

Bijlage 3: Begrippenlijst overheidsbrede visie op generatieve AI

1. *Algoritme*
Een set van regels en instructies die een computer uitvoert.
2. *Artificial general intelligence (AGI)*
Een technologie die intelligentie heeft over een breed scala aan domeinen waaronder redeneren, plannen en leren, en met deze vaardigheden op of boven het menselijke niveau presteert.
3. *Artificial intelligence (AI)-chatbot*
Digitale chatbots die via tekst en afbeeldingen kunnen communiceren op een manier die sterk lijkt op menselijke interactie. Een van de bekendste AI-chatbots op dit moment is ChatGPT.
4. *Artificiële intelligentie (AI)-systeem*
Een op machines gebaseerd systeem dat, voor expliciete of impliciete doelstellingen, afleidt, uit de input die het ontvangt, hoe het output zoals voorspellingen, inhoud, aanbevelingen of beslissingen moet genereren die fysieke of virtuele omgevingen kunnen beïnvloeden. Verschillende AI-systemen variëren in hun mate van autonomie en aanpassingsvermogen na de implementatie (deployment) ervan (OESO 2023).
5. *Bijzondere categorieën persoonsgegevens*
Persoonsgegevens waaruit ras, etnische afkomst, politieke opvattingen, religieuze of levensbeschouwelijke overtuigingen, het lidmaatschap van een vakbond blijken, of de verwerking van genetische gegevens, biometrische gegevens met het oog op de unieke identificatie van een persoon, gegevens over gezondheid, gegevens met betrekking tot iemands seksueel gedrag of seksuele gerichtheid.
6. *Black box-model*
Een (AI-)model waarvan er inzicht ontbreekt in hoe de voorspelling van het model tot stand is gekomen en wat de grondslag voor het gevormde model is.
7. *Dark patterns*
Interfaces, vooral in online gebruikersinterfaces, die de autonomie, besluitvorming en keuze van consumenten kunnen schaden. Ze misleiden, dwingen of manipuleren consumenten vaak, wat waarschijnlijk directe of indirecte schade veroorzaakt. Het kan echter moeilijk zijn om deze schade te meten.
8. *Deepfake*
Een met technologie gecreëerde foto, video of audio waarop te zien of te horen is dat een persoon dingen doet of zegt, die hij of zij niet daadwerkelijk heeft gedaan of gezegd.
9. *Desinformatie*
Desinformatie is het doelbewust, veelal heimelijk, verspreiden van misleidende informatie, met het doel om schade toe te brengen aan het publieke debat, democratische processen, de kenniseconomie of volksgezondheid.
10. *Finetuning*
Tijdens het 'finetunen' van een AI-model wordt een reeds getraind model aangepast aan een specifieke taak of dataset. In plaats van een model 'from scratch' te trainen, wordt gebruik gemaakt van een bestaand model.
11. *Foundation model*
Een foundation model is een basis machine learning model dat als fundament dient voor verdere gespecialiseerde modellen. Een large language model (LLM) is een type foundation model. Een voorbeeld van een foundation model is GPT-4, het foundation model voor ChatGPT.
12. *Generatieve AI*
Een vorm van AI waarbij complexe algoritmes worden ingezet om nieuwe content te genereren zoals tekst, afbeeldingen, computercode of video's. Chatbot ChatGPT vormt hiervan een bekende exponent.
13. *Hallucinatie*
Door een 'large language model' (LLM) gegenereerde informatie die feitelijk onjuist is. Het model genereert antwoorden die niet gebaseerd zijn op de gegeven input of op feitelijke informatie uit de trainingsdata. Een hallucinatie kan worden veroorzaakt door verschillende redenen, zoals een gebrek aan specifieke informatie in de trainingsdata, een gebrek aan context of foutieve en inconsistente informatie in de trainingsdata.
14. *Jailbreaking*
Het ontwerpen van prompts met de intentie model biases uit te buiten om zo output te genereren die niet strookt met het doel van het model. Het model zal bijvoorbeeld antwoord geven op vragen die normaliter door het model niet zouden worden beantwoord.
15. *Large language model (LLM)*
Een gespecialiseerd type AI-model dat getraind is op grote hoeveelheden tekst om bestaande content te begrijpen en content te genereren.
16. *Machine learning*
Een deelgebied van artificiële intelligentie dat computers in staat stelt om te leren van data. Een machine learning algoritme leert van voorbeelden en ervaringen om patronen en regels in data te ontdekken.

17. *Model*
Een AI-model is het resultaat van het trainen van een algoritme op data. Een algoritme is een set instructies, en het model is het specifieke resultaat van het volgen van de set instructies op basis van bepaalde data. GPT-4 is een voorbeeld van een AI-model, in dit geval een large language model.
18. *Model collapse*
Een fenomeen wat kan voorkomen wanneer LLM's worden getraind met 'vervulde' data; een database die door AI-gegenereerde data bevat. Er wordt gesproken van model collapse wanneer een LLM foutieve en minder gevarieerde output genereert, veroorzaakt door vervulde data.
19. *Modelparameters*
Aanpasbare instellingen in AI-modellen die beslissen hoe een LLM output genereert. Modelparameters hebben invloed op de kwaliteit, diversiteit en creativiteit van de output. Modelparameters komen op verschillende manieren tot stand, waaronder wiskundige berekeningen maar ook menselijke inmenging.
20. *Open source*
Het open source publiceren van een generatief AI-model houdt in dat de broncode van het model (en soms ook andere componenten) wordt gepubliceerd om in te zien, te analyseren, te hergebruiken en op voort te bouwen met eigen aanpassingen (*finetuning*).
21. *Persoonsgegevens*
Alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon ('de betrokkene'). Als 'identificeerbaar' wordt beschouwd een natuurlijke persoon die direct of indirect kan worden geïdentificeerd, met name aan de hand van een identificator zoals een naam, een identificatienummer, locatiegegevens, een online identifier of van een of meer elementen die kenmerkend zijn voor de fysieke, fysiologische, genetische, psychische, economische, culturele of sociale identiteit van die natuurlijke persoon.
22. *Pre-training*
Tijdens pre-training wordt een AI-model gevoed met grote hoeveelheden trainingsdata (tekst, audio, afbeeldingen, video) van verschillende bronnen. Het model leert tijdens pre-training patronen te herkennen in de data. Dit vereist aanzienlijke computerkracht en wordt uitgevoerd op gespecialiseerde hardware.
23. *Reinforcement Learning from Human Feedback (RLHF)*
In het geval van RLHF wordt menselijke feedback opgenomen in het trainingsproces van AI-algoritmes om het leren van het AI-algoritme te sturen of te verbeteren. Deze feedback van mensen kan als effect hebben dat het algoritme sneller en effectiever kan leren. Het doel is vaak om menselijke expertise te benutten om AI-algoritmes een bepaalde gewenste richting op te sturen.
24. *Systeem*
Een AI-systeem omvat niet alleen het model, maar ook de gehele infrastructuur eromheen. Dit omvat de hardware, software, gegevensverwerking, input- en outputinterfaces, en alle componenten die nodig zijn om het model effectief te laten werken. Een voorbeeld van een AI-systeem is ChatGPT.
25. *Taakspecifieke AI ('narrow AI')*
AI die geprogrammeerd is voor één specifieke taak, in tegenstelling tot generatieve AI, die kan worden ingezet voor een breed scala aan taken.
26. *Training*
Het proces waarbij een algoritme wordt geleerd om patronen in gegevens te herkennen.
27. *Transparantie*
Een model is transparant wanneer bekend is met welke formules, handelingen en waarden een model output genereert. Een transparant algoritme is het tegenovergestelde van een black box-algoritme.
28. *Uitlegbaarheid*
Een model is uitlegbaar wanneer uit te leggen en te begrijpen is waarom het model bepaalde output genereert. Uitlegbaarheid zorgt ervoor dat een mens kan begrijpen waarom een model iets doet zonder dat een mens de formules, handelingen en waarden binnen het model hoeft te kennen. Een uitlegbaar model is daarom niet per definitie transparant, en een transparant model is niet per definitie uitlegbaar.
29. *Vangrails ('guardrails')*
Beperkingen, richtlijnen of veiligheidsmaatregelen die worden ingesteld om ervoor te zorgen dat het gebruik van LLM's binnen ethische en verantwoorde grenzen blijft.
30. *Web scraping*
Het gebruik van software om informatie van webpagina's te extraheren om deze vervolgens te analyseren.