Rennes SB Summit
10 October 2023

Alignment, Innovation, and Regulation—Paradoxes in AI Governance

IEA PARIS
INSTITUT D'ÉTUDES AVANCÉES DE PARIS

UNIVERSITY of VIRGINIA
SCHOOL of LAW

three paradoxes
a mirage
takeaways for AI governance

Input: I’m making a presentation with the title “Alignment, Innovation, and Regulation— Paradoxes in AI Governance.” In the first slide, I'm going to put in a quote generated by ChatGPT. Please generate an appropriate quote.

> “In navigating AI’s vast potential, we find ourselves at the intersection of creativity’s limitlessness and the essential confines of ethical governance.” — ChatGPT, OpenAI

1. The hype-reality paradox

"The better the tech, the bigger the disappointment."

Input: I'm making a presentation with the title "Alignment, Innovation, and Regulation— Paradoxes in AI Governance." In the first slide, I'm going to put in a quote generated by ChatGPT. Please generate an appropriate quote.

> "In navigating AI's vast potential, we find ourselves at the intersection of creativity's limitlessness and the essential confines of ethical governance." — ChatGPT, OpenAI


200
250
300
°C
50
DISHWASHER SAFE

2. The performance paradox

"The more useful the tech, the harder it is to assess."

Article 28b

1. A provider of a **foundation model** shall, prior to making it available on the market or putting it into service, ensure that it is compliant with the requirements set out in this Article, regardless of whether it is provided as a standalone model or embedded in an AI system or a product, or provided under free and open source licences, as a service, as well as other distribution channels.

2. For the purpose of paragraph 1, the provider of a foundation model shall:

(a) **demonstrate through appropriate design, testing and analysis the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development** with appropriate methods such as with the involvement of independent experts, as well as the documentation of remaining non-mitigable risks after development

(b) process and incorporate only datasets that are subject to appropriate data governance measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation

(c) design and develop the foundation model in order to achieve throughout its lifecycle appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity assessed through appropriate methods such as model evaluation with the involvement of independent experts, documented analysis, and extensive testing during conceptualisation, design, and development;

(d) design and develop the foundation model, making use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency, and the overall efficiency of the system, whithout prejudice to relevant existing Union and national law. This obligation shall not apply before the standards referred to in Article 40 are published. Foundation models shall be designed with capabilities enabling the measurement and logging of the consumption of energy and resources, and, where technically feasible, other environmental impact the deployment and use of the systems may have over their entire lifecycle;

(e) draw up extensive technical documentation and intelligible instructions for use, in order to enable the downstream providers to comply with their obligations pursuant to Articles 16 and 28(1);.

(f) establish a quality management system to ensure and document compliance with this Article, with the possibility to experiment in fulfilling this requirement,

(g) register that foundation model in the EU database referred to in Article 60, in accordance with the instructions outlined in Annex VIII point C. When fulfilling those requirements, the generally acknowledged state of the art shall be taken into account, including as reflected in relevant harmonised standards or common specifications, as well as the latest assessment and measurement methods, reflected in particular in benchmarking guidance and capabilities referred to in Article 58a;

3. Providers of foundation models shall, for a period ending 10 years after their foundation models have been placed on the market or put into service, keep the technical documentation referred to in paragraph 2(e) at the disposal of the national competent authorities 4. Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video (“generative AI”) and providers who specialise a foundation model into a generative AI system, shall in addition

a) comply with the transparency obligations outlined in Article 52 (1),

b) train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression,

c) without prejudice to Union or national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

3. The control paradox

"The more control a jurisdiction asserts, the less it shapes the tech."

Article 28b

1. A provider of a **foundation model** shall, prior to making it available on the market or putting it into service, ensure that it is compliant with the requirements set out in this Article, regardless of whether it is provided as a standalone model or embedded in an AI system or a product, or provided under free and open source licences, as a service, as well as other distribution channels.

2. For the purpose of paragraph 1, the provider of a foundation model shall:

(a) **demonstrate through appropriate design, testing and analysis the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development** with appropriate methods such as with the involvement of independent experts, as well as the documentation of remaining non-mitigable risks after development

(b) process and incorporate only datasets that are subject to appropriate data governance measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation

(c) design and develop the foundation model in order to achieve throughout its lifecycle appropriate levels of performance, predictability, interpretability, corrigibility, safety and cybersecurity assessed through appropriate methods such as model evaluation with the involvement of independent experts, documented analysis, and extensive testing during conceptualisation, design, and development;

(d) design and develop the foundation model, making use of applicable standards to reduce energy use, resource use and waste, as well as to increase energy efficiency, and the overall efficiency of the system, whithout prejudice to relevant existing Union and national law. This obligation shall not apply before the standards referred to in Article 40 are published. Foundation models shall be designed with capabilities enabling the measurement and logging of the consumption of energy and resources, and, where technically feasible, other environmental impact the deployment and use of the systems may have over their entire lifecycle;

(e) draw up extensive technical documentation and intelligible instructions for use, in order to enable the downstream providers to comply with their obligations pursuant to Articles 16 and 28(1);.

(f) establish a quality management system to ensure and document compliance with this Article, with the possibility to experiment in fulfilling this requirement,

(g) register that foundation model in the EU database referred to in Article 60, in accordance with the instructions outlined in Annex VIII point C. When fulfilling those requirements, the generally acknowledged state of the art shall be taken into account, including as reflected in relevant harmonised standards or common specifications, as well as the latest assessment and measurement methods, reflected in particular in benchmarking guidance and capabilities referred to in Article 58a;

3. Providers of foundation models shall, for a period ending 10 years after their foundation models have been placed on the market or put into service, keep the technical documentation referred to in paragraph 2(e) at the disposal of the national competent authorities 4. Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ("generative AI") and providers who specialise a foundation model into a generative AI system, shall in addition

a) comply with the transparency obligations outlined in Article 52 (1),

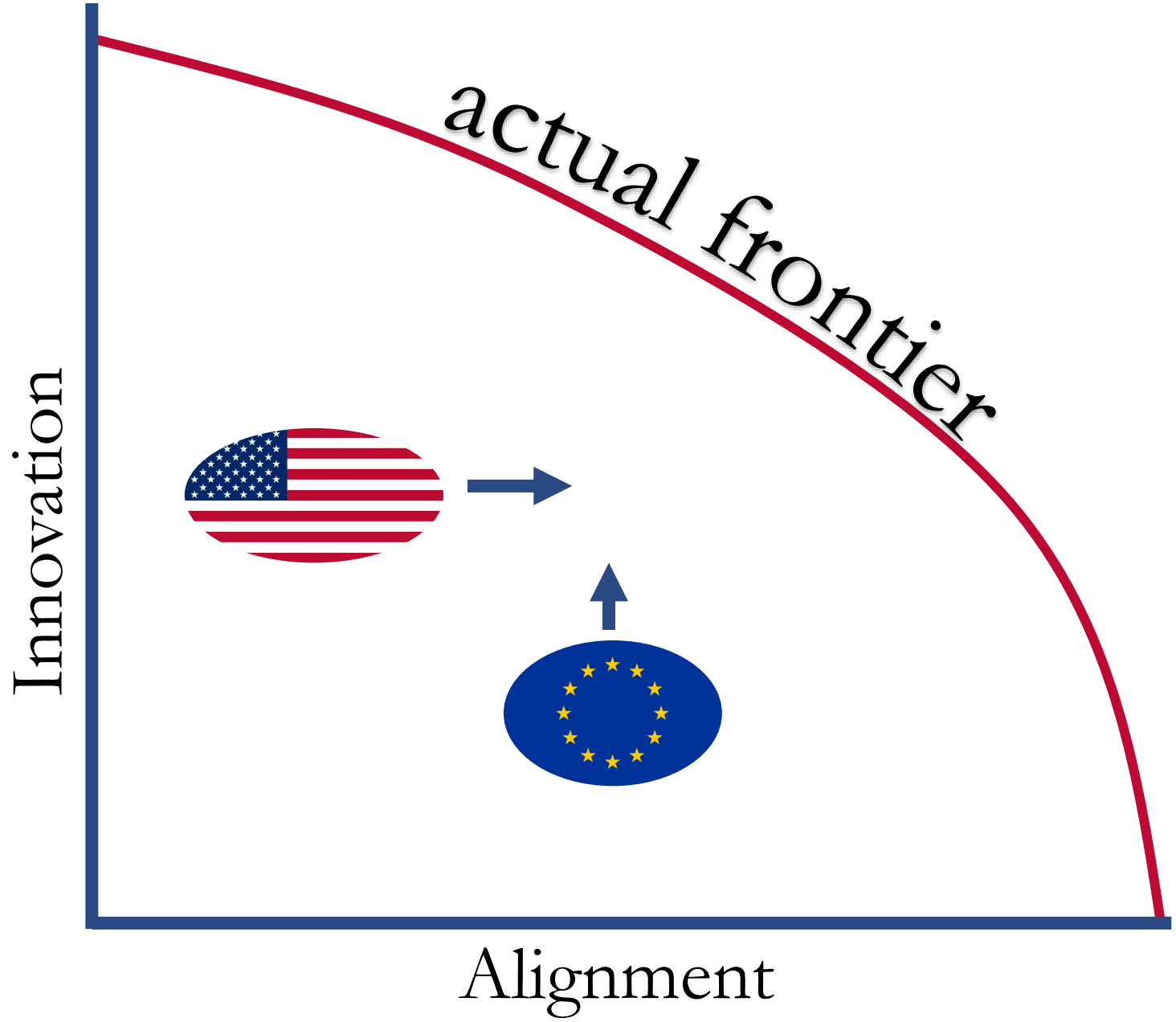
b) train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression,

c) without prejudice to Union or national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law.

The innovation-alignment tradeoff?

Innovation

Alignment


actual frontier
Innovation
Alignment

takeaways AI governance

- real v. imaginary risks
- aligning incentives, not systems
- fostering innovation to retain influence