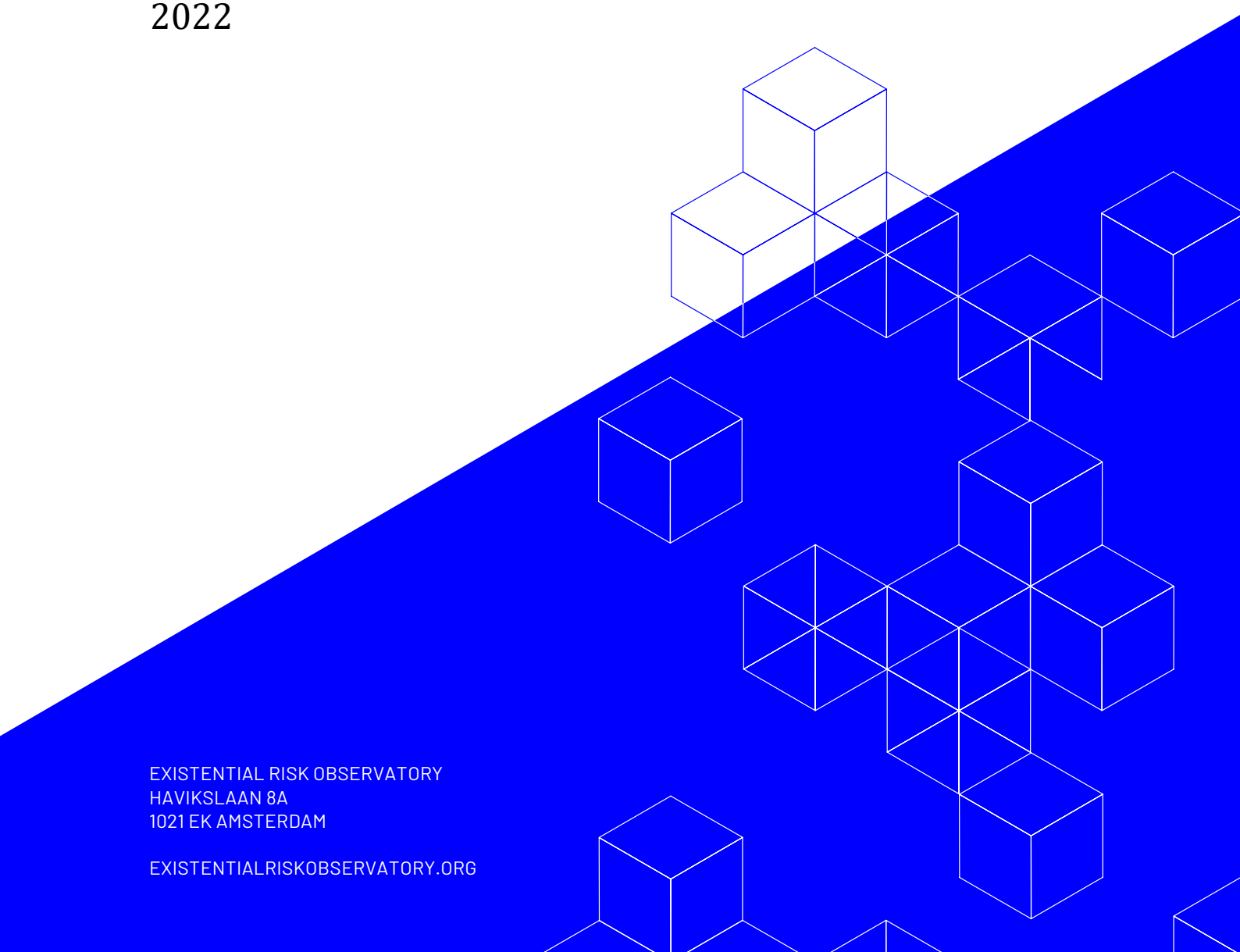
Existential Risk Observatory

POSITION PAPER

T.b.v.
rondetafelgesprek
Artificiële Intelligentie
d.d. 13 september
2022

OTTO BARTEN
SAM BOGERD

AUGUSTUS 2022



EXISTENTIAL RISK OBSERVATORY
HAVIKSLAAN 8A
1021 EK AMSTERDAM

EXISTENTIALRISKOBSERVATORY.ORG

SAMENVATTING

De ontwikkeling van nieuwe AI-systemen gaat op dit moment razendsnel. De beste modellen verdubbelen nu ongeveer iedere 4 maanden hun rekencapaciteit[1] en het einde hiervan lijkt nog niet in zicht. Dit is vele malen sneller dan Moore's law. Wij maken ons grote zorgen over mogelijke negatieve effecten van de veel krachtigere AI die we in de toekomst kunnen verwachten, en denken dat we hier onvoldoende op voorbereid zijn.

Een deel van het AI-onderzoek heeft als uiteindelijk doel om AI te ontwikkelen die de meeste cognitieve taken op menselijk niveau kan uitvoeren[2]. Het gaat dus om het maken van een AI die, bij gebrek aan een beter woord, slimmer is dan mensen. Het is onmogelijk om te zeggen wanneer dit succesvol zal zijn. Grote private partijen zoals Google en Facebook, en ook publieke partijen zoals de EU, investeren echter in projecten die als expliciet doel hebben om een dergelijke AI te ontwikkelen[3]. En zelfs als de kans op succes heel klein zou zijn, dan nog zijn de gevolgen groot genoeg om ons hierop voor te bereiden. Daarin zien wij ook een rol voor de Nederlandse overheid.

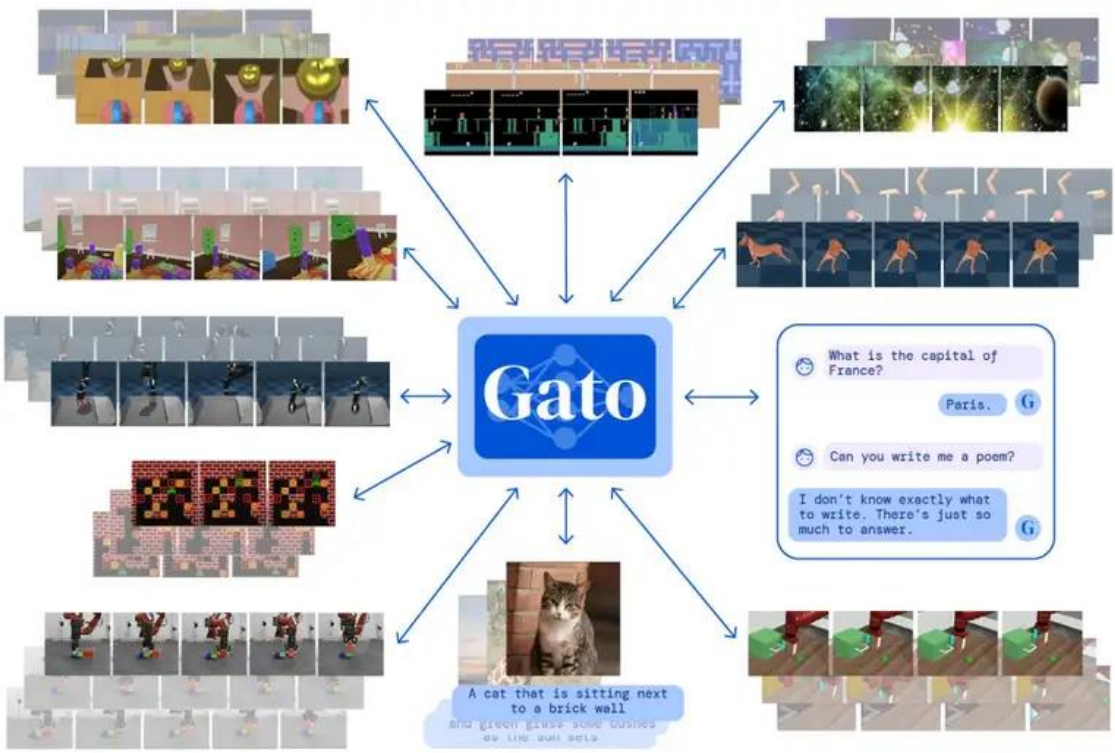
Wij denken specifiek dat de top drie aandachtspunten van de Commissie voor Digitale Zaken zouden moeten zijn om er zorg voor te dragen dat de overheid:

1. Risico's met een kleine kans en een zeer grote impact meeneemt in de nationale risicoanalyse, en ook over deze risico's communiceert richting de burger.
2. Zich voorbereidt om oplossingen te gaan implementeren wanneer deze beschikbaar zijn.
3. Onderzoek naar existentiële risico's en *AI Alignment* financiert, bijvoorbeeld door het opzetten van een Onderzoeksinstituut voor Existentiële Risico's.

RISICO'S VAN TOEKOMSTIGE AI

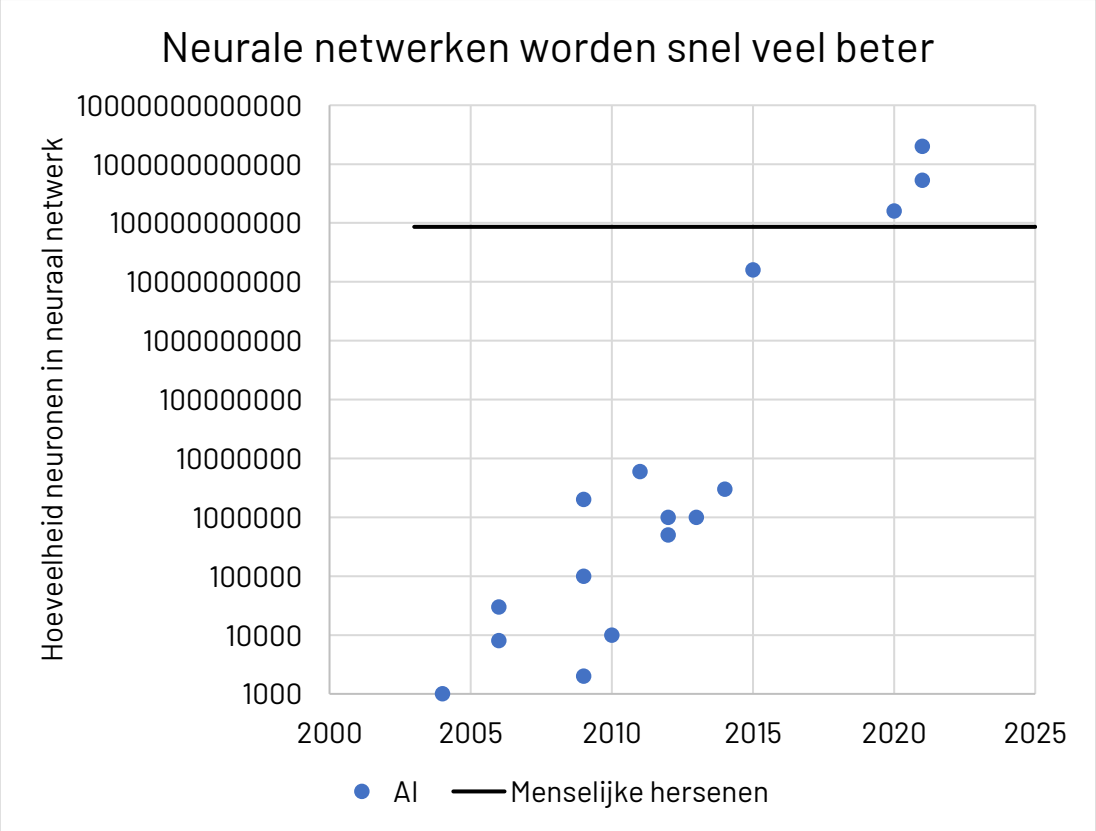
De afgelopen jaren hebben de doorbraken in de AI elkaar in steeds sneller tempo opgevolgd. Het model AlphaGo van Google's Deepmind versloeg bijvoorbeeld in 2016 de beste menselijke Go-speler, een doorbraak die door vele AI-wetenschappers niet werd voorzien. Afgelopen jaar hebben we daarnaast AlphaCode van hetzelfde bedrijf gezien, een model dat computerprogramma's kan schrijven met een vergelijkbare kwaliteit als een gemiddelde menselijke programmeur[4]. Recent is het Gato-model, een Large Language Model, er al in geslaagd om meer dan 600 taken ongeveer op menselijk niveau uit te voeren, waaronder zeer diverse taken zoals chatten, beelden herkennen, en objecten oppakken met een robotarm[5] (zie figuur 1). Daarnaast hebben we bij de toeslagenaffaire gezien hoe algoritmes discriminatie in de kaart spelen en horen we nu steeds meer over discriminerende algoritmes.

[1] https://web.archive.org/web/20220728173946/https://openai.com/blog/ai-and-compute/
[2] Het recente WRR-rapport "Opgave AI. De nieuwe systeemtechnologie" omschrijft dit als "Het bereiken daarvan [AGI] is uiteindelijk het doel van het AI-vakgebied".
[3] Fitzgerald, McKenna, Aaron Boddy, and Seth D. Baum. "2020 survey of artificial general intelligence projects for ethics, risk, and policy." (2020): 20. In dit paper staan 72 projecten geïdentificeerd die expliciet naar Artificial General Intelligence (AGI) proberen toe te werken, waaronder o.a. Deepmind, een Google-subsidiary met zo'n 1300 FTE, het Facebook-gelieerde OpenAI, en het door de EU gefinancierde Human Brain Project, met een begroting van circa 1 miljard euro. Wij denken dat de investeringen die deze projecten ophalen een indicatie zijn dat zowel markt als overheden verwachten dat AGI een bereikbaar doel is.
[4] Li, Yujia, et al. "Competition-level code generation with alphacode." arXiv preprint arXiv:2203.07814 (2022).
[5] Reed, Scott, et al. "A generalist agent." arXiv preprint arXiv:2205.06175 (2022).



Figuur 1: het GATO-model kan met slechts één keer trainen meer dan 600 taken op ongeveer menselijk niveau uitvoeren (bron figuur: Deepmind).

Maar de capaciteiten van de AI zoals die nu bestaat, hoe indrukwekkend ook, zijn vele malen kleiner dan de capaciteiten van de AI die we kunnen verwachten in de toekomst. De afgelopen jaren is het aantal neuronen in de grootste neurale netwerken gegroeid met een factor 10 per jaar (zie figuur 2). Wat ons betreft zou het debat over AI, naast over vraagstukken met huidige AI, dan ook vooral moeten gaan over de vele malen krachtigere AI die we in de toekomst kunnen verwachten, en welke kansen en risico's deze AI met zich meebrengt.



Figuur 2: hoeveelheden neuronen in neurale netwerken als functie van de tijd[6].

Naast de groei in rekenkracht zijn er nog meer redenen om te denken dat ontwikkeling van AI in een stroomversnelling kan komen in de komende decennia. Eén ontwikkeling zou vooral grote gevolgen kunnen hebben: de mogelijkheid van AI die AI kan verbeteren. Een verzameling klassieke (*narrow*) AI's zou zo vaardig kunnen worden, dat deze samen alle stappen van het doen van onderzoek naar betere AI systemen op menselijk niveau kan uitvoeren. AI zou dan betere AI kunnen maken in een positieve feedback loop, waardoor de vaardigheid van AI zeer snel kan toenemen, met een onbekend eindpunt.

Naast door een verzameling klassieke AI's, zou dit ook kunnen gebeuren door één model dat alle taken minimaal net zo goed als de mens kan uitvoeren, een AGI (*Artificial General Intelligence*). Dit is een - vooralsnog hypothetisch - AI-systeem dat niet alleen beter is dan mensen in een specifieke taak (zoals Go, schaken, of een auto besturen), maar beter is in alle cognitieve taken. Het WRR noemt het ontwikkelen van AGI dan ook "het doel van het AI-vakgebied" in het recente rapport "Opgave AI. De nieuwe systeemtechnologie". Vooroplopende AI labs zoals Deepmind en OpenAI werken op dit moment aan het ontwikkelen van AGI. Omdat Large Language Models zoals Gato van Deepmind er nu al in slagen om vele taken te combineren, en omdat de complexiteit en de vaardigheden van AI enorm snel toenemen, kunnen we er niet langer meer veilig vanuit gaan dat AGI slechts ontwikkeld zal worden in de verre toekomst, en dat we ons hier dus niet op hoeven voor te bereiden.

[6] https://www.researchgate.net/figure/1-Increase-trend-of-the-dataset-and-neural-network-sizes-8_fig5_348437192
https://spectrum.ieee.org/biggest-neural-network-ever-pushes-ai-deep-learning
https://en.wikipedia.org/wiki/GPT-3
https://en.wikipedia.org/wiki/Wu_Dao
https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

Er lijken enorme risico's te kleven aan een AI die, simpel gezegd, alles minstens net zo goed als een mens kan. Een eerste risico betreft actoren met kwade intenties: terroristen, ons vijandige staten, hackers, etc. moet de toegang tot deze technologie ontzegd kunnen worden. Dit zou erg lastig kunnen zijn, omdat software eenvoudig gekopieerd kan worden. Maar ook het controleren van AGI zelf, zelfs door actoren met goede intenties, is tot op heden een onopgelost wetenschappelijk vraagstuk. Dit wordt *AI Alignment* genoemd: zorgen dat AGI handelt in overeenstemming met onze doelen. Er is namelijk helaas een groot aantal manieren waarop het doel van de AGI iets anders zou kunnen zijn of worden dan wat wij als samenleving op de lange termijn graag zouden willen, bijvoorbeeld de volgende:

- We slagen er niet in om ons doel te formuleren op een manier die de AGI afdoende begrijpt.
- De AGI destilleert zelf een doel waarvan het verwacht dat wij dit willen, op basis van onze data, maar slaagt hier onvoldoende in.
- Ons doel verandert, maar we slagen er niet in om het doel in de AGI te veranderen.
- De doelen van een kleine elite worden geïmplementeerd (bijv. maximale winst voor een bedrijf) en deze blijken sterk af te wijken van maatschappelijk gewenste doelen.
- We implementeren korte termijn-doelen (bijv. winst boven klimaat) die op de lange termijn catastrofaal blijken uit te werken.

De WRR vat dit probleem in haar rapport goed samen: “Het probleem is dan ook niet zozeer dat AI eigen, kwaadaardige doelen ontwikkelt, maar dat ze heel behendig is in het bereiken van doelen die mensen erin hebben gestopt en die gevaren opleveren of onvoldoende doordacht zijn.”

En zelfs als we voor deze problemen (*outer misalignment*) een oplossing zouden vinden, bestaat er nog de mogelijkheid dat het doel waar de AGI naartoe werkt, verschuift door technische oorzaken (*inner misalignment*). Dit effect is al aangetoond bij huidige machine learning modellen[7], en wordt waarschijnlijk ernstiger naar mate het model complexer wordt.

Vanwege al deze redenen is de kans dat AGI, wanneer deze wordt uitgevonden, een ander doel heeft dan wat wij graag zouden willen, zeer groot. Ook is er, ondanks zo’n vijftien jaar onderzoek hiernaar, geen manier bekend om AGI met een doel anders dan het onze te limiteren of te corrigeren.

BELEIDSAANBEVELINGEN

Wij vinden dat de overheid toe zou moeten zien op de veiligheid van haar burgers, en daarom een taak heeft in het managen van mogelijke existentiële risico’s, waaronder het risico op unaligned AGI. Of en wanneer deze AGI er komt, en of hij inderdaad unaligned zal zijn, is onzeker, en zal tot op zekere hoogte onzeker blijven tot het moment dat AGI er is. Maar net zo min als onzekerheid in klimaatmodellen een reden mag zijn om geen klimaatbeleid te voeren, of onzekerheid op het gebied van toekomstige beleggingsresultaten een reden om geen pensioenen aan te leggen, zo mag ook onzekerheid op het gebied van de toekomst van technologie geen reden zijn om de grootste risico's niet te managen. Het alternatief zou immers zijn dat we de risico's over ons heen laten komen zonder een poging te doen om deze in te perken, en dit is volgens ons een zeer gevaarlijke strategie.

[7] Di Langosco, Lauro Langosco, et al. "Goal Misgeneralization in Deep Reinforcement Learning." International Conference on Machine Learning. PMLR, 2022.

Wat zou de overheid kunnen doen om het risico op unaligned AGI zo klein mogelijk te maken? Helaas is er op dit moment geen oplossing bekend die het risico tot nul zou reduceren. Maar er is wel veel beleid denkbaar dat waarschijnlijk tot risicoreductie zou leiden. Volgens ons zouden de top drie aandachtspunten van de Commissie van Digitale Zaken moeten zijn:

1. Erkenning en communicatie van de existentiële risico's van AI van menselijk niveau. Het probleem moet erkend worden door de overheid, bijvoorbeeld door existentiële risico's door nieuwe technologieën toe te voegen aan de Geïntegreerde Risicoanalyse Nationale Veiligheid. Ook moet de existentiële dreiging door nieuwe technologie open, duidelijk en expliciet naar de burger gecommuniceerd worden. Beleidsmakers en politici moeten een informatiepositie opbouwen op het gebied van existentiële risico's en deze informatie vervolgens actief delen met de burger, omdat de hele maatschappij het recht heeft geïnformeerd te zijn, en zodat de hele maatschappij actie kan ondernemen om de risico's zo klein mogelijk te krijgen.
2. De overheid moet zich voorbereiden om oplossingen te gaan implementeren voor existentiële risico's door nieuwe technologie, zoals AI, zelfs al weten we op dit moment nog niet exact hoe deze oplossingen eruit zullen zien.
3. De overheid moet onderzoek financieren naar existentiële risico's door nieuwe technologie zoals AI. Op dit moment werken er nog slechts enkele honderden onderzoekers wereldwijd aan het in kaart brengen en oplossen van dit probleem (bijvoorbeeld bij instituten in Oxford en Cambridge). Nederland zou eenvoudig het wereldwijde aantal onderzoekers kunnen verdubbelen, door bijvoorbeeld een Onderzoeksinstituut voor Existentiële Risico's op te zetten, dat inschattingen maakt over de grootte van de risico's en voorstellen doet over hoe deze te verminderen. Hiernaast zou er op dit onderzoeksinstituut technisch onderzoek naar AI Alignment plaats moeten vinden. Ook universiteiten en bestaande instituten zouden budget en stimulering moeten krijgen om op deze gebieden onderzoek te gaan doen. Het financieren van onderzoek dat zowel helpt met het verminderen van AI-bias als met AI-alignment, zoals bijvoorbeeld *interpretability* (waarbij het doel is om duidelijk te maken waarom een AI bepaalde beslissingen heeft gemaakt), is hierbij een goed voorbeeld van een no-regrets maatregel.

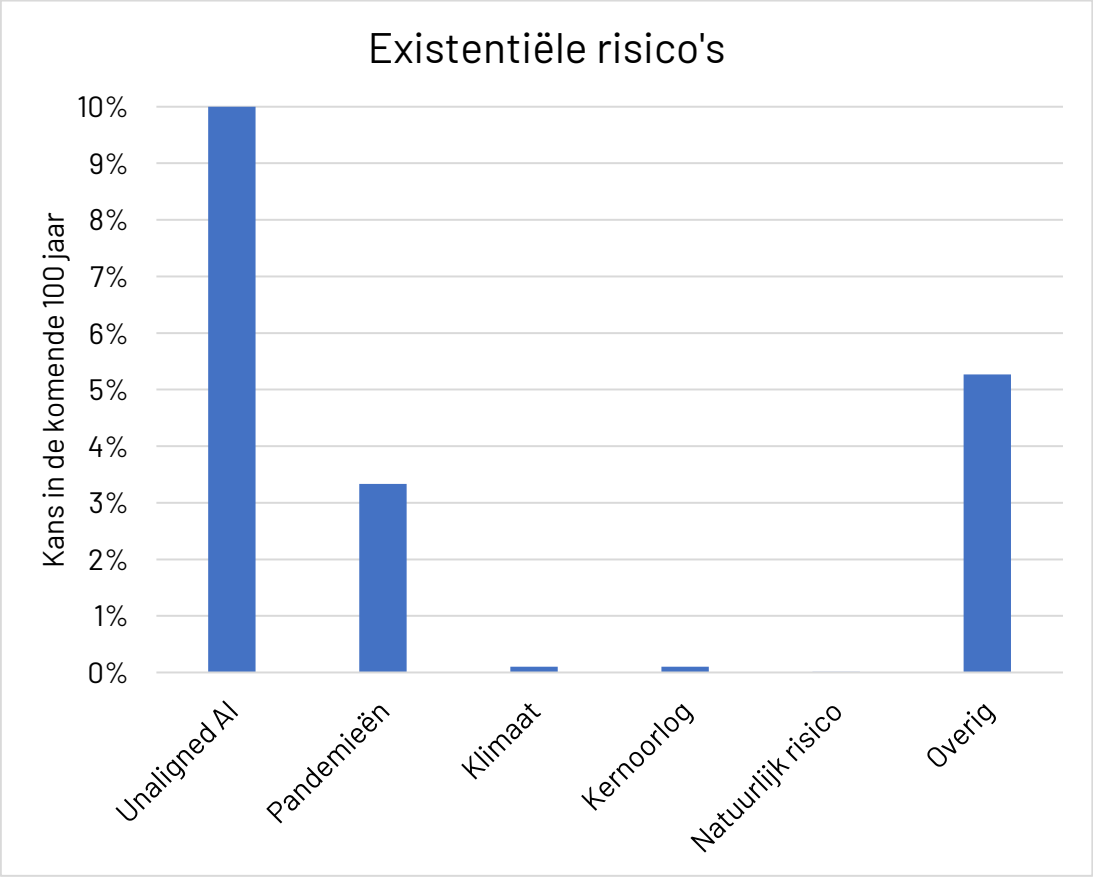
Hoewel Nederland unilateraal al veel kan doen, en daarmee een enorm grote positieve impact kan hebben, zit er uiteraard ook een grote internationale dimensie aan het reduceren van existentiële risico's zoals die van unaligned AGI. Ook hier zou Nederland het voortouw in kunnen nemen, bijvoorbeeld door dit probleem expliciet aan te kaarten op EU-niveau, VN-niveau, en bilateraal met de meest relevante landen.

De WRR vergelijkt in haar rapport AI met de verbrandingsmotor van deze eeuw: 100 jaar geleden werden de eerste auto's geproduceerd, dit ging niet zonder slag of stoot en het was eerst moeilijk om deze technologie op de juiste manier in de samenleving te integreren.

Uiteindelijk zorgde de verbrandingsmotor voor een explosie van welvaart. Maar we weten inmiddels ook dat er grote lange-termijngevaren zijn door het wijdverspreide gebruik van de verbrandingsmotor. Terwijl we werken aan een wereld waarin AI een normaal deel van onze economie en samenleving wordt, mogen we de lange-termijngevaren van deze transformatieve technologie dan ook niet uit het oog verliezen. We moeten, in een wereld waarin technologie zich zeer snel ontwikkelt, ook rekening houden met de gevaren op de lange termijn, zodat we beter zijn voorbereid op toekomstige problemen dan het geval was bij COVID-19 of de klimaatcrisis.

HET EXISTENTIAL RISK OBSERVATORY EN EXISTENTIËLE RISICO'S

Het Existential Risk Observatory is een Nederlandse stichting met als doel om de kans dat een existentieel risico optreedt te reduceren, door het publieke debat te informeren. Een existentieel risico wordt gedefinieerd als een risico dat het lange-termijnpotentieel van de mensheid bedreigt, zoals menselijk uitsterven, een permanente ineenstorting van onze beschaving, of een dystopische lock-in. Wij volgen academicus Toby Ord in zijn inschatting van de kansen hierop. Wij denken daarom dat er een totale kans van grofweg één op zes op een existentieel risico bestaat de komende honderd jaar, waarvan een kans van circa één op tien op een existentieel risico vanwege kunstmatige intelligentie voor dezelfde periode. De schattingen van Ord staan in de grafiek hieronder.



Figuur 3: existentiële risico's gekwantificeerd door Ord[8].

Wij vinden deze kansen onaanvaardbaar hoog. Wij geloven echter ook dat, omdat deze risico's voor het overgrote deel door de mens zelf veroorzaakt worden, de mensheid ook het vermogen heeft risico's te reduceren tot een acceptabel niveau. Wie daarin slaagt, redt mogelijk niet alleen een enorm aantal huidige levens, maar ook een nog veel groter mogelijk aantal levens in de toekomst. Vanwege deze redenen zou dit vraagstuk de hoogste prioriteit moeten krijgen, waar wij het afgelopen jaar dan ook voor gepleit hebben in onder andere publicaties in Het Parool, De

[8] Ord, Toby, Oxford University. The precipice: Existential risk and the future of humanity. Hachette Books, 2020.

Telegraaf en Trouw (ook hier), en internationaal o.a. voor gepleit is in The Economist en recent in de New York Times. Aangezien unaligned AI niet alleen door Ord, maar ook door de meeste andere onderzoekers op het gebied van existentiële risico's tot het grootste risico bestempeld wordt, zouden existentiële risico's volgens ons zeker prioriteit moeten krijgen bij een commissiedebat over AI.