OVER KUNSTMATIGE INTELLIGENTIE BESTAAT DE CONSENSUS DAT ZE WELDRA SLIMMER ZAL ZIJN DAN DE MENS EN ZO ENORME KANSEN ZAL BIEDEN VOOR DE ONTWIKKELING VAN ONZE BESCHAVING. TEGELIJK BESCHOUWEN VELEN HAAR NET OOK ALS EEN BEDREIGING VOOR DIEZELFDE BESCHAVING. IN ELK GEVAL STAAN WE OP EEN KRUISPUNT WAAROP WE EEN ANTWOORD MOETEN ZOEKEN OP EEN AANTAL BELANGRIJKE VRAGEN. WORDEN WE BINNENKORT ALLEMAAL WERKLOOS, WANT VERVANGEN DOOR MACHINES? WIE DRAAGT DE VERANTWOORDELIJKHEID WANNEER ER IETS MISGAAT?

Romantische monsters, spiegels en zwarte dozen: over ethiek en toekomst van kunstmatige intelligentie

Mark Coeckelbergh

De meeste transhumanisten vinden dat de mens een beetje een *mislukkeling* is, een betreurenswaardig foutje van de natuur. Hij is te dom, te weinig geevolueerd. We moeten de mens daarom verbeteren of de rest van de evolutie dan maar overlaten aan de *slimme* machines, die ons in de toekomst sowieso zullen vervangen. Volgens Ray Kurzweil komt er een zogenaamde 'singulariteit': door de accelererende ontwikkeling van technologie staat er een grote verandering op stapel waarna niets meer hetzelfde zal zijn voor ons. Kunstmatige intelligentie is dan een welkome ontwikkeling op weg naar een post-apocalyptische toekomst, de mens voorbij.

Tegelijk luiden diezelfde transhumanisten de alarmbel over kunstmatige intelligentie. Als we hen moeten geloven is de kunstmatige intelligentie niet alleen een enorme kans maar ook een fundamentele bedreiging voor onze beschaving (volgens Elon Musk), of een tikkende (atoom)bom waarmee we aan het spelen zijn (volgens Nick Bostrom). We horen heel wat alarmerende geluiden in de media. Moeten we bang zijn? Of is het daar misschien al te laat voor?

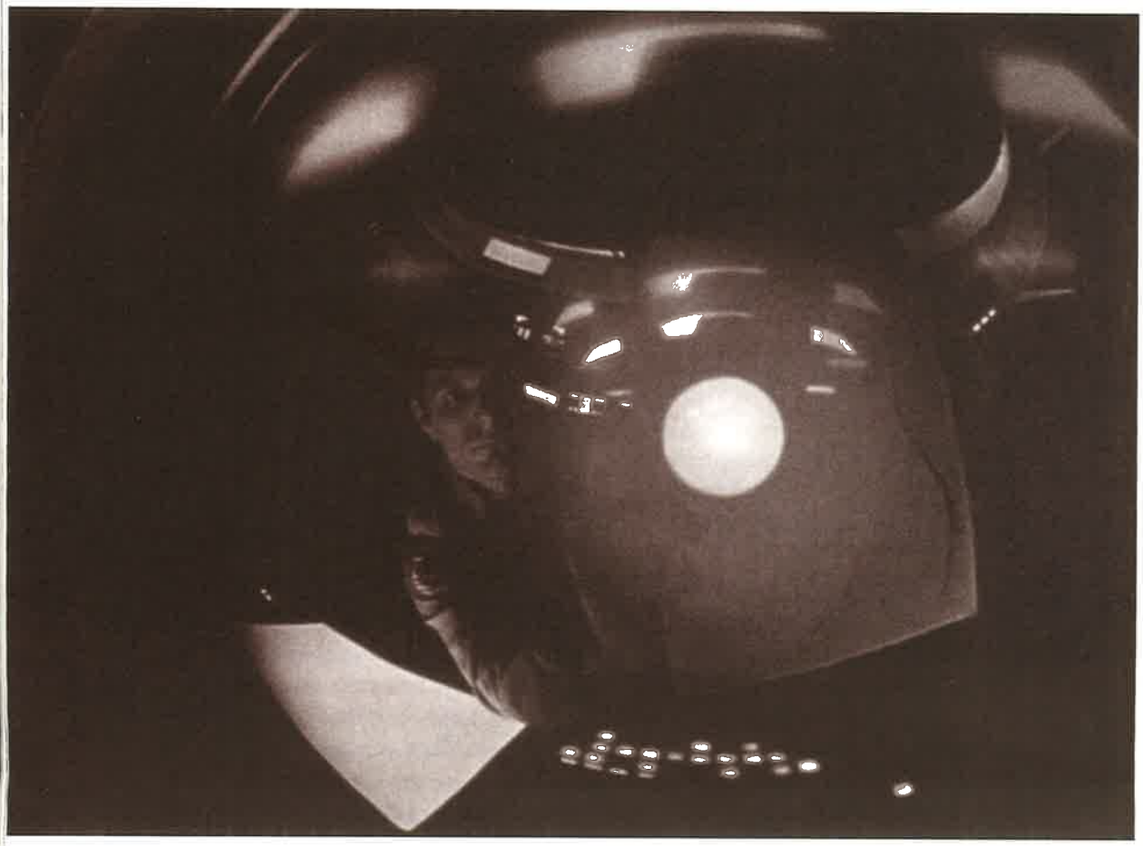
Volgens MIT-professor Max Tegmark, auteur van *Life 3.0 – Being Human in the Age of Artificial Intelligence*, staan we in elk geval op een kruispunt waarop we moeten beslissen over de toekomst van het menselijk leven op aarde. Na een scenario van een aantal slimme computernerds die de wereld gaan domineren met behulp van kunstmatige intelligentie en zo een nieuwe wereldorde zullen vestigen, schetst hij zijn geloof in een toekomst die begint met een intelligentie-explosie en die leidt naar nieuwe, kunstmatige levensvormen voorbij de mens. We komen van een universum van levenloze materie en zijn op weg naar steeds meer intelligentie. Op korte termijn moeten we dat allemaal in goede banen leiden door ons te bekommeren om mogelijke problemen rond hacking, achterlopende juridische systemen, slimme maar dodelijke wapens, en jobs die niet meer zeker zijn. Het moet mogelijk zijn, denkt hij, om kunstmatige intelligentie te ontwikkelen die het leven van mensen beter maakt en om de samenleving zo in te richten dat minder tewerkstelling eerder kansen biedt dan een bedreiging vormt. Maar de race naar de intelligentie-explosie is in elk geval al volop aan de gang. *Fasten your seatbelts*, en maak je klaar om nederig te buigen voor onze kunstmatige 'meerdermensen' die zoveel slimmer en wijzer (of meer sapiens) zullen zijn dan wij, voelende maar achterlijke mensen.

Wat moeten we hiermee? Als filosoof en expert techniekethiek zie ik drie interessante wegen, die allemaal voorbij het alarmisme en wilde toekomstvoorspellingen gaan maar tegelijkertijd toch ook een aantal fenomenen en uitdagingen ernstig nemen. De eerste plaatst het alarmisme in een bredere cultuurhistorische context van denken en verbeelding over techniek, de tweede gaat in op enkele filosofische vragen rond de mens en zijn relatie tot techniek, en de derde verlegt de focus naar de kunstmatige-intelligentietechnologie die nu voorhanden is en ontvouwt daarbij een aantal concrete uitdagingen op korte termijn, ook qua beleid.

Fascinatie en beklemming maken minstens sinds het monster van Frankenstein deel uit van onze houding tegenover techniek

De opwinding en angst die, deels aangezwengeld door de media, het huidige publieke debat beheersen, zijn niet zo verwonderlijk voor wie de geschiedenis van de verbeelding over techniek kent. Fascinatie en beklemming zijn ten laatste sinds het monster van Frankenstein deel geworden van onze houding tegenover techniek. Mary Shelleys roman draait precies om deze twee elementen: het enthousiasme van de (romantische) wetenschapper – vaak een man – die hoopt op nieuw leven en het creëren van kunstmatig leven door de mens, maar tegelijkertijd ook de huivering bij het zien van de gevolgen van de menselijke scheppingen. Het nieuwe wezen blijkt een monster te zijn dat zich ongelukkig en onwillig tegen zijn schepper keert. De schepper van zijn kant kan ook schrikken van het afstotelijke resultaat van zijn creativiteit en daarom de verantwoordelijkheid voor zijn geesteskind willen ontlopen: 'Deze is niet van mij.' Het zijn altijd de anderen die het zover hebben laten komen. Met zo'n houding krijg je natuurlijk problemen. En over vaders en zonen gesproken: in de joods-christelijke traditie is er niet alleen het verhaal van de Golem – ook weer dode materie die leven ingeblazen wordt – maar ook het apocalyptische denken en het geloof in onsterfelijkheid. Allemaal spannend en mysterieus, zo heeft de romanticus het graag.

De nieuwe hersenspinsels van het transhumanisme komen dan ook niet uit de lucht gevallen. Als de natuur het klaarspeelt om dode materie levend te maken en zelfs intelligent, waarom kunnen wij dat dan niet? De romantici van vandaag hebben een nieuwe technologie. Het is niet meer de elektriciteit, die magische kracht die Mary Shelley en veel van haar tijdgenoten zo fascineerde, die het wonder moet verrichten. Het zijn nu de robotica en de kunstmatige intelligentie die daarvoor moeten zorgen. En wie is er niet in de ban van onsterfelijkheid? Het verhaal van de 'singulariteit' klinkt eigenlijk maar al te bekend: het geloof in een hiernamaals, dit keer op aarde weliswaar, of op één van de planeten van ons zonnestelsel zodra die daarvoor gepreppareerd zullen zijn. Wie voldoende geld heeft, kan nu al een plaatsje reserveren in de wachtkamer om



Lens van HAL 9000 in 2001: A Space Odyssey. Foto: Ronald Grant Archive / Mary Evans / Isopix

te worden geüpload of in de diepvriezers van de nieuwe Frankensteins te worden gestopt. Dat uploaden en weer tot leven wekken van het hoofd staat weliswaar nog niet helemaal op punt, maar dat zijn details; de vooruitgang en vooral de voortdurende versnelling van de technologie zorgen er wel voor dat het goedkomt. Intussen dromen we alvast van een nieuwe utopie, een nieuw 'nergens' waarin het helemaal niet erg is als je werkloos bent. Eindelijk tijd om lekker creatief te doen! Geld en zo, dat regelen we wel.

Maar alle gekheid op een stokje: het is niet om te lachen. Dit is een krachtige verbeeldingscocktail die al eeuwen heeft gewerkt en ook invloed heeft op reële ontwikkelingen in de wetenschap en de techniek. Net zoals in de 19de eeuw. Zoals ik eerder al in mijn boek *New Romantic Cyborgs* uit de doeken heb gedaan, hebben cultuur en wetenschap, fictie en techniek, machinebouwers en romantiek altijd al met elkaar geflirt. Denk maar aan de computers en smartphones die ons vandaag vergezellen: die komen niet alleen uit de koker van de computerwetenschapper, maar ook uit die van de hippies. Dat komt omdat het zo'n verleidelijke machines zijn, die drijven op onze romantische verbeelding en onze honger naar een spannend leven voeden, liefst ook na de dood.

Verleidelijke machines als smartphones drijven op onze romantische verbeelding en voeden onze honger naar een spannend leven, liefst ook na de dood

Maar als alles terug te brengen is tot fantasie, zullen we het probleem dan maar gewoon negeren? Ten eerste is het niet *enkel* fantasie. Zoals gezegd heeft onze verbeelding over mens en techniek lange en diepe culturele wortels; het gaat vaak om oeroude verhalen en menselijke dromen waar we serieus over moeten nadenken en die wel degelijk echte, tastbare gevolgen hebben. Daar kom ik zo meteen op terug. Ten tweede werd al duidelijk dat het transhumanisme en de bijhorende verhalen ons doen nadenken over de mens. Dat is mooi meegenomen voor filosofen, want ook de filosofie heeft haar verhalen, denkbeelden, en (denk)experimenten nodig. Zo gebruiken veel filosofen zombies om over bewustzijn na te denken – om maar meteen een relevant thema te noemen voor de discussie over kunstmatige intelligentie. En wat had Descartes kunnen doen zonder zijn machines?

Als we weer naar de geschiedenis van het denken kijken, dan blijkt immers dat we de mens altijd ook al hebben willen begrijpen door hem te vergelijken met wat hij *niet* is. De mens is geen kip, bijvoorbeeld, zoals in de oudheid werd gezegd (dat was Diogenes' antwoord aan Plato toen die de mens als een 'veerloze tweevoeter' omschreef). Hij is ook geen gewone aap. En zeker geen engel, laat staan een god (al denken sommige dichters dat natuurlijk wel in het diepst van hun gedachten). Blijkbaar hebben we andere wezens vooral gebruikt om te zeggen wat we niet zijn. Ik heb dat een 'negatieve antropologie' genoemd, naar het voorbeeld van de negatieve theologie, die ook vond dat men alleen maar kon zeggen wat God niet was. Vandaag is dat niet-menselijke de kunstmatige intelligentie, zij het dan wel in een zo menselijk mogelijke vorm: een robot dus. Wat is de mens? Geen machine. Of, volgens vele transhumanisten maar ook sommige wetenschappers, net wél een machine. Zoals Descartes vergelijken we. Het verschil is dat onze machines nu slimmer zijn dan vroeger, dus de filosofische antropologie wordt wat moeilijker: wat is precies het verschil als die machine ook praat, net zoals wij? Wat is het verschil als die machine ook kunst kan maken? Wat is het verschil als die machine steeds meer op ons lijkt, zoals de robots van Hiroshi Ishiguro? Wat als ze beter Go kan spelen dan mensen? En wat voor soort intelligentie hebben machines eigenlijk, is die wel dezelfde als die van mensen? Kunstmatige intelligentie wordt dus een soort spiegel, een negatieve spiegel van de mens. Door die te gebruiken leren we niet alleen iets over machines, maar ook over onszelf. Dat is boeiend, niet alleen voor filosofen maar ook voor wetenschappers en in feite voor iedereen. Dus toch: lang leve de artificiële intelligentie?

Kunstmatige intelligentie is een soort negatieve spiegel van de mens die ons niet alleen iets leert over machines, maar ook over onszelf

De ideeën van de transhumanisten hebben natuurlijk niet enkel wortels in de cultuurgeschiedenis. Ze zijn ook niet enkel voer voor filosofen. Er zijn ook concrete technologische ontwikkelingen, die terecht leiden tot discussies over de toekomst van de techniek en onze samenleving, en waarop het transhumanisme zich heeft geënt. Voorbij het alarmisme en voorbij het grote verhaal van het transhumanisme vinden we wel degelijk concrete ethische en maatschappelijke uitdagingen rond kunstmatige intelli-

gentie. Daar moet het publieke debat, politiek gezien dan, vooral over gaan. Laat me enkele voorbeelden geven van dergelijke vragen en problemen. Ik zal me daarbij weliswaar moeten beperken, want het is een soort van doos van Pandora waaruit steeds weer nieuwe kwesties oprijzen. Achtereenvolgens bespreek ik de bezorgdheid dat de mens vervangen wordt en geen werk meer zal hebben, vragen rond verantwoordelijkheid, en het probleem van ondoorzichtigheid en de kwestie rond vooroordelen in en door kunstmatige intelligentie.

Er zijn behoorlijk veel geluiden over mensen die vervangen worden door machines, vooral op het werk. Volgens een intussen bekend rapport uit Oxford (*The Future of Employment*) zou dat niet enkel gaan om jobs van arbeiders, zoals in het verleden, maar ook zogenaamde 'white collar jobs', die tot hiertoe veilig leken. Dit is geen exacte wetenschap en de percentages in rapporten lopen nogal uiteen. Maar het idee is dat de slimme machines weldra ook allerlei dienstverlenende taken zullen aankunnen. Denk bijvoorbeeld aan jobs in winkels of in de administratie, maar ook journalistieke banen. Deels is dat een vorm van aandachttrekkerij, maar toch is er met de nieuwe technologie wel degelijk een nieuwe golf van automatisering mogelijk die wellicht vergelijkbaar is met vroegere golven in de industriële revoluties. Eén antwoord daarop is dat machines niet enkel kunnen vervangen, maar ook naast en met mensen ingezet kunnen worden. Het probleem is dan eerder hoe mensen en machines op een veilige en goede manier kunnen samenwerken. Een ander antwoord is dat er wel degelijk banenverlies komt, en dat we ons als samenleving daar toch op moeten voorbereiden. Hoe kunnen we dit doen? Met een basisinkomen? Een andere manier van herverdelen? Dit is vooralsnog zeer onduidelijk. Wat wel zeker is, is dat dit hele debat veel zegt over wat voor jobs mensen nu doen: moeten we ons vragen stellen bij jobs die zo gemakkelijk geautomatiseerd kunnen worden? En hoe voelen mensen zich daarbij, als hun job op de tocht staat?

Een ander zeer belangrijk probleem is verantwoordelijkheid. Slimmere machines betekenen ook vaak autonomere machines, die ook zonder mensen kunnen opereren. Willen we dat wel? En zo ja, wat doen we als er iets misgaat? Wie is er dan verantwoordelijk? Het lijkt niet mogelijk om de machine verantwoordelijk te houden. Maar wie van de vele handen (zoals dat dan heet) die bij de technologie betrokken zijn is moreel verantwoordelijk? Wie moet juridisch aansprakelijk gesteld worden? Het lijkt een goed idee om de verantwoordelijkheid te delen, maar het is niet duidelijk hoe precies. Wat doen we bijvoorbeeld bij zelfrijdende auto's? Is de bestuurder verantwoordelijk – terwijl die eigenlijk geen bestuurder maar een 'operator' is? Is de eigenaar of het bedrijf (bijvoorbeeld een taxibedrijf) dat de auto inzet verantwoordelijk? De autobouwer? Wat doet de verzekering? Deze vragen moeten dringend beantwoord worden als we meegaan in deze ontwikkeling. Gelijkaardige problemen stellen zich in de financiële wereld, waar slimme algoritmes transacties doen, en in de juridische wereld, waar de taak van de rechter wellicht (deels) zal worden overgenomen.

Tenslotte zijn er nog problemen rond vooroordelen en ondoorzichtigheid. Met *machine learning*, de belangrijkste drijfveer van de huidige technologische ontwikkelingen, is het mogelijk om patronen te vinden in grote hoeveelheden data. Maar daardoor sluipen vaak vooroordelen binnen. Bijvoorbeeld algoritmes in rekrutering die vooral blanke mannen uitkiezen. Is dat een probleem van de dataset of van het algoritme? Is het de technologie of de samenleving? Want algoritmes voeden zich met de informatie die ze bijvoorbeeld op internet vinden. Als ze dan racistisch worden, zoals het algoritme Tay, is dat dan een probleem van de ontwikkelaars, van de technologie, of van de samenleving? In elk geval blijkt hier opnieuw dat we door de nieuwe technologie ook weer iets leren over mensen, hier over onze samenleving.

Algoritmes voeden zich met de informatie die ze bijvoorbeeld op internet vinden. Als ze dan racistisch worden, zoals Tay, is dat dan een probleem van de ontwikkelaars, van de technologie, of van de samenleving?

Een ander probleem met *machine learning* is dat deze algoritmes op een heel andere manier tot beslissingen komen dan mensen. Vaak is het zelfs voor het bedrijf dat deze technologie inzet niet duidelijk hoe het algoritme aan een bepaalde uitkomst komt. Het is een soort zwarte doos ('black box'). Is het dan wel verantwoord om dat soort algoritmes te gebruiken? Kunnen ze doorzichtiger of in elk geval wat witter of grijzer gemaakt worden? Kan de technologie uitleggen hoe de beslissing genomen is? In elk geval lijkt het best om ook altijd mensen te hebben die nog oordelen en de (eind)beslissing nemen. Denk opnieuw aan de juridische context of aan medische beslissingen. Misschien willen we toch graag een mens die over belangrijke zaken oordeelt. Maar het is duidelijk dat er steeds meer druk zal komen om de technologie in te zetten: om te besparen (in de gezondheidszorg), om de werkdruk te verlichten (voor rechters), enzovoort.

Elk van deze ethische problemen verdient een grondige discussie op zich. In elk geval is het belangrijk om op tijd een debat te houden. De beste vorm van techniekethiek doet er goed aan niet te wachten tot de technologie ingezet wordt en dan te klagen; beter is het om op een proactieve manier aan ethiek te doen en te proberen de ethiek al een rol te laten spelen tijdens de ontwikkeling van de techniek. Dit vraagt echter een beleid rond innovatie en techniek dat dit soort processen en instituties ondersteunt. Naast regulering is het ook goed om na te denken over ethiekonderwijs voor technische opleidingen en, beter nog, voor iedereen: het wordt steeds dringender om beter om te leren gaan met techniek in het dagelijkse leven. Dat is zo voor kinderen, maar ook voor volwassenen. Ook is het belangrijk om in een vroeg stadium burgers en niet-gouvernementele organisaties te betrekken bij het beleid, om later problemen te vermijden.

Uiteindelijk hangt veel af van de grote vragen rond de inrichting van de maatschappij. Wat voor samenleving en democratie willen we? Zo is bijvoorbeeld vandaag de macht rond techniek vooral in handen van een klein aantal heel grote bedrijven. Willen we dat zo houden, of is een betere situatie denkbaar? Wat moet de rol van de overheid zijn? En moeten we, gezien de globale aard van de technologie, ook niet nadenken over de rol van intergouvernementele en wellicht supragouvernementele organisaties? Globale problemen hebben globale oplossingen nodig.

Filosofie, als ze haar maatschappelijke verantwoordelijkheid opneemt, kan bijdragen aan dit nadenken over techniek en over de toekomst van onze samenleving. Ze kan nadenken over 'de mens', en dat is ook goed zo, maar het gaat ook over mensen die hier en nu leven, over ons en onze kinderen. Filosofische reflectie en verbeelding kunnen pendelen tussen die verschillende werelden, tussen verleden en toekomst, tussen universeel en particulier, tussen grote verhalen en kleine verhalen. Daar kunnen machines niet aan tippen. ●

Max Tegmark, *Life 3.0 – Being Human in the Age of Artificial Intelligence.* (Londen: Penguin Books, 2018).
Carl B. Frey & Michael A. Osborne, *The Future of Employment: How Susceptible are Jobs to Computerisation?.* (Oxford: Oxford University, 2013).

MARK COECKELBERGH is hoogleraar Filosofie van Media en Technologie aan de Universiteit van Wenen, departement Wijsbegeerte. Hij is eveneens voorzitter van de internationale Vereniging voor Filosofie en Technologie.

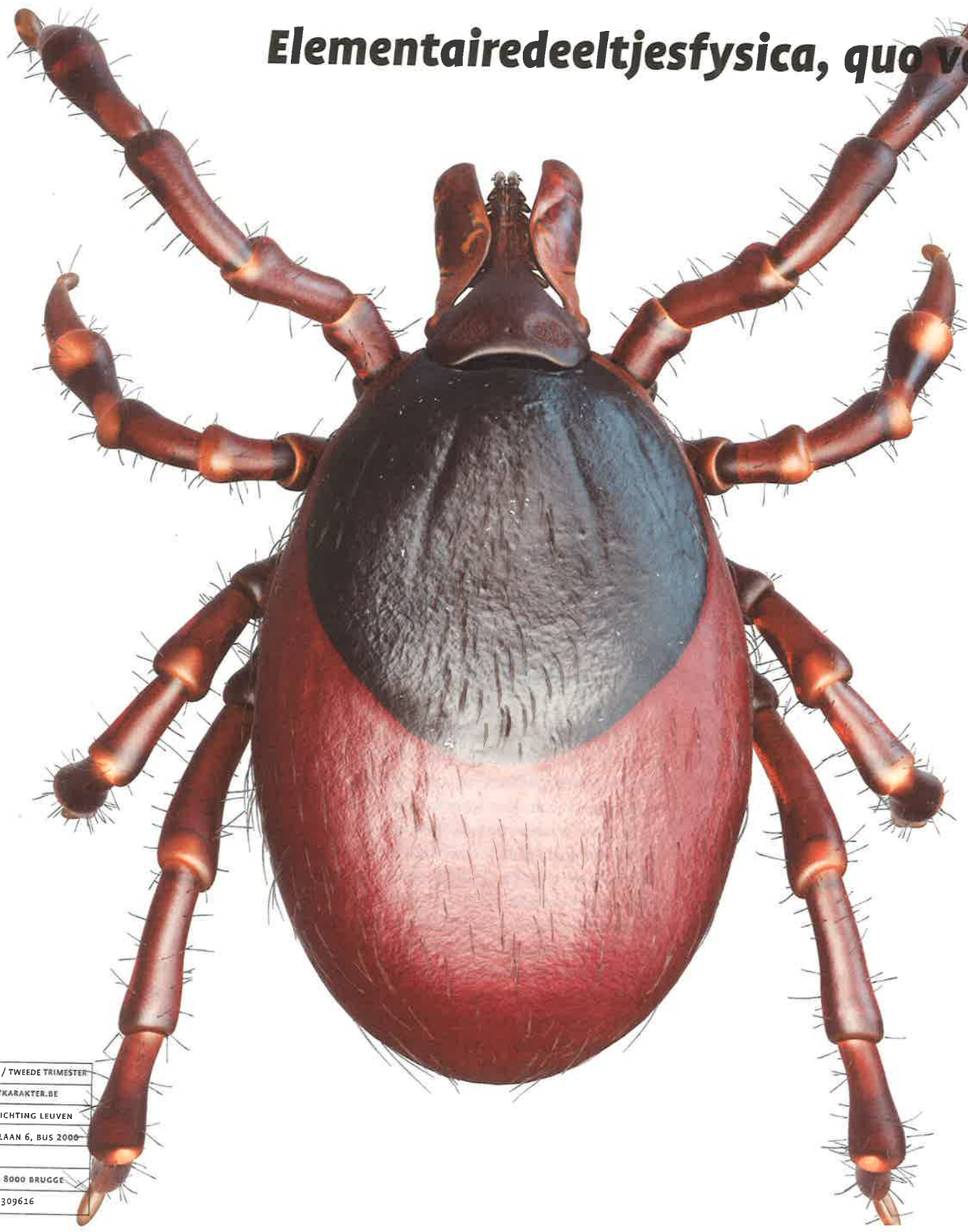
KARAKTER 66

TIJDSCHRIFT VAN WETENSCHAP

De mythe van Lyme

History of his-story?

Elementairedeeltjesfysica, quo vadis?



DRIEMAANDELIJKS / TWEEDE TRIMESTER
WWW.TIJDSCHRIFTKARAKTER.BE
ACADEMISCHE STICHTING LEUVEN
WILLEM DE CROYLAAN 6, BUS 2000
3001 HEVERLEE
AFGIFTEKANTOOR 8000 BRUGGE
1-2° AFDELING/P 309616