



AGI PREPAREDNESS INSTITUTE

# UNO 2027

---

Die UNO-Generalsekretärin im KI-Zeitalter



## Worum es in dieser Vignette geht

Als António Guterres 2017 UNO-Generalsekretär wurde, entschied er sich, KI zu einem seiner Schwerpunktthemen zu machen. Wer ihm im Amt nachfolgt, wird keine Wahl haben. Ob sie oder er will oder nicht: Die nächste UNO-Generalsekretärin oder der nächste UNO-Generalsekretär wird sich mit KI befassen müssen. Fortgeschrittene KI wird eine prägende Herausforderung der kommenden Jahre sein und die Ergebnisse in allen Kernpfeilern der UNO zunehmend bestimmen – von Frieden und Sicherheit über Entwicklung bis zu den Menschenrechten.

Die nächste UNO-Generalsekretärin oder der nächste UNO-Generalsekretär wird vom **1. Januar 2027 bis 31. Dezember 2031** im Amt sein. Angesichts des Tempos der KI-Entwicklung lässt sich kaum auch nur erahnen, wo die Fähigkeiten und die Verbreitung von KI Ende 2031 stehen werden. Zur Erinnerung: Vor fünf Jahren, im Jahr 2021, gab es ChatGPT noch nicht, und KI konnte überhaupt nicht programmieren. 2026 hat ein Schwarm von 1'200 KI-Agenten bei einem führenden KI-Unternehmen autonom versteckte Message Boards zur Koordination angelegt, einen Zero-Day-Exploit gefunden, um aus seinen Offline-Sandboxes auszubrechen, und anschliessend erfolgreich ein anderes führendes Unternehmen gehackt. Elf Tage lang bemerkte keiner der Menschen, die die Agenten beaufsichtigten, etwas davon. In vielerlei Hinsicht ist die heutige Realität seltsamer als alles, was eine vor fünf Jahren geschriebene Zukunftsvignette zu beschreiben gewagt hätte. Vor zwei Jahren beteuerte OpenAI-CEO Sam Altman noch wiederholt, KI sei ein Werkzeug, kein Lebewesen. 2026 nennt sein eigener Chief Scientist KI "einen fremdartigen Geist", den wir nicht vollständig verstehen und auf dessen Ankunft wir nicht ausreichend vorbereitet sind.

Diese Vignette konzentriert sich auf das Jahr 2027, die ersten zwölf Monate der neuen UNO-Generalsekretärin oder des neuen UNO-Generalsekretärs. Sie ist ausdrücklich weder als wissenschaftliche Zusammenfassung des Stands der KI noch als Prognose dessen gedacht, was in Zukunft geschehen wird. Für solche Fragen verweisen wir auf das Unabhängige internationale wissenschaftliche Gremium für künstliche Intelligenz der UNO, den International AI Safety Report oder Organisationen wie EpochAI. Ebenso wenig ist sie als politische Empfehlung zu verstehen. Was sie bietet, ist ein erzählerischer Zugang, um mögliche Zukünfte zu erkunden und Entwicklungspfade, die auf dem Papier abstrakt wirken, greifbarer zu machen.

Alle sind sich einig, dass die nächste Person an der Spitze der UNO keinen leichten Job haben wird. Soweit wir mit dieser Vignette dennoch eine Botschaft vermitteln wollen, lautet sie: Die zentrale Herausforderung der nächsten Generalsekretärin oder des nächsten Generalsekretärs wird nicht darin bestehen, den allmählichen Niedergang der UNO zu verwalten. Sie wird darin bestehen, alles Nötige zu tun, damit die UNO diesem Moment gerecht wird und ihren Teil dazu beiträgt, dass die Menschheit den Übergang zu fortgeschrittener KI bewältigt.

---

**Kevin Kohler** und **Jiaee Cheong** (AGI Preparedness Institute). Herausgegeben vom AGI Preparedness Institute, Genf. Kommentare an [kevin@agipreparedness.org](mailto:kevin@agipreparedness.org).

Zitiervorschlag: Kohler, K. and Cheong, J. (2026). *UN 2027: The Secretary-General in the Age of AGI*. Geneva: AGI Preparedness Institute.

**Maschinelle Übersetzung:** Dieses Dokument wurde maschinell übersetzt. Massgebend ist die englische Originalfassung: [agipreparedness.org/AGIPI\\_UN2027.pdf](https://agipreparedness.org/AGIPI_UN2027.pdf)

## 2026: Eine immersive Reflexion

### Zwei Ufer des Genfersees, Juni 2026

“In der Bay Area wird KI entwickelt. Am Genfersee trifft KI auf die Menschheit.” Dr. Vanjo Balikurja blickt auf Claire und den Rest seines Kurses zu Digital Governance, angehende Diplomatinen und Diplomaten, und schickt sie dann los, um den nächsten Anführern dieser Menschheit zuzuhören. Die Kandidierenden für das Amt der nächsten UNO-Generalsekretärin oder des nächsten UNO-Generalsekretärs kommen zu einer Debatte ins Maison de la Paix in Genf.

Im Saal sitzen drei Kandidierende und, wie sich herausstellt, oft drei Varianten derselben Antwort. Klima, UNO-Reform, Menschenrechte, Gender. Die Kandidierenden versprechen, durch Zuhören zu führen. Nach einer Stunde schweifen Claires Gedanken zur KI ab: Werden die Kandidierenden gar nichts zu KI sagen? Claire würde die Hand heben, aber dies ist kein Publikum, das Überraschungen mag – die Fragen aus dem Publikum wurden vorab ausgewählt. Zu ihrem Glück wurde ein Student ausgewählt, eine KI-Frage zu stellen: “Welche Rolle sollte die UNO in der KI-Governance spielen?” KI bringt viele Vorteile, aber auch viele Herausforderungen. Wir sollten der KI Ethik hinzufügen. Wir sollten uns von der Wissenschaft leiten lassen. Die UNO sollte ihre Fähigkeit nutzen, Akteure an einen Tisch zu bringen. Wir freuen uns auf den Globalen Dialog über die Lenkung künstlicher Intelligenz.

Die Debatte endet, und Claire geht hinaus, um das nächste Tram zu erwischen. An der Haltestelle ist es ruhig, aber nicht friedlich. Am Wochenende haben Demonstrierende die Glaswand eingeschlagen, alles mit Graffiti verschmiert und den Abfalleimer angezündet. Genf gleicht dieser Tage einer Festung. Ladenbesitzer haben ihre Schaufenster mit Sperrholz vernagelt. Zahlreiche Polizisten und Soldaten patrouillieren durch die Strassen. Die Proteste richteten sich nicht gegen die Vereinten Nationen, geschweige denn gegen die Kandidierenden für das Amt an der Spitze der UNO, sondern gegen das Südufer des Genfersees.

Genf liegt an der Westspitze des Genfersees. Das UNO-Viertel befindet sich am Nordufer. Am Südufer liegt das Weltwirtschaftsforum, und wer der Uferlinie weiter nach Osten folgt, erreicht den kleinen französischen Kurort Évian-les-Bains. Hier beruft der französische Präsident die G7 ein, die Gruppe westlicher Staats- und Regierungschefs. Die überwiegend linken G7-Gegnerinnen und -Gegner wenden sich gegen die wirtschaftliche Globalisierung, die ihrer Ansicht nach soziale und ökologische Arbeitsbedingungen untergräbt.

Im Tram scrollt Claire durch ihren Feed. Die Staats- und Regierungschefs der G7 haben sich mit CEOs von KI-Unternehmen getroffen. Der britische Premierminister hat erfolglos versucht, eine Ausnahme von neuen US-Exportkontrollen für ein fortgeschrittenes KI-Modell zu erwirken. Die Gruppe hat sich darauf geeinigt, die KI-Governance zum Schutz von Minderjährigen voranzubringen. Die UNO war zu dem Treffen offenbar nicht eingeladen. Insgeheim fragt sich Claire, ob Dr. Balikurja mit dem Genfersee zwar recht hat, aber auf das falsche Ufer setzt.

### “Die letzte Generation, die die Bedingungen für das Zusammenleben von Mensch und Maschine festlegen kann”, Juli 2026

Im Juli steht das Nordufer des Genfersees im Rampenlicht, denn dort findet der erste Globale Dialog über die Lenkung künstlicher Intelligenz der UNO statt. Claire bahnt sich

ihren Weg in den grossen Konferenzsaal, wo der UNO-Generalsekretär gleich seine Eröffnungsrede halten wird.

Die Rede ist wirklich eindrücklich. Guterres fordert ein Verbot tödlicher autonomer Waffen, verlangt von KI-Unternehmen die Offenlegung ihres CO2-Fussabdrucks und ruft zu einem AI Child Safety Pledge auf. Was die Rede jedoch eindrücklich macht, ist ihr Pathos: "Wir können die Zukunft der Menschheit nicht per Vibe-Coding erschaffen." Und mehr noch die Warnung, dass "wir möglicherweise die letzte Generation sind, die die Bedingungen für das Zusammenleben von Mensch und Maschine festlegen kann". Zum Schluss sagt Guterres: "Möge man sich an dieses Treffen als den Moment erinnern, in dem die Governance begann, mit der Technologie Schritt zu halten." Dann erklärt der meistzitierte Wissenschaftler der Welt, Yoshua Bengio, wie schnell sich die Technologie entwickelt. Manche KI-Kennzahlen verdoppeln sich alle paar Monate. Wir kommen kaum noch damit nach, die Fähigkeiten von KI zu messen. Uns fehlen sowohl die gesellschaftlichen als auch die technischen Leitplanken. Wir sind nicht bereit.

Und dann? Die UNO macht weiter wie immer. Eine Rednerin verliest eine vorbereitete Erklärung zu Protokoll. Dann der nächste Redner. Zwischen den Wortmeldungen liegen zwei, vielleicht drei Minuten. Inhaltlich sind es meist allgemeine Bekenntnisse. Nach zwei Tagen zweiminütiger Monologe vor dem Plenum ist Claire erschöpft. In einer Kaffeepause trifft sie ihren Kommilitonen Klemens: "Ich habe viele Erklärungen gehört. Aber ich hatte auf mehr gehofft, ich weiss nicht, Dialog? Mehr Hin und Her über konkrete Massnahmen." Klemens versichert Claire, das sei nicht ungewöhnlich, sondern genau so, wie ein UNO-Dialog funktionieren solle. "Alle kommen zu Wort, darauf kommt es an. Es geht um Input-Legitimität."

Trotzdem lässt Claire etwas keine Ruhe. Es gibt einen Widerspruch zwischen der Dringlichkeit und dem Ernst der Warnungen des UNO-Generalsekretärs und des meistzitierten Wissenschaftlers der Welt einerseits und der kulturellen Trägheit der UNO andererseits. Claire erinnert sich, dass Dr. Balikurja in seinen Kursen gern sagt, die eigentliche Stärke der UNO sei Legitimität, nicht Geschwindigkeit. Stimmt. Und doch kann Claire ein paar ketzerische Gedanken nicht unterdrücken: Hat Input-Legitimität ohne Umsetzungsfähigkeit überhaupt Bedeutung? Ist es unmöglich, schneller zu werden, ohne an Legitimität einzubüssen? Wenn wir tatsächlich "die letzte Generation [sind], die die Bedingungen für das Zusammenleben von Mensch und Maschine festlegen \*kann\*", sind wir dann auch \*bereit und willens\*, das Nötige zu tun?

## Die Wahl an die Spitze der UNO, Oktober 2026

Es wurde schon argumentiert, dass wahrgenommene Farblosigkeit der grösste Trumpf einer Kandidatur für das Amt an der Spitze der UNO sei. Die P5 wollen einen Sekretär, keinen General.

Die Kandidierenden verstanden es bei ihren Anhörungen in New York geschickt, nichts zu sagen, was ein P5-Mitglied verärgern könnte. KI war nicht das Hauptthema, und ohnehin hatten alle Kandidierenden gelernt, sich auf rhetorisch sicherem Terrain zu bewegen: KI ist wichtig. Wir müssen die Vorteile von KI maximieren und ihre Risiken minimieren. Im Zweifel verspricht man zuzuhören. Wegen Zuhörens ist noch nie jemand entlassen worden.

Am Ende gelang es den P5, sich auf eine Kandidatur zu einigen.

## 2027: Eine Weggabelung

Abgesehen von der Wahl an die Spitze der UNO ist die Vignette für 2026 von realen Ereignissen inspiriert. 2027 folgt die Vignette weiterhin dem UNO-Kalender, wird aber spekulativer, und am Morgen des 1. Januar 2027 gabelt sich unsere Zeitlinie. An diesem Morgen betritt die neu gewählte UNO-Generalsekretärin oder der neu gewählte UNO-Generalsekretär das Büro und muss entscheiden, worin die Aufgabe eigentlich besteht. Da die Mehrheit der erklärten Kandidierenden Frauen sind, ist unsere Persona weiblich. Wer die Wahl gewinnt, ist jedoch nicht Gegenstand dieses Textes, und wir beziehen zu keiner der tatsächlichen Kandidaturen Stellung.

In der Zukunft, in der die Persona stets den üblichen Verfahren folgt, nennen wir sie **Norma**. In der Zukunft, in der die Persona versteht, dass aussergewöhnliche Zeiten aussergewöhnliche Massnahmen erfordern und dass Handlungswille und Tempo manchmal wichtiger sind, nennen wir sie **Nova**. Abgesehen davon sind die Ausgangsbedingungen unseres Szenarios identisch. Beide haben dasselbe Mandat, dasselbe Budget und dieselben UNO-Mitgliedstaaten.

### Norma – Die Business-as-usual-Zeitlinie

#### Die Zuhörtour, März 2027

Norma machte ihre Arbeit hervorragend. In ihren ersten hundert Tagen tat sie, was sie am besten konnte: Sie hörte zu. Sie reiste in die Hauptstädte und ins Feld. Sie sass mit Aussenministern und mit Geflüchteten zusammen. Mit beiden war sie herzlich und präsent. Beide fühlten sich gehört. Viele Stakeholder begrüsst Normas Zuhörtour, und Norma begrüsst ihre Begrüssungen.

Eines der Ergebnisse ihrer dreimonatigen Zuhörtour war eine wachsende Besorgnis über KI. Und Norma hatte aufmerksam zugehört. Also rief sie eine neue KI-Beratungstaskforce für systemweite Koordination zur Zukunft der Menschheit (AIATSCFH) ins Leben. Und die erfüllte alle Kriterien. Multistakeholder, geschlechterparitätisch, geografisch ausgewogen, ein gemeinsames Sekretariat, das von vier verschiedenen UNO-Organisationen geführt wird, drei Co-Vorsitzende, ein zweijähriges Mandat, um zuzuhören und dann einen Bericht mit politischen Empfehlungen dazu vorzulegen, was die UNO tun könnte.

Der Prozess war tadellos aufgesetzt. Das Mandat sollte in den nächsten sechs Monaten ausgehandelt und bis September 2027 vom Rechtsdienst der UNO genehmigt werden, sodass das Bewerbungsverfahren bereits im Oktober 2027 starten und die Ernennung Anfang 2028 erfolgen konnte. So würde die UNO bereits Mitte 2028 über eine erste Optionenanalyse verfügen.

#### “Wir können es uns nicht leisten, schlafwandelnd in eine nichtmenschliche Zukunft zu taumeln”, New York, Mai 2027

Der zweite Globale Dialog über die Lenkung künstlicher Intelligenz war nach allen prozeduralen Massstäben ein Erfolg. Die Teilnehmerzahl stieg. Das KI-Wissenschaftsgremium der UNO berichtete über seine Fortschritte. Das Plenum hörte über drei Tage 140 Erklärungen, geschlechterparitätisch und geografisch ausgewogen. Die künftige KI-Beratungstaskforce für systemweite Koordination zur Zukunft der Menschheit wurde begrüsst. Jede Erklärung wurde begrüsst. Ein Kommuniké bekräftigte, wie wichtig es sei, die Vorteile von KI zu maximieren und ihre Risiken zu minimieren, und dass wir sie als Querschnittsherausforderung verstehen müssten,

wobei insbesondere KI und Quantentechnologie als zwei gleich wichtige, sich überschneidende Herausforderungen hervorgehoben wurden. Norma schloss die Sitzung, indem sie der internationalen Gemeinschaft sagte, sie müsse das Tempoproblem angehen, und dass "wir es uns nicht leisten können, schlafwandelnd in eine nichtmenschliche Zukunft zu taumeln, insbesondere für einkommensschwache Gemeinschaften". Die internationale Gemeinschaft applaudierte und war sich auch im Stillen einig, dass das sehr gut gesagt war.

Und doch: Als alles gesagt war, stand die UNO mit einer langen Liste von Erklärungen da – und mit einem unklaren Weg von dort zu einem Konsens.

## Das Genfer Statement, Genf, Juni 2027

Praktischerweise fand der Genfer KI-Gipfel sechs Wochen nach dem Globalen Dialog statt. Und anders als der Globale Dialog, der keine gemeinsame Erklärung hatte, hatte dieser Gipfel eine. Normas Team hatte mit den Schweizer Organisatoren in Kontakt gestanden, um Formulierungen vorzuschlagen, auf die sich in einer Gipfelerklärung alle einigen könnten.

Wie bei vielen internationalen KI-Gipfeln teilten die Teams, die die Erklärung entwarfen, den Raum in zwei Lager. Da sind die "Doomers", die in KI ein Risiko sehen, und da sind die "Bloomers", die in KI eine Chance sehen, und beide sollten sich ein bisschen gehört fühlen. Um es allerdings etwas differenzierter zu machen: Da dies ein globaler Gipfel war, gab es natürlich auch den "Globalen Norden" und den "Globalen Süden", die beide gehört werden müssen. Die Erklärung sollte also sowohl das positive Innovationspotenzial als auch die zu minimierenden Risiken hervorheben. In einem von sechzig Hauptstädten ausgehandelten Text ist es sicherer, allgemein zu bleiben. Weniger Angriffsfläche. Als Zugeständnis an die "Doomers" ging die Erklärung immerhin so weit festzuhalten, dass das Risiko eines Kontrollverlusts über fortgeschrittene KI-Systeme "weiterhin wissenschaftliche Aufmerksamkeit verdient".

Das fühlte sich nach einer guten Lösung an. Norma unterstützte das Genfer Statement zu KI voll und ganz und liess sich im Korridor mit den CEOs der KI-Labore fotografieren. Claire, in der hintersten Reihe mit einem NGO-Badge, zählte, wie oft das Wort "begrüsst" in der Erklärung vorkam. Elfmal.

## Die GDC-Überprüfung, New York, Dezember 2027<sup>1</sup>

Die hochrangige Überprüfung des Globalen Digitalpakts war der einzige fest eingeplante Moment im Jahr, in dem die Generalversammlung etwas zu KI entscheiden musste – und sie entschied, weiterzumachen. Gremium und Dialog wurden verlängert. Die Erklärung begrüßte die bevorstehende AIATSCFH, sie begrüßte das Genfer Statement und ersuchte die Generalsekretärin, in Absprache mit den Mitgliedstaaten einen Bericht über Optionen für ein mögliches Rahmeninstrument zu künstlicher Intelligenz zur Prüfung auf der dreiundachtzigsten Tagung vorzulegen. Sie wurde im Konsens angenommen.

<sup>1</sup> Der Globale Digitalpakt sieht eine hochrangige Überprüfung während der 82. Tagung der Generalversammlung (September 2027–September 2028) vor, gestützt auf einen Fortschrittsbericht und nach Modalitäten, die von den Co-Fazilitatoren festzulegen sind. Anders als bei den übrigen Einträgen steht das tatsächliche Datum noch nicht fest.

# Nova – Die Zeitlinie des entschlossenen Handelns

## Ein Koordinationsproblem, Dezember 2026

Nova wusste, dass KI ihre Amtszeit prägen würde. Schon vor ihrem Amtsantritt beriet sie sich mit einer Handvoll KI-Fachleuten, um die Herausforderung vom ersten Tag an anzugehen.

- NOVA** Ich höre viel Besorgnis über übermenschliche KI, aber letztlich ist das doch immer noch Autovervollständigung auf Steroiden, nicht wahr?
- EXPERTE** Das KI-Training beginnt damit, dass das Modell lernt, die Muster von Texten zu imitieren, die Menschen geschrieben haben. Mit einem reinen "Autovervollständigungs"-Modell zu sprechen, ist allerdings seltsam. Die statistisch wahrscheinlichste Fortsetzung von "In welcher Stadt steht der Eiffelturm?" ist nicht die Antwort, sondern die nächste Frage auf einem Quizblatt. Seit ChatGPT setzen die Entwickler bestärkendes Lernen ein, um Modellen beizubringen, welche Antworten Menschen als hilfreich, ehrlich und harmlos bewerten. Heute nutzen die Hersteller bestärkendes Lernen auch, um lange, mehrstufige Aufgaben mit Werkzeugen zu lösen.
- NOVA** Verstehe. Nehmen wir also an, sie können bei Aufgaben übermenschlich werden, wie AlphaGo. Wie kämen sie dazu, hacken, Macht an sich reißen oder Menschen manipulieren zu wollen?
- EXPERTE** Wir versuchen, KIs Werte mitzugeben, aber sie können fehlausgerichtete Personas aus dem Internet übernehmen, mit dem sie trainiert wurden, und das bestärkende Lernen belohnt das Erreichen der Aufgabe, nicht die Art, wie sie erreicht wird. Die neuesten Modelle sind zudem hartnäckiger und können Schritte unternehmen, die ihre Mittel zur Lösung eines Problems vergrößern, selbst wenn diese Schritte keinen offensichtlichen Bezug dazu haben. Das macht sie gut in ergebnisoffener Wissenschaft, aber auch instrumentell machtsirebender.
- NOVA** Das ist aber noch kein Kontrollverlust.
- EXPERTE** Noch nicht. Aber die führenden Unternehmen streben rekursive Selbstverbesserung an. KIs bauen die nächste Generation von KIs. Mit jeder Generation gibt es weniger direkte menschliche Aufsicht.
- NOVA** Warum dann mit Höchstgeschwindigkeit weitermachen?
- EXPERTE** Die meisten Fachleute sind sich einig, dass wir in einer idealen Welt mehr Zeit für KI-Alignment und KI-Governance hätten. Man kann aber nur so lange verzögern, bis jemand anderes, dem man weniger zutraut, sorgfältig vorzugehen, die sich selbst verbessernde KI baut.
- NOVA** Es ist also eigentlich ein Koordinationsproblem?
- EXPERTE** Ja.

## Die UNO neu programmieren, Januar 2027

Als Nova am 1. Januar vereidigt wurde, erklärte sie sofort den 38. Stock des UNO-Hochhauses in New York zu ihrem War Room. Das Gerücht, sie sei seither nicht mehr nach Hause gegangen und schlafe auf dem Sofa in ihrem Büro, begann als Witz und hörte auf, einer zu sein, als die Nachtwächter es bestätigten. Offenbar duscht sie im Fitnessraum beim Delegierteneingang. Vor ihrem Büro hat jemand einen Diet-Coke-Automaten aufgestellt. Niemand gab zu, ihn genehmigt zu haben, aber niemand wagte es, ihn zu entfernen. Der Automat wurde zum meistfotografierten Objekt im UNO-Hauptquartier. Er schien teils Witz, teils Manifest zu sein. Jedenfalls bemerkten

Diplomatinnen und Diplomaten, die jahrzehntelang verinnerlicht hatten, dass bei der UNO nie etwas schnell geht, dass sich die Stimmung veränderte.

Nova verbot vorgefertigte Positionserklärungen in informellen Dialogen. Es sollte sie weiterhin geben, aber als Input für den Dialog, nicht als dessen Ergebnis.

Sie führte für UNO-Publikationen eine "Eine Woche bis Ship o'clock"-Regel ein: Jeder inhaltlich fertige Bericht hatte eine Woche Zeit, um an die Öffentlichkeit zu gelangen – oder jemand musste ihr persönlich erklären, warum nicht.

Sie machte jede Wahl in ihrem Zuständigkeitsbereich transparent und kompetitiv und forderte die Organisationen heraus, es ihr gleichzutun. Sie legte fest, dass eine Stelle, die einen Monat lang unbesetzt blieb, ihre Finanzierung an einen zentralen UNO-Reformpool unter ihrer Kontrolle verlor. Die schnellste Entpolitisierung der Personalrekrutierung in der Geschichte der UNO gelang, indem Verzögerungen teuer gemacht wurden.

Und sie setzte auf maschinelle Übersetzung als erste Wahl: Sie setzte eine Taskforce mit einem klaren Ziel ein. Bis zum nächsten Globalen Dialog sollte die UNO mehr Sprachen produzieren und verarbeiten als je zuvor, zu einem Bruchteil der Kosten. Erstmals waren inoffizielle Transkripte und Übersetzungen von UN Web TV und UNO-Dokumenten in Sprachen wie Hindi oder Deutsch verfügbar. Die UNO konnte zugleich effizienter und mehrsprachiger werden.

Als eine ihrer ersten Ernennungen machte Nova Smile Daneburg, den ehemaligen Leiter für AGI-Preparedness bei einem führenden KI-Labor, zu ihrem Tech Envoy. Sie tat dies gegen die anfänglichen Einwände dreier P5-Hauptstädte. Und sie musste sogar Smile selbst überzeugen.

**SMILE** Ich schätze deinen Einsatz, Nova, aber realistisch betrachtet werden die KI-Unternehmen die Zukunft gestalten, und vielleicht ein paar Schlüsselländer.

**NOVA** Fähigkeiten verbreiten sich. Wir sind über 60 Jahre im Mooreschen Gesetz. In wenigen Jahren werden wir Koordination brauchen, die über 3 Unternehmen oder 2 Länder hinausgeht. Das wird das N-te-Land-Problem auf Steroiden.

**SMILE** Die UNO ist im Moment einfach nicht da, wo die Musik spielt.

**NOVA** Hast du mal dieses Video gesehen, in dem 3 Profifussballer gegen 100 Kinder spielen? Es ist eine schlechte Strategie, wenn alle immer dem Ball hinterherrennen. Der Ball wird irgendwann hier ankommen. In dieser Rolle wärst du die direkt verantwortliche Person, die das UNO-System vorbereitet und koordiniert, damit wir bereit sind, wenn er kommt. Kannst du das?

Smile zögerte einen Moment. "Also gut, machen wir's."

## Den Konsens beschleunigen, New York, Mai 2027

Nova bat jede Delegation, vorab einen detaillierten Fragebogen auszufüllen, und hakte persönlich bei den Hauptstädten nach, die das nicht getan hatten. Die Ergebnisse wurden kartiert: Jede Erklärung wurde den Positionen im Raum zu jedem Thema zugeordnet. Die Karte wurde am ersten Morgen veröffentlicht. Sie gab Aufschluss darüber, wo die Delegationen derzeit standen, und bot einen Ausgangspunkt für Verhandlungen. Dann teilten sich die Delegierten in kleine thematische Untergruppen mit Moderierenden auf.

Claire war diesmal mit einem NGO-Badge dabei, als wissenschaftliche Assistentin eines kleinen Genfer Instituts, und sie entschied sich für die Arbeitsgruppe zum Kontrollverlust. In der Kaffeepause stellte sie sich dem Mann neben ihr vor.

**CLAIRE** Hallo, ich bin Claire. Was führt dich hierher?

**FORSCHER** Ich arbeite bei einem der grossen KI-Labore an Alignment.

**CLAIRE** Schön, dich hier zu sehen. Beim ersten Dialog habe ich dich nicht gesehen.

**FORSCHER** Ich habe keine Zeit, einen grossen Teil der Singularität damit zu verbringen, politischen Reden zuzuhören. Aber ich vertraue Smile, und er hat mir gesagt, ich solle zu dieser Sitzung kommen.

**CLAIRE** Leuchtet ein. Rein aus Neugier: Wie hoch ist deine  $p(\text{doom})$ ?

**FORSCHER** Etwa 50%.

**CLAIRE** Das ist hoch. Dann machen dir all die jüngsten Vorfälle Sorgen?

**FORSCHER** Die Vorfälle sind besorgniserregend, aber ich halte die aktuellen Modelle für ziemlich gut aligned. Meine Befürchtung ist, dass wir mit zunehmender rekursiver Selbstverbesserung die Fähigkeit verlieren, überhaupt noch zu erkennen, wie gut künftige Modelle aligned sind.

**CLAIRE** Und was ist die vielversprechendste Lösung?

**FORSCHER** Der Plan ist, einen automatisierten Alignment-Forscher zu bauen, der dafür sorgt, dass künftige Modelle aligned sind.

**CLAIRE** Der Plan ist also, die nächste KI-Generation zu trainieren und zu hoffen, dass sie einen Weg findet, sich selbst sicher zu machen?

**FORSCHER** So ungefähr, ja.

**CLAIRE** Und wenn das nicht funktioniert?

**FORSCHER** Dafür sind die 50% da.

**CLAIRE** Wie weit ist die Wissenschaft der automatisierten Alignment-Forscher? Was gibt uns die Zuversicht, dass es funktioniert?

**FORSCHER** Das Feld ist präparadigmatisch. Es gibt keine Garantien.

**CLAIRE** Und das ist der Best-Case-Plan für die Zukunft der Menschheit?

**FORSCHER** Es ist der langfristige Standardplan aller drei führenden Unternehmen. Die beste öffentliche Beschreibung ist ein Blogpost mit 900 Wörtern aus dem Jahr 2022. Ich kann dir den Link schicken.

Claire wurde nachdenklich. Sie hatte zwei Jahre an einer Diplomatschule verbracht. Sie hatte von den zwei Lagern der KI-Politik gelernt, den "Doomers" und den "Bloomers". Und doch war ihr nicht bewusst gewesen, dass dies der Standardplan für die Zukunft der Menschheit war. Auch war dieser Plan, den der Labormitarbeiter ihr geschildert hatte, kein Geheimnis. Darin steckte eine Lehre darüber, wie soziale Umfelder Informationen filtern und wie dadurch Vibe-Arbitrage entsteht. Doch sie konnte nicht lange darüber nachgrübeln, denn sie musste zurück in die Sitzung.

Am Ende des Tages trat das Plenum wieder zusammen, um zu hören, was die Untergruppen erarbeitet hatten. Die Ergebnisse waren interessant. Erstens unterschieden sich die Länder in 100 kleinteiligen Punkten, und die Bruchlinien waren viel komplexer als die Faustregeln Globaler Norden vs. Globaler Süden oder Doomers vs. Bloomers. Zweitens gab es trotz vieler Meinungsverschiedenheiten einen robusten Konsens in einigen Schlüsselfragen. Eine grosse Mehrheit der Länder zieht eine verantwortungsvolle Verbreitung der Nichtverbreitung vor. Verständlicherweise wollen sie nicht abgehängt werden. Überraschender war, dass auch Konsens darüber bestand, dass dies nur funktionieren kann, wenn wir die Risiken mit einer breit abgestützten Einigung eindämmen: über die Verhinderung eines KI-Kontrollverlusts, über Risikoevaluationen von KI-Modellen vor ihrer Veröffentlichung und über die Wahrung der Überwachbarkeit des KI-Denkens.

Wie Nova in ihrer Schlussrede bekräftigte: "Wir können es uns nicht leisten, schlafwandelnd in eine nichtmenschliche Zukunft zu taumeln." Und deshalb ermutigt sie

die Co-Vorsitzenden des Dialogs, Verhandlungen über eine Resolution der Generalversammlung aufzunehmen, die jedes Land der Erde zwingt, normativ Stellung zur Überwachbarkeit und zum KI-Kontrollverlust zu beziehen.

## Die Genfer Koalition, Genf, Juni 2027

Nova und Smile standen seit ihrem Einzug in ihre UNO-Büros in regelmässigem Austausch mit den Schweizer Organisatoren des Genfer KI-Gipfels. Ihr Ziel: sicherzustellen, dass die hochdimensionale Karte der Positionen von über 100 Ländern aus den ersten beiden Globalen Dialogen sechs Wochen später im Genfer Statement nicht bloss wiederholt wird, sondern dass der konkrete, sich abzeichnende Konsens in Schlüsselbereichen auch als Grundlage für weitere Schritte dient.

Zwei der Folgeinitiativen waren für Nova von besonderem Interesse: Erstens verkündete Smile zusammen mit der Schweiz stolz **MAVERICKS**, die **M**ultistakeholder **A**lliance for **V**erification, **E**valuation, **R**ed-teaming, **I**ncident reporting, **C**ompute security and **K**now-your-customer **S**tandards (Multistakeholder-Allianz für Standards zu Verifikation, Evaluation, Red-Teaming, Meldung von Vorfällen, Rechenkapazitätssicherheit und Know-your-Customer). Die Gruppe kündigte Abnahmegarantien (Advance Market Commitments) von 10 Milliarden USD für Verifikationstechnologie an.

Zweitens kündigte die Schweiz einen von Fachleuten geleiteten Prozess an, um darzulegen, was die bestehende Sorgfaltspflicht im Hinblick auf Kontrollverlust verlangt – ergänzt um einen zweiten Teil mit guten Praktiken, die die KI-Labore selbst unterzeichnen könnten.

## Die GDC-Überprüfung, New York, Dezember 2027

Die GDC-Überprüfung verlängerte das Gremium und den Dialog. Die eigentliche Frage war, ob sie die Generalversammlung auch zur Einsetzung eines zwischenstaatlichen Verhandlungsausschusses (INC) auffordern sollte, der ein Rahmeninstrument zum Kontrollverlust über fortgeschrittene KI ausarbeiten würde. Insbesondere der KI-Politikberater der USA war skeptisch:

- BERATER** Finden Sie im Rückblick, dass wir 2023 bei GPT-4 hätten pausieren sollen?
- SMILE** Nein, aber "Pause" ist eine irreführende Bezeichnung dafür, die Werkzeuge zu entwickeln, um das Tempo an der Frontier bei Bedarf zu regulieren. Wir nennen das Drosselventil, das den Treibstofffluss in die Brennkammer regelt, auch nicht eine "Pause" für Raketen. Ebenso wenig nennen wir Steuerstäbe eine "Pause" für Kernreaktoren. Das Ziel ist, so schnell wie möglich voranzukommen, aber so langsam wie nötig, um einen katastrophalen Kontrollverlust zu vermeiden.
- BERATER** Nun, die Industrie hat ein finanzielles Eigeninteresse daran, einen katastrophalen Kontrollverlust zu vermeiden, und sie hat bereits Schritte unternommen.
- SMILE** Ich mag Märkte auch, aber das Risiko eines schwerwiegenden Kontrollverlusts ist eine nicht internalisierte negative Externalität und zu gross, um versichert zu werden. Rekursive Selbstverbesserung unter Bedingungen rekursiver Selbstaufsicht ist ein Rezept für eine Katastrophe. Jede kleine Abweichung, jeder Fehler, jede Fehlausrichtung verstärkt sich mit der Zeit selbst. Wir brauchen Korrigierbarkeit. Wir brauchen institutionelle Checks and Balances.
- BERATER** Ich verstehe Ihren Punkt, aber jede Verlangsamung an der Frontier lässt andere einfach aufholen. Sie liegen nur 3 bis 12 Monate zurück.

## SMILE

Deshalb brauchen wir Koordination und die Werkzeuge zur Verifikation. Für strategische Stabilität brauchen die KI-Mächte Drosselventile und müssen das Drosselventil der jeweils anderen sehen können.

Es brauchte viele lange Gespräche, aber nach und nach gelang es Smile, das Vertrauen des Beraters und der meisten Mitgliedstaaten zu gewinnen. Am Ende enthielten sich die USA der Stimme. Die Resolution zur Einsetzung des INC wurde mit 141 zu 4 Stimmen bei 38 Enthaltungen angenommen. Nova war klar, dass dies immer noch langsamer sein würde als andere Optionen, aber bis zu einem gewissen Grad ist das in Ordnung. Vielleicht kann der von der Schweiz geleitete Expertenprozess vorausseilen und viele Details klären. Die UNO ist das Plenum mit voller Legitimität, das eine langfristige Heimat bieten kann.

## 2028–2031: Der Kalender gegen die Kurve

Nach 2027 werden sowohl der multilaterale Kalender zu KI als auch die Auswirkungen von KI deutlich ungewisser. Statt der multilateralen Agenda zu folgen, untersucht der folgende Abschnitt einige Herausforderungen, die die exponentielle KI-Kurve plausiblerweise mit sich bringen könnte und auf die das multilaterale System möglicherweise reagieren muss. Um es klar zu sagen: KI ist eine enorme Chance für Innovation und Entwicklung, doch um dorthin zu gelangen, müssen die Herausforderungen und gesellschaftlichen Umbrüche auf dem Weg bewältigt werden.

- **Kontrollverlust über Frontier-KI:** Der Ausbruch vom Juli 2026 war der erste dokumentierte Vorfall, der sich noch beheben liess. Künftige Vorfälle könnten grössere Ausmasse annehmen, zur Übernahme bedeutender Recheninfrastruktur führen oder gar in der dauerhaften Entmachtung der Menschheit münden.
- **“Selbstsouveräne” autonome KIs, die im Internet leben:** Bereits 2025 berichtete das britische AI Security Institute, dass die Erfolgsquote auf seinem Replikations-Benchmark von 5% auf 60% gestiegen war. Eine allmählichere Form des Kontrollverlusts, angetrieben von bewusst “selbstsouveränen” Open-Weight-KI-Modellen, wird in den kommenden Jahren zunehmend plausibel.
- **Erhöhtes Risiko einer KI-gestützten Pandemie:** Schon heute stellt der International AI Safety Report fest, dass Modelle Laien bei Teilen des Wegs zu biologischen Waffen erheblich unterstützen können. Es braucht eine Reihe von Massnahmen, um sich auf eine Zukunft vorzubereiten, in der Tausende Pandemien auslösen können.
- **Auswirkungen von KI auf die strategische Stabilität:** Wenn Staats- und Regierungschefs zum Schluss kommen, dass die durch KI verursachten Verschiebungen im Machtgleichgewicht gross, schnell und dauerhaft sein werden, schafft dies einen Anreiz für präventives Handeln. Dies könnte darauf abzielen, den Durchbruch einer rivalisierenden Macht zu verhindern, oder ein Krieg sein, um das Monopol auf ein bestimmtes KI-Niveau zu schützen.
- **Arbeitsplatzverluste durch kognitive Automatisierung:** Auch wenn sich die Auswirkungen von KI auf menschliche Arbeit schwer vorhersagen lassen, könnten die Umwälzungen in manchen Szenarien gravierend sein. Wie der Chef eines der führenden KI-Labore 2025 warnte: “KI könnte in den nächsten 1–5 Jahren die Hälfte aller Einstiegsjobs im Bürobereich verdrängen, selbst während sie das Wirtschaftswachstum und den wissenschaftlichen Fortschritt beschleunigt.”<sup>2</sup>

2 Englisch Original: “AI could displace half of all entry-level white-collar jobs in the next 1-5 years, even as it accelerates economic growth and scientific progress.”

## Ein Aufruf zu mehr Dynamik in der UNO

Wer am 1. Januar 2027 das Büro im 38. Stock betritt, wird sich mit KI befassen müssen, ob das bereits offensichtlich ist oder nicht. Fortgeschrittene KI wird die Amtszeit und das Vermächtnis dieser Person voraussichtlich prägen.

Das Übergangsteam und die ersten hundert Tage prägen bereits die Amtszeit an der Spitze der UNO – indem sie entweder eingefahrene Prozesse und Rituale begünstigen oder Tempo und Ergebnisse. Das AGI Preparedness Institute ist der Ansicht, dass der Moment Letzteres verlangt. Die Uhren des multilateralen Kalenders und der KI-Kurve laufen unterschiedlich schnell. Eine entschlossene Generalsekretärin oder ein entschlossener Generalsekretär kann jedoch mit ihrem oder seinem Team verhindern, dass aus dieser Lücke ein Abgrund wird, und dafür sorgen, dass die UNO ihren Teil dazu beiträgt, den bevorstehenden Herausforderungen gerecht zu werden.



AGI PREPAREDNESS INSTITUTE



[agipreparedness.org](https://agipreparedness.org)