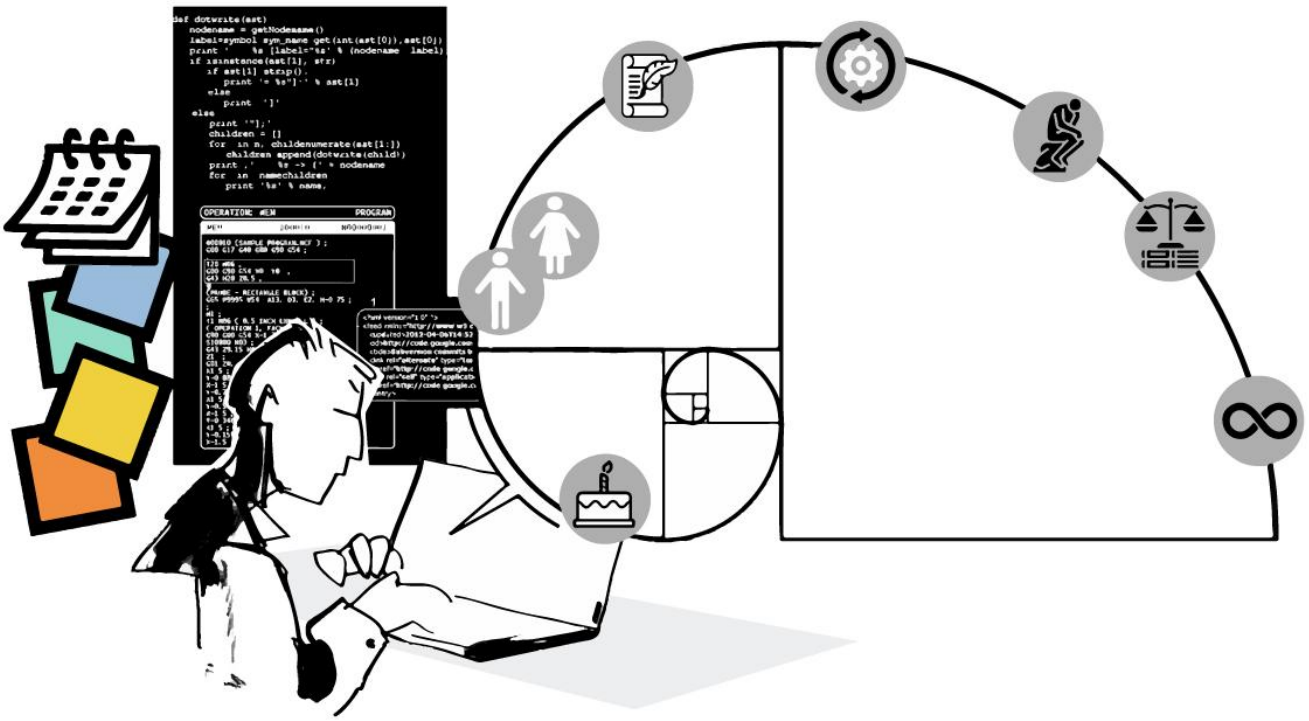
GENEVA INTERNET PLATFORM
digwatch

NEWSLETTER

Issue 85, **December 2023**



The four seasons of AI

Page 4

DIGITAL POLICY SNAPSHOT

AI governance and content policy dominated news headlines in November.

Pages 2-3

IN FOCUS

The EU lawmakers are warring on the EU AI Act, with foundation models and biometric uses of AI being the biggest stumbling stones in negotiations.

Page 6

IN FOCUS

Have OpenAI's Q* and Google's Gemini raised stakes in the race for artificial general intelligence?

Page 7

LAUNCH

The Geneva Manual clarifies the roles and responsibilities of these non-state stakeholders in implementing cyber norms.

Page 8

Snapshot: What's making waves in digital policy?

AI governance

Google and Anthropic have announced an expanded partnership, encompassing joint efforts on AI safety standards, committing to the highest standards of AI security, and using TPU chips for AI inference.

Google has unveiled 'The AI Opportunity Agenda,' offering policy guidelines for policymakers, companies, and civil societies to collaborate in embracing AI and capitalising on its benefits.

The OECD launched The AI Incidents Monitor, which offers comprehensive policy analysis and data on AI incidents, shedding light on AI's impacts to help shape informed AI policies. US President Joe Biden and Chinese President Xi Jinping, on the sidelines of the Asia-Pacific Economic Cooperation's (APEC) Leaders' Week, agreed on the need 'to address the risks of advanced AI systems and improve AI safety through USA-China government talks.

The Italian Data Protection Authority (DPA) initiated a fact-finding inquiry to assess whether online platforms have implemented sufficient measures to stop AI platforms from scraping personal data for training AI algorithms.

Switzerland's Federal Council has tasked the Department of the Environment, Transport, Energy, and Communications (DETEC) with providing an overview of potential regulatory approaches for AI by the end of 2024. The council aims to use the analysis as a foundation for an AI regulatory proposal in 2025.

Technologies

Yangtze Memory Technologies Co (YMTC), China's largest memory chip manufacturer, filed a lawsuit against Micron Technology and its subsidiary for violating eight patents. Under the EU-India Trade and Technology Council (TTC) framework, the EU and India signed a Memorandum of Understanding on working arrangements in the semiconductor ecosystem, its supply chain, and innovation. Joby Aviation and Volocopter air taxi manufacturers showcased their electric aircraft in New York. Amazon introduced Q, an AI-driven chatbot tailored for its Amazon Web Services, Inc. (AWS) customers, serving as a versatile solution catering to business intelligence and programming needs.

Security

The UK, the USA, and 16 other partners have released the first global guidelines to enhance cybersecurity throughout the life cycle of an AI system. The guidelines span four key areas within the life cycle of the development of an AI system: secure design, secure development, secure deployment, and secure operation and maintenance.

The EU Parliament and the EU Council reached a political agreement on the Cyber Resilience Act. The agreement will now be subject to formal approval by the parliament and the council.

Infrastructure

The EU's Gigabit Infrastructure Act (GIA) is undergoing significant alteration as the 'tacit approval principle,' designed to expedite the deployment of broadband networks, has been excluded from the latest compromise text circulated by the Spanish presidency of the EU Council. ICANN launched its Registration Data Request Service (RDRS) to simplify requests for access to nonpublic registration data related to generic top-level domains (gTLDs).

The International Telecommunication Union (ITU) has adopted ITU-R Resolution 65, which aims to guide the development of a 6G standard. This resolution enables studies on the compatibility of current regulations with potential 6th-generation international mobile telecommunications (IMT) radio interface technologies for 2030 and beyond.

The Indian government has launched its Global Digital Public Infrastructure Repository and created the Social Impact Fund to advance digital public infrastructure in the Global South as part of its G20 initiatives.

Legal

The EU Council adopted the Data Act, setting principles of data access, portability, and sharing for users of IoT products. OpenAI has initiated the Copyright Shield, a program specifically covering legal expenses for its business customers who face copyright infringement claims stemming from using OpenAI's AI technology.

Internet economy

Apple, TikTok, and Meta appealed against their gatekeeper classification under the EU Digital Markets Act (DMA), which aims to enable user mobility between rival services like social media platforms and web browsers. Conversely, Microsoft and Google have opted not to contest the gatekeeper label. The US Treasury reached a record $4.2 billion settlement with Binance – the world's largest virtual currency exchange, for violating anti-money laundering and sanctions laws, mandating a five-year monitoring period and rigorous compliance measures. Australia's regulator called for a new competition law for digital platforms due to their growing influence.

Digital rights

The Court of Justice of the EU (CJEU) ruled that data subjects have the right to appeal the decision of the national supervisory authority regarding the processing of their personal data.

Content policy

Nepal decided to ban TikTok, citing the disruption of social harmony caused by the misuse of the popular video app. YouTube introduced a new policy that requires creators to disclose the use of Generative AI. OpenAI and Anthropic have joined the Christchurch Call to Action, a project started by French President Emmanuel Macron and then New Zealand Prime Minister Jacinda Ardern to suppress terrorist content. X (formerly Twiter) is on the EU Commission's radar for having significantly fewer content moderators than its rivals.

Development

ITU's Facts and Figures 2023 report reveals uneven progress in global internet connectivity, which exacerbates the disparities of the digital divide, particularly in low-income countries. Switzerland announced plans for a new state-run digital identity system, slated for launch in 2026, after voters rejected a private initiative in 2021 due to personal data protection concerns. Indonesia's Ministry of Communication and Information introduced a new policy on digital identity, which will later require all citizens to have a digital ID.

THE TALK OF THE TOWN – GENEVA

The World Economic Forum (WEF) held its Annual Meeting on Cybersecurity 2023 on 13–15 November, assembling over 150 leading cybersecurity experts. Based on WEF's Global Security Outlook 2023 report released in January 2023, the annual meeting provided a space for experts to address growing cyber risks with strategic, systemic approaches and multistakeholder collaborations.

The 12th UN Forum on Business and Human Rights took place from 27 to 29 November, focusing on the actual changes that have been made by states and businesses to implement the UN Guiding Principles on Business and Human Rights (UNGPs) standards. Among the discussed topics was the improvement in disability rights implementation via advancements in assistive technologies, AI and digitalisation, and other care and support systems.

Held in conjunction with the 12th UN Forum on Business and Human Rights, the UN B-Tech Generative Summit on 30 November explored the undertaking of due diligence in human rights when putting AI into practice. The full-day summit presented the B-Tech Project's papers on human rights and generative AI and provided a platform for all stakeholders to discuss the practical uses of the UN Guiding Principles on Business and Human Rights (UNGP) and other human-rights-based approaches in analysing the impacts of generative AI.

Four seasons of AI

ChatGPT, a revolutionary creation by OpenAI launched on 30 November 2022, has not only captivated the tech world but also shaped the narrative around AI. As ChatGPT marks its first anniversary, it prompts a collective step back to reflect on the journey so far and consider what lies ahead.

A symbolic journey through the seasons has been a compelling backdrop to AI's trajectory since last November. The *winter of excitement* saw rapid user adoption, surpassing even social media giants with its pace. Within 64 days, ChatGPT amassed an astounding 100 million users, a feat that Instagram, for instance, took 75 days to achieve. The sudden surge in interest in generative AI has taken major tech companies by surprise. In addition to ChatGPT, several other notable generative AI models, such as Midjourney, Stable Diffusion, and Google's Bard, have been released.

The subsequent *spring of metaphors* ushered in a wave of imaginative comparisons and discussions on AI governance. Anthropomorphic descriptions and doomsday scenarios emerged, reflecting society's attempts to grapple with the implications of advanced AI.

As ChatGPT entered its contemplative *summer of reflection*, a period of introspection ensued. Drawing inspiration from ancient philosophies and cultural contexts, the discourse broadened beyond mere technological advancements. The exploration of wisdom from Ancient Greece to Confucius, India, and the Ubuntu concept in Africa sought answers to the complex challenges posed by AI, extending beyond simple technological solutions.

Now, in the *autumn of clarity*, the initial hype has subsided, making room for precise policy formulations. AI has secured its place on the agendas of national parliaments and international organisations. In policy documents from various groups like the G7, G20, G77, and the UN, the balance between opportunities and risks has shifted towards a greater focus on risks. The long-term existential threats of AI have taken centre stage in conferences like the London AI Summit, with governance proposals drawing inspiration from entities like the International Atomic Agency (IAAA), CERN, and the International Panel on Climate Change (IPCC).

What lies ahead? We should focus on the two main issues at hand: how to address AI risks and what aspects of AI should be governed.

In managing AI risks, a comprehensive understanding of three categories – immediate knowns, looming unknowns, and long-term unknowns – is crucial for shaping effective regulations. While short-term risks like job loss and data protection are familiar and addressable with existing tools, mid-term risks involve potential monopolies controlling AI knowledge, demanding attention to avoid dystopian scenarios. Long-term risks encompassing existential threats dominate public discourse and policymaking, as seen in the Bletchley Declaration. Navigating the AI governance debate requires transparently addressing risks and prioritising decisions based on societal responses.

Regarding the governance of AI aspects, current discussions revolve around computation, data, algorithms, and applications. Computation aspects involve the race for powerful hardware, with geopolitical implications between the USA and China. The data, often called the *oil* of AI, demands increased transparency regarding its usage. Algorithmic governance, which is focused on long-term risks, centres on the relevance of weights in AI models. At the apps and tools level, the current shift from algorithmic to application-focused regulations may significantly impact technological progress. Debates often overlook data and app governance, areas detailed in regulation but not aligned with tech companies' interests.

This text is inspired by Dr J. Kurbalija's Recycling Ideas blog series. It's a collection of concepts, traditions, and thoughts aimed at constructing a social contract suitable for the AI era.

EU lawmakers warring over the bloc's AI Act

After more than 22 hours of the initial trilogue negotiations in the EU on 6 and 7 December, encompassing an agenda of 23 items, agreement on the AI Act remained elusive. Here's what reports pointed to.

Foundation models. The negotiations hit a significant snag when France, Germany, and Italy spoke out against the tiered approach initially envisioned in the EU AI Act for foundation models (base models for developers). The tiered approach would mean categorising AI into different risk bands, with more or less regulation depending on the risk level. What France, Germany, and Italy want is to regulate only the use of AI rather than the technology itself, because they want to ensure that AI innovation in the EU is not stifled. They proposed 'mandatory self-regulation through codes of conduct' for foundation models. European Parliament officials walked out of a meeting to signal that leaving foundation models out of the law was not politically acceptable.

According to a compromise document seen by Euractiv, the tiered approach was retained in the text of the act. However, the legislation would not apply to general-purpose AI (GPAI) systems offered under free and open-source licenses. This exemption can be nullified if the open-source model is put into commercial use. At the same time, lawmakers agreed that the codes of conduct would serve as supplementary guidelines until technical standards are harmonised.

According to the preliminary agreement, any model that was trained using computing power greater than 10^25 floating point operations (FLOPs) will be automatically categorised as having systemic risks. These models would face extensive obligations, including evaluation, risk assessment, cybersecurity, and energy consumption reporting.

An EU AI office will be established within the commission to enforce foundational model rules, with national authorities overseeing AI systems through the European Artificial Intelligence Board (EAIB) for consistent application of the law. An advisory forum will gather feedback from stakeholders. A scientific panel of independent experts will advise on enforcement, identify systemic risks, and contribute to the classification of AI models.

Contentious issues. While approximately ten issues remain unresolved on the agenda, the primary obstacles revolve around prohibitions, remote biometric identification, and the national security exemption.

Prohibitions. So far, lawmakers have tentatively agreed on prohibiting manipulative techniques, systems exploiting vulnerabilities, social scoring, and indiscriminate scraping of facial images. At the same time, the European Parliament has proposed a much longer list of prohibited practices and is facing a strong pushback from the council.

Remote biometric identification. Members of the European Parliament (MEPs) are pushing for a blanket ban on biometric categorisation systems based on sensitive personal traits, including race, political opinions, and religious beliefs. At the same time, member states are pushing for exemptions to use biometric surveillance when there is a threat to national security.

National security exemption. France, leading the EU countries, advocates for a broad national security exemption in AI regulations, emphasising member states' discretion in military, defence, and national security issues. However, this will likely face resistance from progressive lawmakers, who will likely advocate for an outright ban.

Update: After 36 hours of negotiations over three days (22 of which were consecutive), a provisional agreement was finally reached. Explore the details in our Weekly #139.

Higher stakes in the race for AGI?

The buzz around OpenAI's November saga has been nothing short of gripping, and we've been right in the thick of it, following every twist and turn.

In summary, OpenAI CEO Sam Altman was ousted from the company because he 'was not consistently candid in his communications' with the board. Most of OpenAI's workforce, approximately 700 out of 750 employees, expressed their intention to resign and join Altman at Microsoft, prompting his reinstatement as CEO. Additionally, OpenAI's board changed some of its members.

Reports (and speculation) of Q* swiftly broke through. Reuters reported that Altman was dismissed partly because of Q*, an AI project allegedly so powerful that it could threaten humanity.

Q* can supposedly solve certain math problems. Although its mathematical prowess is on the level of grade-school students (the first 6 or 8 grades), this could be a potential breakthrough in artificial general intelligence (AGI), as it suggests a higher reasoning capacity. OpenAI sees AGI as AI that aims to surpass human capabilities in economically valuable tasks.

Upon his return as CEO, Altman's comment about Q* was: 'No particular comment on that unfortunate leak.'

The news has caused quite a stir, with many wondering what exactly Q* is, if it even exists. Some savvy observers think Q* might be tied to a project from OpenAI in May, bragging about 'process supervision' – a technique that trains AI models to crack problems step-by-step.

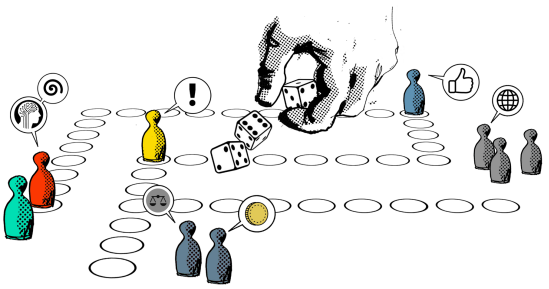
Some theorise the Q* project might blend Q-learning (i.e. a type of reinforcement learning where a model iteratively learns and improves over time by being rewarded for taking the correct action) with an algorithm that computers can use to figure out how to get somewhere between two places quickly (A* search).

Others posited that the name Q* might reference the Q* search algorithm, which was developed to control deductive searches in an experimental system.

Google joins the race. The beginning of December saw the launch of Google's Gemini, an AI model that, according to Google, has outperformed human experts on massive multitask language understanding, a measurement designed to measure AI's knowledge of math, history, law, and ethics. This model reportedly can outperform GPT-4 in grade school math. However, Google has declined to comment on Gemini's parameter counts.

Is this all really about AGI? Well, it's hard to tell. On the one hand, AI surpassing human capabilities sounds like a dystopia (why does no one ever think it might be a utopia?) is ahead. On the other hand, experts say that even if an AI could solve math equations, it wouldn't necessarily translate to broader AGI breakthroughs.

What are all these speculations really about? Transparency – and not only at OpenAI and Google. We need to understand who (or what) will shape our future. Are we the leading actors or just audience members waiting to see what happens next?



Balancing online speech: Combating hate while preserving freedom

The ongoing battle about preventing and combating online hate speech while ensuring that freedom of expression is protected has had the EU Agency for Fundamental Rights (FRA) calling for 'appropriate and accurate content moderation'.

FRA has published a report on the challenges in detecting online hate speech against people of African descent, Jewish people, and Roma people and others on digital platforms, including Telegram, X (formerly known as Twitter), Reddit, and YouTube. Data were collected from Bulgaria, Germany, Italy, and Sweden to provide a comparative analysis based on their current national policies. FRA called regulators and digital platforms to ensure a safer space for people of African descent, Jews, and Roma because it was found that they experience very high levels of hate speech and cyber harassment. Additionally, FRA drew attention to effective content moderation regulation for women as there are higher levels of incitement to violence against them compared to other groups.

Is the DSA enough to ensure content moderation in the EU? While the Digital Security Act (DSA) is considered a big step in moderating online hate speech, FRA claims its effect is yet to be seen. According to FRA, clarification is needed about what is regarded as hate speech, including training for law enforcement, content moderators, and flaggers about legal thresholds for the identification of hate speech. This training should also ensure that platforms do not over-remove content.

UNESCO's guidelines. UNESCO's Director-General, Audrey Azoulay, sounded an alarm about the surge in online disinformation and hate speech, labelling them a 'major threat to stability and social cohesion'. In response, UNESCO published guidelines for the governance of digital platforms to combat online disinformation and hate speech while protecting freedom of expression. The guidelines include establishing independent public regulators in countries worldwide, ensuring linguistically diverse moderators on digital platforms, prioritising transparency in media financing, and promoting critical thinking.

The importance of civil society. Since the Israeli-Palestinian war began, posts about Palestine and content removals reached an 'unprecedented scale' said Jillian York, the director for international freedom of expression at the Electronic Freedom Foundation (EFF). Thus, several Palestinian human rights advocacy groups initiated the 'Meta: Let Palestine Speak' petition calling on the tech giant to address the unfair removal of Palestinian content.

And, of course, AI. As found in FRA's report, human-based content assessment often uses biassed and discriminatory parameters. This, however, does not mean that AI could be prevented from doing this, as seen in Meta's auto-translation, which applied the term 'terrorist' to Palestinian users who had an Arabic phrase in their bios, for which Meta publicly apologised in October 2023.

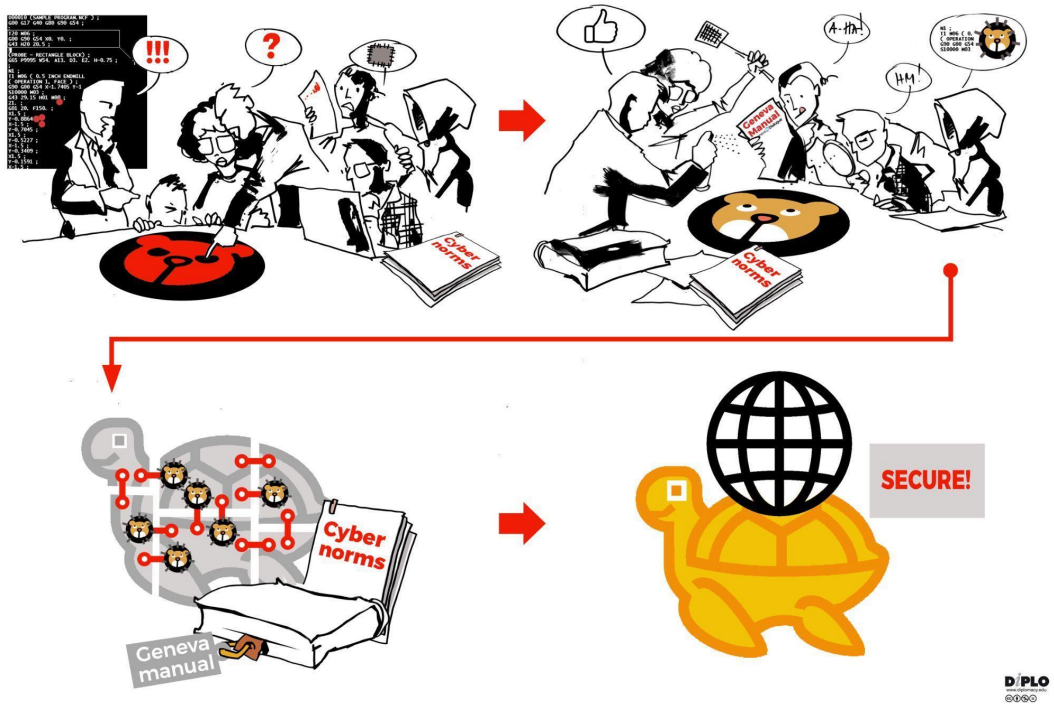
Launch of the Geneva Manual on Responsible Behaviour in Cyberspace

The recently launched Geneva Manual focuses on the roles and responsibilities of non-state stakeholders in implementing two UN cyber norms related to supply chain security and responsible reporting of ICT vulnerabilities.

The manual was drafted by the Geneva Dialogue on Responsible Behaviour in Cyberspace, an initiative established by the Swiss Federal Department of Foreign Affairs and led by DiploFoundation with the support of the Republic and state of Geneva, the Center for Digital Trust (C4DT) at the Swiss Federal Institute of Technology in Lausanne (EPFL), Swisscom, and UBS.

The Geneva Dialogue plans to expand the manual by discussing the implementation of additional norms in the coming years.

The manual is a living document, open for engagement and enrichment. Visit genevadialogue.ch to contribute your thoughts, ideas, and suggestions, as we chart a course toward a more secure and stable cyberspace.



About this issue: Issue 85 of the Digital Watch monthly newsletter, published on 8 December 2023 by the Geneva Internet Platform and DiploFoundation. For more coverage and analysis, visit the Digital Watch observatory
Team: Andrijana Gavrilović (author), Boris Begović (contributor), Anastasiya Kazakhova (contributor), Bojana Kovač (contributor), Yung-Hsuan Wu (contributor), Ginger Paque (editor), Diplo's CreativeLab (design) | Get in touch: digitalwatch@diplomacy.edu
On the cover: *The four seasons of AI.* Credit: Vladimir Veljasević DiploFoundation (2023)

The Geneva Internet Platform is an initiative of:
GENEVA INTERNET PLATFORM digwatch
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
REPUBLIQUE ET CANTON DE GENEVE
DiPLO www.diplomacy.edu