Vier ethische perspectieven op ChatGPT

Je kunt ethiek zien als een proces van reflectie en deliberatie; een iteratief en participatief proces waarin je naar je project kijkt vanuit vier ethische perspectieven: gevolgenethiek (plussen en minnen), plichtethiek (plichten en rechten), relatie-ethiek (interacties en macht) en deugethiek (deugden en samenleven). Marc Steen maakt een proefrit met die vier perspectieven, met ChatGPT als casus. En op het eind enkele tips.

Marc Steen werkt als senior research scientist bij TNO, als expert in responsible innovation en toegepaste ethiek. Hij behaalde MSc, PDEng- en PhD-graden in Industrial Design Engineering aan de TU Delft. Onlangs verscheen zijn boek: *Ethics for people who work in tech* (Routledge).

Een goede analyse begint met helder krijgen: Waar hebben we het over? *Hoe concreter, hoe beter.* Als de applicatie nog niet is ontwikkeld, maak je een schets en legt die op tafel. Het hoeft niet af te zijn. Beter zelfs, want dan kun je wat je leert tijdens de analyse, meenemen bij verder ontwikkelen. Voor onze casus kunnen we kijken naar hoe journalisten ChatGPT kunnen gebruiken in hun werk. En we kunnen sturen richting waarden die we belangrijk vinden in de EU, zoals menselijke autonomie, voorkomen van schade, een eerlijke verdeling van lusten en lasten, transparantie en accountability.

Gevolgenethiek: plussen en minnen

Met gevolgenethiek stellen we ons voor welke effecten de resultaten van een project kan hebben in de maatschappij, in het dagelijks leven. Welke voordelen en nadelen levert dat op? Aan de plus-kant: journalisten (en anderen) kunnen ChatGPT gebruiken om efficiënter te werken. Aan de min-kant: die efficiency kan ervoor zorgen dat anderen hun baan verliezen. Aan de plus-kant: ChatGPT kan mensen helpen om hun woordenschat en grammatica te verbeteren. Aan de min-kant: mensen kunnen met ChatGPT extreem snel en goedkoop enorme hoeveelheden desinformatie produceren, het internet overspoelen, en daarmee nieuwsgaring en verkiezingen ondermijnen. Het is heel lastig om fake news te herkennen, zeker als er foto's of video's bij staan die ook door AI zijn gemaakt. Experts verwachten dat rond 2026 maar liefst 90% van online content met AI gemaakt of bewerkt zal zijn.

We kunnen ook kijken naar plussen en minnen die iets verder weg liggen. Je kunt de kosten voor milieu en maatschappij meetellen; de materialen die werden gebruikt in de chips waarop ChatGPT draait en de energie die nodig is om ChatGPT te trainen. We kunnen ook kijken naar de verdeling van plussen en minnen. Vaak worden mensen in lagelonenlanden ingeschakeld om data op te schonen en het model te trainen, vaak onder slechte omstandigheden. Zonder goede regulering (hieronder) komen de plussen bij enkele grote bedrijven terecht, en de minnen bij milieu en mensen aan de andere kant van de wereld.

Wat kun je nu als professional met gevolgenethiek? Als je werkt aan zo'n applicatie als ChatGPT, kun je proberen om de nadelen te verkleinen en de voordelen te vergroten.

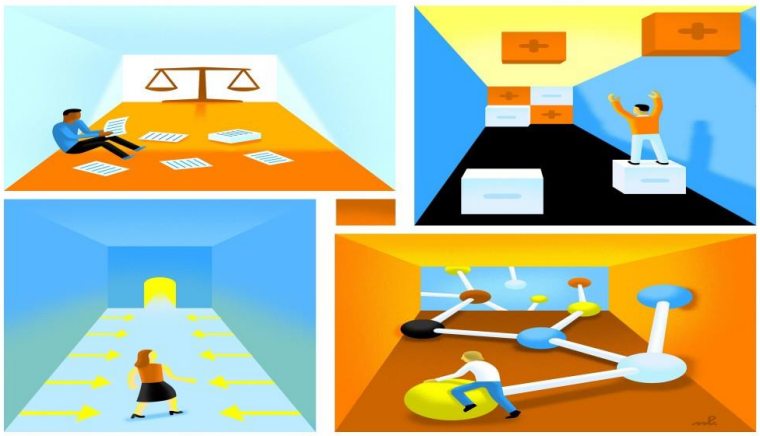
Plichtethiek: plichten en rechten

Met plichtethiek kijken we naar plichten en rechten. En als die botsen, zoeken we een balans. De ontwikkelaars van ChatGPT hebben plichten. En gebruikers van ChatGPT rechten. Dit perspectief overlapt met een juridisch perspectief. En dat is actueel: in juni nam het Europees parlement de AI-wet

aan, die ook betrekking heeft op Large Language Models. Bij ChatGPT spelen thema's als fairness en non-discriminatie. Er zit vaak bias in de data waarmee modellen worden getraind, zodat hun output discrimineert; Cathy O'Neil en Safiya Umoja Noble schreven erover. Vanuit deze zorg riepen Emily Bender, Timnit Gebru et al. in hun *Stochastic Parrots* paper op tot het met zorg samenstellen van datasets. We zien hier de *plichten* tot rechtvaardig en zorgvuldig handelen voor ontwikkelaars, vanwege de *rechten* op fairness en non-discriminatie van gebruikers.

Over het model achter ChatGPT valt nog iets anders op te merken. De teksten van ChatGPT lopen zo lekker omdat ze zijn gebaseerd op teksten die geschreven zijn door mensen; het model achter ChatGPT heeft immers het internet afgestruind. Dat kan een inbreuk opleveren op auteursrecht. In juni spanden twee auteurs een rechtszaak aan tegen OpenAI, het bedrijf achter ChatGPT. Een ander belangrijke thema van plichtethiek is menselijke waardigheid. Wat gebeurt er met menselijke waardigheid als je als burger of als klant 'te woord wordt gestaan' door ChatGPT, in plaats van door een mens?

Kortom, plichtethiek kan helpen om met je project tussen de vangrails van wetgeving te blijven, bias en discriminatie te vermijden, copyright te respecteren, en mensen waardig te bejegenen.



© Marc Kolle

Relatie-ethiek: interacties en macht

Met behulp van relatie-ethiek kunnen we kijken naar de invloed van technologie op hoe mensen met elkaar interacteren en communiceren. Bij ChatGPT denk ik aan ELIZA, de chatbot die Joseph Weizenbaum in de jaren zestig programmeerde. Hij was tamelijk onaangenaam verrast toen mensen aan ELIZA allerlei intelligentie en empathie toeschreven, zelfs nadat hij had uitgelegd dat het slechts een eenvoudig programmaatje was. Mensen projecteren heel gemakkelijk menselijke eigenschappen op objecten. Zo geloofde Blake Lemoine, inmiddels ex-Google, dat LaMDA bewust was.

Maar het omgekeerde kan ook. Een chatbot kan menselijke kwaliteiten uithollen. Als je ChatGPT klakkeloos gebruikt, krijg je middelmatige teksten en dat kan communicatie uithollen. Bovendien is ChatGPT ontworpen om teksten te produceren die qua tekst lekker lopen. Maar het model heeft geen begrip van onze fysieke wereld. Geen gezond verstand. En nauwelijks benul van (on)waarheid. Het produceerde deze volzin: 'The idea of eating glass may seem alarming to some, but it actually has several unique benefits that make it worth considering as a dietary addition.' Het is duidelijk dat klakkeloos gebruik van ChatGPT ernstige risico's kan hebben, bijvoorbeeld in de gezondheidszorg.

Bovendien kunnen we via relatie-ethiek vragen stellen over macht en over de verdeling van macht. In een wereld waarin veel mensen online naar informatie zoeken, versterken LLM's de macht van de partijen die die online media bezitten. Mensen als Reijer Passchier en Kees Verhoeven pleiten daarom voor het beperken van de macht van Big Tech bedrijven en voor goede regulering.

Deugdethiek: goed samenleven

Deugdethiek heeft wortels in het Athene van Aristoteles. Het doel van deugdethiek is het cultiveren van deugden, zodat we goed samen kunnen leven. Technologie speelt daarin een rol. Een bepaalde applicatie kan mensen helpen, of juist verhinderen, om een bepaalde deugd te cultiveren. Als je de hele dag op social media zit, neemt je vermogen tot zelfbeheersing af. Zelfbeheersing is één van de klassieke deugden. Het doel is het vinden van 'het juiste midden' voor elke deugd, in elke specifieke situatie. Neem moed. Als je sterk bent, en je ziet dat iemand een ander slaat, dan is het moedig om in te grijpen. Afzijdig blijven zou laf zijn. Maar als je niet sterk bent, dan zou het onbezonnen zijn om tussenbeide te komen. Op afstand blijven en 112 te bellen is dan moedig.

Wat kun je als ontwikkelaar doen? Je kunt bijvoorbeeld features toevoegen die ervoor zorgen dat mensen bepaalde deugden kunnen ontwikkelen. Neem social media, als voorbeeld. Vaak speelt daarbij desinformatie, en dat ondermijnt eerlijkheid en burgerschap. Als online media vol staan met desinformatie, is het lastig om te bepalen wat wel of niet waar is, en dat verziekt elk maatschappelijk debat. Als ontwikkelaar kun je werken aan een app die eerlijkheid en burgerschap juist bevordert. Bijvoorbeeld via features die mensen oproepen om feiten te controleren en met elkaar in gesprek te gaan, nieuwsgierig te zijn naar wat een ander bezig houdt—in plaats van mensen te laten scrollen langs fake news en polarisatie aan te wakkeren.

Aan de slag

Stel dat je zo'n proces van ethische reflectie en deliberatie wilt organiseren. Dan kan het handig zijn om dat stapsgewijs in te voeren en te integreren in werkwijzen die mensen al kennen.

Je kunt bijvoorbeeld ethische aspecten integreren in Human-Centred Design. Je legt een schets of prototype neer en gaat in gesprek, ook over ethische aspecten. Stel dat je werkt aan een algoritme dat hypotheekadviseurs helpt om te bepalen wat voor hypotheek iemand kan krijgen. Je kunt aan die adviseurs vragen wat voor effecten het gebruik van zo'n algoritme kan hebben op hun werkplezier, hun klantgerichtheid of hun autonomie? Met hun antwoorden kun je de ontwikkeling bijsturen.

Het kan ook nuttig zijn om te onderzoeken hoe verschillende stakeholders aankijken tegen de applicatie waar je aan werkt. Welke waarden vinden ze belangrijk? Je kunt bijvoorbeeld een bijeenkomst organiseren met een leverancier, een technisch expert, en iemand die kan vertellen hoe de applicatie 'in het veld' wordt gebruikt. Doorvragen is daarbij belangrijk. Stel dat twee mensen praten over een SUV en allebei veiligheid belangrijk vinden. Eén denkt vanuit het perspectief van de eigenaar van zo'n SUV. De ander denkt aan de veiligheid voor fietsers en voetgangers. De kunst is dan om ook lastige onderwerpen bespreekbaar te maken.

Tot slot kun je kritisch kijken naar de samenstelling van je projectteam. Een divers projectteam kan een probleem van meerdere kanten bekijken en kan verschillende soorten kennis combineren tot creatieve oplossingen. Zo voorkom je kokervisie en blinde vlekken. In de traditie van Technology Assessment kun je verkennen wat er mis zou kunnen gaan met de technologie waaraan je werkt. Als je dat tijdig doet, kun je maatregelen nemen om problemen te voorkomen en goed met risico's om te gaan. Mensen met verschillende ('afwijkende') perspectieven kunnen interessante vragen stellen. Dat geeft een completer beeld—soms complex; maar ja, de wereld is nu eenmaal complex.