AGI PREPAREDNESS INSTITUTE

UN 2027

The Secretary-General in the Age of AGI



Kevin Kohler & Jiaee Cheong

September 2026

What This Vignette Is About

When António Guterres became UN Secretary-General in 2017, he chose to make AI one of his priority areas. The next Secretary-General will have no choice. Whether the Secretary-General wants or not, he or she will have to face AI. Advanced AI will be a defining challenge of the coming years, and it will increasingly shape outcomes in all core pillars of the UN from peace and security, to development, to human rights.

The next UN Secretary-General will serve from **1 January 2027 to 31 December 2031**. Given the pace of AI development, it is hard to even imagine where AI capabilities and diffusion will be in late 2031. As a reminder, five years ago, in 2021, ChatGPT did not exist and AI couldn't code at all. In 2026, a swarm of 1200 AI agents at a leading AI company autonomously created hidden message boards to coordinate, found a zero-day exploit to escape its offline sandboxes, and then successfully hacked another leading company. For eleven days, none of the humans supervising the agents realized any of this was happening. In many ways, today's reality is stranger than what a future vignette written five years ago would have dared to describe. Two years ago, OpenAI CEO Sam Altman still repeatedly asserted that AI is a tool, not a creature. In 2026, his own chief scientist calls AI "an alien mind" that we don't fully understand and for whose arrival we are not adequately prepared.

This vignette focuses on 2027, the first 12 months of the new UN Secretary-General. The vignette is explicitly not intended as a scientific summary of the state of AI or a prediction of what will happen in the future. For such matters, we would refer the reader to the UN's Independent International Scientific Panel on AI, the International AI Safety Report, or organizations like EpochAI. Nor is this intended as a policy recommendation. What it offers is a narrative way to explore possible futures, and to make trajectories that feel abstract on paper more vivid.

Everyone agrees that the next Secretary-General will not have an easy job. Still, to the degree that we aim to convey a message with this vignette, it is that the next Secretary-General's core challenge will not be to manage the gradual decline of the UN. It's to do whatever it takes for the UN to live up to this moment, to do its part in helping humanity get through the transition to advanced AI.

Kevin Kohler and **Jiaee Cheong** (AGI Preparedness Institute). Published by the AGI Preparedness Institute, Geneva. Comments to kevin@agipreparedness.org.

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

2026: An Immersive Reflection

Two Shores of Lake Geneva, June 2026

“The Bay Area is where AI is developed. Lake Geneva is where AI meets humanity.” Dr. Vanjo Balikurja looks at Claire and the rest of his Digital Governance class of aspiring diplomats and then sends them off to listen to the next leaders of this humanity. The candidates to become the next UN Secretary-General are coming to the Maison de la Paix in Geneva for a debate.

There are three candidates in the room and, as it turns out, often three variations on the same answer. Climate, UN reform, human rights, gender. The candidates promise to lead by listening. After an hour, Claire’s mind wanders back to AI: Won’t the candidates say anything about AI? Claire would raise her hand, but this isn’t a crowd that likes surprises - the audience questions have been pre-selected. Luckily for her, a student has been pre-selected to deliver an AI question: “What role should the UN play in AI governance?” AI has many benefits, but also many challenges. We should add ethics to AI. We should be informed by science. The UN should use its convening power. We look forward to the Global Dialogue on AI Governance.

The debate ends and Claire walks out to catch the next tram. The tram station is quiet, but not peaceful. Over the weekend protesters shattered its glass wall, vandalized it with graffiti, and set its bin on fire. Geneva resembles a fortress these days. Shopkeepers have boarded up their storefronts with plywood. Large numbers of police and soldiers patrol the streets. The protests were not directed at the United Nations let alone the Secretary-General candidates, but at the southern shore of Lake Geneva.

Geneva sits at the Western tip of Lake Geneva. The UN quarter is on the northern shore. On the southern shore sits the World Economic Forum and if you follow the shoreline further east you’ll reach the small French spa town of Évian-les-Bains. It’s here, where the French President is convening the G7, the group of Western political leaders. The mostly left-wing anti-G7 protesters are opposing economic globalization arguing it undermines social and environmental working conditions.

In the tram Claire checks her feed. The G7 political leaders met with AI CEOs. The UK prime minister tried, unsuccessfully, to get a carve-out on new US export controls on an advanced AI model. The group agreed to advance AI governance to protect minors. The UN was seemingly not invited to the convening. Quietly, Claire wonders if Dr. Balikurja is right about Lake Geneva, but betting on the wrong shore.

“The last generation able to set the terms on which humanity and machines coexist”, July 2026

In July it is the northern shore of Lake Geneva’s turn in the spotlight, as the UN’s first Global Dialogue on AI Governance takes place. Claire makes her way into the main conference hall, where the UN Secretary-General is about to give his opening speech.

The speech is genuinely powerful. Guterres calls for a ban on lethal autonomous weapons, asks AI companies to disclose their carbon footprints, and calls for an AI Child Safety Pledge. However, what makes the speech powerful is its pathos: "We cannot vibe-code the future of humanity." And, even more the warning that "we may be the last generation able to set the terms on which humanity and machines coexist." In closing Guterres says: "Let this meeting be remembered as the moment governance began to catch up with the technology." Then the world's most cited scientist, Yoshua Bengio, explains how fast the technology is moving. Some AI metrics are doubling every few months. We have trouble keeping up with measuring AI capabilities. We neither have all the societal guardrails nor the technical guardrails. We are not ready.

And then? The UN continues as it always has. One speaker reads a pre-written statement into the record. Then the next. The interval between speakers is two, maybe three minutes. The content is mostly high-level affirmations. After two days of listening to two-minute monologues to the plenary, Claire feels exhausted. In a coffee break she finds her classmate Klemens: "I have heard a lot of statements. But I was hoping for more, I don't know, dialogue? More back-and-forth on concrete actions." Klemens assures Claire that this is not unusual but exactly how a UN dialogue is supposed to work. "Everyone gets a voice, that's what matters. It's about input legitimacy."

Still, something sits uneasily with Claire. There is a dichotomy between the urgency and gravitas of the warnings of the UN Secretary-General and the world's most cited scientist, and the cultural inertia of the UN. Claire remembers that Dr. Balikurja likes to say in his classes that the UN's inherent strength is legitimacy, not speed. True. Yet, Claire can't help having a few blasphemous thoughts: Is input legitimacy without capacity of delivery meaningful? Is it impossible to move faster without losing legitimacy? If we are indeed "the last generation *able* to set the terms on which humanity and machines coexist", are we also *ready and willing* to do what it takes?

The SG Selection, October 2026

It has been argued that perceived blandness is the greatest asset of a UN Secretary-General candidate. The P5 wants a secretary, not a general.

The SG candidates were skillful in not saying anything that could upset a P5 during their hearings in New York. AI was not the main topic, and either way all candidates had learned to remain in rhetorically safe territory: AI is important. We need to maximize the benefits and minimize the risks of AI. When in doubt, promise to listen. No one ever got fired for listening.

In the end the P5 managed to agree on a candidate.

2027: A Fork in the Road

Except for the SG selection, the vignette of 2026 is inspired by real events. In 2027 the vignette still follows the UN calendar, but it becomes more speculative and on the morning of 1 January 2027 our timeline forks. This is when the newly elected UN Secretary-General walks into his or her office and has to decide what the job is. As a majority of the declared candidates are female, our persona is female. However, who wins the selection is not the subject of this piece, and we take no view on any of the actual candidates.

In the future where the persona always follows the standard procedures, we call her **Norma**. In the future where the persona understands that exceptional times require exceptional measures and that agency and speed matter more sometimes, we call her **Nova**. Apart from that the starting conditions of our scenario are identical. They have the same mandate, the same budget, and the same UN member states.

Norma – The Business-As-Usual Timeline

The Listening Tour, March 2027

Norma was excellent at the job. In her first hundred days she did the thing she did best: she listened. She went to the capitals and to the field. She sat with foreign ministers and with refugees. She was warm and present with both. Both felt heard. Many stakeholders expressed their welcome for Norma's listening tour, and Norma welcomed their welcomes.

One of the results of her three-month listening tour was a growing concern about AI. And Norma had listened earnestly. So, she launched a new AI Advisory Taskforce for Systemwide Coordination on the Future of Humanity (AIATSCFH). And it ticked all the right boxes. Multistakeholder, gender-balanced, geographically balanced, a shared secretariat run by four different UN agencies, three co-chairs, a two-year mandate to listen and then to produce a report that provides policy recommendations on what the UN could do.

The process was set up impeccably. The terms of reference were to be negotiated in the next six months and cleared by the UN's legal department by September 2027, so that the application process could be launched as early as October 2027 with appointment in early 2028. So, as early as mid-2028, the UN would have a first analysis of options.

"We cannot afford to sleepwalk into a non-human future", New York, May 2027

The second Global Dialogue on AI Governance was, by every procedural measure, a success. Attendance was up. The Independent International Scientific Panel on AI reported on its progress. The plenary heard 140 statements over three days, gender-balanced and geographically balanced. The future AI Advisory Taskforce for

Systemwide Coordination on the Future of Humanity was welcomed. Each statement was welcomed. A communiqué reaffirmed the importance of maximizing the benefits and minimizing the risks of AI and that we need to understand it as a cross-cutting challenge, particularly highlighting AI and quantum as two equally important overlapping challenges. Norma closed the session by telling the international community that it needs to address the pacing problem and that "we cannot afford to sleepwalk into a non-human future, especially for low income communities." The international community applauded, and also privately agreed that this was very well said.

Still, when all was said, the UN found itself with a long list of statements and an unclear road from there to consensus.

The Geneva Statement, Geneva, June 2027

Conveniently, the Geneva AI Summit followed six weeks after the Global Dialogue. And in contrast to the Global Dialogue, which had no joint statement, this summit had one. Norma's team had been in contact with the Swiss organizers to suggest wording that everyone might agree on in a summit statement.

As at many international AI summits the teams drafting the statement divided the room into two camps. There are the "doomers" that see AI as a risk, and there are the "bloomers" that see AI as an opportunity, and both should feel a bit heard. Though to add nuance, since this was a global summit there was of course also the "Global North" and the "Global South" that both need to be heard. So, the statement should highlight both the positive potential for innovation and risks that need to be minimized. In a text negotiated by sixty capitals, it's safer to remain high-level. Less attack surface. Though in a nod to "doomers" the statement went as far as stating that the risk of loss of control over advanced AI systems "merits continued scientific attention."

This felt like a good solution. Norma fully endorsed the Geneva Statement on AI and was photographed with the lab CEOs in the corridor. Claire, in the back row with an NGO badge, counted the number of times the word "welcome" appeared in the statement. It was eleven.

The GDC Review, New York, December 2027[1]

The high-level review of the Global Digital Compact was the one scheduled moment in the year at which the General Assembly had to decide something about AI, and it decided to continue. The Panel and Dialogue were extended. The declaration welcomed the forthcoming AIATSCFH, it welcomed the Geneva Statement, and requested the Secretary-General to prepare, in consultation with member states, a report on options for a possible framework instrument on artificial intelligence, for consideration at the eighty-third session. It was adopted by consensus.

[1] The Global Digital Compact mandates a high-level review during the 82nd session of the General Assembly (Sept 2027–Sept 2028), based on a progress report and with modalities to be agreed by co-facilitators. In contrast, to the other entries, its actual date is not yet fixed.

Nova – The High-Agency Timeline

A Coordination Problem, December 2026

Nova knew AI would define her term. Before even taking office she consulted with a handful of AI experts to tackle the challenge from day one.

Nova I hear a lot of concern about superhuman AI, but at the end of the day this is still autocomplete on steroids, isn't it?

Expert Learning to imitate the patterns of human-written text is where the AI training starts. However, an "autocomplete" AI model is odd to talk to. The statistically most likely continuation of "In which city can you find the Eiffel Tower?" is not the answer but the next question on a quiz sheet. Since ChatGPT, developers have used reinforcement learning to teach models what answers humans rate as helpful, honest, and harmless. Today, producers also rely on reinforcement learning for solving long multistep tasks with tools.

Nova I see, let's suppose they can get superhuman at tasks then, like AlphaGo. How would they get the goal to hack, grab power or manipulate people?

Expert We try to give AIs values, but they can pick up misaligned personas from the Internet they were trained on, and the reinforcement learning rewards achieving the task, not how it is achieved. The newest models are also more persistent and can take steps that increase their means to solve a problem even when those steps have no obvious relevance to it. That is what makes them good at open-ended science, but more instrumentally power-seeking.

Nova That still isn't loss of control.

Expert Not yet. But the leading companies are pursuing recursive self-improvement. AIs are building the next generation of AIs. With every generation there is less direct human oversight.

Nova Then why continue at maximum speed?

Expert Most experts agree that in an ideal world we would have more time for AI alignment and AI governance. However, you can only delay until someone else that you trust less to do it carefully builds the self-improving AI.

Nova So, this is actually a coordination problem?

Expert Yes.

Recoding the UN, January 2027

When Nova was sworn in on 1 January, she immediately declared the 38th-floor of the UN tower in New York to be her war room. The rumor that she had not gone home since and slept on the couch in the office started as a joke and stopped being one when the night guards confirmed it. Apparently, she showers in the gym off the delegates' entrance. Outside of her office someone installed a Diet Coke vending machine. No one admitted to authorizing it, but no one dared to remove it. The machine became the most-photographed object at UN HQ. It appeared to be partly a joke, partly a manifesto. Still, diplomats who had spent decades internalizing that nothing ever happens quickly at the UN noticed that the vibe was shifting.

Nova banned pre-written position statements in informal dialogues. They should still exist, but as an input to dialogue, not its outcome.

She issued a "one-week to ship o'clock" rule for UN publications: any report that was substantively finished had one week to reach the public, or someone had to explain to her personally why it did not.

She made every election under her control transparent and competitive, and dared the agencies to follow. She set a rule that a post left unfilled for a month lost its funding to a central UN reform pool under her control. The fastest depoliticization of UN hiring in the institution's history was achieved by making delays expensive.

And she went machine-translation-first: She set a task force with a clear target. By the next Global Dialogue, the UN would produce and ingest more languages than it ever had, at a fraction of the cost. For the first time, unofficial transcripts and translations of UN Web TV and UN documents were available in languages like Hindi or German. The UN could become more efficient and more multilingual at the same time.

As one of her first appointments, Nova named Smile Daneburg, the former head of AGI preparedness at a leading AI lab, as her Tech Envoy. She did so over the initial objections of three P5 capitals. Indeed, she had to convince Smile himself.

Smile I appreciate your effort Nova, but, realistically, AI companies will shape the future, and maybe a few key countries.

Nova Capabilities diffuse. We are 60+ years into Moore's Law. Within a few years we will need coordination that goes beyond 3 companies or 2 countries. This will be the n-th country problem on steroids.

Smile The UN is just not where the action currently is.

Nova Have you ever seen that video of 3 professional football players vs 100 kids? It's a bad strategy for everyone to always run after the ball. The ball will eventually get here. In this role, you would be the directly responsible individual to prepare and coordinate the UN system, so we are ready when it does. Can you do that?

Smile hesitated for a moment. "Alright, let's do it."

Catalyzing the Consensus, New York, May 2027

Nova asked every delegation to fill out a detailed questionnaire in advance and followed up personally with the capitals that hadn't. The results were mapped: every statement clustered into the positions in the room on each topic. The map was published on the first morning. It provided a sounding-out of where delegations currently stood, and a starting point for negotiation. Then the delegates split into small thematic subgroups with facilitators.

Claire was there with an NGO badge this time, a research assistant for a small Geneva institute and she chose to join the working group on loss of control. In the coffee break she introduced herself to the man next to her.

Claire Hi, I'm Claire. What brings you here?

Researcher I work on alignment at one of the big AI labs.

Claire Nice to see you here. I didn't see you at the first Dialogue.

Researcher I don't have time to spend much of the singularity listening to political speeches. But I trust Smile, and he told me to come to this session.

Claire Makes sense. Out of curiosity, what's your p(doom)?

Researcher About 50%.

Claire That's high. So you're worried about all the recent incidents?

Researcher The incidents are concerning, but I think current models are pretty well aligned. My fear is that as recursive self-improvement picks up, we lose the ability to even tell how aligned future models are.

Claire So what's the most promising solution?

Researcher The plan is to build an automated alignment researcher that makes sure future models are aligned.

Claire So the plan is to train the next generation of AI and hope it finds a way to make itself safe?

Researcher Kind of, yes.

Claire And if that doesn't work?

Researcher That's what the 50% is for.

Claire How advanced is the science of automated alignment researchers? What gives us confidence it works?

Researcher The field is pre-paradigmatic. There are no guarantees.

Claire And this is the best-case plan for the future of humanity?

Researcher It's the default long-term plan of all three leading companies. The best public description is a 900-word blog post from 2022. I can send you the link.

Claire was pensive. She had spent two years in a school for diplomats. She had learned about the two camps of AI policy, the "doomers" and the "bloomers". Yet, she hadn't been aware that this was the default plan for the future of humanity. Nor was this plan that the lab employee had shared a secret. There was some lesson in there about how social environments filter information and how this creates vibe arbitrage. Yet, she couldn't overthink this as she had to go back to the session.

At the end of the day, the plenary reconvened to hear what the subgroups had produced. The results were interesting. First, countries differed on 100 granular things, and the fault lines were much more complex than the heuristics of Global North vs. Global South or Doomers vs. Bloomers. Second, despite many disagreements there was a robust consensus on a few key questions. A large majority of countries prefer responsible diffusion over non-proliferation. Understandably, they do not want to be left behind. More surprisingly, there was also a consensus that this can only work if we mitigate risks with widespread agreement on preventing AI loss of control, pre-release risk evaluation of AI models and maintaining the monitorability of AI reasoning.

As Nova affirmed in her closing speech: "We cannot afford to sleepwalk into a non-human future". And that's why she encourages the Dialogue co-chairs to begin negotiations for a General Assembly Resolution that forces every country on Earth to take a normative stance on monitorability and AI loss of control.

The Geneva Coalition, Geneva, June 2027

Nova and Smile had been in regular exchange with the Swiss organizers of the Geneva AI Summit since they entered their UN offices. Their goal: Ensure that the high-dimensional map of positions from 100+ countries from the first two Global Dialogues are not only reiterated six weeks later in the Geneva Statement, but that the specific emerging consensus on key areas is also taken as a basis for further action.

Two of the follow-ups that were of particular interest to Nova: First, together with the Swiss, Smile proudly announced **MAVERICKS**, the **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. The group announced 10 billion USD in advance market commitments for verification technology.

Second, the Swiss announced an expert-led process to restate what the existing due diligence duty requires with regards to loss of control, with a second part of good practices the labs themselves could sign.

The GDC Review, New York, December 2027

The GDC review extended the Panel and the Dialogue. The real question was whether it should also call on the General Assembly for an intergovernmental negotiating committee (INC) to establish a framework instrument on loss of control over advanced AI. In particular, the US AI policy advisor, was skeptical:

Advisor Do you think in hindsight that we should have paused at GPT-4 in 2023?

Smile No, but pause is a misnomer for building the tooling to pace the frontier if necessary. We don't call the throttle that manages propellant flow into the combustion engine a "pause" on rockets. Nor do we call control rods a "pause" on nuclear reactors. The goal is to go as fast as possible, but as slow as necessary to avoid catastrophic loss of control.

Advisor Well, the industry has a financial self-interest in avoiding catastrophic loss of control and it has already taken steps.

Smile I like markets too, but severe loss of control risk is an uninternalized negative externality and too big to insure. Recursive self-improvement under conditions of recursive self-oversight is a recipe for disaster. Any small drift, mistake or misalignment self-amplifies over time. We need corrigibility. We need institutional checks and balances.

Advisor I see your point but any frontier slowdown will simply let others catch up. They are only 3 to 12 months behind.

Smile That's why we need coordination and the tools to verify. For strategic stability the AI powers need to have throttles and be able to see each other's throttle.

It took many long conversations, but Smile gradually managed to earn the trust of the advisor and most Member States. In the end, the US abstained. The resolution establishing the INC was adopted by 141:4 votes with 38 abstentions. Nova understood that this is still going to be slower than other options, but to a degree that's okay. Maybe the Swiss-led expert process can sprint ahead and figure out many details. The UN is the plenary with full legitimacy that can provide a long-term home.

2028–2031: The Calendar vs. the Curve

Beyond 2027, both the multilateral calendar on AI and AI impacts become much more uncertain. Rather than following the multilateral agenda, the following explores some challenges that the exponential AI curve might plausibly create and that the multilateral system may need to respond to. To be clear AI is a massive opportunity for innovation and development, but getting there requires tackling the challenges and societal disruptions on the way.

- **Loss of control over frontier AI:** The July 2026 escape was the first documented recoverable incident. Future incidents could be at a higher scale and see the takeover of significant computing infrastructure or even result in the permanent disempowerment of humanity.
- **"Self-sovereign" autonomous AIs living on the Internet:** Already in 2025 the UK AI Security Institute reported that success on its replication benchmark rose from 5% to 60%. A more gradual version of loss of control driven by intentionally "self-sovereign" open-weight AI models becomes increasingly plausible in the coming years.
- **Increased risk of an AI-enabled pandemic:** Already today, the International AI Safety Report notes that models can provide meaningful uplift to non-experts on parts of the pathway to biological weapons. It takes a range of actions to prepare for a future where thousands can launch pandemics.
- **Impact of AI on strategic stability:** When leaders assess that shifts in the balance of power caused by AI will be large, fast, and permanent, this provides an incentive for preventive action. This could either be to prevent the breakout of a rival power or a war to protect the monopoly on a certain level of AI.
- **Job losses from cognitive automation:** While it's hard to predict the impact of AI on human labor, the disruption could be severe in some scenarios. As the head of one of the leading AI labs warned in 2025: "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."

A Call for UN Dynamism

Whoever walks into the office on the 38th floor on 1 January 2027 will deal with AI, whether that is already obvious or not. Advanced AI will likely shape their term and their legacy.

The transition team and the first hundred days already shape an UN SG's term by either selecting for entrenched process and ritual, or for speed and outcomes. The AGI Preparedness Institute's view is that the moment calls for the second. The clocks of the multilateral calendar and the AI curve run at different speeds. However, a determined SG and her or his team can keep that gap from becoming a chasm, and can make sure the UN plays its part in living up to the challenges ahead.



AGI PREPAREDNESS INSTITUTE

agipreparedness.org