs verergeren. Je kunt rechtvaardigheid cultiveren door bias in trainingsdata te vermijden. Dat deed Google niet toen het bedrijf in 2015 een foto van twee donkergekleurde tieners labelde als gorilla’s. Het verantwoordelijke algoritme was getraind op foto’s van lichtgekleurde mensen.

Ten vierde bescheidenheid. Stel dat je werkt aan een algoritme dat fraudeurs opspoort. Dat algoritme produceert ook foute voorspellingen: soms wijst het iemand aan als fraudeur die geen fraudeur is (false positive), soms wijst het iemand aan als niet-fraudeur die wel fraudeert (false negative). Je kunt bescheidenheid cultiveren door een proces in te richten waarmee mensen dit soort fouten kunnen opsporen en corrigeren. Of je kunt de onzekerheid in beeld brengen, bijvoorbeeld door de mate van (on)zekerheid bij een uitkomst af te beelden.

Hoe kun je nu aan de slag met deugdethiek? Je kunt starten met de vraag: welke deugden heb ik nodig in dit project? Daarna ga je die deugden oefenen. Je kunt ‘nee’ zeggen tegen een feature (zelfbeheersing). Of aan de bel trekken als je ziet dat een algoritme onrechtvaardige resultaten produceert (moed). Ook kun je ervaringen uitwisselen. En je kunt leren van mensen die een of meer deugden belichamen in hun werk; op de site DearEngineer.eu staan voorbeelden. Deugdethiek is bij uitstek een ethiek voor ingenieurs; zij willen meehelpen bouwen aan een rechtvaardige samenleving waarin mensen kunnen floreren.

Dr.ir. Marc Steen werkt als senior onderzoeker bij TNO, onder meer in het programma Verantwoorde Waardecreatie met Big Data.