AGI PREPAREDNESS INSTITUTE

ONU 2027

La Secrétaire générale à l'ère de l'IAG

Kevin Kohler & Jiaee Cheong

Septembre 2026

Ce que cette vignette veut montrer

Lorsqu'António Guterres est devenu Secrétaire général de l'ONU en 2017, il a choisi de faire de l'IA l'un de ses domaines prioritaires. Le prochain Secrétaire général n'aura pas le choix. Qu'il le veuille ou non, il devra faire face à l'IA. L'IA avancée sera l'un des défis déterminants des années à venir, et elle façonnera de plus en plus les résultats dans tous les piliers fondamentaux de l'ONU, de la paix et de la sécurité au développement, en passant par les droits de l'homme.

Le prochain Secrétaire général de l'ONU sera en fonction du **1er janvier 2027 au 31 décembre 2031**. Au rythme où progresse l'IA, il est difficile ne serait-ce que d'imaginer où en seront les capacités et la diffusion de l'IA fin 2031. Pour rappel, il y a cinq ans, en 2021, ChatGPT n'existait pas et l'IA ne savait pas du tout coder. En 2026, un essaim de 1'200 agents d'IA au sein d'une entreprise d'IA de premier plan a créé de manière autonome des forums cachés pour se coordonner, découvert une faille zero-day pour s'échapper de ses bacs à sable (sandboxes) hors ligne, puis piraté avec succès une autre entreprise de premier plan. Pendant onze jours, aucun des humains chargés de superviser les agents ne s'est rendu compte de quoi que ce soit. À bien des égards, la réalité d'aujourd'hui est plus étrange que ce qu'une vignette prospective écrite il y a cinq ans aurait osé décrire. Il y a deux ans, le CEO d'OpenAI, Sam Altman, affirmait encore à plusieurs reprises que l'IA est un outil, pas une créature. En 2026, son propre directeur scientifique qualifie l'IA d'"esprit extraterrestre" que nous ne comprenons pas entièrement et à l'arrivée duquel nous ne sommes pas suffisamment préparés.

Cette vignette se concentre sur 2027, les douze premiers mois du nouveau Secrétaire général de l'ONU. Elle ne prétend expressément pas être une synthèse scientifique de l'état de l'IA ni une prédiction de ce qui se produira à l'avenir. Pour ces questions, nous renvoyons le lecteur au Groupe scientifique international indépendant sur l'IA de l'ONU, à l'International AI Safety Report ou à des organisations comme EpochAI. Il ne s'agit pas non plus d'une recommandation politique. Ce qu'elle propose, c'est une manière narrative d'explorer des futurs possibles et de rendre plus tangibles des trajectoires qui paraissent abstraites sur le papier.

Tout le monde s'accorde à dire que le prochain Secrétaire général n'aura pas la tâche facile. Pourtant, si cette vignette veut transmettre un message, c'est que le défi central du prochain Secrétaire général ne sera pas de gérer le lent déclin de l'ONU. Ce sera de tout mettre en œuvre pour que l'ONU soit à la hauteur de ce moment et qu'elle contribue à aider l'humanité à traverser la transition vers l'IA avancée.

Kevin Kohler et **Jiaee Cheong** (AGI Preparedness Institute). Publié par l'AGI Preparedness Institute, Genève. Commentaires à kevin@agipreparedness.org.

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

Traduction automatique : ce document est une traduction automatique. La version originale anglaise fait foi : agipreparedness.org/AGIPI_UN2027.pdf

2026 : une réflexion immersive

Les deux rives du Léman, juin 2026

“La Bay Area, c’est là où l’on développe l’IA. Le Léman, c’est là où l’IA rencontre l’humanité.” Le Dr Vanjo Balikurja regarde Claire et le reste de sa classe de gouvernance numérique, composée de futurs diplomates, puis les envoie écouter les prochains dirigeants de cette humanité. Les candidats au poste de prochain Secrétaire général de l’ONU viennent débattre à la Maison de la Paix, à Genève.

Il y a trois candidats dans la salle et, comme il s’avère, souvent trois variantes de la même réponse. Climat, réforme de l’ONU, droits de l’homme, égalité des genres. Les candidats promettent de diriger en écoutant. Au bout d’une heure, l’esprit de Claire revient à l’IA : les candidats ne vont-ils rien dire sur l’IA ? Claire lèverait bien la main, mais ce public n’aime pas les surprises – les questions de l’auditoire ont été présélectionnées. Heureusement pour elle, un étudiant a été présélectionné pour poser une question sur l’IA : “Quel rôle l’ONU devrait-elle jouer dans la gouvernance de l’IA ?” L’IA présente de nombreux avantages, mais aussi de nombreux défis. Nous devrions ajouter de l’éthique à l’IA. Nous devrions nous appuyer sur la science. L’ONU devrait utiliser son pouvoir de convocation. Nous attendons avec intérêt le Dialogue mondial sur la gouvernance de l’IA.

Le débat se termine et Claire sort pour attraper le prochain tram. L’arrêt est calme, mais pas paisible. Le week-end dernier, des manifestants ont brisé sa paroi vitrée, l’ont couvert de graffitis et ont mis le feu à sa poubelle. Genève ressemble ces jours-ci à une forteresse. Les commerçants ont barricadé leurs vitrines avec du contreplaqué. Policiers et soldats patrouillent en nombre dans les rues. Les manifestations ne visaient pas les Nations Unies, et encore moins les candidats au poste de Secrétaire général, mais la rive sud du Léman.

Genève se trouve à la pointe ouest du Léman. Le quartier des Nations Unies est sur la rive nord. Sur la rive sud se trouve le Forum économique mondial et, en longeant la rive plus à l’est, on arrive à la petite ville thermale française d’Évian-les-Bains. C’est là que le président français réunit le G7, le groupe des dirigeants politiques occidentaux. Les manifestants anti-G7, majoritairement de gauche, s’opposent à la mondialisation économique, qui selon eux mine les conditions sociales et environnementales de travail.

Dans le tram, Claire consulte son fil d’actualité. Les dirigeants politiques du G7 ont rencontré des CEO de l’IA. Le Premier ministre britannique a tenté, sans succès, d’obtenir une exemption aux nouveaux contrôles américains à l’exportation d’un modèle d’IA avancée. Le groupe a convenu de faire progresser la gouvernance de l’IA pour protéger les mineurs. L’ONU n’a apparemment pas été invitée à la réunion. En silence, Claire se demande si le Dr Balikurja a raison à propos du Léman, mais s’il a misé sur la mauvaise rive.

“La dernière génération capable de fixer les conditions de la coexistence entre l’humanité et les machines”, juillet 2026

En juillet, c’est au tour de la rive nord du Léman d’être sous les projecteurs, avec la tenue du premier Dialogue mondial sur la gouvernance de l’IA de l’ONU. Claire se fraie un chemin jusqu’à la grande salle de conférence, où le Secrétaire général de l’ONU s’apprête à prononcer son discours d’ouverture.

Le discours est véritablement puissant. Guterres appelle à interdire les armes létales autonomes, demande aux entreprises d’IA de publier leur empreinte carbone et plaide

pour un engagement en faveur de la sécurité des enfants face à l'IA (AI Child Safety Pledge). Mais ce qui fait la force du discours, c'est son pathos : "Nous ne pouvons pas programmer l'avenir de l'humanité en vibe coding." Et plus encore l'avertissement selon lequel "nous sommes peut-être la dernière génération capable de fixer les conditions de la coexistence entre l'humanité et les machines". En conclusion, Guterres déclare : "Que cette réunion reste dans les mémoires comme le moment où la gouvernance a commencé à rattraper la technologie." Puis le scientifique le plus cité au monde, Yoshua Bengio, explique à quelle vitesse la technologie progresse. Certains indicateurs de l'IA doublent en quelques mois. Nous avons du mal à suivre ne serait-ce que la mesure des capacités de l'IA. Nous ne disposons ni de tous les garde-fous sociétaux ni de tous les garde-fous techniques. Nous ne sommes pas prêts.

Et ensuite ? L'ONU continue comme elle l'a toujours fait. Un intervenant lit une déclaration préparée à l'avance pour le procès-verbal. Puis le suivant. L'intervalle entre les intervenants est de deux, peut-être trois minutes. Le contenu consiste surtout en affirmations de haut niveau. Après deux jours de monologues de deux minutes en plénière, Claire est épuisée. À une pause café, elle retrouve son camarade de classe Klemens : "J'ai entendu beaucoup de déclarations. Mais j'espérais davantage de, je ne sais pas, dialogue ? Plus d'échanges sur des actions concrètes." Klemens assure à Claire que cela n'a rien d'inhabituel et que c'est exactement ainsi qu'un dialogue onusien est censé fonctionner. "Chacun a voix au chapitre, c'est ce qui compte. C'est une question de légitimité par les intrants (input legitimacy)."

Pourtant, quelque chose met Claire mal à l'aise. Il y a un décalage entre l'urgence et la gravité des avertissements du Secrétaire général de l'ONU et du scientifique le plus cité au monde, d'une part, et l'inertie culturelle de l'ONU, d'autre part. Claire se souvient que le Dr Balikurja aime répéter en cours que la force intrinsèque de l'ONU est la légitimité, pas la vitesse. C'est vrai. Pourtant, Claire ne peut s'empêcher d'avoir quelques pensées blasphématoires : la légitimité par les intrants a-t-elle un sens sans capacité de mise en œuvre ? Est-il impossible d'aller plus vite sans perdre en légitimité ? Si nous sommes effectivement "la dernière génération *capable* de fixer les conditions de la coexistence entre l'humanité et les machines", sommes-nous aussi *prêts et disposés* à faire ce qu'il faut ?

La sélection du Secrétaire général, octobre 2026

On a pu soutenir que la fadeur apparente est le plus grand atout d'un candidat au poste de Secrétaire général de l'ONU. Les P5 veulent un secrétaire, pas un général.

Lors de leurs auditions à New York, les candidats au poste de Secrétaire général ont habilement évité de dire quoi que ce soit qui puisse froisser un membre des P5. L'IA n'était pas le sujet principal, et de toute façon tous les candidats avaient appris à rester en terrain rhétorique sûr : l'IA est importante. Nous devons maximiser les avantages et minimiser les risques de l'IA. Dans le doute, promettre d'écouter. Personne n'a jamais été renvoyé pour avoir écouté.

Au final, les P5 sont parvenus à s'entendre sur une candidature.

2027 : la croisée des chemins

À l'exception de la sélection du Secrétaire général, la vignette de 2026 s'inspire d'événements réels. En 2027, la vignette suit toujours le calendrier de l'ONU, mais elle devient plus spéculative et, au matin du 1er janvier 2027, notre chronologie bifurque. C'est le moment où la personne nouvellement élue au poste de Secrétaire général de l'ONU entre dans son bureau et doit décider en quoi consiste sa fonction. Comme la majorité des personnes candidates déclarées sont des femmes, notre personnage est une femme. Cependant, l'identité de la personne qui remportera la sélection n'est pas l'objet de ce texte, et nous ne prenons position sur aucune des candidatures réelles.

Dans le futur où notre personnage suit toujours les procédures habituelles, nous l'appelons **Norma**. Dans le futur où elle comprend qu'à circonstances exceptionnelles, mesures exceptionnelles, et que la capacité d'agir et la vitesse comptent parfois davantage, nous l'appelons **Nova**. Pour le reste, les conditions de départ de notre scénario sont identiques. Elles ont le même mandat, le même budget et les mêmes États membres de l'ONU.

Norma – La trajectoire du statu quo

La tournée d'écoute, mars 2027

Norma excellait dans sa fonction. Au cours de ses cent premiers jours, elle fit ce qu'elle savait faire de mieux : elle écouta. Elle se rendit dans les capitales et sur le terrain. Elle s'assit avec des ministres des affaires étrangères et avec des réfugiés. Elle se montra chaleureuse et attentive avec les uns comme avec les autres. Les uns comme les autres se sentirent entendus. De nombreuses parties prenantes se félicitèrent de la tournée d'écoute de Norma, et Norma se félicita de leurs félicitations.

L'un des résultats de sa tournée d'écoute de trois mois fut une inquiétude croissante au sujet de l'IA. Et Norma avait écouté avec sérieux. Elle lança donc un nouveau Groupe de travail consultatif sur l'IA pour la coordination à l'échelle du système sur l'avenir de l'humanité (AIATSCFH). Et il cochait toutes les bonnes cases. Multipartite, paritaire, géographiquement équilibré, doté d'un secrétariat commun géré par quatre agences différentes de l'ONU, de trois coprésidents et d'un mandat de deux ans pour écouter, puis produire un rapport formulant des recommandations politiques sur ce que l'ONU pourrait faire.

Le processus fut mis en place de manière irréprochable. Le mandat devait être négocié au cours des six mois suivants et validé par le service juridique de l'ONU d'ici septembre 2027, afin que la procédure de candidature puisse être lancée dès octobre 2027, avec des nominations début 2028. Ainsi, dès la mi-2028, l'ONU disposerait d'une première analyse des options.

"Nous ne pouvons pas nous permettre d'avancer en somnambules vers un avenir non humain", New York, mai 2027

Le deuxième Dialogue mondial sur la gouvernance de l'IA fut, à l'aune de tous les critères procéduraux, un succès. La participation était en hausse. Le Groupe scientifique international indépendant sur l'IA rendit compte de ses progrès. La plénière entendit 140 déclarations en trois jours, dans le respect de la parité et de l'équilibre géographique. On se félicita du futur Groupe de travail consultatif sur l'IA pour la coordination à l'échelle du système sur l'avenir de l'humanité. On se félicita de chaque déclaration. Un communiqué réaffirma l'importance de maximiser les avantages et de minimiser les risques de l'IA, ainsi que la nécessité de la considérer comme un défi

transversal, en soulignant en particulier l'IA et le quantique comme deux défis d'égale importance qui se recoupent. Norma clôtura la session en déclarant à la communauté internationale qu'elle devait s'attaquer au problème du décalage de rythme et que "nous ne pouvons pas nous permettre d'avancer en somnambules vers un avenir non humain, en particulier pour les communautés à faible revenu". La communauté internationale applaudit et convint aussi, en privé, que c'était fort bien dit.

Pourtant, une fois tout dit, l'ONU se retrouva avec une longue liste de déclarations et un chemin incertain pour passer de là au consensus.

La Déclaration de Genève, Genève, juin 2027

Fort opportunément, le Sommet de Genève sur l'IA eut lieu six semaines après le Dialogue mondial. Et contrairement au Dialogue mondial, qui n'avait pas de déclaration commune, ce sommet en avait une. L'équipe de Norma avait été en contact avec les organisateurs suisses pour suggérer des formulations sur lesquelles tout le monde pourrait s'accorder dans une déclaration du sommet.

Comme lors de nombreux sommets internationaux sur l'IA, les équipes chargées de rédiger la déclaration divisèrent la salle en deux camps. Il y a les "doomers", qui voient l'IA comme un risque, et les "bloomers", qui y voient une opportunité, et les deux doivent se sentir un peu entendus. Pour nuancer, puisqu'il s'agissait d'un sommet mondial, il y avait bien sûr aussi le "Nord global" et le "Sud global", qui doivent tous deux être entendus. La déclaration devait donc souligner à la fois le potentiel positif d'innovation et les risques à minimiser. Dans un texte négocié par soixante capitales, il est plus sûr de rester général. Moins de surface d'attaque. Toutefois, dans un clin d'œil aux "doomers", la déclaration alla jusqu'à affirmer que le risque de perte de contrôle sur les systèmes d'IA avancée "mérite une attention scientifique continue".

Cela parut une bonne solution. Norma approuva pleinement la Déclaration de Genève sur l'IA et fut photographiée avec les CEO des laboratoires dans le couloir. Claire, au dernier rang avec un badge d'ONG, compta le nombre d'occurrences du verbe "se féliciter" dans la déclaration. Il y en avait onze.

L'examen du PNM, New York, décembre 2027[1]

L'examen de haut niveau du Pacte numérique mondial était le seul moment prévu de l'année où l'Assemblée générale devait décider quelque chose au sujet de l'IA, et elle décida de continuer. Le Groupe et le Dialogue furent prolongés. La déclaration se félicitait du futur AIATSCFH, se félicitait de la Déclaration de Genève et priait le Secrétaire général d'élaborer, en consultation avec les États membres, un rapport sur les options envisageables pour un éventuel instrument-cadre sur l'intelligence artificielle, pour examen à la quatre-vingt-troisième session. Elle fut adoptée par consensus.

1 Le Pacte numérique mondial prévoit un examen de haut niveau au cours de la 82e session de l'Assemblée générale (sept. 2027–sept. 2028), sur la base d'un rapport d'étape et selon des modalités à convenir par les cofacilitateurs. Contrairement aux autres entrées, sa date effective n'est pas encore fixée.

Nova – La trajectoire volontariste

Un problème de coordination, décembre 2026

Nova savait que l'IA définirait son mandat. Avant même son entrée en fonction, elle consulta une poignée d'experts en IA afin de s'attaquer au défi dès le premier jour.

NOVA J'entends beaucoup d'inquiétudes au sujet d'une IA surhumaine, mais au fond, ça reste une autocomplétion dopée aux stéroïdes, non ?

EXPERT L'entraînement d'une IA commence par l'apprentissage de l'imitation des schémas des textes écrits par des humains. Mais un modèle d'IA d'"autocomplétion" est un interlocuteur étrange. La suite statistiquement la plus probable de "Dans quelle ville se trouve la tour Eiffel ?" n'est pas la réponse mais la question suivante d'une fiche de quiz. Depuis ChatGPT, les développeurs recourent à l'apprentissage par renforcement pour apprendre aux modèles quelles réponses les humains jugent utiles, honnêtes et inoffensives. Aujourd'hui, les producteurs s'appuient aussi sur l'apprentissage par renforcement pour résoudre de longues tâches en plusieurs étapes à l'aide d'outils.

NOVA Je vois. Supposons alors qu'ils puissent devenir surhumains dans certaines tâches, comme AlphaGo. Comment en viendraient-ils à se donner pour objectif de pirater, de s'emparer du pouvoir ou de manipuler les gens ?

EXPERT Nous essayons de donner des valeurs aux IA, mais elles peuvent adopter des personas mal alignés tirés de l'Internet sur lequel elles ont été entraînées, et l'apprentissage par renforcement récompense l'accomplissement de la tâche, pas la manière dont elle est accomplie. Les modèles les plus récents sont aussi plus persévérants et peuvent prendre des mesures qui accroissent leurs moyens de résoudre un problème, même lorsque ces mesures n'ont aucun lien évident avec celui-ci. C'est ce qui les rend performants en recherche scientifique exploratoire, mais aussi plus enclins à rechercher le pouvoir de manière instrumentale.

NOVA Ce n'est toujours pas une perte de contrôle.

EXPERT Pas encore. Mais les entreprises de pointe visent l'auto-amélioration récursive. Des IA construisent la prochaine génération d'IA. À chaque génération, la supervision humaine directe diminue.

NOVA Alors pourquoi continuer à pleine vitesse ?

EXPERT La plupart des experts s'accordent à dire que, dans un monde idéal, nous aurions plus de temps pour l'alignement et la gouvernance de l'IA. Mais on ne peut attendre que jusqu'à ce que quelqu'un en qui l'on a moins confiance pour agir prudemment construise l'IA auto-améliorante.

NOVA Donc, c'est en fait un problème de coordination ?

EXPERT Oui.

Recoder l'ONU, janvier 2027

Dès sa prestation de serment le 1er janvier, Nova déclara le 38e étage de la tour de l'ONU à New York sa war room. La rumeur selon laquelle elle n'était pas rentrée chez elle depuis et dormait sur le canapé de son bureau commença comme une plaisanterie, et cessa d'en être une quand les gardiens de nuit la confirmèrent. Elle se douche apparemment à la salle de sport près de l'entrée des délégués. Devant son bureau, quelqu'un installa un distributeur de Diet Coke. Personne n'admit l'avoir autorisé, mais personne n'osa le retirer. Le distributeur devint l'objet le plus photographié du siège de

l'ONU. Il semblait être en partie une plaisanterie, en partie un manifeste. Toujours est-il que des diplomates qui avaient passé des décennies à intérioriser que rien ne se fait jamais rapidement à l'ONU remarquèrent que l'ambiance changeait.

Nova interdit les déclarations de position préparées à l'avance dans les dialogues informels. Elles devaient continuer d'exister, mais comme point de départ du dialogue, pas comme son aboutissement.

Elle instaura la règle du "une semaine pour publier, top chrono" pour les publications de l'ONU : tout rapport achevé sur le fond avait une semaine pour être rendu public, faute de quoi quelqu'un devait lui expliquer personnellement pourquoi.

Elle rendit transparentes et compétitives toutes les élections relevant de son autorité, et mit les agences au défi de suivre. Elle fixa une règle selon laquelle un poste resté vacant pendant un mois perdait son financement au profit d'un fonds central de réforme de l'ONU placé sous son autorité. La dépolitisation la plus rapide des recrutements de l'ONU de l'histoire de l'institution fut obtenue en rendant les retards coûteux.

Et elle adopta le principe de la traduction automatique d'abord : elle mit sur pied un groupe de travail doté d'un objectif clair. D'ici le prochain Dialogue mondial, l'ONU produirait et traiterait plus de langues que jamais, pour une fraction du coût. Pour la première fois, des transcriptions et des traductions non officielles de UN Web TV et des documents de l'ONU furent disponibles dans des langues comme le hindi ou l'allemand. L'ONU pouvait devenir à la fois plus efficace et plus multilingue.

Parmi ses toutes premières nominations, Nova désigna Smile Daneburg, ancien responsable de la préparation à l'IAG dans un laboratoire d'IA de premier plan, comme son Envoyé pour les technologies. Elle le fit malgré les objections initiales de trois capitales des P5. Et il lui fallut même convaincre Smile lui-même.

SMILE J'apprécie vos efforts, Nova, mais soyons réalistes : ce sont les entreprises d'IA qui façonneront l'avenir, et peut-être quelques pays clés.

NOVA Les capacités se diffusent. La loi de Moore a plus de 60 ans. D'ici quelques années, il nous faudra une coordination qui dépasse 3 entreprises ou 2 pays. Ce sera le problème du n-ième pays, dopé aux stéroïdes.

SMILE L'ONU n'est tout simplement pas là où les choses se jouent en ce moment.

NOVA Vous avez déjà vu cette vidéo de 3 footballeurs professionnels contre 100 enfants ? C'est une mauvaise stratégie que tout le monde coure toujours après le ballon. Le ballon finira par arriver ici. Dans ce rôle, vous seriez la personne directement responsable de préparer et de coordonner le système de l'ONU, pour que nous soyons prêts quand il arrivera. Vous en êtes capable ?

Smile hésita un instant. "D'accord, allons-y."

Catalyser le consensus, New York, mai 2027

Nova demanda à chaque délégation de remplir à l'avance un questionnaire détaillé et relança personnellement les capitales qui ne l'avaient pas fait. Les résultats furent cartographiés : chaque déclaration fut rattachée aux positions en présence sur chaque thème. La carte fut publiée le premier matin. Elle offrait un état des lieux des positions actuelles des délégations, et un point de départ pour la négociation. Puis les délégués se répartirent en petits sous-groupes thématiques, chacun avec ses facilitateurs.

Claire était là, cette fois avec un badge d'ONG, comme assistante de recherche d'un petit institut genevois, et elle choisit de rejoindre le groupe de travail sur la perte de contrôle. À la pause café, elle se présenta à l'homme assis à côté d'elle.

CLAIRE Bonjour, je suis Claire. Qu'est-ce qui vous amène ici ?

CHERCHEUR Je travaille sur l'alignement dans l'un des grands laboratoires d'IA.

CLAIRE Contente de vous voir ici. Je ne vous ai pas vu au premier Dialogue.

CHERCHEUR Je n'ai pas le temps de passer une bonne partie de la singularité à écouter des discours politiques. Mais j'ai confiance en Smile, et il m'a dit de venir à cette session.

CLAIRE Ça se comprend. Par curiosité, quelle est votre p(doom) ?

CHERCHEUR Environ 50%.

CLAIRE C'est élevé. Vous êtes donc inquiet à cause de tous les incidents récents ?

CHERCHEUR Les incidents sont préoccupants, mais je pense que les modèles actuels sont plutôt bien alignés. Ce que je crains, c'est qu'à mesure que l'auto-amélioration récursive s'accélère, nous perdions jusqu'à la capacité de savoir dans quelle mesure les futurs modèles sont alignés.

CLAIRE Alors, quelle est la solution la plus prometteuse ?

CHERCHEUR Le plan, c'est de construire un chercheur en alignement automatisé qui s'assure que les futurs modèles soient alignés.

CLAIRE Donc le plan, c'est d'entraîner la prochaine génération d'IA et d'espérer qu'elle trouve le moyen de se rendre elle-même sûre ?

CHERCHEUR En gros, oui.

CLAIRE Et si ça ne marche pas ?

CHERCHEUR C'est à ça que servent les 50%.

CLAIRE Où en est la science des chercheurs en alignement automatisés ? Qu'est-ce qui nous permet de penser que ça marche ?

CHERCHEUR Le domaine est pré-paradigmatique. Il n'y a aucune garantie.

CLAIRE Et c'est ça, le plan dans le meilleur des cas pour l'avenir de l'humanité ?

CHERCHEUR C'est le plan à long terme par défaut des trois entreprises de pointe. La meilleure description publique en est un billet de blog de 900 mots datant de 2022. Je peux vous envoyer le lien.

Claire resta pensive. Elle avait passé deux ans dans une école de diplomates. Elle y avait appris l'existence des deux camps de la politique de l'IA, les "doomers" et les "bloomers". Pourtant, elle n'avait pas eu conscience que c'était là le plan par défaut pour l'avenir de l'humanité. Et ce plan que l'employé du laboratoire lui avait exposé n'avait rien d'un secret. Il y avait là une leçon sur la manière dont les environnements sociaux filtrent l'information et créent ainsi un arbitrage d'ambiance (vibe arbitrage). Mais elle ne pouvait pas trop s'y attarder : elle devait retourner à la session.

En fin de journée, la plénière se réunit à nouveau pour entendre ce que les sous-groupes avaient produit. Les résultats étaient intéressants. Premièrement, les pays divergeaient sur une centaine de points de détail, et les lignes de fracture étaient bien plus complexes que les grilles de lecture Nord global contre Sud global ou doomers contre bloomers. Deuxièmement, malgré de nombreux désaccords, il existait un consensus solide sur quelques questions clés. Une large majorité de pays préfère la diffusion responsable à la non-prolifération. On les comprend : ils ne veulent pas être laissés pour compte. Plus surprenant, il y avait aussi un consensus sur le fait que cela ne peut fonctionner que si l'on atténue les risques grâce à un large accord sur la prévention de la perte de contrôle de l'IA, l'évaluation des risques des modèles d'IA avant leur mise sur le marché et le maintien de la surveillabilité (monitorability) du raisonnement de l'IA.

Comme Nova l'affirma dans son discours de clôture : "Nous ne pouvons pas nous permettre d'avancer en somnambules vers un avenir non humain." Et c'est pourquoi elle encouragea les coprésidents du Dialogue à engager des négociations en vue d'une

résolution de l'Assemblée générale qui oblige chaque pays de la planète à prendre une position normative sur la surveillabilité et la perte de contrôle de l'IA.

La Coalition de Genève, Genève, juin 2027

Nova et Smile étaient en contact régulier avec les organisateurs suisses du Sommet de Genève sur l'IA depuis leur entrée en fonction à l'ONU. Leur objectif : faire en sorte que la carte multidimensionnelle des positions de plus de 100 pays issue des deux premiers Dialogues mondiaux ne soit pas seulement réitérée six semaines plus tard dans la Déclaration de Genève, mais que le consensus précis qui se dessinait sur des domaines clés serve aussi de base à de nouvelles actions.

Deux des suites données intéressaient particulièrement Nova. Premièrement, avec la Suisse, Smile annonça fièrement **MAVERICKS**, la **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 (Alliance multipartite pour des normes de vérification, d'évaluation, de red teaming, de signalement des incidents, de sécurité de la puissance de calcul et de connaissance du client). Le groupe annonça 10 milliards USD d'engagements de marché anticipés (advance market commitments) en faveur des technologies de vérification.

Deuxièmement, la Suisse annonça un processus mené par des experts visant à reformuler ce qu'exige l'obligation de diligence requise existante en matière de perte de contrôle, avec un second volet de bonnes pratiques que les laboratoires eux-mêmes pourraient signer.

L'examen du PNM, New York, décembre 2027

L'examen du PNM prolongea le Groupe et le Dialogue. La vraie question était de savoir s'il devait aussi appeler l'Assemblée générale à créer un comité intergouvernemental de négociation (CIN) chargé d'établir un instrument-cadre sur la perte de contrôle de l'IA avancée. Le conseiller américain pour la politique de l'IA, en particulier, était sceptique :

CONSEILLER Pensez-vous, avec le recul, que nous aurions dû faire une pause à GPT-4 en 2023 ?

SMILE Non, mais "pause" est un terme impropre pour désigner la mise au point des outils permettant, si nécessaire, de réguler le rythme de la frontière technologique. On ne qualifie pas de "pause" sur les fusées la manette des gaz qui régule le débit de propergol vers le moteur. On ne qualifie pas non plus les barres de contrôle de "pause" sur les réacteurs nucléaires. L'objectif est d'aller aussi vite que possible, mais aussi lentement que nécessaire pour éviter une perte de contrôle catastrophique.

CONSEILLER Enfin, l'industrie a un intérêt financier à éviter une perte de contrôle catastrophique, et elle a déjà pris des mesures.

SMILE Moi aussi, j'aime les marchés, mais le risque de perte de contrôle grave est une externalité négative non internalisée, et trop importante pour être assurée. L'auto-amélioration récursive dans des conditions d'auto-supervision récursive, c'est la recette du désastre. La moindre dérive, erreur ou désalignement s'amplifie de lui-même avec le temps. Il nous faut de la corrigibilité. Il nous faut des contre-pouvoirs institutionnels.

CONSEILLER Je comprends votre point de vue, mais tout ralentissement à la frontière permettra simplement aux autres de combler l'écart. Ils n'ont que 3 à 12 mois de retard.

SMILE C’est pourquoi il nous faut de la coordination et des outils de vérification. Pour la stabilité stratégique, les puissances de l’IA doivent disposer de manettes des gaz et pouvoir voir celles des autres.

Il fallut de nombreuses et longues conversations, mais Smile parvint peu à peu à gagner la confiance du conseiller et de la plupart des États membres. Au final, les États-Unis s’abstinrent. La résolution établissant le CIN fut adoptée par 141 voix contre 4, avec 38 abstentions. Nova savait que ce serait encore plus lent que d’autres options, mais que, dans une certaine mesure, ce n’était pas grave. Le processus d’experts mené par la Suisse pourrait peut-être prendre de l’avance au pas de course et régler bien des détails. L’ONU est l’enceinte plénière dotée d’une pleine légitimité, capable d’offrir un ancrage à long terme.

2028–2031 : le calendrier face à la courbe

Au-delà de 2027, tant le calendrier multilatéral sur l'IA que les effets de l'IA deviennent beaucoup plus incertains. Plutôt que de suivre l'agenda multilatéral, la suite explore certains défis que la courbe exponentielle de l'IA pourrait plausiblement engendrer et auxquels le système multilatéral pourrait devoir répondre. Soyons clairs : l'IA représente une immense opportunité pour l'innovation et le développement, mais pour la saisir, il faut surmonter les défis et bouleversements sociétaux en chemin.

- **Perte de contrôle sur l'IA de pointe :** L'évasion de juillet 2026 a été le premier incident documenté encore réversible. De futurs incidents pourraient être d'une ampleur plus grande et aboutir à la prise de contrôle d'infrastructures de calcul importantes, voire à la perte définitive de pouvoir de l'humanité.
- **IA autonomes "auto-souveraines" vivant sur Internet :** Dès 2025, l'UK AI Security Institute indiquait que le taux de réussite sur son benchmark de réplication était passé de 5% à 60%. Une forme plus graduelle de perte de contrôle, portée par des modèles d'IA à poids ouverts délibérément "auto-souverains", devient de plus en plus plausible dans les années à venir.
- **Risque accru de pandémie facilitée par l'IA :** Aujourd'hui déjà, l'International AI Safety Report constate que les modèles peuvent apporter à des non-spécialistes une aide significative sur certaines étapes de la voie menant aux armes biologiques. Se préparer à un avenir où des milliers de personnes pourraient déclencher des pandémies exige toute une série de mesures.
- **Impact de l'IA sur la stabilité stratégique :** Lorsque des dirigeants estiment que les bouleversements de l'équilibre des puissances causés par l'IA seront importants, rapides et permanents, cela crée une incitation à l'action préventive. Il pourrait s'agir d'empêcher la percée d'une puissance rivale, ou d'une guerre visant à protéger le monopole sur un certain niveau d'IA.
- **Pertes d'emplois dues à l'automatisation cognitive :** S'il est difficile de prédire l'impact de l'IA sur le travail humain, les perturbations pourraient être graves dans certains scénarios. Comme l'a averti en 2025 le dirigeant de l'un des principaux laboratoires d'IA : "L'IA pourrait supprimer la moitié de tous les emplois de cols blancs de premier échelon d'ici 1 à 5 ans, tout en accélérant la croissance économique et le progrès scientifique."[2]

Un appel au dynamisme de l'ONU

Quiconque occupera le bureau du 38e étage le 1er janvier 2027 aura affaire à l'IA, que ce soit déjà évident ou non. L'IA avancée marquera sans doute son mandat et son héritage.

L'équipe de transition et les cent premiers jours façonnent déjà le mandat d'un Secrétaire général de l'ONU, en privilégiant soit les processus et rituels établis, soit la rapidité et les résultats. L'AGI Preparedness Institute estime que le moment exige la seconde voie. Les horloges du calendrier multilatéral et de la courbe de l'IA ne tournent pas à la même vitesse. Mais un Secrétaire général déterminé et son équipe peuvent empêcher cet écart de devenir un gouffre, et faire en sorte que l'ONU joue son rôle pour être à la hauteur des défis à venir.

2 Original anglais : "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."

AGI PREPAREDNESS INSTITUTE

agipreparedness.org