Background Paper for the
Transatlantic Leadership Network's Transatlantic Technology and Trade Forum

Towards Strengthening the Transatlantic Tech Diplomacy: Trustworthy AI in the EU-U.S. Trade and Technology Council

Raluca Csernatoni, PhD
Research Fellow and EU Cyber Direct Team Leader on New Technologies, Carnegie Europe
Guest Professor, Centre for Security, Diplomacy and Strategy, Vrije Universiteit Brussel

Abstract

The EU-U.S. Trade and Technology Council (TTC) is a valuable diplomatic initiative and a socialising forum for enabling stronger and sharper transatlantic cooperation, even if the outcomes of the negotiations are not always as concrete and fast-paced as expected. At the third TTC held outside Washington DC on 5 December 2022, the EU and U.S. policymakers issued joint initiatives on Artificial Intelligence (AI) including the first AI Roadmap, which will 'enable trustworthy AI systems that enhance innovation, lower barriers to trade, bolster market competition, operationalise common values, and protect the universal human rights and dignity of our citizens.'[1] While the aspirational language is to be applauded, the roadmap should pave the way for concrete joint actions bringing the EU and the U.S. on the same page concerning the technical dimensions of AI and setting global technological standards. A first step should be pinning down common definitions on key buzzworthy terms, such as trustworthy, risk-based, ethical, robust, safe, and human-centric AI systems.

Introduction

With the European Union's (EU) Artificial Intelligence (AI) Act expected to be finalised in 2023 and the growing hype around ChatGPT-3 at the end of 2022 and early 2023, the governance of AI will likely be high on the EU's and the United States' (U.S.) policy agendas this year. ChatGPT, heralded as the most advanced AI chatbot in 2022, is a large language model developed by OpenAI that can be used for natural language processing tasks such as text generation and language translation.[2] Touted as the start of the AI revolution[3] and an AI Cambrian Explosion, ChatGPT is indeed an extraordinary accomplishment. And as the technology advances, it will become a remarkable source for researchers, policymakers, developers, academics, and journalists.

[1] https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management
[2] https://openai.com/blog/chatgpt/
[3] https://www.washingtonpost.com/business/is-chatgpt-the-start-of-the-ai-revolution/2022/12/09/9eb9dfec-77b5-11ed-a199-927b334b939f_story.html

Yet, it also raises substantive **ontological, technical, ethical, and regulatory questions**, especially regarding its profound effects on the knowledge economy, disinformation practices, copyright infringements, and other malicious cyber activities.[4] Even with enormous benefits for the so-called 'better good' of humanity, ChatGPT is only one example of the many generative AI systems, and the battle against its legislative aspects has just started. In view of such rapid developments, stronger and more coordinated cooperation on AI between the EU and the U.S. in key regulatory areas is ever more necessary.

This has been made clear during the **U.S.-EU Trade and Technology Council (TTC)**[5] **third ministerial meeting on 5 December 2022**, held at the University of Maryland in College Park.[6] Indeed, building on the close and strategic ties between the EU and the U.S., the TTC, since its initial launch at the June 2021 EU-U.S. Summit, has advanced cooperation on several critical economic and technology initiatives in line with shared democratic values and to meet recent geopolitical challenges. The TTC is now the preferred transatlantic political body servings as diplomatic forum to coordinate tech and trade policy between the U.S. and the EU, comprising of 10 **TTC Working Groups**, which are chaired by relevant U.S. agencies and European Commission services.

Specifically, the TTC Working Group on Technology Standards and the work of its TTC subgroup dedicated to sharing expertise on AI are valuable frameworks for transatlantic collaboration on AI. When it comes to AI, both parties have endorsed the OECD Principles on AI and are founding members of the Global Partnership on Artificial Intelligence (GPAI), an international and multistakeholder initiative developed within the G7 open to those who endorse the OECD Principles.[7] As AI systems increasingly transform all spheres of life, both the EU and the U.S. should indeed jointly sketch an ambitious and risk-based vision for the promotion of trustworthy AI, by articulating opportunities, risks, and ethical challenges that accompany current AI developments.

Trustworthy AI and Collaborating on New and Emerging Technologies

The third TTC meeting from December 2022 was set against the backdrop of a sober international landscape: from geopolitical rifts, the ongoing war in Ukraine, the rise of technological sovereignty across the world,[8] disruptions in critical supply chains, trade tensions, to the proliferation of emerging and disruptive technologies (EDTs).[9] This meeting was also held amid trade disputes and European discontent at the U.S.'s recently passed 'Buy America' provisions in the Inflation Reduction Act (IRA),[10] which, among other things, would subsidise U.S.-manufactured electric vehicles, thus putting EU-based automotive manufacturers at a disadvantage in the race for the global electric market. Likewise,

[4] https://carnegieeurope.eu/2022/09/15/artificial-intelligence-and-cybersecurity-nexus-taking-stock-of-european-union-s-approach-pub-87886
[5] https://www.whitehouse.gov/briefing-room/statements-releases/2022/05/16/fact-sheet-u-s-eu-trade-and-technology-council-establishes-economic-and-technology-policies-initiatives/
[6] https://www.whitehouse.gov/briefing-room/statements-releases/2022/12/05/fact-sheet-u-s-eu-trade-and-technology-council-advances-concrete-action-on-transatlantic-cooperation/
[7] https://www.transatlantic.org/wp-content/uploads/2022/03/TTC-Artificial-Intelligence.pdf
[8] https://eucyberdirect.eu/research/digital-sovereignty-narrative-policy
[9] https://www.lse.ac.uk/ideas/Assets/Documents/reports/LSE-IDEAS-FNF-Beyond-Autonomy.pdf#page=28
[10] https://www.congress.gov/bill/117th-congress/house-bill/5376/text

addressing U.S. concerns about the EU's quest for digital and technological sovereignty will require EU policymakers to nuance their approach.

Notwithstanding such tensions, at the occasion of the meeting, the EU and U.S. policymakers discussed further controls on emerging technologies, including work done be the TTC subgroup dedicated to AI. Importantly, they adopted a Joint Statement agreeing on the EU-U.S. TTC Joint AI Roadmap to develop common tools and standards for trustworthy AI; the Pilot Project on Privacy-Enhancing Technologies; the Collaboration on AI and Computing Research for the Public Good. Additionally, a joint study was issued on 'The Impact of Artificial Intelligence on the Future of Workforces in the EU and the U.S.'.

The **TTC Joint Roadmap for Trustworthy AI and Risk Management**[11] aims to 'advance shared terminologies and taxonomies, but also to inform our approaches to AI risk management and trustworthy AI on both sides of the Atlantic', which would equally support ongoing cooperative work in other contexts such as the OECD and GPAI. As a next collaborative step on AI, the goal of the roadmap is to build a common repository of metrics for measuring AI trustworthiness and risk management methods. It also has the potential to inform and advance collaborative approaches in international standards bodies related to AI. The roadmap further envisages to enable trustworthy AI systems that 'enhance innovation, lower barriers to trade, bolster market competition, operationalise common values, and protect the universal human rights and dignity of our citizens'.[12] However, instead of substantively zooming in on human rights, it could be observed that the roadmap is largely focused on getting the EU and the U.S. on the same page concerning the technical dimensions of AI, and setting global technological standards, rather than outright prohibiting certain worrying uses like facial recognition in public spaces.

Arguably, **'trustworthy'** AI is yet another buzzworthy term in a series of notions used to describe AI that is human-centric, responsible, lawful, ethical, and technically robust. Establishing trust can be extremely difficult, especially since the AI value chain is incredibly complicated and often hard to untangle. The underlying idea behind this framing is that AI will only reach its full potential when trust can be established in every stage of its lifecycle, from research, design to development, fielding, and uses. Yet, the OpenAI's Chat GPT case and its likelihood to produce potentially dangerous content within hours, despite ex-ante measures to develop and test the technology by the world's leading researchers to avoid such responses, showcases that it is nearly impossible to achieve trustworthiness, robustness, and transparency, as well as anticipate malicious uses, as the domains it is applied in become more complex.

Under the banner of the TTC, both the EU and the U.S. have agreed to develop and implement 'trustworthy AI' as part of a commitment to promote a 'human-cantered' and a risk-based approach.[13] This becomes even more relevant as a new crop of generative AI tool could be used to further infringe upon human rights and disrupt democratic processes. One big question is how regulators will approach the technology. Building on a trustworthy AI approach, the EU has been already working on

[11] https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management
[12] https://ec.europa.eu/commission/presscorner/detail/en/statement_22_7516
[13] https://www.transatlantic.org/wp-content/uploads/2022/03/TTC-Artificial-Intelligence.pdf

the **AI Act** (AIA) since its introduction in April 2021,[14] which aspires to establish the first sweeping regulatory scheme for governing such technology. By putting forward a risk-based approach along four levels of risk in AI, from unacceptable, high, limited, to minimal or no risks, the EU's AI Act will be the first-ever horizontal legislation to regulate AI systems.

Following multiple amendments and discussions, notably concerning definitional issues surrounding **general purpose AI** and requirements for high-risk systems, the Council of the EU adopted on 6 December 2022 its common position or 'general approach' on the Act. Its aim is to ensure that AI systems placed on the EU market and used across the bloc are trustworthy, safe, and respect existing laws on fundamental rights and EU values. Accordingly, general purpose AI covers systems that could handle tasks such as language processing, image and speech recognition, pattern detection, and question answering. More specifically, in the last iteration of the Act, many of the requirements for high-risk AI systems have been clarified and fine-tuned in order to be more technically feasible and less burdensome for stakeholders to comply with, for instance with regard to issues related to the quality of data.

In terms of general-purpose AI systems (GPAIS) and with a view of regulating large generative AI models (LGAIMs), new provisions have been added to account for instances in which AI systems can be used for many different purposes and where general-purpose AI technology is consequently integrated into another high-risk system. Increasingly popular models such as ChatGPT should fall into this category, as they can generate content-based on human inputs, and may take fake news, manipulation, and hate speech to unprecedented levels if not properly governed, thus marking a new frontline for 'trustworthy' AI. Conversely, the U.S. administration has warned that placing risk-management requirement on general purpose AI systems providers could prove 'very burdensome, technically difficult and in some cases impossible', while leading providers of such systems being large American companies.[15]

Still, U.S. policymakers also share the EU's interest in mitigating risks associated with AI and in fostering trustworthy AI systems. Contrasting to the EU's substantive and horizontal efforts to put forward a biding AI regulation, and in order to avoid regulatory interventions that would hamper AI innovation, the U.S. does not favour federal-level action and leaves such tasks to independent regulatory agencies.[16] What is more, given the fast-paced AI developments and the serious implications of the technological giants' failed voluntary and self-regulatory practices, U.S. policymakers have reached the conclusion that more proactive action is needed to safeguard fundamental rights and democratic societies.[17]

In this regard, reflecting the Biden Administration's approach to algorithmic regulation, the White House Office of Science and Technology Policy has published the '**Blueprint for and AI Bill of**

[14] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[15] https://www.euractiv.com/section/digital/news/the-us-unofficial-position-on-upcoming-eu-artificial-intelligence-rules/
[16] https://www.transatlantic.org/wp-content/uploads/2022/03/TTC-Artificial-Intelligence.pdf
[17] https://www.transatlantic.org/wp-content/uploads/2021/12/12-13-2021-Hajdu_Sawiris_Transatlantic-Approach-to-AI.pdf

Rights. Making Automated Systems work for the American People' (AIBoR).[18] While it is not meant to be a universal AI guidance and it is nonbinding blueprint, with is set of five principles and associated practices to help guide the design, deployment, and uses of automated systems, it aims to guarantee protections from biased or inaccurate algorithms, ensure transparency, and safeguard citizens from pervasive or discriminatory surveillance.[19]

Yet, the degree to which the blueprint's principles will be implemented largely depends on the actions and practices of U.S. federal agencies. Hence, the U.S. still approaches AI regulations with a light touch. When it comes to the risk management of AI systems, the National Institute for Standards and Technology (NIST) is creating an **AI Risk Management Framework (AI RMF)**.[20] To be published at the end of January 2023 and putting forward an approach inclusive of all stakeholders, the framework aims to 'better manage risks to individuals, organisations, and society associated with artificial intelligence (AI)'.

Contrasting to the EU AI Act, the AI RMF is designed to facilitate the management of risks from any type of AI system, not just those classified as high-risk.[21] It is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. Despite differing regulatory approaches, noteworthy is the fact that both the EU and the U.S. are aligned around the goal to ensure the innovation, development, deployment, and use of trustworthy AI. Furthermore, the TTC is indeed the right forum to address differences in policy approaches between the two parties, by increasing synergies, mitigating risks associated with AI, while also respecting the regulatory autonomy of each of the two parties.

The third TTC meeting from December 2022 also explored the need to recognise the importance of privacy in advancing responsible AI development, with the EU and the U.S. working together on a Pilot Project[22] to assess the use of privacy-enhancing technologies and synthetic data in healthcare and medicine, in line with applicable data protection rules. Besides, a joint study on '**The Impact of Artificial Intelligence on the Future of Workforces in the EU and the U.S.'**[23] was finalised, intending to highlight the economics behind AI-driven technological change with a particular focus on the institutional and policy decisions that will shape the AI's future (disruptive) impact on the workforce.

Interestingly, to illustrate the transformative power of AI technology, the study highlights tools such as the OpenAI's ChatGPT (p.5) and DALL-E, a model trained to generate images from a text description provided by a user (p.7). In the study and in the example on ChatGPT, it was asked to both provide its own definition of AI, which was proficiently articulated in one paragraph, as well as assess the AI's positive and negative impacts on the future of the workforce. Among the negative effects,

[18] https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[19] https://www.brookings.edu/2022/11/21/the-ai-bill-of-rights-makes-uneven-progress-on-algorithmic-protections/
[20] https://www.nist.gov/itl/ai-risk-management-framework
[21] https://futureoflife.org/wp-content/uploads/2022/08/Lessons_from_NIST_AI_RMF-v2.pdf
[22] https://ec.europa.eu/commission/presscorner/detail/en/statement_22_7516
[23] https://digital-strategy.ec.europa.eu/en/library/impact-artificial-intelligence-future-workforces-eu-and-us

ChatGPT listed job losses, inequality, security risks as malicious actors begin to use AI technology, and last but not the least, it emphasised ethical concerns (p. 16-18).

With OpenAI's ChatGPT-4, AnthropicAI's Claude,[24] DeepMind's Sparrow,[25] and Text-to-Video fast coming, the challenge is to untangle the impact of such tools on a so-called post-rational and post-scientific world view, 'a belief that if you gather enough data and have enough computing power, you can "create" authoritative knowledge […] in which intellectual authority rests not with subject matter experts but with those who can create and manipulate digital data.'[26]

Undeniably, the challenge is not only to make sense of the hype surrounding such technologies, but also to understand just how such technologies with human-like traits can be trusted, as well as how they will change how we live, create knowledge, and mediate work. Here, the bone of contention centres around the fact that the private sector is driven by a profit maximising logic when it comes to developing and adopting AI systems for automating work, which may not necessarily lead to socially desirable outcomes for the so-called 'good of humanity'.

Conclusion

As there are very distinctive visions on the governance and regulation of AI globally, it is crucial for the EU and the U.S. to clearly spell out how various AI systems should be aligned with human rights and democratic principles. While regulatory alignment between the EU and U.S. is still to be achieved, and differences between the AI governance approaches persist due to a different hierarchy of values, fundamental rights, and strategic interests, the TTC Joint Roadmap is a promising step towards achieving more transatlantic convergence.

The next step is to ensure even more concreteness in the roadmap's stated objectives and when it comes to coordinated action. According to the Centre for AI and Digital Policy, a Washington DC-based think tank, the TTC has made too little advancements on ensuring that AI cannot be used for unethical purposes.[27] Despite such challenges, it is undeniable that both the EU and the U.S. are facing an insecure geopolitical landscape, coupled by the proliferation of dual-use EDTs.

Hence, the case for transatlantic cooperation on the governance of trustworthy new and emerging technologies has never been stronger. Comprehensive political buy-in on both sides of the Atlantic is a necessary piece in the puzzle to deepen transatlantic unity and to overcome short-term disputes or divergencies. Now is indeed the time to move from rhetoric to action.

[24] https://techcrunch.com/2023/01/09/anthropics-claude-improves-on-chatgpt-but-still-suffers-from-limitations/
[25] https://www.deepmind.com/blog/building-safer-dialogue-agents
[26] https://www.cigionline.org/articles/chatgpt-strikes-at-the-heart-of-the-scientific-world-view/?utm_source=cigi_newsletter&utm_medium=email&utm_campaign=chatgpt-whats-at-stake
[27] https://sciencebusiness.net/news/eu-and-us-set-out-plan-create-rules-road-artificial-intelligence