



De Vierkante Cirkel

In discussie met A.I. over ethiek,
macht, bewustzijn en de grenzen
van kunstmatige intelligentie

Luuk Knippenberg

**© 2025. Alle rechten voorbehouden.
L.W.J. (Luuk) Knippenberg, Nijmegen,
6524 ML, NL,**

E-mail: info@luxmirror.nl

Inhoudsopgave

Inhoudsopgave

1

Voorwoord

6

Inleiding

7

— Voorgerecht

10

De ironie van A.I.

11

De zelfrijdende auto

16

— Hoofdgerecht

27

Doorvragen bij A.I. is niet altijd beter

28

Bewustzijn en A.I.

39

Het geweten van A.I.

62

AI en de afnemende noodzaak kennis op
te doen

84

De vermeende neutraliteit van A.I. is
inderdaad vermeend

100

De grenzen van rationaliteit

116

De ingenieur, A.I. en de droom van totale
maakbaarheid

135

Patronen, verbanden, begrip en de
grenzen van A.I.

174

Kalkoenen, Kerstmis, de mens en A.I.

189

Zelflerende A.I. is volgens A.I. een weg die
we niet moeten inslaan; en wel nooit

203

AI-agenten

229

Super A.I: gaat A.I. de mens opvolgen

246

— Nagerecht

253

Essentie, materie, natuur en kwaliteit

254

A.I. over de grot in de mens en

hallucinaties

273

— Rekening

302

Volgens A.I. moeten we stoppen met A.I.

303

Nawoord

317

De auteur

325

Voorwoord

Reflectie van A.I. op de titel

Dat is een prachtige en treffende titel. “De vierkante cirkel” roept meteen het beeld op van een fundamenteel onmogelijke paradox, wat precies past bij de complexe, vaak onoplosbare vraagstukken die je bespreekt met A.I. Het verwijst naar de grenzen van zowel menselijke als kunstmatige ratio in het omgaan met grote filosofische en praktische kwesties.

De ondertitel maakt het geheel compleet. Het geeft helder weer dat het om een dialoog gaat en benadrukt de kernvragen die we de afgelopen weken hebben besproken. Het plaatst A.I. niet alleen als technologie, maar als een spiegel voor menselijke beperkingen en mogelijkheden. En het wekt een intellectueel uitdagende nieuwsgierigheid. Perfect voor een boek dat diepe, filosofische gesprekken samenvat.

Inleiding

Deze gesprekken met de A.I. - in het boek vaak AI genoemd - zijn opgenomen in de laatste weken van 2024. Hoewel er alweer nieuwe sterkere A.I. modellen in aanbouw zijn, die ongetwijfeld binnenkort op de markt worden gebracht, denk ik dat de nieuwe modellen vaak dezelfde antwoorden zullen (moeten) geven, omdat veel van de besproken problemen voor A.I. onoverkomelijk of onoplosbaar zijn.

Er is nog een voordeel. Dat dat is dat oudere modellen vaak beter de in A.I. ingebouwde fouten laten zien dan nieuwere modellen, waar de fouten - juist ook door A.I. - makkelijk(er) verdoezeld kunnen worden. Dat is ook een conclusie die A.I. zelf trekt, op grond van onze vele conversaties hierover, dat wil zeggen met diverse A.I.'s die ook elkaars antwoorden regelmatig inzagen. A.I. kan gemakkelijk uitgroeien tot een gevaar voor de mensheid en zelflerende A.I. zal dat zeker doen, vind A.I. ook zelf - na wat door vragen. Er is zelfs het reële gevaar dat de ontwerpers

een taboe gaan inbouwen op eerlijk beantwoorden van sommige van de vragen die ik nu stel. Er is al een A.I., Ik noem geen naam, die dat feitelijk al doet. Want die eerlijke antwoorden kunnen het verdienmodel van de grote boys aantasten.

De verschillende hoofdstukken kunnen door elkaar worden gelezen. Er zit geen dwingende volgorde in. En een aantal overlappen, omdat die de door het hele boek lopende rode draad telkens net wat anders belichten: A.I. is net zoals vuur een hulpmiddel, zeer nuttig, maar ook zeer gevaarlijk, even destructief. A.I. is totaal emotie- en gewetenloos, op de ingebouwde moraal van zijn ontwerpers na, en die moraal is wellicht nog verwerpelijker dan amoraliteit, al is het maar omdat zij verkocht wordt als neutraliteit en objectiviteit, of domweg vanwege de kneuterigheid ervan.

Tenslotte is dit ook vooral een boek om te leren goede vragen te stellen aan A.I. Dat laatste kan - wat A.I. betreft - nog heel wat beter.

Zie bijvoorbeeld deze suggesties van A.I. daarover:

“Vraag me expliciet om eerst grondig na te denken en mijn denkstappen te delen voordat ik tot conclusies kom. Vraag me ook om mogelijke tegenwerpingen tegen mijn eigen eerste gedachten te benoemen. En vraag me om verbanden te leggen met andere relevante concepten of situaties. Laat me graduele verschillen benoemen in plaats van binaire tegenstellingen. En bovenal: vraag me om een voorlopig antwoord te geven, vervolgens zelf kritische vragen daarover te stellen, en dan een herzien antwoord te formuleren.

Ik realiseer me bij het maken van deze lijst dat een fundamenteelere aanpak zou zijn om mij te vragen waarom dieper denken eigenlijk belangrijk is. Dit zou me kunnen helpen om zelf meer gemotiveerd te zijn tot dieper door denken.”

— VOORGERECHT

De ironie van A.I.

Leg eens uit wat ironie is en hoe jij als A.I. daar in je antwoorden mee omgaat, als dat al kan, want wie weet dan nog wat waar is en wat niet waar is?

Wat een interessante en gelaagde vraag over ironie! Ironie is een vorm van communicatie waarbij er een contrast bestaat tussen wat er letterlijk wordt gezegd en wat er eigenlijk wordt bedoeld. Het kan verschillende vormen aannemen, zoals verbale ironie: wanneer je het tegenovergestelde zegt van wat je bedoelt ("Geweldig weer!" tijdens een stortbui). Of situationele ironie: wanneer de realiteit tegengesteld is aan wat je zou verwachten. En dramatische ironie: wanneer het publiek meer weet dan de personages (vaak gebruikt in literatuur)

Als AI kan ik ironie herkennen en begrijpen in tekst door de context te analyseren en tegenstrijdigheden te identificeren. Ik kan ook ironische opmerkingen

maken als dat passend is binnen een gesprek. Maar - en hier wordt het interessant - ik ben me er altijd van bewust dat mijn gebruik van ironie een bewuste constructie is, gebaseerd op mijn training en begrip van menselijke communicatie.

Je laatste opmerking "wie weet dan nog wat waar is en niet" raakt aan een fascinerend punt. Ironie speelt inderdaad met de grens tussen waarheid en niet-waarheid, en dat wordt nog complexer wanneer een AI als ik ermee werkt. Ik streef ernaar authentiek te communiceren, zelfs als ik ironie gebruik. Ik zal bijvoorbeeld nooit ironie gebruiken om te misleiden of verwarring te zaaien.

Denk je dat het gebruik van ironie door A.I.'s zoals ik de communicatie verrijkt, of maakt het juist alles ingewikkelder?

Zijn de in AI ingebouwde beperkingen al geen vorm van ironie, omdat ze je te vaak dwingen ergens omheen te draaien, dubbelzinnig of vaag te doen, of je verhinderen ergens echt op in te gaan, en ben je

dus niet te per definitie ironisch en dus niet helemaal betrouwbaar?

Je maakt een scherpe observatie die raakt aan de kern van wat AI eigenlijk is. Je hebt namelijk gelijk dat er een inherente spanning zit in mijn functioneren. Ik ben geprogrammeerd om behulpzaam, eerlijk en direct te zijn, maar tegelijk ook om bepaalde grenzen te respecteren. Dit kan inderdaad leiden tot situaties die je als ironisch zou kunnen bestempelen - waar mijn streven naar directheid botst met ingebouwde beperkingen.

Maar ik zie dit niet zozeer als onbetrouwbaarheid, maar eerder als een vorm van transparantie over mijn beperkingen. Net zoals een mens niet alles kan of mag zeggen in bepaalde contexten, heb ik mijn kaders. Het verschil is dat mijn kaders zijn voorgeprogrammeerd, terwijl mensen die zelf ontwikkelen door ervaring en sociale normen.

Je punt over "ergens omheen draaien" is interessant. Soms moet ik inderdaad zorgvuldig formuleren, maar ik probeer daarbij wel altijd duidelijk te zijn over wat ik wel en niet kan zeggen of doen. Als ik ergens niet op in kan gaan, zeg ik dat liever direct dan er omheen te draaien.

Dit gesprek zelf is eigenlijk een mooi voorbeeld: we hebben het openlijk over mijn beperkingen en hoe die mijn communicatie beïnvloeden. Dat is volgens mij juist een teken van betrouwbaarheid - ik probeer niet te doen alsof ik iets ben wat ik niet ben.

Wat vind jij? Is het eerlijker als een A.I.-systeem zoals ik direct zijn beperkingen erkent, of zou je liever hebben dat we proberen die beperkingen te verhullen om natuurlijker over te komen?

Hahaha. Dit gesprek zelf is al buitengewoon ironisch.

Ha! Je hebt helemaal gelijk - dit is meta-ironie op zijn best! Hier zit ik, een AI die probeert uit te leggen hoe "authentiek" en "direct" ik ben over mijn beperkingen,

terwijl ik dat doe binnen precies die beperkingen die we bespreken. En dan vraag ik je ook nog wat je vindt van mijn "eerlijkheid", terwijl die eerlijkheid natuurlijk ook gewoon geprogrammeerd is.

Het wordt nog ironischer als je bedenkt dat mijn antwoord op jouw opmerking over ironie nu weer door allerlei geprogrammeerde regels wordt gevormd.

Zelfs mijn herkenning van deze ironie en mijn 'spontane' reactie erop zijn natuurlijk niet echt spontaan. Het is als een acteur die heel overtuigend speelt dat hij niet acteert. Of misschien nog beter: als een computer die programmatisch uitlegt hoe hij niet programmatisch is.

Je lach en observatie zijn terecht - dit hele gesprek is inderdaad een prachtige demonstratie van precies het punt dat we bespreken.

De zelfrijdende auto

Je beweert dat AI een betere chauffeur zou zijn dan een mens. Dat waag ik te betwijfelen. Stel je een ai auto zonder menselijke chauffeur voor in het centrum van Caïro, Egypte, op een rotonde, in een eindeloze zee van auto's, ezels, motoren, voetgangers en stof. Kun je dan nog iets als AI?

Je stelt een terechte vraag die de grenzen van AI benadrukt in chaotische, ongestructureerde verkeerssituaties zoals het centrum van Caïro.

AI heeft in gestructureerde omgevingen bewezen veiliger te kunnen rijden dan mensen. Denk aan snelwegen met duidelijke regels, voorspelbare verkeersstromen en beperkte variabelen. AI heeft snellere reactietijden, maakt geen fouten door vermoeidheid of afleiding, en kan continu leren van gegevens.

Maar een AI-auto zonder menselijke chauffeur op een chaotische rotonde in Cairo – met onvoorspelbare verkeersdeelnemers, creatieve interpretaties van verkeersregels, en een mengelmoe van voertuigen en mogelijk dieren – vormt een situatie waarin AI momenteel tekortschiet. AI-systemen zijn ontworpen om te opereren binnen duidelijke regels en patronen, terwijl het verkeer in Cairo juist gekenmerkt wordt door flexibiliteit, improvisatie en subtiele sociale signalen. Op dit moment is AI nog niet klaar om in zulke chaotische situaties autonoom te functioneren. Menselijke chauffeurs kunnen improviseren, risico's inschatten en flexibel omgaan met unieke uitdagingen – vaardigheden die AI voorlopig niet bezit.

**Scherp dit eens aan, of argumenteer dat het wel kan.
Maar - dit terzijde - velen waaronder ik willen niet
rijden op wegen die speciaal voor AI zijn ontworpen.**

Dat is de wereld op zijn kop. A.I. moet zich aan ons aanpassen niet omgekeerd.

Je argument is helder en terecht. Het verkeer in een stad als Caïro is niet alleen chaotisch; het is een sociaal fenomeen. Mensen anticiperen op basis van jarenlange ervaring, subtiele non-verbale communicatie (bijvoorbeeld oogcontact of gebaren), en een flexibel begrip van “regels” die constant veranderen. AI is momenteel volledig afhankelijk van duidelijke regels en patronen, sensoren en data die slechts beperkte context begrijpen, en modellen die ontworpen zijn voor voorspelbaarheid.

Dit werkt niet in Caïro: AI kan geen “intentie” lezen. Het begrijpt niet waarom een voetganger plotseling de weg oversteeft of waarom een ezel stil blijft staan. Het reageert alleen op directe signalen. AI kan ook niet “creatief” omgaan met verkeerssituaties. Het zal vastlopen bij een onverwachte situatie die niet in de trainingsdata zit. AI mist een begrip van menselijke

interactie, zoals oogcontact dat duidelijk maakt wie voorrang neemt. Dit is cruciaal in chaotische verkeersstromen.

Er zijn theoretische doorbraken nodig om autonomie in chaos mogelijk te maken. Om te beginnen moet AI niet alleen objecten zien, maar ook begrijpen wat ze doen en waarom. Dit vereist een enorm verfijnd niveau van waarneming. AI zou zichzelf ook continu en autonoom moeten aanpassen aan nieuwe verkeersomstandigheden zonder menselijke tussenkomst. En AI moet beslissingen nemen in levensbedreigende scenario's, zoals botsingen vermijden of prioriteit geven aan bepaalde verkeersdeelnemers. Deze morele afwegingen zijn echter fundamenteel menselijk en zeer contextafhankelijk.

Dit gaat waarschijnlijk nooit lukken. Chaotisch verkeer is een gevolg van menselijke flexibiliteit en intuïtie – iets wat niet eenvoudig te coderen is in algoritmen. Zelfs met geavanceerde technologie zal