MORATORIUM OF LARGE-SCALE AI CAPABILITIES *Niklas V. Lehmann*\*

ABSTRACT

The structure of this document closely resembles that of a genuine treaty. Part I, titled the ”Protocol on the Development and Deployment of Artificial Intelligence,” serves as the treaty itself. Part II, known as ”Comments on the Protocol,” offers valuable insights into the treaty’s design and highlights crucial nuances for its successful implementation. I suggest reading Part II before delving into Part I.

TABLE OF CONTENTS

PREAMBLE 2

I PROTOCOL ON THE DEVELOPMENT AND DEPLOYMENT OF ARTI FICIAL INTELLIGENCE 2 *A Article 1: Definitions* . . . . . . . . . . . . . . . . . . . . . . . . 2 *B Article 2: International Committee on Artificial Intelligence* . . . 2 *C Article 3: Financing of the protocol and the Committee* . . . . . . 3 *D Article 4: Compute capacity restrictions* . . . . . . . . . . . . . . 3 *E Article 5: Windfall clause* . . . . . . . . . . . . . . . . . . . . . . 4 *F Article 6: Entry into force* . . . . . . . . . . . . . . . . . . . . . 4 *G Article 7: Successor to this protocol* . . . . . . . . . . . . . . . . 5 *H Article 8: Withdrawal and Non-compliance* . . . . . . . . . . . . 5 *I Article 9: Reporting of data, public awareness and exchange of information* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

COMMENTS ON THE PROTOCOL 6 *J Safe Artificial Intelligence as a public good* . . . . . . . . . . . . 6 *K Early licensing and registration* . . . . . . . . . . . . . . . . . . 7 *L Windfall clause* . . . . . . . . . . . . . . . . . . . . . . . . . . . 7 *M Re-negotiation* . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

\*PhD candidate in Economics at Technical University Bergakademie Freiberg. Contact: niklasl.2306@gmail.com

1

2 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 15-Jul-23

PREAMBLE

This protocol shall ensure the safe development of artificial intelligence, to benefit hu mankind in the present and protect future generations from harm. This protocol aims to regulate the development of artificial intelligence on a global scale, so as to allow coordi nation and cooperation to triumph over unsafe use of this powerful tool.

I. PROTOCOL ON THE DEVELOPMENT AND DEPLOYMENT OF ARTIFICIAL INTELLIGENCE

*A. Article 1: Definitions*

• ”Artificial Intelligence” (AI) refers to a computer system that receives and processes information from its environment to produce an output based on its function.

• ”Model” refers to a complete specification, i.e. blueprint, of an artificially intelligent system.

• ”Parameters” refer to the variables that are set and adjusted during the training of the AI to optimize performance for a specific task. Any part of the system that is used to control and adjust the behavior of the system in order to achieve desirable outcomes is considered a parameter.

• ”Committee” refers to the *International Committee on artificial intelligence*, an in ternational body that shall oversee this protocol.

*B. Article 2: International Committee on Artificial Intelligence*

The International Committee on Artificial Intelligence, hereinafter referred to as the ”Com mittee,” shall be established as a globally recognized institution. The Committee shall be comprised of one representative from each party to this protocol. Its primary responsibility shall be to monitor and supervise the development of artificial intelligence in all member countries. In order to fulfill this mandate, the Committee shall collaborate closely with national authorities to ensure adherence to the specified compute capacity restrictions, as outlined in Article 4.

The Committee shall possess the authority to conduct investigations within member countries to the extent deemed necessary. The local authorities within each member coun try shall provide full cooperation and assistance in facilitating these investigations.

15-Jul-23 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 3

*C. Article 3: Financing of the protocol and the Committee*

Funding for the Committee shall be provided in an equitable manner by all parties to the protocol. The specific amount of funding required for the effective functioning of the Committee shall be determined through consensus among its members.

*D. Article 4: Compute capacity restrictions*

The local authorities of member countries to this protocol shall bear the responsibility of overseeing both private and public endeavors related to the construction of large-scale computer infrastructure. Any entity seeking to operate computer infrastructure with spec ifications exceeding the following threshold must obtain a valid license:

• An instance of artificial intelligence containing more than 250 million parameters.

The Committee reserves the right to modify these specifications through consensus based decision making in the future. Entities intending to deploy compute infrastructure meeting the aforementioned specifications must acquire a license from the appropriate local governmental body prior to use. Local authorities are obligated to prosecute any infringements of these regulations using the most stringent measures available to them. The Committee shall oversee the prosecution process conducted by local authorities and, if deemed necessary, notify member countries to this protocol of any inappropriate actions taken.

Additionally, no entity within member countries is permitted to operate any computer system meeting one or more of the following criteria:

1. An instance of artificial intelligence with over 500 million parameters.

2. An instance of artificial intelligence that relies on input data generated by other artificially intelligent systems, with a cumulative parameter count exceeding 100 million.

Furthermore, all member parties to the protocol shall strictly refrain from engaging in or providing assistance to non-members in:

• Utilizing quantum computers for AI-related activities

• Conducting research and development with the aim to build artificial general intel ligence

To ensure compliance with the aforementioned agreements, all member countries must enact national laws within their jurisdictions. These laws shall be designed to prevent any entity from violating the provisions described above.

4 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 15-Jul-23

*E. Article 5: Windfall clause*

In order to obtain a license for operating an artificial intelligence system with more than 250 million parameters, the entity must provide the local authorities with the complete model, while retaining all rights to it. Additionally, the entity must present a detailed plan outlining the intended deployment and usage of the system. The deployment of the system may only proceed upon official authorization by the local authorities.

All financial gains derived from the deployment of the licensed artificial intelligence model shall be subject to local taxation. It is required that a minimum of 80% of the resulting profits be allocated for tax payment and shared locally.

*F. Article 6: Entry into force*

The protocol shall be deemed binding and ratified only upon the fulfillment of the follow ing conditions:

1. A minimum of 38 member countries have signed the protocol and formally ex pressed their agreement to be a party to it.

2. The parties to the protocol include all of the nations specified below: • United States

• Great Britain

• France

• Germany

• Russia

• China

• Japan

• India

• Israel

• South Korea

Once the protocol is ratified, it shall be immediately considered legally binding.

15-Jul-23 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 5

*G. Article 7: Successor to this protocol*

The protocol shall cease to be binding if any of the two conditions stated in Article 6 are no longer met. Additionally, the protocol shall lose its binding status if the Committee declares, ratifies, and establishes a successor protocol. Alternatively, if all parties unani mously agree that the development of artificial intelligence can be responsibly conducted without international regulation, the Committee may terminate this protocol with a 60-day notice period.

*H. Article 8: Withdrawal and Non-compliance*

Any party has the right to withdraw from this protocol at any time without needing to provide a reason. The withdrawal shall be communicated to the Committee at least 14 days prior to its effective date.

In the event that the Committee determines a country to be in non-compliance with the protocol, meaning that the party has knowingly permitted or is currently allowing the potential use of compute infrastructure that exceeds the specifications outlined in Article 4, the Committee shall promptly inform all member parties to this protocol. Moreover, the Committee shall suspend indefinitely all obligations of member countries to this protocol until the non-compliance issue is resolved.

*I. Article 9: Reporting of data, public awareness and exchange of information*

Parties are obligated to regularly submit reports to the Committee containing all relevant data related to their licensed computer infrastructure and its utilization, in accordance with the specifications outlined in Article 4. These reports shall be provided on a quarterly ba sis. Additionally, members of this protocol must inform the Committee of their endeavors aimed at detecting fraud and illegal handling of compute infrastructure.

The Committee is strictly prohibited from disclosing individual data to other parties. However, in cases where the Committee identifies a specific system or its usage as a poten tial threat, it may request the assistance of independent interpretability experts to analyze the model. The Committee may publicly announce this request to all parties of the proto col.

To promote transparency, the Committee shall annually publish an openly accessible report detailing its activities. This report should include a section dedicated to the current state of research in AI safety, written by independent researchers. Moreover, a separate report shall be prepared for sharing with all parties to the protocol.

6 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 15-Jul-23

COMMENTS ON THE PROTOCOL

*J. Safe Artificial Intelligence as a public good*

While the advantages of advancing AI technology are significant, the risks associated with its development must not be overlooked.

The benefits of AI development are primarily enjoyed by individuals or organizations, while the potential harm caused by misaligned AI is not limited to a specific group and is thus a shared concern. Although every country may desire to ensure the safe development of AI, they are faced with a dilemma as they have two unfavorable options to choose from.

1. Countries can either impose regulations on AI development, which may hinder their progress in achieving more powerful AI. Although this approach may not signifi cantly reduce the global risk associated with AI (since most of the risk comes from abroad), it means that the country will lose out on the advantages of advanced AI due to its own regulations.

2. Countries can choose to develop AI without prioritizing safety, allowing them to reap its benefits while also sharing the potential risks with the rest of the world.

The protocol addresses this dilemma by offering countries a means to coordinate *with out incurring significant costs*. By establishing a framework for collaboration, the proto col encourages developed nations to commit to the safe development of AI. If all relevant countries endorse the protocol, they can be assured of maintaining their competitive posi tion in the global race for artificial intelligence. Countries can consider this protocol as a superior cooperative alternative to the two unilateral options mentioned.

Therefore, it is crucial that countries have the option to join the treaty without incurring any costs or being legally obligated to adopt safety measures until all other participants do so. The treaty only activates when all major AI developer countries are onboard and a majority of developing countries have signed up. The specific limit of 38 members may be subject to change as needed. The key objective is to prevent a scenario where a sufficient number of non-member countries can collaborate to create more powerful AI. Until the membership requirement is met, countries can join the treaty without impeding their own AI progress. They must only transition to safer and slower development once everyone can reasonably reduce their own risk exposure by collaborating. Additionally, the treaty ensures the protection of each country’s information through the international body that collects state information but refrains from redistributing it, except in cases of fraudulent conduct.

The termination mechanism of the treaty operates in a similar manner. If a party de cides to withdraw, signaling their intent to engage in unsafe AI development, it would elevate the risk for all participants while allowing the withdrawing party to potentially

15-Jul-23 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 7

gain an advantage. To prevent this scenario, the treaty provides for immediate termination in such cases. As a result, the withdrawing party can anticipate that other participants will also resume unsafe AI development if they choose to withdraw. This is likely to be a less appealing prospect for them. Paradoxically, the swift termination of the treaty in such instances reduces the incentives to withdraw and enhances the overall stability of the coordination efforts.

The general dynamics of AI development, which have been the basis for modeling this treaty, are detailed in:

Stuart Armstrong, Nick Bostrom, and Carl Shulman. “Racing to the precipice: a model of artificial intelligence development”. In: *AI & society* 31 (2016), pp. 201–206

*K. Early licensing and registration*

In addition to imposing compute capacity restrictions, the treaty includes an early regis tration provision. This requires organizations planning to deploy AI models that approach the specified thresholds to obtain licenses. Consequently, local authorities gain visibility into the corporations and institutions operating substantial compute infrastructure and AI systems. This facilitates monitoring and ensures compliance with the established thresh olds.

*L. Windfall clause*

The licensed AI systems are subjected to significant taxation, which ensures that only highly profitable (i.e. fewer) ventures would pursue obtaining the license. This idea has been developed in:

Cullen O’Keefe et al. “The windfall clause: Distributing the benefits of AI for the common good”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 327–331

Rather than relying on a rigid threshold for AI systems, the regulation of AI deploy ment is divided into two stages:

• AI systems with less than 250 million parameters: There are no restrictions on their use.

• AI systems with 250-500 million parameters: A license is required, and profits gen erated from these systems are subject to taxation.

• AI systems with 500 million parameters or more: These systems are not permitted.

8 MORATORIUM OF LARGE-SCALE AI CAPABILITIES 15-Jul-23

The licensing process allows for some flexibility in accommodating systems close to the limit of 500 million parameters. Moreover, the generated tax revenue serves as an incentive for policymakers to endorse and join the treaty.

*M. Re-negotiation*

Given the rapid evolution of computer systems in recent times, the treaty acknowledges the need to adapt to new circumstances. To accommodate such changes, the treaty incor porates multiple mechanisms for consensus-based modifications.