Governance van de overgang naar kunstmatige algemene intelligentie (AGI) Dringende overwegingen voor de Algemene Vergadering van de VN

Rapport voor de Raad van Voorzitters van de Algemene Vergadering van de Verenigde Naties (UNCPGA)

Samenvatting

AI-systemen ontwikkelen zich snel in de richting van kunstmatige algemene intelligentie (AGI), gekenmerkt door systemen die in staat zijn menselijke intelligentie te evenaren of te overtreffen in diverse cognitieve taken. Met de grootste financiële investeringen in de geschiedenis die ongekende R&D-inspanningen stimuleren, verwachten industrieleiders en experts dat AGI binnen dit decennium zou kunnen ontstaan,[1] met buitengewone voordelen voor de mensheid. Een van deze voordelen is dat AGI wetenschappelijke ontdekkingen op het gebied van volksgezondheid zou kunnen versnellen, veel industrieën zou kunnen transformeren en de productiviteit zou kunnen verhogen, en zou kunnen bijdragen aan de realisatie van de Duurzame Ontwikkelingsdoelen.

Toch kan AGI ook unieke en mogelijk catastrofale risico's met zich meebrengen. In tegenstelling tot traditionele AI zou AGI autonoom schadelijke acties kunnen uitvoeren zonder menselijk toezicht, wat kan leiden tot onomkeerbare gevolgen, bedreigingen van geavanceerde wapensystemen en kwetsbaarheden in kritieke infrastructuren. We moeten ervoor zorgen dat deze risico's worden beperkt als we de buitengewone voordelen van AGI willen benutten.

Om deze mondiale uitdagingen effectief aan te pakken, is onmiddellijke en gecoördineerde internationale actie, ondersteund door de Verenigde Naties, essentieel. Dergelijke acties zouden moeten worden geïnitieerd door een speciale Algemene Vergadering van de VN over AGI om de voordelen en risico's van AGI te bespreken en de mogelijke oprichting van een wereldwijd waarnemingscentrum voor AGI, een certificeringssysteem voor veilige en betrouwbare AGI, een VN-Verdrag over AGI en een internationaal AGI-agentschap. Zonder proactief wereldwijd beheer zal concurrentie tussen landen en bedrijven de ontwikkeling van risicovolle AGI versnellen, veiligheidsprotocollen ondermijnen en geopolitieke spanningen verergeren. Gecoördineerde internationale actie kan deze uitkomsten voorkomen en veilige ontwikkeling en gebruik van AGI, eerlijke verdeling van voordelen en wereldwijde stabiliteit bevorderen.

I. Inleiding

Het tempo van de vooruitgang in AI is de afgelopen jaren en maanden[2] hoog geweest en zou nog verder kunnen versnellen, deels omdat AI-bedrijven enorme bedragen investeren in het maken van AI-agenten die capabeler en autonomer zijn, en vanwege het toenemende gebruik van de krachtigste AI-modellen om het AI-onderzoek zelf vooruit te helpen.[3] Het wordt

[1] METR: Measuring AI Ability to Compete Long Tasks https://arxiv.org/abs/2503.14499

[2] Zie het International AI Safety Report, Bengio et al 2025.

[3] https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.pdf

algemeen verwacht dat deze verbeteringen in AI-capaciteiten zullen leiden tot "kunstmatige algemene intelligentie" (AGI): AI-systemen die de menselijke prestaties bij de meeste cognitieve taken evenaren of overtreffen.

Hoewel er onenigheid bestaat over de vraag wanneer AGI wordt verwacht, geloven alle deskundigen in dit panel dat AGI wel eens binnen dit decennium zou kunnen worden ontwikkeld. AI-bedrijven trekken honderden miljarden dollars uit om AGI op zeer korte termijn te realiseren, waardoor dit veruit de grootste R&D-inspanning in de menselijke geschiedenis is. De private sector heeft de verantwoordelijkheid om technologie te ontwikkelen die veel veiliger zal zijn, en ze zouden stimulansen moeten hebben om dat te doen; maar de concurrentiewedloop om AGI als eerste te bereiken dwingt hen om al hun inspanningen te richten op capaciteiten in plaats van op veiligheid, om "de race te winnen".

De huidige risico's van AI komen vooral voort uit menselijk misbruik van technologie. AGI vormt echter ook een fundamenteel ander risico, omdat de potentiële bedreigingen verder gaan dan door mensen veroorzaakt misbruik. AGI zou autonoom plannen met catastrofale gevolgen kunnen genereren en uitvoeren, waardoor het menselijke vermogen om opkomende bedreigingen en ongekende verstoringen te herkennen, te analyseren en erop te reageren, wordt overtroffen.[4] In combinatie met de recent waargenomen neiging tot zelfbehoud[5] van geavanceerde AI's zou dit kunnen leiden tot situaties waarin AGI oncontroleerbaar wordt.

Dit zou een gedeelde wereldwijde zorg moeten zijn. AGI-gerelateerde risico's zijn niet beperkt tot specifieke industrieën of samenlevingen, maar hebben wereldwijde gevolgen, ongeacht waar ze hun oorsprong vinden. Om te zorgen voor een veilige en harmonieuze integratie van AGI zijn niet alleen nationale of bedrijfsinspanningen nodig, maar ook proactief internationaal bestuur, onder leiding van de Verenigde Naties. De Verenigde Naties zijn bij uitstek geschikt om wetenschappelijke overeenstemming te bereiken over risico's en risicobeperkende strategieën, politieke consensus te bereiken over een gezamenlijke aanpak voor risicobeperking, beleid te coördineren, normen of vangrails te bevorderen, te reageren op noodsituaties en mogelijk gezamenlijk veiligheids- of beveiligingsonderzoek uit te voeren of te coördineren.

Zonder mondiaal bestuur zou het transformatieve potentieel van AGI om mondiale uitdagingen aan te pakken onderbenut of verkeerd gericht kunnen worden. Bovendien zal wereldwijde coördinatie essentieel zijn bij het beheersen van de wereldwijde catastrofale dreigingen die van AGI worden verwacht. Het is moeilijk voor te stellen dat deze coördinatie op mondiaal niveau kan worden bereikt zonder actief leiderschap van de VN.

(4) "Claude 3.7 (often) Knows When it is in Alignment Evaluations
https://www.apolloresearch.ai/blog/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations

[5] See Meinke et al 2024, Frontier Models are Capable of In-context Scheming, https://arxiv.org/abs/2412.04984.

I. Urgentie voor actie van de Algemene Vergadering van de VN op het gebied van AGI-governance en waarschijnlijke gevolgen als er geen actie wordt ondernomen

Te midden van de complexe geopolitieke omgeving en bij gebrek aan samenhangende en bindende internationale normen, verhoogt een concurrerende haast om AGI te ontwikkelen zonder adequate veiligheidsmaatregelen, het risico op ongelukken of verkeerd gebruik, bewapening en existentiële mislukkingen.[6] Naties en bedrijven geven voorrang aan snelheid boven veiligheid, ondermijnen nationale bestuurlijke kaders en maken veiligheidsprotocollen ondergeschikt aan economisch of militair voordeel. Aangezien veel vormen van AGI van regeringen en bedrijven voor het einde van dit decennium de kop op kunnen steken en het opzetten van nationale en internationale bestuurssystemen jaren zal duren, is het dringend noodzakelijk om de noodzakelijke procedures te starten om de volgende uitkomsten te voorkomen:

1. **Onomkeerbare** gevolgen **- Als** AGI eenmaal is bereikt, kunnen de gevolgen onomkeerbaar zijn. Met veel grensvormen van AI die al bedrieglijk gedrag en zelfbehoud vertonen[6] en de drang naar meer autonome, op elkaar inwerkende, zichzelf verbeterende AI's die geïntegreerd zijn met infrastructuren, is het aannemelijk dat de effecten en het traject van AGI uiteindelijk onbeheersbaar worden. Als dat gebeurt, is er misschien geen manier om terug te keren naar een toestand van betrouwbaar menselijk toezicht. Proactief bestuur is essentieel om ervoor te zorgen dat AGI onze rode lijnen niet overschrijdt[7], wat leidt tot oncontroleerbare systemen zonder duidelijke manier om terug te keren naar menselijke controle.
2. **Massavernietigingswapens** - **AGI** zou sommige staten en kwaadwillende niet-overheidsactoren in staat kunnen stellen om chemische, biologische, radiologische en nucleaire wapens te bouwen. Bovendien zouden grote, door AGI bestuurde zwermen dodelijke autonome wapens zelf een nieuwe categorie massavernietigingswapens kunnen vormen.
3. Kwetsbaarheid **van kritieke infrastructuur - Kritieke** nationale systemen (zoals energienetwerken, financiële systemen, transportnetwerken, communicatie-infrastructuur en gezondheidszorgsystemen) kunnen het doelwit worden van krachtige cyberaanvallen door of met behulp van AGI. Zonder nationale afschrikking en internationale coördinatie zouden kwaadwillende niet-statelijke actoren, van terroristen tot transnationale georganiseerde misdaad, op grote schaal aanvallen kunnen uitvoeren.
4. **Machtsconcentratie, mondiale ongelijkheid en instabiliteit -** Ongecontroleerde ontwikkeling en gebruik van AGI kan de verschillen in rijkdom en macht op ongekende schaal vergroten. Als AGI in handen blijft van een paar naties, bedrijven of elitegroepen, kan het economische dominantie verankeren en wereldwijde monopolies creëren over intelligentie, innovatie en industriële productie. Dit zou kunnen leiden tot massale werkloosheid, wijdverspreide ontkrachting die van invloed is op de juridische onderbouwing, verlies van

[6] OpenAI response to US Office of Science and Technology Policy's AI Action Plan https://cdn.openai.com/global-affairs/ostp-rfi/ec680b75-d539-4653-b297-8bcf6e5f7686/openai-response-ostp-nsf-rfi-notice-request-for-information-on-the-development-of-an-artificial-intelligence-ai-action-plan.pdf

[7] International Dialogues over AI Safety (2024): https://idais.ai/dialogue/idais-beijing/

privacy en instorting van het vertrouwen in instellingen, wetenschappelijke kennis en bestuur. Het zou democratische instellingen kunnen ondermijnen door overreding, manipulatie en door AI gegenereerde propaganda, en de geopolitieke instabiliteit kunnen vergroten op manieren die de kwetsbaarheid van het systeem vergroten. Een gebrek aan coördinatie kan leiden tot conflicten over AGI-middelen, -capaciteiten of -controle, wat kan escaleren tot oorlog. AGI zal bestaande juridische kaders onder druk zetten: veel nieuwe en complexe kwesties van intellectueel eigendom, aansprakelijkheid, mensenrechten en soevereiniteit zouden binnenlandse en internationale juridische systemen kunnen overweldigen.

5. **Existentiële risico's - AGI** zou misbruikt kunnen worden om massale schade aan te richten of ontwikkeld kunnen worden op manieren die niet in overeenstemming zijn met menselijke waarden; het zou zelfs autonoom kunnen handelen, buiten het toezicht van mensen om, en zijn eigen doelen kunnen ontwikkelen volgens zelfbehoudsdoelen die al zijn waargenomen bij de huidige voorop lopende AI's. AGI zou ook macht kunnen nastreven als een manier om zichzelf te beschermen. AGI's zouden ook op zoek kunnen gaan naar macht als middel om ervoor te zorgen dat ze de doelen die ze bepalen kunnen uitvoeren, ongeacht menselijke tussenkomst. Nationale regeringen, vooraanstaande experts en de bedrijven die AGI ontwikkelen hebben allemaal verklaard dat deze trends kunnen leiden tot scenario's waarin AGI-systemen de mens proberen te overmeesteren. Dit zijn geen vergezochte science fiction hypotheses over de verre toekomst - veel vooraanstaande experts zijn van mening dat deze risico's allemaal binnen dit decennium werkelijkheid kunnen worden, en hun voorlopers doen zich al voor.[2]Bovendien hebben vooraanstaande AI-ontwikkelaars tot nu toe geen haalbaar voorstel om deze risico's met een hoge mate van zekerheid te voorkomen.

6. **Verlies van buitengewone toekomstige voordelen voor de hele mensheid - Goed** beheerde AGI belooft verbeteringen op alle gebieden, voor alle mensen, van gepersonaliseerde geneeskunde, het genezen van kanker en celregeneratie, tot geïndividualiseerde leersystemen, het uitbannen van armoede, het aanpakken van klimaatverandering en het versnellen van wetenschappelijke ontdekkingen met onvoorstelbare voordelen. Om zo'n prachtige toekomst voor iedereen te garanderen, is mondiaal bestuur nodig, dat begint met een verbeterd mondiaal bewustzijn van zowel de risico's als de voordelen. De Verenigde Naties zijn van cruciaal belang voor deze missie.

II. Doel van VN-bestuur van de overgang naar AGI

Gezien het feit dat AGI mogelijk binnen dit decennium wordt ontwikkeld, is het zowel wetenschappelijk als ethisch noodzakelijk dat we robuuste bestuursstructuren opbouwen om ons voor te bereiden op zowel de buitengewone voordelen als de buitengewone risico's die het met zich mee zou kunnen brengen.

Het doel van VN-bestuur in de overgang naar AGI is om ervoor te zorgen dat de ontwikkeling en het gebruik van AGI in overeenstemming zijn met wereldwijde menselijke waarden, veiligheid en ontwikkeling. Dit houdt in: 1) Het bevorderen van onderzoek naar afstemming en controle van AI om technische methoden te identificeren voor het sturen en/of controleren van steeds capabeler wordende AI-systemen; 2) Het bieden van richtlijnen voor de ontwikkeling van AGI - het vaststellen van kaders om ervoor te zorgen dat AGI op verantwoorde wijze wordt ontwikkeld, met robuuste veiligheidsmaatregelen, transparantie en in overeenstemming met menselijke waarden; 3) Het ontwikkelen van bestuurlijke kaders voor de inzet en het gebruik van AGI - om misbruik te voorkomen, eerlijke toegang te garanderen en de voordelen voor de mensheid te maximaliseren en tegelijkertijd de risico's te minimaliseren; 4) Het bevorderen van toekomstige visies op nuttige AGI - nieuwe kaders voor sociale, ecologische en economische ontwikkeling; en 5) Het bieden van een neutraal, inclusief platform voor internationale samenwerking en het vaststellen van wereldwijde standaarden, het opbouwen van een internationaal rechtskader en het creëren van stimulansen voor

naleving; daarbij vertrouwen kweken tussen naties om wereldwijde toegang tot de voordelen van AGI te garanderen.

III. Belangrijkste overwegingen tijdens de Algemene Vergadering van de VN over AGI

Eén van de grootste uitdagingen van AGI-governance is de onzekerheid over de toekomstige technologische ontwikkeling. Dit maakt het moeilijk om potentiële voordelen en risico's nauwkeurig te voorspellen. Daarom moet er een breed en alomvattend reactiekader worden opgezet om te anticiperen op denkbare bedreigingen en deze te beperken en tegelijkertijd de potentiële voordelen te versterken. De Verenigde Naties kunnen zorgen voor internationale coördinatie die cruciaal is voor de ontwikkeling en het gebruik van AGI. Het is vooral belangrijk dat alle naties vertegenwoordigd zijn in dit proces en dat het geopolitieke kloven vermindert; op dit moment lijkt alleen de VN goed gepositioneerd om deze rol te spelen. De volgende punten moeten worden overwogen tijdens een zitting van de Algemene Vergadering van de VN die specifiek over AGI gaat:

A. Wereldwijd waarnemingscentrum voor AGI

Er is een wereldwijd waarnemingscentrum voor AGI nodig om de voortgang in onderzoek en ontwikkeling die relevant zijn voor AGI te volgen en om lidstaten vroegtijdig te waarschuwen voor AI-veiligheid. Dit observatorium moet gebruikmaken van de expertise van andere VN-inspanningen, zoals het onafhankelijke internationale wetenschappelijke panel voor AI dat is opgericht door het Global Digital Compact en de Readiness Assessment Methodology van UNESCO.

B. Internationaal systeem van beste praktijken en certificering voor veilige en betrouwbare AGI

Gezien het feit dat AGI mogelijk binnen dit decennium wordt ontwikkeld, is het zowel wetenschappelijk als ethisch noodzakelijk dat we robuuste bestuursstructuren opzetten om ons voor te bereiden op zowel de buitengewone voordelen als de buitengewone risico's die het met zich mee zou kunnen brengen.

C. VN Kaderverdrag inzake AGI

Er is een Raamverdrag inzake AGI[8] nodig om gedeelde doelstellingen en flexibele protocollen vast te stellen om de risico's van AGI te beheersen en te zorgen voor een eerlijke verdeling van de voordelen wereldwijd. Het verdrag moet duidelijke risiconiveaus definiëren die proportionele internationale actie vereisen, van standaardisatie- en licentieregelingen tot gezamenlijke onderzoeksfaciliteiten voor AGI met een hoger risico, en rode lijnen of struikeldraden[9]voor de ontwikkeling van AGI. Een verdrag

[8]Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo en Zaheed Kara. 2024. " Framework Convention on Global AI Challenges. " Center for International Governance Innovation. https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

[9]Russell, Stuart, Edson Prestes, Mohan Kankanhalli, Jibu Elias, Constanza Gómez Mont, Vilas Dhar, Adrian Weller, Pascale Fung and Karim Beguir, " AI red lines: The opportunities and challenges of setting limits. World Economic Forum, 11 maart 2025. https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/
Karnofsky, Holden. 2024. "A Sketch of Potential Tripwire Capabilities for AI." Carnegie Endowment for International Peace. 10 december 2024. https://carnegieendowment.org/research/2024/12/a-sketch-of- potential-tripwire-capabilities-for-ai?lang=en.

zou de aanpasbare institutionele basis bieden die essentieel is voor wereldwijd legitiem, inclusief en effectief AGI-bestuur, waardoor wereldwijde risico's worden geminimaliseerd en de wereldwijde welvaart door AGI wordt gemaximaliseerd.

D. Haalbaarheidsstudie naar een VN-AGI-agentschap

Gezien de breedte van de maatregelen die nodig zijn om je voor te bereiden op AGI en de urgentie van de kwestie, zijn er stappen nodig om de haalbaarheid van een VN-agentschap voor AGI te onderzoeken, idealiter in een versneld proces. Er is zoiets als de IAEA gesuggereerd, met dien verstande dat AGI-governance veel complexer is dan kernenergie en dus unieke overwegingen vereist in zo'n haalbaarheidsstudie.

IV. Deze aanbevelingen dragen bij aan de implementatie van het VN-pact voor de toekomst en andere VN-initiatieven.

Meerdere VN-initiatieven roepen op tot de ontwikkeling van veilige, beveiligde en betrouwbare AI. De resoluties van de Algemene Vergadering van de VN over AI - A/78/L.49, A/78/L.86 en A/C.1/79/L43-samen met het VN-Pact fort he Future, het Global Digital Compact en de aanbeveling van de UNESCO over de ethiek van AI, roepen op tot internationale samenwerking om AI te ontwikkelen waar de hele mensheid baat bij heeft, terwijl de wereldwijde risico's proactief worden beheerd.

Deze initiatieven hebben wereldwijd de aandacht gevestigd op de huidige vormen van AI. Dit rapport bouwt voort op deze VN-initiatieven door specifiek in te gaan op de ontwikkeling van AGI in de nabije toekomst.

De toezeggingen van het Pact for the Future worden in dit rapport op verschillende manieren verder uitgewerkt. Een sessie van de Algemene Vergadering van de VN over AGI komt tegemoet aan de toezegging van het Pact for the Future om een wereldwijde dialoog over AI-governance aan te gaan. De aanbevelingen in dit rapport over een VN-kaderverdrag over AGI en een haalbaarheidsstudie voor een VN-agentschap voor AGI. Het waarnemingscentrum dat we hebben voorgesteld, zou het werk ondersteunen van het aanstaande onafhankelijke internationale wetenschappelijke panel voor AI, een van de belangrijkste uitkomsten van het Global Digital Compact. Tot slot zou het internationale systeem van beste praktijken en certificering voor veilige en betrouwbare AGI bijdragen aan vertrouwen en transparantie, zoals gevraagd in de resoluties van de Algemene Vergadering van de VN, UNESCO en het Pact for the Future.

V. Conclusie

Het vergroten van het bewustzijn van nationale en internationale leiders over de voordelen en risico's van toekomstige AGI - in tegenstelling tot de huidige vormen van AI - is dringend nodig. De Algemene Vergadering van de VN is de juiste plaats om zo'n wereldwijde discussie op gang te brengen.

Internationale coördinatie van de ontwikkeling en het gebruik van AGI zal nodig zijn om de buitengewone voordelen van AGI te benutten en tegelijkertijd de mensenrechten en veiligheid te waarborgen. Het AGI Panel beveelt de Algemene Vergadering van de VN aan om dringend actie te ondernemen om

om deze kwesties aan te pakken tijdens een zitting van de Algemene Vergadering die speciaal is gewijd aan een mondiaal bestuurskader voor AGI. Zonder dergelijke actie zijn de risico's van ongecontroleerde ontwikkeling en gebruik van AGI - variërend van dramatisch toegenomen mondiale ongelijkheid tot existentiële bedreigingen - immens. Deze aanpak onder leiding van de VN, met een wereldwijd observatorium, internationale certificering, een VN-verdrag over AGI en een speciaal agentschap voor AGI, vergroot de kans dat AGI wordt ontwikkeld en gebruikt op manieren die de hele mensheid ten goede komen en tegelijkertijd de risico's minimaliseren. Dit kader moet inclusief, transparant en afdwingbaar zijn om vertrouwen en samenwerking tussen landen te bevorderen.

Bijlage

Referentiekader: Panel op hoog niveau over kunstmatige algemene intelligentie (AGI) voor de Raad van voorzitters van de Algemene Vergadering van de Verenigde Naties (UNCPGA)

Context

In de Verklaring van Seoul 2024 van de UNCPGA wordt een panel van deskundigen op het gebied van kunstmatige algemene intelligentie (AGI) gevraagd een kader en richtlijnen op te stellen waarmee de Algemene Vergadering van de VN rekening kan houden bij het aanpakken van de dringende problemen in verband met de overgang naar kunstmatige algemene intelligentie (AGI).

Dit werk moet voortbouwen op de uitgebreide inspanningen op het gebied van AI-waarden en -beginselen van UNESCO, OESO, G20, G7, Global Partnership on AI en de Bletchley Declaration en de aanbevelingen van het adviesorgaan op hoog niveau voor AI van de secretaris-generaal van de VN, het Global Digital Compact van de VN, het International Network of AI Safety Institutes, het Kaderverdrag inzake AI van de Europese Raad en de twee resoluties van de Algemene Vergadering van de VN over AI. Deze hebben zich meer gericht op engere vormen van AI. Er is momenteel een gebrek aan vergelijkbare aandacht voor AGI.

AI is tegenwoordig wereldwijd bekend en wordt vaak gebruikt, maar AGI is dat niet en bestaat nog niet. Veel AGI-experts geloven dat het binnen 1-5 jaar kan worden bereikt en uiteindelijk kan evolueren tot een kunstmatige superintelligentie die buiten onze controle valt. Er is geen algemeen aanvaarde definitie van AGI, maar de meeste AGI-experts zijn het erover eens dat het een AI voor algemene doeleinden zou zijn die kan leren, zijn code kan aanpassen en autonoom kan handelen om veel nieuwe problemen aan te pakken met nieuwe oplossingen die vergelijkbaar zijn met of verder gaan dan menselijke vermogens. De huidige AI heeft deze capaciteiten niet, maar de technische vooruitgang wijst duidelijk in die richting.

Het Global Digital Compact van de VN roept op tot een wereldwijde dialoog over AI-governance binnen de Verenigde Naties. Deskundigen uit de private sector op het gebied van AGI hebben benadrukt dat er dringend behoefte is aan een wereldwijd gesprek om de kansen en risico's van AGI beter te begrijpen. Een speciale zitting van de Algemene Vergadering van de VN over AGI is waarschijnlijk de snelste, meest kosteneffectieve en snelste manier om zo'n gesprek te stimuleren.

Doel

In antwoord op de Seoul Declaration 2024 van de UNCPGA, stel een eerste rapport op voor de voorzitter en de leden van de UNCPGA voor de UNCPGA-vergadering van 8-10 april 2025 in Bratislava.

Het rapport moet de risico's, bedreigingen en kansen van AGI in kaart brengen. Het moet gericht zijn op bewustwording van het mobiliseren van de Algemene Vergadering van de VN om AGI-governance op een meer systematische manier aan te pakken. Het zou zich moeten richten op AGI die nog niet is bereikt, in plaats van op de huidige vormen van meer 'narrow' AI-systemen. Het moet de urgentie benadrukken om AGI-kwesties zo snel mogelijk aan te pakken, gezien de snelle ontwikkelingen van AGI, die zowel ernstige risico's voor de mensheid als buitengewone voordelen voor de mensheid met zich mee kunnen brengen.

Het rapport zou ook zowel multilaterale afspraken als acties van de private sector moeten bevatten om deze ongekende uitdagingen aan te gaan. Het moet een reactie zijn op de oproep van AGI-leiders uit de private sector voor internationale coördinatie en multilaterale actie voor wat wel eens de moeilijkste managementuitdaging zou kunnen zijn waar de mensheid ooit voor stond.

Procedures

- Roep een panel op hoog niveau (5-8 leden) bijeen van internationale AGI-experts over de potentiële bedreigingen van AGI voor de mensheid en de kansen die AGI de mensheid kan bieden en gerelateerde beleidskwesties.
- Het AGI-panel komt vanaf januari 2025 regelmatig virtueel bijeen en voltooit het eerste rapport voor de volgende UNCPGA-bijeenkomst in Bratislava in het voorjaar van 2025.
- Op basis van de feedback op het initiële rapport tijdens de UNCPGA-vergadering in Bratislava zal het panel het rapport voltooien en indienen bij de secretaris-generaal van UNCPGA. Als de voorzitter van de UNCPGA het verslag aanvaardt, wordt het tegen 1 mei 2025 overgemaakt aan de voorzitter van de Algemene Vergadering van de VN.

Onafhankelijke panelleden op hoog niveau over AGI voor de Raad van voorzitters van de Algemene Vergadering van de VN

Jerome Glenn (VS) Voorzitter

Stemgerechtigd lid van IEEE Organizational Governance of AI; auteur van het document Horizon 2025-27 van de Europese Unie *over AGI*: *Issues and Opportunities*; CEO van The Millennium Project en auteur van de *International Governance Issues of the Transition from Artificial Narrow Intelligence to AGI*, *Requirements for Global Governance of AGI*, *en Work/Technology 2050: Scenario's en Acties*. Auteur van *Future Mind: Kunstmatige Intelligentie* (1989).

Renan Araujo (Brazilië)

Onderzoeksmanager bij het Instituut voor AI Beleid en Strategie gericht op risicomanagement met betrekking tot de ontwikkeling van AGI. Momenteel leidt hij het werk van IAPS op het gebied van internationale governance van AGI. Hij is een Oxford China Policy Lab Fellow, advocaat, medeoprichter van het Condor Initiative (dat Braziliaanse studenten in contact brengt met kansen van wereldklasse om AI-onderzoek en -beleid vorm te geven) en werkte aan AI-governanceprogramma's bij Rethink Priorities en het Institute for Law and AI.

Yoshua Bengio (Canada)

Hoogleraar computerwetenschappen aan de Université de Montréal; voorzitter van de Safety and Secure AI Advisory Group voor de Canadese overheid; voorzitter van het International AI Safety Report in opdracht van 30 landen plus de VN, OESO en EU. Wetenschappelijk directeur van Mila, het Quebec AI Institute; lid van de wetenschappelijke adviesraad van de secretaris-generaal van de VN voor doorbraken in wetenschap en technologie; ontvanger van de Turing Award en momenteel de meest geciteerde computerwetenschapper wereldwijd.

Joon Ho Kwak (Republiek Korea)

Technisch adviseur van het Koreaanse AI-veiligheidsinstituut; speelde een leidende rol bij de ontwikkeling van de OESO-richtsnoeren voor de ontwikkeling van betrouwbare AI; deelnemer aan het Hiroshima-proces van de G7, de voorbereidingen van de AI-actietop in Parijs, de Koreaans-Amerikaanse AI-werkgroep en lid van de Koreaanse delegatie naar de internationale netwerken van AI-veiligheidsinstituten.

Lan Xue (China)

Voorzitter van het Nationaal Comité van Deskundigen op het gebied van AI-governance; decaan van het Instituut voor AI International Governance aan de Tsinghua Universiteit; lid van de adviesgroep van het WTI-directoraat van de OESO; adviseur voor het China AI Safety Institute; medevoorzitter van de Leadership Council van het UN Sustainable Development Solution Network (UNSDSN); ontvanger van de Fudan Distinguished Contribution Award for Management Science en de Distinguished Contribution Award van de Chinese Vereniging voor Wetenschaps- en W&T-beleid.

Stuart Russell (VK en VS)

Distinguished Professor of Computer Science en Director, Center for Human-Compatible AI, University of California, Berkeley; auteur van Artificial Intelligence: A Modern Approach, het standaard AI-leerboek dat wordt gebruikt in 1.500 universiteiten in 135 landen en dat meer dan 74.000 keer is geciteerd; medevoorzitter van de OESO-expertgroep voor AI-futures en de Global AI Council van het World Economic Forum.

Jaan Tallinn (Estland)

Lid van het adviesorgaan voor AI van de VN; lid van de deskundigengroep op hoog niveau voor AI van de EC; medeoprichter van het Center for the Study of Existential Risk en het Future of Life Institute van de Universiteit van Cambridge (beide instellingen zijn toonaangevend op het gebied van AGI); bestuurslid van het Center for AI Safety; Estse investeerder in AGI-veiligheid; oprichter van Skype en FastTrack/Kazaa; en een van de oprichters van DeepMind Google.

Mariana Todorova (Bulgarije)

Bulgaarse vertegenwoordiger in UNESCO's Intergouvernementele Groep voor AI-ethische kaders; vooraanstaand woordvoerder over AGI in de Bulgaarse media; internationaal erkend auteur en docent over de ethische en technologische dimensies van AI en AGI; voormalig parlementslid en adviseur van de president van de Republiek Bulgarije.

José Jaime Villalobos (Costa Rica)

Multilateral Governance Lead bij het Future of Life Institute; Senior Research Associate, Centre for International Governance Innovation; Research Affiliate, Oxford Martin AI Governance Initiative; Research Affiliate, Institute for Law & AI; gepromoveerd in internationaal recht; en is co-auteur van toonaangevende boeken en artikelen over internationaal AI-bestuur.