AI regulation in the EU, the US and China: An NLP quantitative and qualitative lexical analysis of the official documents

Kristjan Prenga

University of Milan

This research presents a comprehensive lexical analysis of AI regulatory frameworks in the European Union, the United States, and China, utilizing a blend of quantitative and qualitative Natural Language Processing (NLP) techniques and methods. By means of a systematic examination of official documents, the study identifies the lexical and semantic variances and statistical distributions that delineate each region's approach to AI governance. By deploying methods such as word frequency analysis, lexical distribution, and co-occurrence metrics, the research unveils how key concepts such as 'risk' and 'security' are interpreted and prioritized differently across jurisdictions. The analysis reveals distinct strategic directions and interests: the EU's regulatory focus on market stability and consumer protection, the US's emphasis on maintaining technological supremacy and national security, and China's approach to harnessing AI for state-led innovation and development. The paper argues that these divergent approaches reflect underlying national priorities and strategic interests, which are crucial for understanding the global AI regulatory landscape. The insights from this study not only enhance understanding of international AI regulations but also inform ongoing policy discussions, advocating for adaptive regulatory measures that accommodate rapid technological advancements and complex global interactions in AI development.

Keywords: AI Regulation and Governance, Regulatory Frameworks, Natural Language Processing (NLP) Analysis

Introduction

Even though AI is definitely not a new concept in the scientific arena (Muthukrishnan et al., 2020), development and progress in the field, since the release of ChatGPT 3 in November 2022 by OpenAI, have brought an urgent call for regulation (Dral & Ullah, 2023). The European Union (EU), the US and China have already taken some steps towards that direction, even though each of them has decided to follow a different approach. The need for regulation comes both after the huge potential of this new technology and its innate risks for some fundamental human rights and even the survival of the human race itself (Muthukrishnan et al., 2020).

More specifically, deployment and use of AI technologies can threat, if used inappropriately and without oversight, various sectors of society. Those sectors vary from personal privacy to national security and democratic processes (Manheim & Kaplan, 2019). Furthermore, AI systems specialised in the creation of deepfakes (Pawelec, 2022) can prove to be particularly dangerous in the realm of elections in democratic countries which could have devastating results in the fundamental principles those countries are built on. This last point is especially pertinent in 2024 where at least 64 countries or 49% of the global population is deciding, by means of democratic elections, their political fates (Time, 2023) . Therefore, the future of the global political panorama is at high stakes.

That brings governments to consider solutions that give them the necessary power over the proliferation, uncontrolled development and deployment of AI technologies that could pose risks to their citizens. There are, however, several reasons that make the implementation of any kind of regulation over AI a very tough and complex endeavour. The majority of these reasons come from the unique characteristics of the technology itself such as its extremely rapid and sometimes unpredictable evolution, growth and expansion and its broad applicability across different sectors. Given these intrinsic properties of the most advanced AI technologies, governments and institutions have been very cautious in taking any steps forward to regulate something that no one really comprehends totally, especially people who are mostly involved in the regulation and creation of laws, namely politicians and legislators.

Nevertheless, the sole complex and multifaceted nature of AI does not constitute a valid reason not to regulate it and keep the risks it involves under control. Probably, the single most important reason that pulls governments back from creating and implementing regulations is the premises of immense growth, prosperity and wealth that this technology could bring to their countries (Suleyman, 2023). Moreover, owning and controlling the most cutting edge technological advancements in AI, gives the owner, thus the country, technological supremacy and huge economic, social, political and military advantage over the rest of the world.

As a result, it gets increasingly complicated to regulate AI innovation as both fear of missing out and lagging behind other countries could be catastrophic, especially for countries such as the US that in the last few years have always been leaders in many technological innovations, including AI. On the other hand, as it will be furtherly discussed in the next paragraphs, China, and its huge growth in the last 40-50 years, does not seem to be willing to risk the economic miracle it has created by regulating what could make them surpass the US and become the first world economy in the race for economic primacy and technological supremacy. The EU, on the other hand, has taken, as will be seen, yet another path.

2. EU, USA and China: three different approaches

The EU, the US and China have been at the forefront of technological advancements for some time now. Undoubtedly, the US has been the undisputed leader in the creation of companies that have introduced and distributed technological innovations all over the world. Companies such as Apple, Microsoft, Amazon and more recently Google and Meta have disrupted many industries and have altered the technological landscape globally with their bold innovations and revolutionary products both in the area of software and hardware. However, in the quest for AI supremacy, it seems that the other two actors – the EU and China – could have a chance of closing the gap and even surpassing the US in this unpredictable race of power and prestige.

The US is definitely the most privileged one in this race. This is due to the presence of leading companies that invest heavily in AI such as Meta, Microsoft, Google, and Amazon. These companies are not only pioneers in AI research and development but they also have vast amounts of resources and influence that contribute to the US's prominent position in the global tech race. This concentration of tech giants in one geographical area fosters innovation, attracts talent, and drives advancements in AI and other related technologies such as robotics. To put this into perspective, based on Tricot, 2021 *"US VC investors were the most active investors in AI firms, representing 43% of the worldwide value of VC investments in AI in 2020, followed by Chinese investors (20%) and then EU27 investors (9%)."* . That means that, at least in the field of Venture Capital (VC), the US invests more than double and almost 5 times the amount that China and the EU respectively do. And this trend does not seem to slow down given Microsoft's $10 billion investment on OpenAI in January 2023 (Badr, 2023). Top of Form

Recent investments in US AI companies (e.g. OpenAI and Anthropic) have widened the gap between the EU's and the US's relative share of private investment in AI (Atomico, 2023). As can be seen in *Figure 1*, the US is the clear leader in VC investment in Generative AI with a peak in 2021 of 120$ billion followed by China with almost $50 billion and the EU27 with less than $20 billion. This obvious dominance in attracting equity investments is a good indicator of the privileged position that the US has over China and the EU. All this puts the US in a condition where regulation could have a stalling effect in the development of the next generation AI systems. This has led the US government to release an Executive Order on October 30th, 2023 (The White House, 2023). This presidential action is a clear message that the US has no intention – at least not in the near future and/or not with this government in power – of issuing any heavy, restrictive laws that build straightforward rules and put companies under direct scrutiny in order to manage and control risk in AI development and deployment. While Executive Orders have the force of law, they can be easily altered and cancelled anytime by the next government in power. This does not happen with regulations which are issued by federal agencies based on the authority derived from statutes enacted by Congress (Mello et al., 2023). It is precisely this Presidential Executive Order that will be the subject of linguistic research and analysis in this paper, and just for the sake of this paper's practicality and simplicity will be called a "regulation" even if it is actually just a Presidential Executive Order.

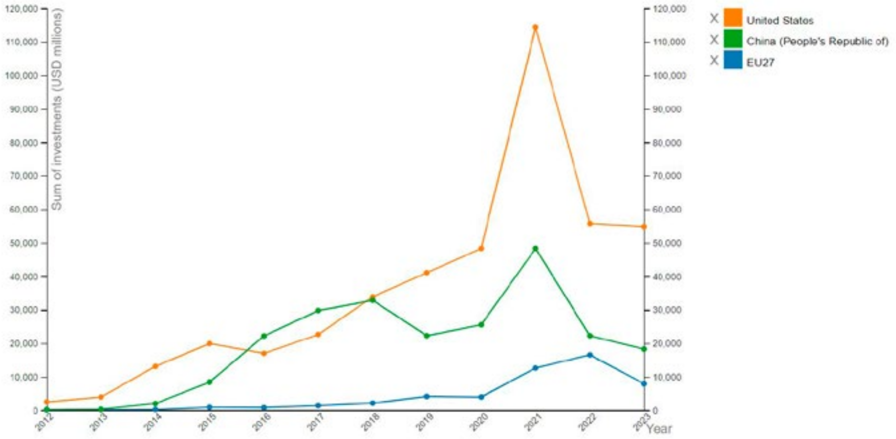


Figure 1 - Venture capital investments in AI in USD millions by country from 2012 onwards. Source: OECD.AI, 2023.

While the US seems to have the upper hand in the AI race, China is keeping up with a rather good pace both from a VC investments attraction (*Figure 1*) and from a regulatory point of view. In particular, the worth of the Chinese AI market was estimated at $23 billion in 2021 and it is expected to triple by 2025 (Xiao, 2024). Moreover, the Chinese government expects AI to cross $150 billion in annual revenue by 2030 (Larson, 2018). From a regulatory point of view, China is the first country in the world to issue a regulation on Generative AI by means of China's Provisional Administrative Measures of Generative Artificial Intelligence Services, which came into effect in August 2023 (Pi, 2024). These measures are the result of a series of national, regional and local level regulatory measures that China has been taking since 2021, when the Personal Information Protection Law (PIPL) passed in August 2021 and the New Generation Artificial Intelligence Codes of Ethics was published in September of the same year (Roberts et al., 2021).

On January 10, 2023, China's Administrative Provisions on Deep Synthesis in Internet-based Information Services (Deep Synthesis Provisions) entered into effect (Pi, 2024). In these provisions, deep synthesis is defined as *"technology utilising generative and/or synthetic algorithms, such as deep learning and virtual reality, to produce text, graphics, audio, video, or virtual scenes."* (Deep Synthesis Provisions, 2023). Right after this publication, the People's Republic of China (PRC) issued the regulation on Generative AI that will be the subject of this study. This regulation has been criticised of lacking distinction and clear legal status between different players in the AI value chain which may lead to ambiguity in accountability, potentially undermining the governance and overall success of AI services (Pi, 2024). This lack of clarity between providers and deployers which translates into notable weaknesses in transparency and accountability maybe a political strategy to keep innovation flourishing without a substantial and well-defined regulation that limits AI deployment and development and punishes legally reckless actors.

China's intentions to promote research and development over strict regulation can also be observed by the government's decision to exempt from the final regulations sectors like the scientific research and industrial applications. Those sectors were initially present in the draft measures but were eventually excluded from the final version of the published regulation (Xiao, 2024). By including those fields, China would have risked to suppress with overregulation important sectors that could provide innovation capable of competing with US's pioneering companies. Another point that China has in common with the US is the vast amount of technology giants working on the field of AI and creating innovation. These companies have been researching and developing cutting edge technological and digital products for as far as American companies have done so. Companies such as Baidu, Tencent, Huawei and Alibaba have a huge influence both nationally and out of the Chinese borders (Hyne and Floridi, 2024). Therefore, neither the US nor China have a strong – especially economic and geopolitical – incentive to issue any regulation that could eventually compromise the advantaged position they have acquired in the last years placing them in the podium of the most technologically advanced countries. This is particularly true considering the immense transformational effect and power that AI will have globally in the coming years.

As far as the EU is concerned, its member states' approach is totally different from the two previous ones. The European equity landscape does not include names comparable to the American or the Chinese ones, apart from some very few exceptions like the German multinational software company System Applications and Products (SAP). This puts the EU in a rather disadvantageous position in the AI revolution where it has to compete with technology conglomerates such as Microsoft, Amazon or Google in the US and the Chinese Baidu, Tencent or Huawei (Hyne and Floridi, 2024). Although, there have been some startups like MistralAI and Contents that try to compete with OpenAI's Generative models like ChatGPT and other similar technologies (Ono and Morita, 2024), the funding that these startups have raised is decisively lower than the respective American and Chinese companies. To put this into perspective, OpenAI received $10 billion in funding by Microsoft alone in 2023 (Badr, 2023) and Anthropic a $4 billion investment by Amazon in 2024 (Amazon, 2024). The French startup MistralAI, which is considered the best of what Europe has to offer in the field of Generative AI, received $414.41 million in a second funding round in December 2023 (Reuters, 2023). In other terms, Anthropic managed to attract 10 times more investment than MistralAI and OpenAI around 25 times more.

These are just a few examples that demonstrate the magnitude of capital investments between European companies and American ones. This is because the European market is prevalently made of Small and Medium Enterprises (SMEs). More specifically, only 0,2% of enterprises in the EU are large (>250 employees), 0,9% are medium (50-249 employees) and 93% are small (<49 employees) (Eurostat, 2019). Therefore, the EU has no large and mature companies with economical resources and talents at their disposal to compete in a global scale with the US and China. On the other hand, mainstream use of these AI technologies like ChatGPT is unavoidable whether the technology is produced in the European soil or not. The EU is well aware of the current situation and has been taking actions to ensure that the right measures are taken in order to gain some competitiveness with respect to very tough and privileged competitors. The EU response once again has been regulation.

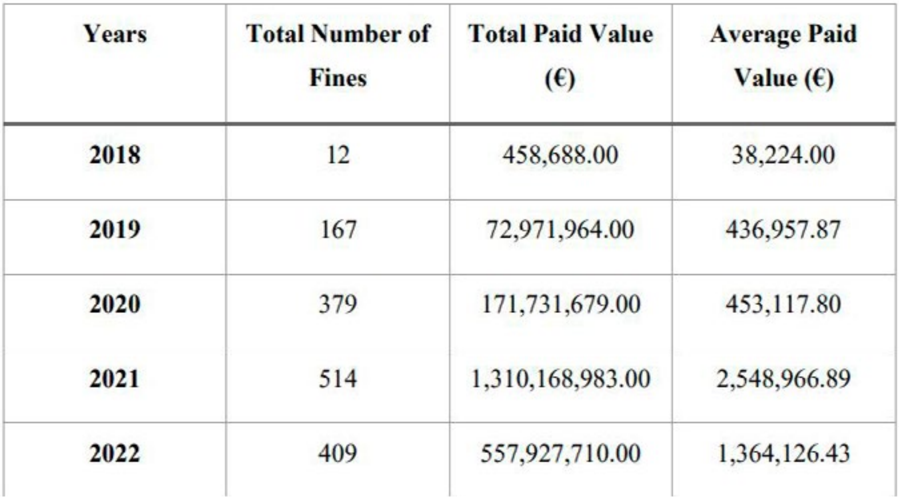
It is not the first time that the EU tries to contain the influence of foreign companies in the technological/digital sector active on the European territory. In fact, the EU has issued several regulations throughout the years in order to regulate the collection, processing and storage of its citizen's personal data. One of its first attempts was the General Data Protection Regulation (GDPR) which *"is a legal framework established by the European Union to protect the privacy and personal data of individuals within the EU and the European Economic Area (EEA)"* (Eur-lex, 2022). GDPR went into force in 2018. In 2022, the EU issued and put into force another regulation, the Digital Markets Act (DMA) which *"is one of the first regulatory tools to comprehensively regulate the gatekeeper power of the largest digital companies."*. And more recently, in 2024, the AI Act which *"is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally."* (European Commission, 2024), was approved and it is expected to enter into force in the next 2 years. It is specifically the AI Act that will be the subject of linguistic research and analysis in this paper.

The way the EU has tried to contain foreign digital and technological companies' influence and personal data collection and processing is through regulation. Throughout these years there have been many instances of companies such as Meta, Amazon, Google etc that have been fined heavily for law violation of the GDPR (Dias et al., 2023). As can be seen from *Figure 2*, in 2021 alone the EU emitted 514 fines to companies that violated GDPR for a total of €1,310,168,983.00. According to Dias et al., 2023:

> *"the three most expensive fines are distributed as follows:*

1. *Amazon Europe Core S.à.r.l. - (CMS process number 778) (746 million euros), Non-compliance with general data processing principles, Article: Unknown; (AMAZON.COM, INC., 2021)*
2. *Meta Platforms, Inc. - (CMS process number 1373) (405 million euros), Noncompliance with general data processing principles, Article: Art. 5 (1) a), c) GDPR, Art. 6 (1) GDPR, Art. 12 (1) GDPR, Art. 24 GDPR, Art. 25 (1), (2) GDPR, Art. 35 GDPR; (Binding Decision, 2/2022)*
3. *WhatsApp Ireland Ltd. - (CMS process number 820) (225 million euros), Insufficient fulfilment of information obligations, Article: Art. 5 (1) a) GDPR, Art. 12 GDPR, Art. 13 GDPR, Art. 14 GDPR. (Decision of the Data Protection Commission, 2021)"*

While many scholars and researchers believe that the AI Act is in the right direction in regulating the future of AI, being such a fast evolving and unpredictable field there are still many challenges that need to be tackled (Hacker, 2023). The greatest of which is, of course, overregulation that may lead to technological stagnation in the EU zone (Much et al., 2023). While the EU is trying to assist SMEs in the AI race against well-established technology giants coming from overseas, some critics sustain that the opposite effect may occur. That is, excessively stringent regulatory regime could deter SMEs from engaging in AI research and development, favouring established entities with greater resources (Much et al., 2023). Moreover, the strict regulations of EU's AI Act could result in a scenario where AI technologies developed in the EU are at a competitive disadvantage compared to those from areas with less restrictive regulations. This imbalance might hinder the development of the EU's AI sector, potentially causing talent loss and diminishing the region's capacity for innovation (Much et al., 2023). If this last scenario were to be realised, it would result in a disastrous choice from the EU that would compromise economically and geopolitically future EU generations. Lagging behind in AI advancement and innovation constitutes the worst case scenario for Europe, and taking the courageous step of being the first to regulate thoroughly AI development could also mean taking a huge risk.



Years	Total Number of Fines	Total Paid Value (€)	Average Paid Value (€)
2018	12	458,688.00	38,224.00
2019	167	72,971,964.00	436,957.87
2020	379	171,731,679.00	453,117.80
2021	514	1,310,168,983.00	2,548,966.89
2022	409	557,927,710.00	1,364,126.43

Figure 2- Fines from the EU for GDPR violations during the period 2018-2022. Source (Dias et al., 2023)

Determining which is the right approach and tell with certainty whether the regulatory decisions taken by the three actors are the right ones is early to say. What is clear, though, is that with one way or another, each governmental body is trying to use at best the assets they possess in order to lead what could be the single most important technological revolution of the 21st century.

3. Aim of the lexical analysis

In this paper, the three regulations will be analysed and researched using Natural Language Processing (NLP) techniques in order to retrieve lexical and semantic information in a statistical and analytical manner. NLP techniques are able to identify patterns and models in the way language was implemented in the drafting of these regulations. Furthermore, it can automate the analysis of large volumes of text, in this case the vast amount of linguistic content especially in the AI Act. This reduces the time for manual analysis, leading to faster insights and information retrieval in a rather objective and systematic way. What is more, NLP can be used for complex relationship detection, meaning that it can uncover complex relationships and dependencies between different regulatory provisions. Last but not least, by analysing the language across different regulations, NLP can help in identifying trends and predicting future regulatory directions.

4. Methods used in the processing of the three documents

The three documents that have been processed by means of NLP techniques in this investigation are:

1. The US's Presidential Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (The White House, 2023).
2. China's Provisional Administrative Measures of Generative Artificial Intelligence Services (Pi, 2024).
3. The EU's Artificial Intelligence Act (European Commission, 2024).

Each document was transformed into a processable linguistic corpus that was processed using AntConc (Version 4.2.4), a corpus analysis toolkit for researchers, teachers and learners (Laurence, 2024). More specifically, AntConc was used for methods 1, 2 and 5 (following in the next paragraphs). Moreover, in some methods (3, 4 and 6) Python (Version 3.11.5) was used for data processing and visualizations. For the sake of practicality and simplicity, the three corpora will be called as follows: the US Corpus (USC), the Chinese Corpus (CC) and the EU Corpus (EUC), linking to their respective regulations as enlisted above.

As far as the processing of linguistic corpora is concerned, the dimension of the corpus is an important metric to keep in mind. With regard to this, the EUC is the largest one with 88.814 tokens, the USC comes second with 22.081 tokens and last the CC with just 1.901 tokens. While the EUC and the USC were originally drafted and published in English, the CC was published in Mandarin Chinese in the PRC's official website. Its English version was downloaded from the China Law Translate website and transformed into the final corpus (China Law Translate, 2023).

One of the first aspects that strikes the attention the most is the significant dimension differences between the three corpora, thus regulations. As commented in the previous section, while the Chinese was the first one to be issued and put into force, it is not by any means the most exhaustive one. On the contrary, being the shortest one, it is constituted by just 24 quite brief articles that touch almost superficially certain aspects of AI without getting very deep into details. This aspect and the ambiguous accountability and responsibility assigned have been the main reasons of criticism by many experts of the field (Xiao, 2024). On the other hand, the USC is more than 10 times bigger than the Chinese one. It can be considered much more exhaustive and detailed, yet it lacks the precision, meticulousness and completeness of the EU's AI Act.

This linguistic research will investigate words in context taking into consideration six specific NLP methods, which constitute six different lexical and statistical metrics of the regulations in question:

1. Word Frequency (WF). Here, the most frequently used words in each corpus will be sorted by descending order, analysed and commented. Typically, the more a term is used in a text, the more importance it acquires within that text.
2. Keywords measured by Keyness (KK). The KK is a method that sorts words based on how much more they appear in a target corpus compared to a general purpose corpus (aka the reference corpus). By means of some statistical calculations, the keyness value is attributed to each word. This value identifies and ranks the degree to which words in a target corpus appear unusually frequently compared to their occurrence in a reference corpus (Laurence, 2023). For example, in a general purpose corpus like the Corpus of Contemporary American English (COCA), the term "AI" is expected to appear decisively less than in the corpora we created. Thus, the word (or in this case the acronym) "AI" will be very high ranked in this wordlist, namely the list of words sorted by descending order based on their keyness. In our case, the COCA will be used as a reference corpus. COCA consists of 560+ million words and it is nearly evenly divided (20% in each genre) between spoken, fiction, popular magazines, newspaper, and academic languages (English Corpora, 2024). While WF counts occurrences, KK highlights contextually significant words by comparing their frequency to a reference corpus, emphasizing domain relevance.

As a first step, an indicative sample of the top 10 most frequent words by frequency and keyness will be analysed and compared between the 3 corpora. These words are a good representation of the content of each corpus. In more concrete terms, by studying the absolute and relative frequency of the most used terms, documented inferences and assertions can be made about the intentions and the points that the regulators consider of greater importance regarding these AI regulations.

3. Regulations' similarity based on common terms. In order to reach a statistical metric of similarity between the three regulations, it was decided to identify the common terms that all regulations shared together and, then, individually with one another. This could shed light on similarities and differences in the use of words in the drafting of these regulations, which could translate into differences and similarities in the regulators' intentions and approaches to AI governance.
4. Part-Of-Speech (POS) distribution by frequency. Using frequency as a classification factor, the first 100 words of the WF wordlist will be analysed and classified into a specific POS in order to extract meaningful statistical information about the distribution of the wordlist's POS. Analysing the distribution of POS can provide valuable insights into the linguistic characteristics and the underlying intentions of the documents.

For example, a higher frequency of nouns often indicates that the text is content-heavy, focusing on specific subjects, entities, and concepts. In the context of AI regulations, the use of high number of nouns might suggest a detailed focus on specific technologies, processes, actors (like developers, users), and regulatory areas (such as data privacy, accountability). Verbs usually reflect actions and processes. A high usage of verbs could imply a focus on the actions to be taken, such as enforcement measures, compliance requirements, and operational processes. It could also indicate prescriptive content, focusing on what should be done in various scenarios. A high number of adjectives may suggest that the regulation is attempting to be very detailed and specific. This could be to ensure clarity in the definitions and scope of the regulation, aiming to cover as many scenarios as possible and leaving little room for ambiguity.

5. N-grams Frequency (NGF). The NGF is the frequency of two or more words (grams) that occur together in a rather frequent way. For example, the 2-gram "Artificial Intelligence" is a collocation of two words that, as it is obvious, will be very frequent in all three corpora. In this study, 2-grams, 3-grams and 4-grams will be analysed and presented in lists sorted by descending order. As with the WF, the NGF is yet another lexical and statistical tool to measure word importance by being more specific in how certain expressions and collocations of words are used quantitatively and qualitatively in context.
6. Frequent Co-occurrences (FC). FCs are a good way to identify words that co-occur in the same context but not necessarily one right after the other (as

in the n-grams). In this method, a specific term is taken as a reference, e.g. "AI", and then the occurrences of words are measured in a predetermined span of words before and after the reference word. The context span is usually decided by the researcher and in this study it was decided to be 7. This means that the algorithm will investigate and count word occurrences in the previous and following 7 words starting from the reference word. What is achieved using this method is a list of the words that occur the most in the same context with a given term. The 7-word context span was chosen because a smaller span (e.g., 3-4 words) might miss longer-range dependencies, while a larger span (e.g., 10+ words) could introduce unrelated noise. Seven strikes a balance, offering a focused yet comprehensive context.

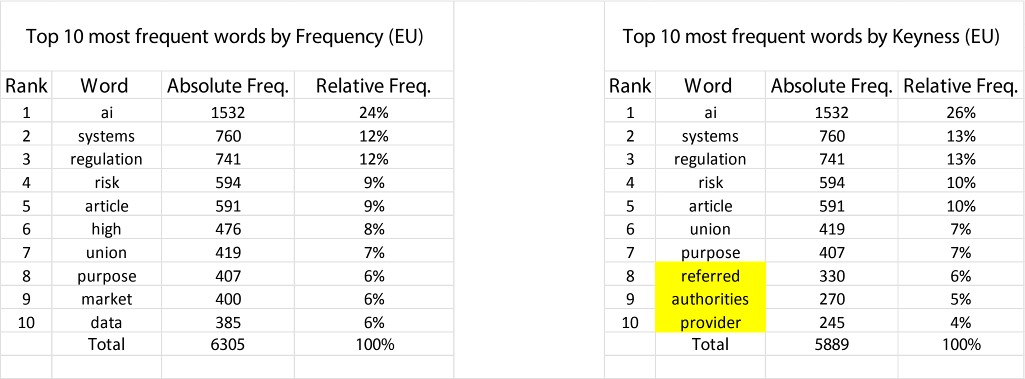
As it is common practice in linguistic research, all wordlists will be cleaned out from any linguistic noise, meaning from any words that do not convey any meaning. These words are the so-called stop words and they include prepositions, pronouns, connectors etc.. This is a procedure that aims at working on and analysing parts of discourse that could convey semantically important information such as nouns, adjectives, verbs and adverbs. Therefore, the words in the wordlists that will be provided in this paper will be already accurately and attentively selected.

Lemmatization, while a potentially valuable preprocessing technique, was not employed in this study to preserve specific grammatical and contextual features essential to the analysis. A key consideration was maintaining distinctions between singular and plural forms, as well as the part-of-speech roles of certain terms. For example, the term "risk" can function both as a noun and as an adjective depending on the context (e.g., "high-risk technology"), whereas "risks" unequivocally serves as a noun. Lemmatizing these terms would obscure such distinctions, potentially limiting the ability to draw nuanced grammatical and contextual insights. By retaining these variations, the study ensures a more precise and meaningful analysis of the regulatory texts. Moreover, this is also emphasised by the main scope of this study which is prevalently the contextual relationship of the analysed terms.

5. Results

Word Frequency and Keywords measured by Keyness

In *Figure 3*, the top 10 most frequent words by frequency and keyness in the EUC are presented. Besides the ranking, there are two other columns with data, the "Absolute Frequency" (AF) and the "Relative Frequency" (RF). Just for clarification, the absolute frequency represents the number of instances found in the whole document of that specific term, while the relative frequency indicates the number of instances in percentage points taking into consideration the first 10 words. That is, how much a term is used compared to the other 9 terms of the list. The RF is also added in order to make feasible the comparisons between the different regulations that, as we know, have different scales of dimension. On the left table the list shows the words ranked by frequency while the right table those by keyness. This kind of comparison will be conducted for all 3 corpora.



Top 10 most frequent words by Frequency (EU)

Rank	Word	Absolute Freq.	Relative Freq.
1	ai	1532	24%
2	systems	760	12%
3	regulation	741	12%
4	risk	594	9%
5	article	591	9%
6	high	476	8%
7	union	419	7%
8	purpose	407	6%
9	market	400	6%
10	data	385	6%
	Total	6305	100%

Top 10 most frequent words by Keyness (EU)

Rank	Word	Absolute Freq.	Relative Freq.
1	ai	1532	26%
2	systems	760	13%
3	regulation	741	13%
4	risk	594	10%
5	article	591	10%
6	union	419	7%
7	purpose	407	7%
8	referred	330	6%
9	authorities	270	5%
10	provider	245	4%
	Total	5889	100%

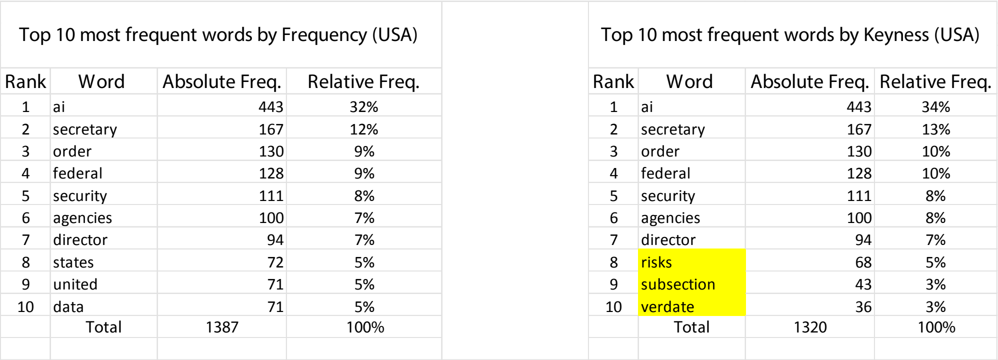
Figure 3 – Top 10 most frequent words (by frequency and keyness) EU

Starting with the largest corpus, the EUC, it can be affirmed that while the term "AI" can be quite intuitively expected to be among the most used words, it results being not just the most used term but the number of its instances is twice the number of the second term in the list "systems" (*Figure 3*). Together with "systems", "AI" sums up to 36% of the most used terms referred to the technology semantic field. And if we also add the 6% of the 10th term, "data", it can be claimed that 42% of the terms in the EU regulation refers to technology related terminology. Being a regulation, the identical same term is 3rd in the list with a RF of 12% of the whole terms. This is undoubtedly a word that is to be expected from a regulatory document as this one. In the 4th place, "risk" is found. This is definitely one of the most interesting insights. This is because, it seems that the EU points out a lot the risky aspects of AI related technologies. And if connected to the 9th most used term, "market", it can be asserted that the EU, indeed, wants to protect its market from the "high" (6th with a RF of 8%) risk posed in the European "union" (7th with a RF of 7%) by general "purpose" (8th with a RF of 6%) AI technologies. This can be furtherly reinforced by direct references from the AI Act stating the following:

> *"it is appropriate to establish a methodology for the classification of general purpose AI models as general purpose AI model with systemic*

> *risks. […] High-impact capabilities in general purpose AI models mean capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models. […] This threshold should be adjusted over time to reflect technological and industrial changes”* (European Commission, 2024, 60n)

As far as KK is concerned, *Figure 3* provides some interesting insights, as well. In this table, the first 7 terms remain the same. However, in the last 3 ones there is the introduction of 3 new terms; “referred”, “authorities” and “provider”. While “referred” is not particularly informative, the same does not happen with the other 2 ones. More specifically, “authorities” and other terms of the same semantic field, seem to obtain around 5% of what is written in the document. Considering its regulatory nature, it comes as no surprise that public authorities and governmental bodies are heavily involved in the adoption and correct implementation of this new regulation. Regarding “provider” (10th with a RF of 4%), it is yet another term which indicates any entity that provides AI technologies both to the public directly by means of products and to other companies indirectly. This role is crucial in the whole ecosystem as it is the figure of major liability and accountability when it comes to risks posed by AI.



Top 10 most frequent words by Frequency (USA)

Rank	Word	Absolute Freq.	Relative Freq.
1	ai	443	32%
2	secretary	167	12%
3	order	130	9%
4	federal	128	9%
5	security	111	8%
6	agencies	100	7%
7	director	94	7%
8	states	72	5%
9	united	71	5%
10	data	71	5%
	Total	1387	100%

Top 10 most frequent words by Keyness (USA)

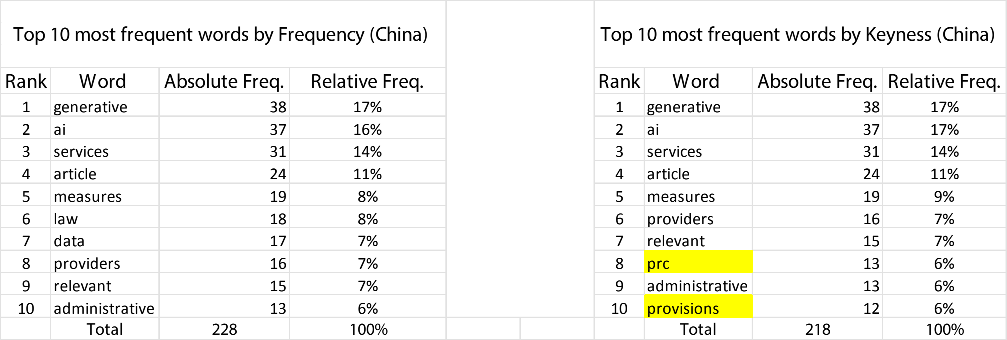
Rank	Word	Absolute Freq.	Relative Freq.
1	ai	443	34%
2	secretary	167	13%
3	order	130	10%
4	federal	128	10%
5	security	111	8%
6	agencies	100	8%
7	director	94	7%
8	risks	68	5%
9	subsection	43	3%
10	verdate	36	3%
	Total	1320	100%

Figure 4 – Top 10 most frequent words (by frequency and keyness) USA

As in the case of the EUC, the USC, is also overwhelmed by the use of the term “AI”. As can be seen *by Figure 4*, 32% of frequency and 34% of keyness is occupied just by this term. And if combined by the 5% of the word “data”, it can be stated that around 37% of the terms used in the US Executive Order is of technological semantic background. Another observation to be made here is the amount of words and their frequency coming from a semantic background indicating national sentiment. From the collocation “united” and “states” to “federal” and “secretary”, there is quite an obvious emphasis on national aspects and issues. And if combined with “security” (5th with a RF of 8%), it can be inferred that national security is considered one of the most essential element of this regulation. The specific assertion regarding "national security" is mainly addressed in Section 4.3 of the U.S. Executive Order:

> "To ensure the protection of critical infrastructure, the following actions shall be taken: (a) Within 90 days of the date of this order, and at least annually thereafter, the head of each agency with relevant regulatory authority over critical infrastructure [...] shall evaluate and provide to the Secretary of Homeland Security an assessment of potential risks related to the use of AI in critical infrastructure sectors involved, including ways in which deploying AI may make critical infrastructure systems more vulnerable to critical failures, physical attacks, and cyber attacks, and shall consider ways to mitigate these vulnerabilities". (The White House, 2024)

In fact, as indicated at the beginning of this paper, the US is trying to maintain its leading position to the development of AI technologies in order not to put at risk its global economic influence and geopolitical security. Apart from "AI" and "data", all the other terms can be classified as concepts referred to national sentiment and legal/regulatory issues, thus 63% of the most frequently used words are around the notion of protection of national interests. As a matter of fact, in the KK table, there is the insertion of the term "risks" (8th with a RF of 5%). Contrary to the EU, though, those risks are not strictly connected to the "market" (previously found in the EUC) but also to national "security" which definitely includes much more than just the economical aspect.



Top 10 most frequent words by Frequency (China)

Rank	Word	Absolute Freq.	Relative Freq.
1	generative	38	17%
2	ai	37	16%
3	services	31	14%
4	article	24	11%
5	measures	19	8%
6	law	18	8%
7	data	17	7%
8	providers	16	7%
9	relevant	15	7%
10	administrative	13	6%
	Total	228	100%

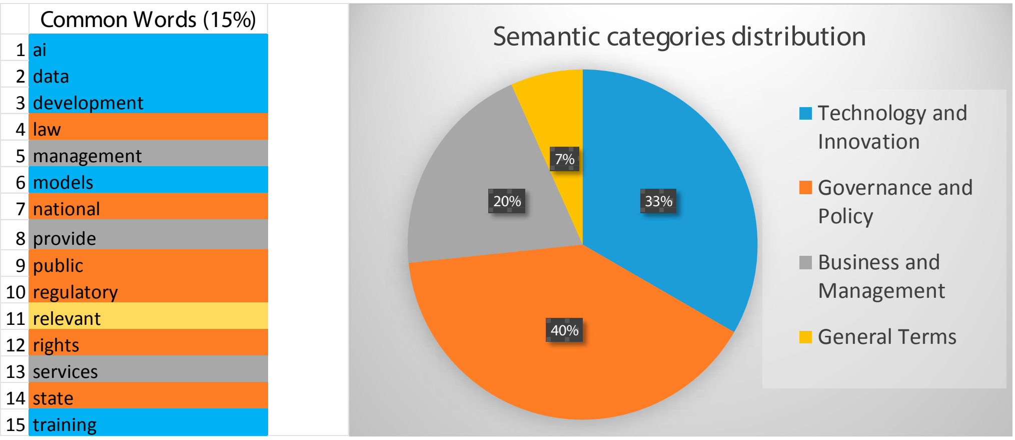
Top 10 most frequent words by Keyness (China)

Rank	Word	Absolute Freq.	Relative Freq.
1	generative	38	17%
2	ai	37	17%
3	services	31	14%
4	article	24	11%
5	measures	19	9%
6	providers	16	7%
7	relevant	15	7%
8	prc	13	6%
9	administrative	13	6%
10	provisions	12	6%
	Total	218	100%

Figure 5 – Top 10 most frequent words (by frequency and keyness) China

Comparatively to the other two regulations, the Chinese one is the briefest and less exhaustive one. What emerges from *Figure 5*, is that the Chinese regulation is much more focused on "generative" "AI", being these exact terms the most used ones and occupying together 33% of the most used terms distribution. It is apparent that AI is considered a service, given that the word "services" is the 3rd most used term. In contrast to the EUC and USC, there is no mention on risks or security, at least not a frequent one. Interestingly enough, the only mention to AI agents is "providers" (8th in the frequency table with a RF of 7%). Providers are companies that are usually considered those who offer AI technologies by possessing the foundation models. These companies are definitely the ones that could be of most concern to China, as it would prefer them to be Chinese rather than foreign ones. Moreover, in contrast to the EUC that uses the term "regulation" and the USC "order", here the most used term to describe the main goal of the documents is "measures" and then "law".

Regulations' similarity based on common terms

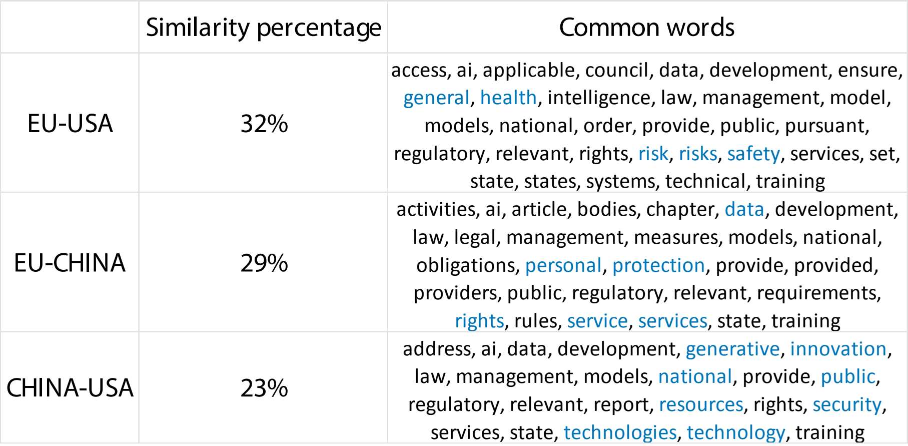


	Common Words (15%)
1	ai
2	data
3	development
4	law
5	management
6	models
7	national
8	provide
9	public
10	regulatory
11	relevant
12	rights
13	services
14	state
15	training

Figure 6 – Common Words and Semantic Categories Distribution

Taking as a sample the 100 most frequent terms in each corpus, it was found that 15% of those were in common in all three regulations. This is a statistical indication of similarity that reveals, at least from a linguistic perspective, that the three approaches do not share much common ground in the way they conduct regulatory documents and the way they intend to approach AI governance. In order to give a better idea of the main aspects that these three regulatory bodies approached, these 15 terms were furtherly categorised into semantic categories, as seen in *Figure 6*. More specifically, it was observed that 40% of these terms fall under the "Governance and Policy" semantic category, 33% were of "Technology and Innovation" semantic nature and 20% were of "Business and Management". While this distribution can be subject to minor changes, it reveals some interesting insights regarding the width and some specificities of the common ground shared by the three regulations.

While similarity between all three of them does not exceed 15%, direct comparison between each regulation with one another reveals some remarkable insights. In particular, EU's AI Act seems to share common language with both the USC (32%) and the CC (29%), whereas US and China have a roughly 23% rate of similarity (*Figure 7*). Common language could be a good indicator of how regulators perceive the field they intend to regulate. Although the present institutional bodies have different objectives, as mentioned above, linguistic metrics show that some regulations are more closely related to others. In the "Common Words" column of *Figure 7*, it can be observed that EU and USA share many terms that emphasize the risks of AI with words such as "risks", "risk", "health", "rights" etc..



	Similarity percentage	Common words
EU-USA	32%	access, ai, applicable, council, data, development, ensure, general, health, intelligence, law, management, model, models, national, order, provide, public, pursuant, regulatory, relevant, rights, risk, risks, safety, services, set, state, states, systems, technical, training
EU-CHINA	29%	activities, ai, article, bodies, chapter, data, development, law, legal, management, measures, models, national, obligations, personal, protection, provide, provided, providers, public, regulatory, relevant, requirements, rights, rules, service, services, state, training
CHINA-USA	23%	address, ai, data, development, generative, innovation, law, management, models, national, provide, public, regulatory, relevant, report, resources, rights, security, services, state, technologies, technology, training

Figure 7 – Similarity between regulations in % and their common words

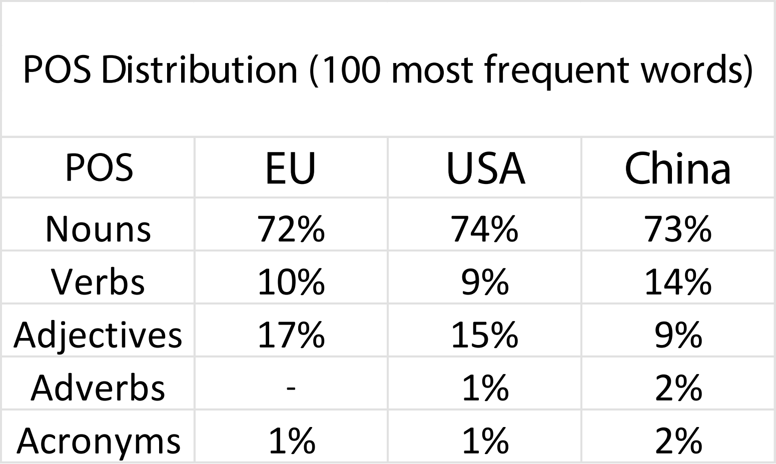
On the other hand, in the EU-China comparison combination terms such as "service", "services", "rights", "personal", "data" and "protection" could be good indicators of terms that focus on the AI services and, thus, the need to protect personal data from automated processing. This should not come as a surprise since both the EU with the General Data Protection Regulation (GDPR) and China with its Personal Information Protection Law (PIPL) are quite aligned with respect to personal information rights. Therefore, the fact that the EU and China have similar terms in common also indicates that both regulations have been heavily influenced by already existing regulations on personal data protection which might result quite intuitive given the nature of AI technologies, namely the use of huge amounts of data – part of which also personal data – that need to be used for model training and fine-tuning.

Last but not least, the China-US comparison combination shows the least similarity ratio with just 23% of common words used in their respective regulatory documents. Terms that these regulations seem to share come from the technology semantic sphere like “technology”, “technologies”, “generative”, “innovation” “technical”, “training”, “resources” and “development” which point out the main objective of these regulations, i.e. their ambitious goals of balancing development and innovation over emphasizing risks as in the EU. This can be also seen directly on the single documents. For example, the CC emphatically balances security and development:

> *“The state is to adhere to the principle of placing equal emphasis on development and security, merging the promotion of innovation with governance in accordance with law; employing effective measures to encourage innovation and development in generative AI, and carrying out tolerant and cautious graded management by category of generative AI services.”* (China Law Translate, 2023)

Then, “security” is also underlined combined with words from national sentiment such as “national”, “public”, “state”. Once again, this confirms USA’s and China’s priorities which focus on getting or maintaining the lead in AI innovation while ensuring economic and geopolitical national security.

POS Distribution by Frequency



POS Distribution (100 most frequent words)

POS	EU	USA	China
Nouns	72%	74%	73%
Verbs	10%	9%	14%
Adjectives	17%	15%	9%
Adverbs	-	1%	2%
Acronyms	1%	1%	2%

Figure 8 – POS Distribution (100 most frequent words)

Analysing the POS distribution of the 100 most frequently used terms provides some interesting insights into the linguistic and regulatory focus of each document. As can be seen by *Figure 8*, across all regulations, nouns dominate the regulatory texts, comprising 72-74% of the terms. This indicates that the regulations are heavily focused on defining and detailing specific entities, concepts, and responsibilities. The next most common category is adjectives (9%-17%), followed by verbs (9%-14%), suggesting that the regulations also emphasize the characteristics of these nouns and the actions associated with them.

Specifically, from a single-term analysis of the EU's regulation it emerged that there is prevalence of nouns (72%) like "system", "data", "risk", and "authority" which points to a focus on systemic aspects of AI, oversight mechanisms, and data-related concerns. Moreover, in the group of adjectives (17%), descriptors like "high", "specific", "systemic", and "fundamental" suggest a detailed approach to regulation, emphasizing the importance of tailored and principled guidelines. Furthermore, verbs (10%) such as "ensure", "provide", and "apply" reflect obligations and procedural aspects of the regulation.

As far as the US is concerned, similar to the EU, nouns (74%) indicate an emphasis on governance structures ("secretary", "president", "agencies") and national security concepts like "security" and "privacy", as was also previously underlined. In addition, adjectives (15%) with terms like "national", "federal", "critical" and "presidential" underscore the regulation's national scope and, again, they highlight the regulatory emphasis on national security and governance. Verbs (9%) like "develop", "ensure", and "address" suggest a proactive and prescriptive regulatory attitude.

China, on the other hand, has a similar percentage ratio of nouns (73%) with respect to the other two regulations. However, the focus here is on entities ("providers", "departments") and practices ("security", "training"). Furthermore, the mention of "internet", "services" and "cybersecurity" points to specific concerns in digital realms. China also has the highest percentage rate of verbs (14%) among the three regulations which may suggest that its regulatory style shows a slightly more directive approach.

N-grams

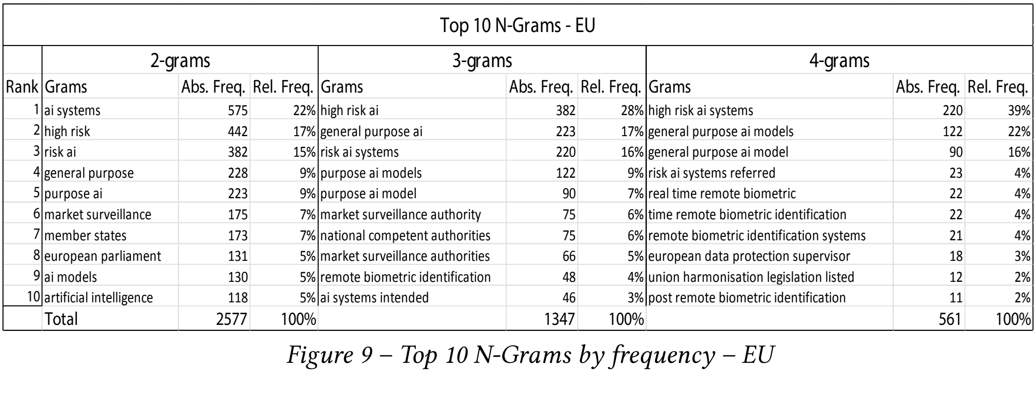
N-grams are a good source of information especially for corpora with a high number of tokens, such as the EUC. They can identify patterns that just single words sometimes cannot. In *Figure 9*, there are 3 tables with the most frequent 2-grams, 3-grams and 4-grams in the EUC, together with their Absolute Frequency (AF) and Relative Frequency (RF). It can be observed that the main topic that dominates all n-grams is the presence of the word "risk" that can be found best in the 4-gram "high risk ai systems" which constitutes 39% of the total RF of the respective table (*Figure 9*). Another topic that seems to dominate discussion in the EU regulations is the high rate of mentions regarding "general purpose AI models" which, again, in the 4-gram wordlist comprises a whopping 38% of the total RF. While the concept of real general purpose AI models is still something that has not been materialised technologically, it seems that the EU wants to be ahead of the game and try to regulate it before it truly becomes a reality.

In particular, the AI Act defines general purpose AI as follows:

> *"general purpose AI model' means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications. This does not cover AI models that are used before release on the market for research, development and prototyping activities."* (European Commission, 2024)

Although the regulation provides a definition of what constitutes general-purpose AI, it does not explicitly reference specific instances of existing technologies or identify companies utilizing these technologies. On the other hand, neither the U.S. Executive Order nor the Chinese regulation explicitly defines general purpose AI.

Next, an additional theme that seems to be discussed, based on the results given by *Figure 9*, concerns the market. As it was pointed out at the beginning of this paper, the EU is making an effort to protect its internal financial market and its own SMEs from the threat of foreign large technology companies that could pose a real threat to competitiveness and lead to monopolies. In the 3-grams wordlist, 11% of the total RF is dedicated to "market surveillance authority" (6%) and "market surveillance authorities" (5%) which might be indications of this effort to keep market development under surveillance and control. Last but not least, in the 4-grams we can get an idea of what is one the highest risks that the general purpose AI systems can pose to EU citizens according to the EU regulation, and that is "real time remote biometric identification systems". Around 12% of the total RF of the 4-grams is based on word collocations highlighting this issue.



Top 10 N-Grams - EU

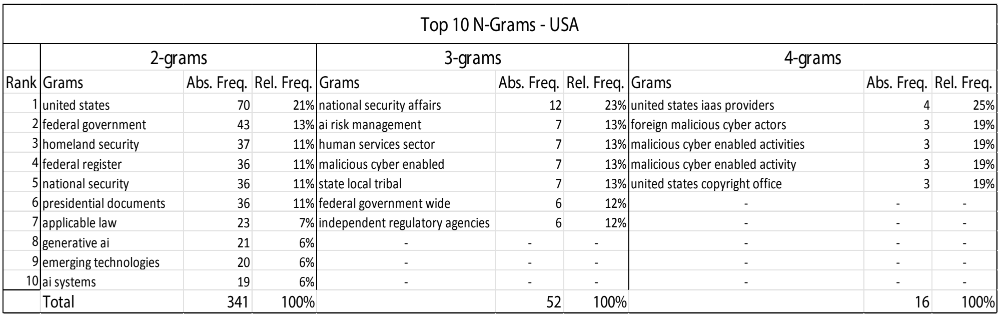
	2-grams			3-grams			4-grams		
Rank	Grams	Abs. Freq.	Rel. Freq.	Grams	Abs. Freq.	Rel. Freq.	Grams	Abs. Freq.	Rel. Freq.
1	ai systems	575	22%	high risk ai	382	28%	high risk ai systems	220	39%
2	high risk	442	17%	general purpose ai	223	17%	general purpose ai models	122	22%
3	risk ai	382	15%	risk ai systems	220	16%	general purpose ai model	90	16%
4	general purpose	228	9%	purpose ai models	122	9%	risk ai systems referred	23	4%
5	purpose ai	223	9%	purpose ai model	90	7%	real time remote biometric	22	4%
6	market surveillance	175	7%	market surveillance authority	75	6%	time remote biometric identification	22	4%
7	member states	173	7%	national competent authorities	75	6%	remote biometric identification systems	21	4%
8	european parliament	131	5%	market surveillance authorities	66	5%	european data protection supervisor	18	3%
9	ai models	130	5%	remote biometric identification	48	4%	union harmonisation legislation listed	12	2%
10	artificial intelligence	118	5%	ai systems intended	46	3%	post remote biometric identification	11	2%
	Total	2577	100%		1347	100%		561	100%

Figure 9 – Top 10 N-Grams by frequency – EU

In the USC, as underlined also in the previous sections, national sentiment and security seem to be the main topics of discussion in this regulation. This can be supported by data found starting from the 2-grams where the single most frequent 2-gram is, indeed, "United States" with a RF as high as 21%. Then, from the 2nd to the 5th position of the next most frequently used 2-grams are word collocations such as "federal government" (with a RF of 13%), "national security" and "federal register" each with 11% of RF. In total, 56% of the RF of the top 10 most frequent 2-grams belong to this semantic field, namely the national sentiment and security (*Figure 10*). This confirms, once again, the previous lexical analyses conducted in this paper.

In the 3-grams apart from the "national security affairs" (RF of 23%), there is explicit mention both on the risk AI can pose and how this risk can be contained and managed. In particular, 3-grams such as "ai risk management" (RF of 13%) and "independent regulatory agencies" (RF of 12%) are instances that highlight the need to manage AI technologies through a management systems and independent agencies (*Figure 10*). In the 4-grams, it becomes quite clear what US regulators consider one of the most prominent risks deriving from AI. According to the most frequent 4-grams, "foreign malicious cyber actors" (19% RF), "malicious cyber enabled activities" (19% RF) and "malicious cyber enabled activity" (19% RF) are the number one threat to national security.

As illustrated in *Figures 10* and *11*, the brevity of the USC and CC resulted in numerous 3-grams and 4-grams with identical absolute frequencies, making them less suited for inclusion in the tables limited to 10 instances. Consequently, for practical reasons, only the most representative 3-grams and 4-grams were included, while those with the same absolute frequency beyond the top 10 were excluded to maintain clarity and focus in the analysis.

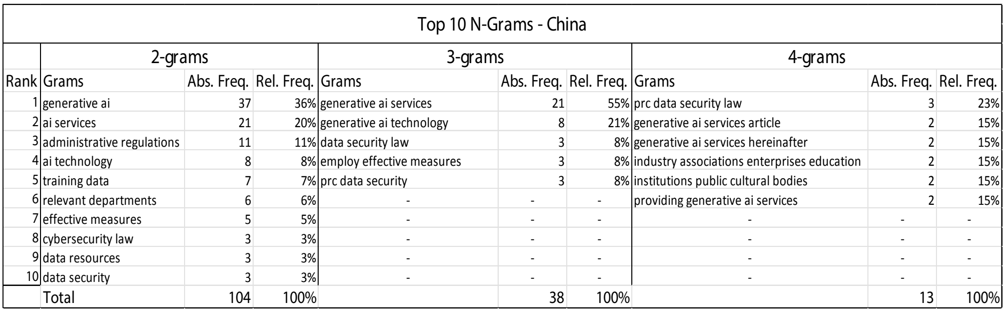


Top 10 N-Grams - USA

	2-grams			3-grams			4-grams		
Rank	Grams	Abs. Freq.	Rel. Freq.	Grams	Abs. Freq.	Rel. Freq.	Grams	Abs. Freq.	Rel. Freq.
1	united states	70	21%	national security affairs	12	23%	united states iaas providers	4	25%
2	federal government	43	13%	ai risk management	7	13%	foreign malicious cyber actors	3	19%
3	homeland security	37	11%	human services sector	7	13%	malicious cyber enabled activities	3	19%
4	federal register	36	11%	malicious cyber enabled	7	13%	malicious cyber enabled activity	3	19%
5	national security	36	11%	state local tribal	7	13%	united states copyright office	3	19%
6	presidential documents	36	11%	federal government wide	6	12%	-	-	-
7	applicable law	23	7%	independent regulatory agencies	6	12%	-	-	-
8	generative ai	21	6%	-	-	-	-	-	-
9	emerging technologies	20	6%	-	-	-	-	-	-
10	ai systems	19	6%	-	-	-	-	-	-
	Total	341	100%		52	100%		16	100%

Figure 10 – Top 10 N-Grams by frequency - USA

Even though China's regulation is relatively a small corpus, there still are some noteworthy insights that can be retrieved by n-grams (*Figure 11*). Beginning with 2-grams, 36% of this table's RF belongs to "generative AI", followed by a 20% of "AI services" and some more 2-grams that in their majority refer to technical/technological term strictly linked to AI. In the 3-grams, the term "security" and "data" appear in word collocations such as "data security law" (8% RF) and "PRC data security" (8% RF). As in the case of the US regulation, cybersecurity issues seem to be something regulators worry about here, as well. On the other hand, "data security" mentions reminds EU's data protection instances found in *Figure 9* above. Furthermore, the frequent use of the "PRC" acronym is yet another evidence of the national sentiment involved in the AI governance.



Top 10 N-Grams - China

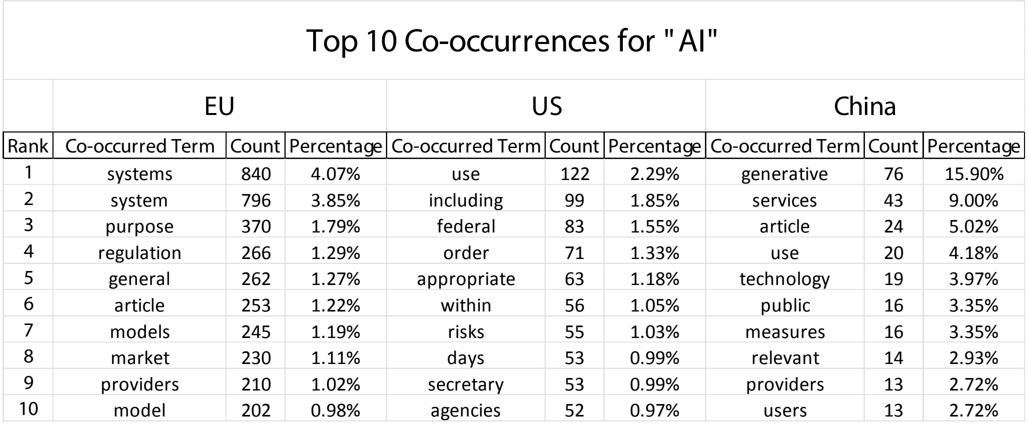
	2-grams			3-grams			4-grams		
Rank	Grams	Abs. Freq.	Rel. Freq.	Grams	Abs. Freq.	Rel. Freq.	Grams	Abs. Freq.	Rel. Freq.
1	generative ai	37	36%	generative ai services	21	55%	prc data security law	3	23%
2	ai services	21	20%	generative ai technology	8	21%	generative ai services article	2	15%
3	administrative regulations	11	11%	data security law	3	8%	generative ai services hereinafter	2	15%
4	ai technology	8	8%	employ effective measures	3	8%	industry associations enterprises education	2	15%
5	training data	7	7%	prc data security	3	8%	institutions public cultural bodies	2	15%
6	relevant departments	6	6%	-	-	-	providing generative ai services	2	15%
7	effective measures	5	5%	-	-	-	-	-	-
8	cybersecurity law	3	3%	-	-	-	-	-	-
9	data resources	3	3%	-	-	-	-	-	-
10	data security	3	3%	-	-	-	-	-	-
	Total	104	100%		38	100%		13	100%

Figure 11 – Top 10 N-Grams by frequency – China

Co-occurrences

Measuring the Co-occurrences Frequency (CF) of certain terms, i.e. the frequency of words that co-occur in the same context but not necessarily one directly right after the other (as in the n-grams), allows a better understanding of how regulators perceive key concepts in the drafting of these regulations. In this section, a CF research on the following four key terms is made: "AI", "data", "security"/"safety" and "innovation". There are many reasons why these terms were selected to be furtherly investigated. One of the most important ones is the fact that they constitute fundamental concepts in all three regulations both from a pragmatic point of view and a statistical one (see previous sections). Furthermore, "AI", "data" and "security"/"safety" were almost always present in the top 10 most frequent terms in all regulations as underlined in the previous sections, while "innovation" results having a catalytic power over all regulations not only because all of them mention it but also because it is a concept that comes hand by hand with AI technologies.

While other concepts would have been interesting to investigate as well, such as "risk(s)", "market" or "economy", it was surprisingly discovered that in the CC the term "risk" was mentioned only once in its singular form, whereas the other above mentioned terms did not appear at all. This indicates that, although China's regulation discusses issues regarding security and data protection (as seen above), no mentions of the word "risk(s)" are used whatsoever. In the CC, this term seems to be substituted in some very few instances by the words "endangered" and "endangering" under Article 4, which is a prescriptive article focusing on the "respect" that AI deployers should always show with respect to a list of rights provided by the regulators.



Top 10 Co-occurrences for "AI"

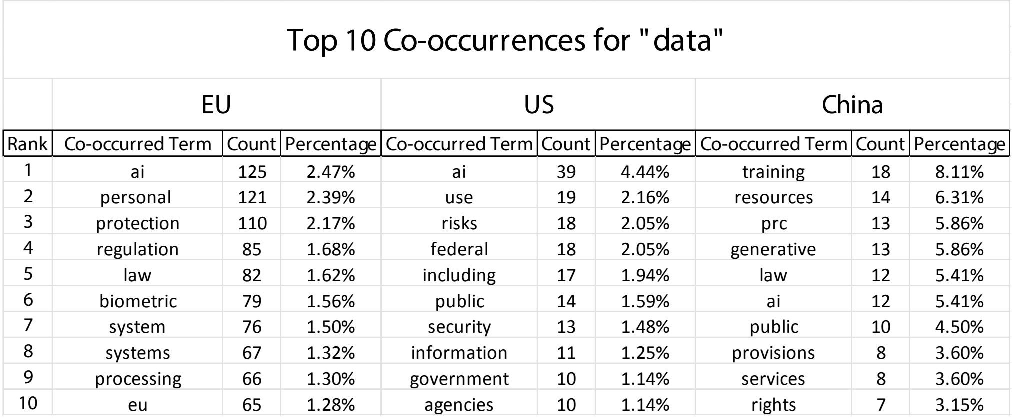
	EU			US			China		
Rank	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage
1	systems	840	4.07%	use	122	2.29%	generative	76	15.90%
2	system	796	3.85%	including	99	1.85%	services	43	9.00%
3	purpose	370	1.79%	federal	83	1.55%	article	24	5.02%
4	regulation	266	1.29%	order	71	1.33%	use	20	4.18%
5	general	262	1.27%	appropriate	63	1.18%	technology	19	3.97%
6	article	253	1.22%	within	56	1.05%	public	16	3.35%
7	models	245	1.19%	risks	55	1.03%	measures	16	3.35%
8	market	230	1.11%	days	53	0.99%	relevant	14	2.93%
9	providers	210	1.02%	secretary	53	0.99%	providers	13	2.72%
10	model	202	0.98%	agencies	52	0.97%	users	13	2.72%

Figure 12 – Top 10 Co-occurrences for "AI"

Starting with "AI", *Figure 12* shows that in the EU the terminology leans towards systemic and regulatory aspects, with terms like "systems," "system," and "regulation" being predominant. This suggests a focus on the infrastructure and the broader framework within which AI operates. Moreover, contrary to the other two regulations, in the EUC, "market" seems to point to the AI market regulation which could ensure the maintenance of its position in global competition while safeguarding its foundational values and interests.. To furtherly support this assertion, the following excerpt from the AI Act emphatically states that:

> *"The purpose of this Regulation is to improve the functioning of the internal market and promoting the uptake of human-centric and trustworthy artificial intelligence, while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, rule of law, and environmental protection against harmful effects of artificial intelligence systems in the Union and supporting innovation"* (European Commission, 2024, Art. 1)

In the US, AI co-occurrences are more about the application and management within specific contexts, with words like "use," "federal," and "order." This indicates a focus on how AI is utilized within certain regulatory or operational boundaries. "Risks" also consists of a term accompanying quite often AI mentions which points out US regulators views on AI and probable outcomes. The Chinese terms are strongly associated with the generative capabilities of AI and its service applications, as seen with "generative" and "services." This could point towards a strong interest in the innovative output from AI technologies rather than restrictive rules and laws. This is even more enforced by the only word from the regulative spectrum, namely "measures", which is far from the more restrictive EU's use of "regulation" and US's "order", alluding to "executive presidential order".



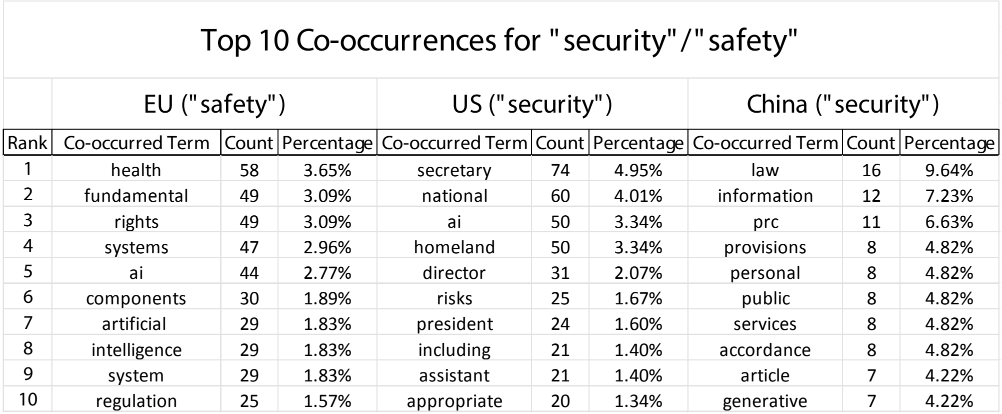
Top 10 Co-occurrences for "data"

	EU			US			China		
Rank	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage
1	ai	125	2.47%	ai	39	4.44%	training	18	8.11%
2	personal	121	2.39%	use	19	2.16%	resources	14	6.31%
3	protection	110	2.17%	risks	18	2.05%	prc	13	5.86%
4	regulation	85	1.68%	federal	18	2.05%	generative	13	5.86%
5	law	82	1.62%	including	17	1.94%	law	12	5.41%
6	biometric	79	1.56%	public	14	1.59%	ai	12	5.41%
7	system	76	1.50%	security	13	1.48%	public	10	4.50%
8	systems	67	1.32%	information	11	1.25%	provisions	8	3.60%
9	processing	66	1.30%	government	10	1.14%	services	8	3.60%
10	eu	65	1.28%	agencies	10	1.14%	rights	7	3.15%

Figure 13 – Top 10 Co-occurrences for "data"

In Figure 13 a strong emphasis on protection and regulation in the "data" co-occurrences of the EUC can be observed. The high frequency of terms like "personal", "protection," "regulation," "law," and "biometric" highlights a comprehensive legal framework aimed at safeguarding personal data. This is cohesive and consistent with the EU's strong position on data privacy, evidenced by regulations like GDPR. In the US, while protection is a concern (terms like "risks," "security"), the emphasis is more on the operational and governance aspects of data usage ("federal," "government," "agencies"). This suggests a lighter – compared to the EU – approach to data regulation, focusing on specific applications and risks associated with data use. In the Chinese regulation, the focus shifts towards utilization and infrastructure development of data ("training," "resources," "generative"). There is, however, a specific mention on the legal aspects ("law," "rights"), indicating a growing awareness and possible implementation of data and AI governance frameworks.

As a whole, the EU and, to a lesser extent, the US demonstrate a robust regulatory discourse. China's focus is less on regulation and more on the application and infrastructure, although legal terms are still significant. All regulations discuss technological aspects but from different angles. The EU and the US seem to take a more cautious approach while China a more capability-enhancement perspective.

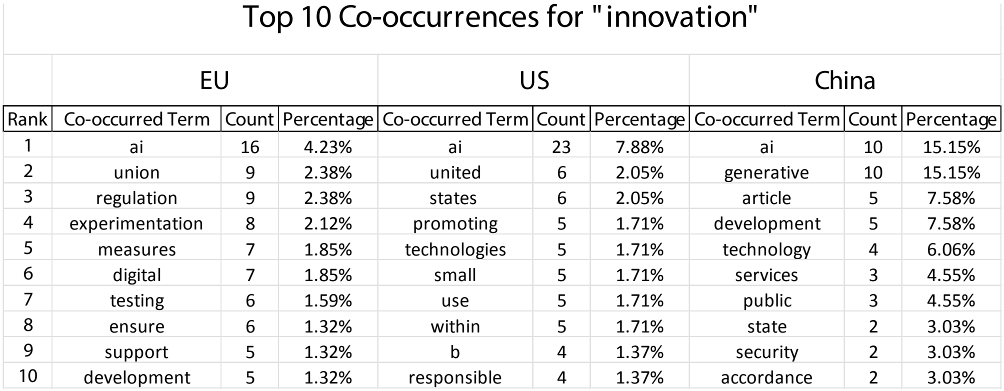


Top 10 Co-occurrences for "security"/"safety"

	EU ("safety")			US ("security")			China ("security")		
Rank	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage
1	health	58	3.65%	secretary	74	4.95%	law	16	9.64%
2	fundamental	49	3.09%	national	60	4.01%	information	12	7.23%
3	rights	49	3.09%	ai	50	3.34%	prc	11	6.63%
4	systems	47	2.96%	homeland	50	3.34%	provisions	8	4.82%
5	ai	44	2.77%	director	31	2.07%	personal	8	4.82%
6	components	30	1.89%	risks	25	1.67%	public	8	4.82%
7	artificial	29	1.83%	president	24	1.60%	services	8	4.82%
8	intelligence	29	1.83%	including	21	1.40%	accordance	8	4.82%
9	system	29	1.83%	assistant	21	1.40%	article	7	4.22%
10	regulation	25	1.57%	appropriate	20	1.34%	generative	7	4.22%

Figure 14 – Top 10 Co-occurrences for "security"/"safety"

Figure 14 shows the distribution of the most frequent co-occurrences for "safety" or "security" in the three corpora. From *Figure 14* emerges that the EU takes a more human-centric focus regarding safety. That is to say, it emphasizes "health", "fundamental" "rights", and "regulation", reflecting a strong concern for individual safety and "rights". The presence of terms like "health" and "rights" suggests an approach where safety is aligned closely with human welfare and social systems. In the US, terms such as "secretary," "national," "homeland," "president," and "assistant" are prominent, suggesting that security is often discussed in a governmental and institutional context with a high national sentiment involvement. It also points towards a national security position that involves various levels of governance. Again, the presence of "risks" indicates a focus on assessing and mitigating potential threats on national security. In the CC, terms like "law," "provisions," and "accordance" could be evidence of a strong legal approach to security, focusing on compliance with laws and regulations. This reflects its regulators' intentions to create security measures within legal frameworks. What is more, words like "information," "prc" (People's Republic of China), and "public" show an orientation towards state control and the importance of managing "information" and "generative" AI "services".



Top 10 Co-occurrences for "innovation"

	EU			US			China		
Rank	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage	Co-occurred Term	Count	Percentage
1	ai	16	4.23%	ai	23	7.88%	ai	10	15.15%
2	union	9	2.38%	united	6	2.05%	generative	10	15.15%
3	regulation	9	2.38%	states	6	2.05%	article	5	7.58%
4	experimentation	8	2.12%	promoting	5	1.71%	development	5	7.58%
5	measures	7	1.85%	technologies	5	1.71%	technology	4	6.06%
6	digital	7	1.85%	small	5	1.71%	services	3	4.55%
7	testing	6	1.59%	use	5	1.71%	public	3	4.55%
8	ensure	6	1.32%	within	5	1.71%	state	2	3.03%
9	support	5	1.32%	b	4	1.37%	security	2	3.03%
10	development	5	1.32%	responsible	4	1.37%	accordance	2	3.03%

Figure 15 – Top 10 Co-occurrences for "innovation"

As for "innovation" co-occurrences, by observing *Figure 15* it can be inferred that the EU emphasizes regulatory and supportive frameworks for "AI" innovation, with terms like "regulation", "experimentation", "measures", "testing", "ensure" and "support" frequently mentioned. This suggests a structured approach to innovation, focusing on creating a supportive environment for technological advancement within a regulatory framework. Indeed, the EU promotes experimentation in "digital" controlled environments or sandboxes before any high risk AI technological innovations are released to the general public. In addition, the mention of "union" underscores a collaborative approach across EU member states, aiming to foster innovation collectively rather than in isolation. The US, on the other hand, advocates for "responsible" "use" of "AI" "technologies" "promoting" development within its national borders ("united" "states"). In the CC, the high number of co-occurrences with words such as "generative" and "technology" alongside "public" and "state" highlight a state-driven approach to innovation, particularly in cutting-edge innovations. This suggests significant government investment and involvement in pushing the boundaries and "development" of AI technologies.

6. Conclusion

With the use of NLP techniques, this research has systematically explored the emerging regulatory landscape of AI through a detailed lexical and semantic analysis of the most important legal texts and regulations of the three most active and prominent actors in the AI revolution: the EU, the US and China. By means of a series of NLP methods such as word frequency, lexical distributions, co-occurrence lexical metrics etc., we achieved to retrieve interesting information and insights of how each institution and government perceives AI governance and what intentions and ambitions they evoke through the drafting of their specific regulations.

More specifically, starting from the word frequency, it was revealed that "risk", as far as AI technologies are concerned, was a term that is perceived differently by each regulation. The EU seemed to link the concept of risk not only to human rights and health but also "markets". This suggests a defensive strategy aimed at protecting the EU's internal market from potentially disruptive external AI forces, which also translates into assisting EU-based SMEs in the global AI race. However, to this realm there is the risk of overregulation which could create technological stagnation, placing the EU at a competitive disadvantage—a scenario that might lead to a significant talent drain and reduced innovative capacity within the region. In contrast, the US and China show a more aggressive regulatory focus based on national security and technological leadership. The prominence of terms such as 'federal' and 'national security' in US documents indicates a strategic emphasis on maintaining technological supremacy and safeguarding economic and geopolitical interests. China, on the other hand, discussed almost superficially the notion of risk. Instead, the frequent mention of "generative AI", "PRC" and "technology" underscores their focus on harnessing AI for state-led developmental and innovative purposes.

In the comparative analysis of the three regulations it was found that they shared little common ground which was estimated at 15% based on the lexical similarity of the most frequent terms. Moreover, EU's regulation seemed to be 32% similar to the US and 29% to the Chinese one. China and the US shared the least amount of common terms reaching a mere 23% of similarity. Moreover, it was shown that across all regulations, nouns dominate the regulatory texts, comprising 72-74% of the terms. This indicates that the regulations are heavily focused on defining and detailing specific entities, concepts, and responsibilities.

Furthermore, the findings from the n-grams analysis illuminated key thematic preoccupations within each jurisdiction. In the EU corpus, the recurring 4-gram "high risk AI systems" highlighted the significant focus on risk management in AI. This reflects the EU's cautious position towards AI, emphasizing regulatory oversight to mitigate potential threats and maintain market stability which might translate into protecting the internal market from the dominance of foreign tech giants and fostering a competitive environment for EU-based SMEs. Conversely, the US corpus prominently featured 2-grams like "United States", "federal government", and "national security" which indicate a strong national focus, with AI viewed as a pivotal element in maintaining economic security and technological supremacy. In China, the prevalent n-grams such as "generative AI" and "AI services" point towards a strategy focused on leveraging AI for technological advancement and state-led innovation, while trying to maintain a balance between development and security, as specifically and explicitly stated in its regulation.

Last but not least, in the co-occurrences analysis our understanding of the regulatory intent and focus across regions was furtherly enriched. In the EU, terms related to “systems”, “regulation”, and “market” frequently intersect with AI, suggesting a comprehensive framework aimed at integrating AI within a tightly regulated environment. This is consistent with the EU’s strong emphasis on data privacy and individual rights, as evidenced by the frequent co-occurrence of the term “data” with terms like “personal”, “protection”, and “regulation”. In the US, the co-occurrence of terms like “risks”, “security”, and “federal” alongside AI reflects a more segmented regulatory approach, emphasizing specific risks and operational concerns over broader systemic regulations. This points to a prioritization of national security and economic interests in the US’s AI strategy. In the Chinese regulation, the high number of co-occurrences of the term “AI” with words such as “generative” and “technology” alongside “public” and “state” highlight a state-driven approach to innovation, particularly in cutting-edge innovations.

The comparative linguistic analysis across different AI regulatory frameworks reveals a complex landscape where strategic national interests influence the development and implementation of AI regulations. While the EU’s approach is characterized by a protective and risk-averse position aimed at protecting market competitiveness and consumer rights, the US and China are more focused on using AI for national security and innovation. As AI continues to evolve, both AI innovation and regulation will need continuous adjustments and updated frameworks. The challenge for policymakers will be to promote and encourage technological innovation, yet ensuring it aligns with national interests and ethical standards. This fundamental balance will be crucial in determining the global trajectory of AI development and its integration into society.

The insights from this study provide a thorough understanding of current AI regulations which could assist future research and policymaking, and could suggest that an adaptive and responsive approach to AI regulation is essential to navigate the complexities of a rapidly evolving technological landscape.

References

Amazon, *Amazon and Anthropic deepen their shared commitment to advancing generative AI*, accessed 18 September 2024, https://www.aboutamazon.com/news/company-news/amazon-anthropic-ai-investment.

Badr M., *Unleashing the power of AI: The Microsoft and OpenAI partnership*, 2023.

China Law Translate, *Interim Measures for the Management of Generative Artificial Intelligence Services*, 2023, https://www.chinalawtranslate.com/en/generative-ai-interim/.

Deep Synthesis Provisions: Administrative Provisions on Deep Synthesis in Internet-based Information Services, accessed 18 September 2024, https://perma.cc/JE3W-PF26.

Dias J. C., Martins A., Pinto P., *An Analysis of Infractions and Fines in the Context of the GDPR*, in *International Journal of Marketing, Communication and New Media* 12, 2023.

Dral P. O., Ullah A., *Call for urgent regulations on artificial intelligence*, available at SSRN 4418449, 2023.

Atomico, *State of the European Tech 2023*, accessed 18 September 2024, https://prismic-io.s3.amazonaws.com/atomico-2023/b598f20b-3e6a-4556-bfbd-9b2d71a72183_Atomico-state+of+european+tech+report+2023+%281%29.pdf.

English Corpora, *The Corpus of Contemporary American English (COCA) and the British National Corpus (BNC)*, accessed 18 September 2024, https://www.english-corpora.org/coca/compare-bnc.asp.

Eur-lex, *General data protection regulation (GDPR)*, accessed 18 September 2024, https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html.

Eurostat, *Large enterprises generate just over one third of employment*, 2019, https://ec.europa.eu/eurostat/cache/digpub/european_economy/bloc-3b.html?lang=en.

European Commission, *AI Act*, accessed 18 September 2024, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Hacker P., *AI Regulation in Europe: From the AI Act to Future Regulatory Challenges*, in *ArXiv*, abs/2310.04072, 2023, https://doi.org/10.48550/arXiv.2310.04072.

Hine E., Floridi L., *Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies*, in *AI & SOCIETY* 39, 2024, 1, pp. 257-278.

Larson C., *China's AI imperative*, in *Science* 359, 2018, 6376, pp. 628-630, https://doi.org/10.1126/science.359.6376.628.

Laurence A., *Tokyo, Japan: Waseda University*, accessed 18 September 2024, http://www.laurenceanthony.net/.

Laurence A., *Common statistics used in corpus linguistics*, accessed 18 September 2024, https://www.laurenceanthony.net/resources/statistics/common_statistics_used_in_corpus_linguistics.pdf.

Manheim K., Kaplan L., *Artificial intelligence: Risks to privacy and democracy*, in *Yale Journal of Law & Technology* 21, 2019, pp. 106.

Mello M. M., Shah N. H., Char D. S., *President Biden's Executive Order on Artificial Intelligence—Implications for Health Care Organizations*, in *JAMA*, 2023.

Musch S., Borrelli M., Kerrigan C., *The EU AI Act: A Comprehensive Regulatory Framework for Ethical AI Development*, available at SSRN 4549248, 2023.

Muthukrishnan N., Maleki F., Ovens K., Reinhold C., Forghani B., Forghani R., *Brief history of artificial intelligence*, in *Neuroimaging Clinics of North America* 30, 2020, 4, pp. 393-399.

Tricot R., *Venture capital investments in artificial intelligence*, in *OECD Digital Economy Papers*, 2021, https://doi.org/10.1787/f97beae7-en.

OECD.AI, *VC investments in AI by country*, visualisations powered by JSI using data from Preqin, accessed 18 September 2024, https://oecd.ai/en/data?selectedArea=investments-in-ai-and-data&selectedVisualization=vc-investments-in-ai-by-country.

Ono K., Morita A., *Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked Against MMLU*, in *Authorea Preprints*, 2024.

Pawelec M., *Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions*, in *Digital Society* 1, 2022, 2, pp. 19.

Pi Y., *Missing value chain in generative AI governance China as an example*, in *ArXiv preprint* arXiv:2401.02799, 2024.

Reuters, *Mistral AI raises 385 mln euros in second round in seven months*, accessed 18 September 2024, https://www.reuters.com/technology/mistral-ai-raises-385-mln-euros-second-round-seven-months-2023-12-11/.

Roberts H., Cowls J., Morley J., Taddeo M., Wang V., Floridi L., *The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation*, in *Ethics, Governance, and Policies in Artificial Intelligence*, 2021, pp. 47-79.

Suleyman M., *The coming wave: Technology, power, and the twenty-first century's greatest dilemma*, Crown, 2023.

Time, *The Ultimate Election Year: All the Elections Around the World in 2024*, accessed 18 September 2024, https://time.com/6550920/world-elections-2024/.

The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, accessed 18 September 2024, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Xiao B., *Agile and Iterative Governance: China's Regulatory Response to AI*, available at SSRN 4705898, 2024.