MASTER THESIS | MASTER'S THESIS

Titel | Title

Regulating AI: A Comparative Analysis of EU, US, and Chinese Law regarding the Protection of Human Rights, Data Privacy, and Safety

verfasst von | submitted by

Vladislav Vdovin

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Laws (LL.M.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 992 548
Universitätslehrgang lt. Studienblatt | Postgraduate programme as it appears on the student record sheet:	Europäisches und Internationales Wirtschaftsrecht (LL.M.) [online Englisch]
Betreut von | Supervisor:	ao. Univ.-Prof. Mag. Dr. Siegfried Fina

ABSTRACT (ENGLISH)

This master thesis contains a comparative analysis of AI regulation on human rights, personal data, and safety in the EU, the US, and China. The study includes the newly developed legislation in recent months and the regulatory policy. Each of the compared jurisdictions has the ambition to lead the development of AI technology. The research confirms the necessity of AI-specific regulation, including setting out AI-specific rights and protections for individuals. The jurisdictions are balancing between development and control and applying a risk-based approach. The scope of the rights and the risks assigned to AI use cases depend on the policy. The safety framework of the compared jurisdictions has much in common and mostly relates to high-risk use cases. Appendix 1 of the thesis contains a detailed comparison of regulatory features.

ABSTRACT (DEUTSCH)

Diese Masterarbeit enthält eine vergleichende Analyse der Regulierung von Künstlicher Intelligenz (KI) in Bezug auf Menschenrechte, personenbezogene Daten und Sicherheit in der EU, den USA und China. Die Studie umfasst die in den letzten Monaten neu entwickelten Gesetzgebungen sowie die regulatorische Politik. Jede der verglichenen Jurisdiktionen verfolgt das Ziel, führend in der Entwicklung von KI-Technologien zu sein. Die Forschung bestätigt die Notwendigkeit einer KI-spezifischen Regulierung, einschließlich der Festlegung von KI-spezifischen Rechten und Schutzmaßnahmen für Einzelpersonen. Die Jurisdiktionen balancieren zwischen Entwicklung und Kontrolle und wenden einen risikobasierten Ansatz an. Der Umfang der Rechte und die den KI-Anwendungsfällen zugeordneten Risiken hängen von der jeweiligen Politik ab. Der Sicherheitsrahmen der verglichenen Jurisdiktionen weist viele Gemeinsamkeiten auf und bezieht sich hauptsächlich auf Hochrisikoanwendungsfälle. Anhang 1 der Arbeit enthält einen detaillierten Vergleich der regulatorischen Merkmale.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

2023 AI RD Plan	NSTC, Select Committee on Artificial Intelligence: Report 'National Artificial Intelligence Research and Development Strategic Plan 2023 Update'
AGI	artificial general intelligence
AGI	artificial general intelligence
AI	artificial intelligence
AI Act	European Parliament and of the Council Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)
AI Bill of Rights	Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People
AI Convention	Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Vilnius, 5 September 2024) Council of Europe Treaty Series No. 225
AI HLEG	High-Level Expert Group on Artificial Intelligence
AI HLEG Ethics Guidelines	AI HLEG, 'Ethics Guidelines for Trustworthy AI' (European Commission, 8 April 2019)
AI Principles for Workers	U.S. Department of Labor, 'Artificial Intelligence And Worker Well-being: Principles And Best Practices For Developers And Employers'
AILD	Commission, 'Proposal for a Directive of The European Parliament and of The Council on adapting non-contractual civil

	liability rules to artificial intelligence (AI Liability Directive)' COM(2022) 496 final
Algorithmic Recommendations Provisions	Cyberspace Administration of China, the Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China, the State Administration for Market Regulation, 'Provisions on the Management of Algorithmic Recommendations in Internet Information Services' (31 December 2021)
ALTAI	Assessment List for Trustworthy Artificial Intelligence of 17 July 2020 by the AI HLEG
ASI	artificial superintelligence
CAIO	chef AI officer
CBRN	chemical, biological, radiological and nuclear materials and agents that could potentially harm society through their accidental or deliberate release, dissemination, or impacts
CCTV	closed-circuit television
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
Charter of Fundamental Rights	Charter of Fundamental Rights of the European Union 2000
China Safety Governance Framework	National Information Security Standardization Technical Committee (TC260), 'AI Safety Governance Framework' (September 2024)
China WP on Trustworthy AI	China Academy of Information and Communications Technology and JD Explore Academy, 'White Paper on Trustworthy Artificial Intelligence,' July 2021

CSET	Center for Security and Emerging Technology
Cyber Resilience Act	European Parliament and the Council Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act) (Text with EEA relevance)
Deep Synthesis Provisions	Cyberspace Administration of China, the Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China, 'Provisions on the Administration of Deep Synthesis Internet Information Services' (25 November 2022)
Digital Markets Act	European Parliament and the Council Regulation (EU) 2022/1925 of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance)
Digital Services Act	European Parliament and the Council Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance)
DODR	European Declaration on Digital Rights and Principles for the Digital Decade
Draft AI Law	Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars) of 16 March 2024
Draft PLD	Commission, 'Proposal for a Directive of the European Parliament and of the Council on liability for defective products' COM/2022/495 final
ENISA	European Union Agency for Cybersecurity

EO 13960	Executive Order No. 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, 3 December 2020
EO 13859	Executive Order No. 13859: Maintaining American Leadership in Artificial Intelligence, 11 February 2019
EO 13859	Executive Order No. 13859 ‘Maintaining American Leadership in Artificial Intelligence’
EO 14110	Executive Order No. 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 30 October 2023
ETSI	European Telecommunications Standards Institute
EUDPR	European Parliament and the Council Regulation (EU) 2018/1725 of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC
GAI	generative artificial intelligence
GDPR	European Parliament and the Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)
Generative AI Measures	Cyberspace Administration of China, National Development and Reform Commission of the People's Republic of China, Ministry of Education of the People's Republic of China, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology of the People's Republic of China, Ministry of Public Security of the People's Republic of China, State Administration of Radio and Television,

	'Interim Measures for the Management of Generative Artificial Intelligence Services' (10 July 2023)
GPAI	general purpose artificial intelligence
GPNGAI	National Governance Committee for the New Generation Artificial Intelligence, 'Governance Principles for the New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence'
GPT	generative pre-trained transformer
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
ITU	the International Telecommunications Union
LED	European Parliament and the Council Directive (EU) 2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA
LLM	large language model
MICA	European Parliament and the Council Regulation (EU) 2023/1114 of 31 May 2023 on markets in crypto-assets, and amending Regulations (EU) No 1093/2010 and (EU) No 1095/2010 and Directives 2013/36/EU and (EU) 2019/1937 (Text with EEA relevance)
NIST	National Institute of Standards and Technology, U.S. Department of Commerce

NIST 2019 Plan	NIST, 'U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools' (submitted on 9 August 2019)
NIST AI 100-4	NIST AI 100-4. Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency (Draft for Public Comment)
NIST AI 100-5	NIST AI 100-5. A Plan for Global Engagement on AI Standards
NIST AI 600-1	NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. July 2024
NIST AI 800-1	NIST AI 800-1. Managing Misuse Risk for Dual-Use Foundation Models (Initial Public Draft)
NIST AI RMF	National Institute of Standards and Technology, U.S. Department of Commerce (NIST). NIST.AI.100-1 Artificial Intelligence Risk Management Framework of January 2023
NSCAI	National Security Commission on Artificial Intelligence
NSM AI	National Security Memorandum on Artificial Intelligence
NSM AI Framework	Framework to Advance AI Governance and Risk Management in National Security
NSTC	National Science and Technology Council
NSTC 2016 AI Plan	National Science and Technology Council (NSTC), Networking and Information Technology Research and Development Subcommittee, 'The national artificial intelligence research and development strategic plan' (October 2016)
NSTC AI Report	National Science and Technology Council (NSTC), Committee on Technology, 'Preparing for the future of artificial intelligence' (October 2016)

OECD	the Organisation for Economic Co-operation and Development
OMB M-21-06	Executive Office of the President, Office of Management and Budget (OMB), 'Memorandum M-21-06 for the heads of executive departments and agencies: Guidance for Regulation of Artificial Intelligence Applications' (17 November 2020)
OMB M-24-10	Executive Office of the President, Office of Management and Budget, 'Memorandum No. M-24-10 for the Heads of Executive Departments and Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence' (28 March 2024)
PET	privacy-enhancing technology
PIPL	Personal Information Protection Law of the People's Republic of China of 20 August 2021
PLD	Product Liability Directive – Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products
TEVV	testing, evaluating, verification and validation
UNESCO	the United Nations Educational Scientific and Cultural Organization
UDHR	Universal Declaration of Human Rights
Updated Coordinated Plan on AI	Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence' COM/2021/205, Annex 'Coordinated Plan on Artificial Intelligence 2021 Review'
WIPO	World Intellectual Property Organization

TABLE OF LEGISLATION

International

Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Vilnius, 5 September 2024) Council of Europe Treaty Series No. 225

OECD AI Principles of 2019 (revised in 2024) <www.oecd.org/en/topics/ai-principles.html> accessed 12 December 2024

Policy Paper: the Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023 (Gov.UK, 1 November 2023) <www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023> accessed 4 December 2024

UN General Assembly, 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (11 March 2024) A/78/L.49

Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR)

The European Union

'A definition of AI: Main capabilities and scientific disciplines.' A report/study by the High-Level Expert Group on Artificial Intelligence (AI HLEG) of 8 April 2019 (European Commission, 18 December 2018) <https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines> accessed 5 December 2024

AH HLEG, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment' (European Commission, 17 July 2020) <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment> accessed 6 December 2024

AI HLEG, 'Ethics Guidelines for Trustworthy AI' (European Commission, 8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 6 December 2024 (AI HLEG Ethics Guidelines)

AI HLEG, 'Policy and investment recommendations for trustworthy Artificial Intelligence' (European Commission, 26 June 2019) <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence> accessed 6 December 2024

Charter of Fundamental Rights of the European Union 2000 (Charter of Fundamental Rights)

Commission staff working document, 'Impact assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts' SWD(2021) 84 final

Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence' COM/2021/205 final

Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human Centric Artificial Intelligence' COM(2019)168

Commission, ‘Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Coordinated Plan on Artificial Intelligence’ COM/2018/795 final

Commission, ‘Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions artificial intelligence for Europe’ COM/2018/237 final

Commission, ‘Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Coordinated Plan on Artificial Intelligence’ COM/2018/795 final

Commission, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence’ COM/2021/205

Commission, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence’ COM/2021/205, Annex ‘Coordinated Plan on Artificial Intelligence 2021 Review’

Commission, ‘Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European democracy action plan’ COM/2020/790 final

Commission, ‘Cyber Resilience Act’ (European Commission, 15 September 2022) <https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act> accessed 7 December 2024

Commission, ‘Decision of 24.1.2024 establishing the European Artificial Intelligence Office’ C(2024) 390 final

Commission, ‘Joint Declaration on the EU's legislative priorities for 2018-19’ (European Commission, 14 December 2017) <https://commission.europa.eu/publications/joint-declaration-eus-legislative-priorities-2018-19_en> accessed 6 December 2024

Commission, ‘Proposal for a Directive of The European Parliament and of The Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)’ COM(2022) 496 final

Commission, ‘Proposal for a Directive of the European Parliament and of the Council on liability for defective products’ COM/2022/495 final

Commission, ‘Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence act) and amending certain Union legislative acts’ COM/2021/206 final

Commission, ‘Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts’ COM(2021) 206 final

Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on machinery products’ COM/2021/202 final

Commission, ‘Proposal for a Regulation of the European Parliament and of the Council on machinery products’ COM(2021) 202 final

Commission, ‘The Cybersecurity Strategy’ (European Commission) <https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-strategy > accessed 7 December 2024

Commission, ‘White Paper On Artificial Intelligence - A European approach to excellence and trust’ COM(2020) 65 final

Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L210

Data Protection Working Party Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 of 3 October 2017 (last revised on 6 February 2018) WP251rev.01

Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety (Text with EEA relevance) [2001] OJ L11

Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast) [2006] OJ L 157/24

European Declaration on Digital Rights and Principles for the Digital Decade [2023] OJ C 23/1

European Parliament and of the Council Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) [2024] OJ L2024/1689

European Parliament and the Council Directive (EU) 2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L119/89

European Parliament and the Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) [2016] OJ L119/1

European Parliament and the Council Regulation (EU) 2018/1725 of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC [2018] OJ L 295/39

European Parliament and the Council Regulation (EU) 2019/1020 of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011 (Text with EEA relevance.) [2019] OJ L169/1.

European Parliament and the Council Regulation (EU) 2022/1925 of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) [2022] OJ L295/1

European Parliament and the Council Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) [2022] OJ L277/1

European Parliament and the Council Regulation (EU) 2023/1114 of 31 May 2023 on markets in crypto-assets, and amending Regulations (EU) No 1093/2010 and (EU) No 1095/2010 and Directives 2013/36/EU and (EU) 2019/1937 (Text with EEA relevance) [2023] OJ L150/50

European Parliament and the Council Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act) (Text with EEA relevance) [2024] (not yet in force, pending notification)

European Parliament and the Council, ‘Joint communication to the European Parliament and the Council the EU's Cybersecurity Strategy for the Digital Decade’ JOIN/2020/18 final

European Parliament resolution 2015/2103(INL) of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics [2018] OJ C252/239

The United States

Consolidated Appropriations Act, 2021, H.R.113, Public Law No. 116-260, 134 stat 1182, Division U, Title I – AI in Government Act of 2020.

Defense Innovation Board, ‘AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (2019) <https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF> accessed 8 December 2024

Executive Office of the President, Office of Management and Budget, ‘Memorandum No. M-24-10 for the Heads of Executive Departments and Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence’ (28 March 2024) <www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf> accessed 9 December 2024

Executive Office of the President, Office of Management and Budget, ‘Memorandum M-21-06 for the heads of executive departments and agencies: Guidance for Regulation of Artificial Intelligence Applications’ (17 November 2020) <www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf> accessed 8 December 2024

Federal AI Governance and Transparency Act, H.R.7532 (Draft) <www.congress.gov/bill/118th-congress/house-bill/7532/all-actions> accessed 9 December 2024

James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, 23 December 2022, H.R.7776, Public Law 117-263, Title LXXIII, Subtitle A – Global Catastrophic Risk Management Act of 2022

James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, 23 December 2022, H.R.7776, Public Law 117-263, Title LXXII, Subtitle B – Advancing American AI Act

National Artificial Intelligence Initiative Act of 2020, 15 U.S.C. § 9401 (2022).

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST.AI.100-1 Artificial Intelligence Risk Management Framework of January 2023 <https://doi.org/10.6028/NIST.AI.100-1> accessed 6 December 2024 (NIST AI RMF), s 3.

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST AI 100-4. Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency (Draft for Public Comment) <https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf> accessed 9 December 2024

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST AI 100-5. A Plan for Global Engagement on AI Standards <https://doi.org/10.6028/NIST.AI.100-5> accessed 9 December 2024

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. July 2024 <https://doi.org/10.6028/NIST.AI.600-1> accessed 9 December 2024

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST AI 800-1. Managing Misuse Risk for Dual-Use Foundation Models (Initial Public Draft) <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf> accessed 9 December 2024

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management, version 1.0 (16 January 2020) <https://doi.org/10.6028/NIST.CSWP.01162020> accessed 9 December 2024

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST SP 800-218A. Secure Software Development Practices for Generative AI and Dual-Use Foundation Models ipd, April 2024 <https://doi.org/10.6028/NIST.SP.800-218A.ipd> accessed 9 December 2024

National Institute of Standards and Technology, U.S. Department of Commerce (NIST), NIST, 'U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools' (submitted on 9 August 2019) <www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf> accessed 8 December 2024.

National Science and Technology Council (NSTC), Committee on Technology, 'Preparing for the future of artificial intelligence' (October 2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf> accessed 8 December 2024

National Science and Technology Council (NSTC), Networking and Information Technology Research and Development Subcommittee, 'The national artificial intelligence research and development strategic plan' (October 2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf> accessed 8 December 2024

National Science and Technology Council (NSTC), Select Committee on Artificial Intelligence: Report 'National Artificial Intelligence Research and Development Strategic Plan 2023 Update' <www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf> accessed 9 December 2024.

National Security Commission on Artificial Intelligence, 'The Final Report' <https://reports.nscai.gov/final-report/> accessed 8 December 2024, 11.

President, Communication to Federal agencies 'Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence' (the White House, 24 October 2024) <www.whitehouse.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/> accessed 9 December 2024.

President, Executive Order No. 13859: Maintaining American Leadership in Artificial Intelligence, 11 February 2019, 84 FR 3967

President, Executive Order No. 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, 3 December 2020, 85 FR 78939

President, Executive Order No. 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 30 October 2023, 88 FR 75191

U.S. Department of Labor, 'Artificial Intelligence And Worker Well-being: Principles And Best Practices For Developers And Employers' <www.dol.gov/general/AI-Principles> accessed 9 December 2024

U.S. Department of State, Bureau of Cyberspace and Digital Policy, 'Risk Management Profile for Artificial Intelligence and Human Rights' (25 July 2024) <www.state.gov/risk-management-profile-for-ai-and-human-rights/> accessed 9 December 2024

US Government, 'Framework to Advance AI Governance and Risk Management in National Security <https://ai.gov/wp-content/uploads/2024/10/NSM-Framework-to-Advance-AI-Governance-and-Risk-Management-in-National-Security.pdf> accessed 9 December 2024.

White House Office of Science and Technology Policy (OSTP), 'The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People' (White House, October 2022) <www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf> accessed 1 December 2024

William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, H.R.6395, Public Law 116-283, Division E – National Artificial Intelligence Initiative Act of 2020.

China

Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars) of 2 May 2024 by the Center for Security and Emerging Technology (CSET) (Translation) <https://cset.georgetown.edu/publication/china-ai-law-draft/> accessed 11 December 2024

China Academy of Information and Communications Technology and JD Explore Academy, 'White Paper on Trustworthy Artificial Intelligence,' July 2021 <http://www.caict.ac.cn/english/research/whitepapers/202110/P020211014399666967457.pdf> accessed 6 December 2024

Cybersecurity Law of the People's Republic of China of 7 November 2016 (Chairman's Order No. 53).

Cyberspace Administration of China, 'Notice on Launching the Clear 2022 Special Action on the Comprehensive Governance of Algorithms' (China Internet Information Office, 8 April 2022) <www.cac.gov.cn/2022-04/08/c_1651028524542025.htm> accessed 9 December 2024

Cyberspace Administration of China, 'Provisions on the Security Assessment of Internet Information Services with Public Opinion Properties or Social Mobilization Capacity' (15 November 2018) <www.cac.gov.cn/2018-11/15/c_1123716072.htm> accessed 9 December 2024

Cyberspace Administration of China, 'Measures for Labelling of AI-Generated Synthetic Content (draft for solicitation of comments)' (14 September 2024) <www.cac.gov.cn/2024-09/14/c_1728000676244628.htm> accessed 9 December 2024

Cyberspace Administration of China, National Development and Reform Commission of the People's Republic of China, Ministry of Education of the People's Republic of China, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology of the People's Republic of China, Ministry of Public Security of the People's Republic of China, State Administration of Radio and Television, 'Interim Measures for the Management of Generative Artificial Intelligence Services' (10 July 2023) Document No. 15 [2023]

Cyberspace Administration of China, the Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China, the State Administration for Market Regulation, 'Provisions on the Management of Algorithmic Recommendations in Internet Information Services' (31 December 2021) Document No. 9 [2022]

Cyberspace Administration of China, the Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China, 'Provisions on the Administration of Deep Synthesis Internet Information Services' (25 November 2022) Document No. 12 [2022]

Data Security Law of the People's Republic of China of 10 June 2021 (Chairman's Order No. 84).

National Committee for the Governance of New Generation Artificial Intelligence, 'The Code of Ethics for the Next Generation of Artificial Intelligence' (25 September 2023) <www.most.gov.cn/kjbgz/202109/t20210926_177063.html> accessed 4 December 2024

National Cyber Security Standardization Technical Committee, 'TC260-003 "Basic Requirements for Security of Generative Artificial Intelligence Services" (29 February 2024) <www.tc260.org.cn/front/postDetail.html?id=20240301164054> accessed 9 December 2024

National Governance Committee for the New Generation Artificial Intelligence, 'Governance Principles for the New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence' (China Daily, 17 June 2019) <www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html> accessed 9 December 2024 (GPNGAI).

National Governance Committee for the New Generation Artificial Intelligence, 'Ethical Code for the New Generation of Artificial Intelligence' (26 September 2021) <http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html> accessed 9 December 2024

National Information Security Standardization Technical Committee, 'Cybersecurity Standard Practice Guide - Generative Artificial Intelligence Service Content Identification Method' (25 August 2023) Document No. 124 [2023]

People's Republic of China State Council, 'Internet Information Service Management Measures' (Decree No. 292 of 25 September 2000).

Personal Information Protection Law of the People's Republic of China of 20 August 2021 (Chairman's Order No. 91)

State Council of China, 'A Next Generation Artificial Intelligence Development Plan' (20 July 2017) <www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm> accessed 4 December 2024

State Internet Information Office, Central Propaganda Department, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Culture and Tourism, State Administration for Market Regulation, State Administration of Radio and Television, 'Guiding Opinions on Strengthening the Overall Governance of Internet Information Service Algorithms' (17 September 2021) Document No. 7 [2021]

State Internet Information Office, Ministry of Culture and Tourism, State Administration of Radio and Television, 'Provisions on the Management of Online A/V Information Services' (18 November 2019) Document No. 3 [2019].

The National Information Security Standardization Technical Committee (TC260), 'AI Safety Governance Framework' (September 2024) <www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf> accessed 9 December 2024

The State Council, 'A Next Generation Artificial Intelligence Development Plan' (20 July 2017) Document No. 35 [2017].

Self-governance

‘Pre-Deployment Evaluation of Anthropic’s Upgraded Claude 3.5 Sonnet: The U.S. AI Safety Institute and the UK AI Safety Institute conducted joint pre-deployment testing of Anthropic's latest model’ (NIST, 19 November 2024) <www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet> accessed 9 December 2024

‘Statement of AI Risk’ (Center for AI Safety) <www.safe.ai/work/statement-on-ai-risk#open-letter> accessed 6 December 2024

‘U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI’ (NIST, 29 August 2024) <www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research> accessed 9 December 2024

Future of Life Institute, ‘Asilomar AI Principles’ (11 August 2017) <https://futureoflife.org/open-letter/ai-principles/> accessed 5 December 2024

Future of Life Institute, ‘Pause Giant AI Experiments: An Open Letter’ (22 March 2023) <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> accessed 6 December 2024

OpenAI, ‘Best practices for deploying language models’ (2 June 2022) <https://openai.com/index/best-practices-for-deploying-language-models/> accessed 9 December 2024

OpenAI, ‘OpenAI Charter’ <https://openai.com/charter/> accessed 9 December 2024

1 INTRODUCTION

Any sufficiently advanced technology is indistinguishable from magic.

A. Clarke

The idea of artificial intelligence (AI) dates back to the 1940—1950s, with the works of McCulloch and Pitts on 'artificial neurons' and the famous work of Alan Turing 'Computing Machinery and Intelligence' of 1950, where he expressed the possibility of machine learning and proposed the widely known Turing Test.[1] For decades the idea of AI was mainly discussed in computer science and science fiction novels. The commercial application of expert systems (a computer system that emulates human decisions) in the 1980s boosted commercial interest in the technology, bringing new funding.[2] By 1985, the market of AI exceeded a billion dollars.[3] In 2012, the deep learning method began to dominate the industry and the new achievements (e.g., DeepMind developed AlphaGo that beat the world champion) increased the interest and funding.[4]

In 2020, OpenAI released GPT-3, a very successful large language model that can generate high-quality human-like texts, including poetry and computer code, digest information and keep dialogue with the user.[5] It gave vast commercial applications of the technology, boosted the research[6], and sky-rocketed investments that reached $100 billion a year. This signified the AI boom (or AI spring).[7] According to Statista, the global market size reached $184 billion in 2024 and will reach $826 billion by 2030.[8]

---

1 Stuart J. Russell, Peter Norvig, *Artificial intelligence: a modern approach* (4th edn, Pearson 2021) 17.

2 Ibid 23.

3 'Artificial Intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Artificial_intelligence#History> accessed 4 December 2024.

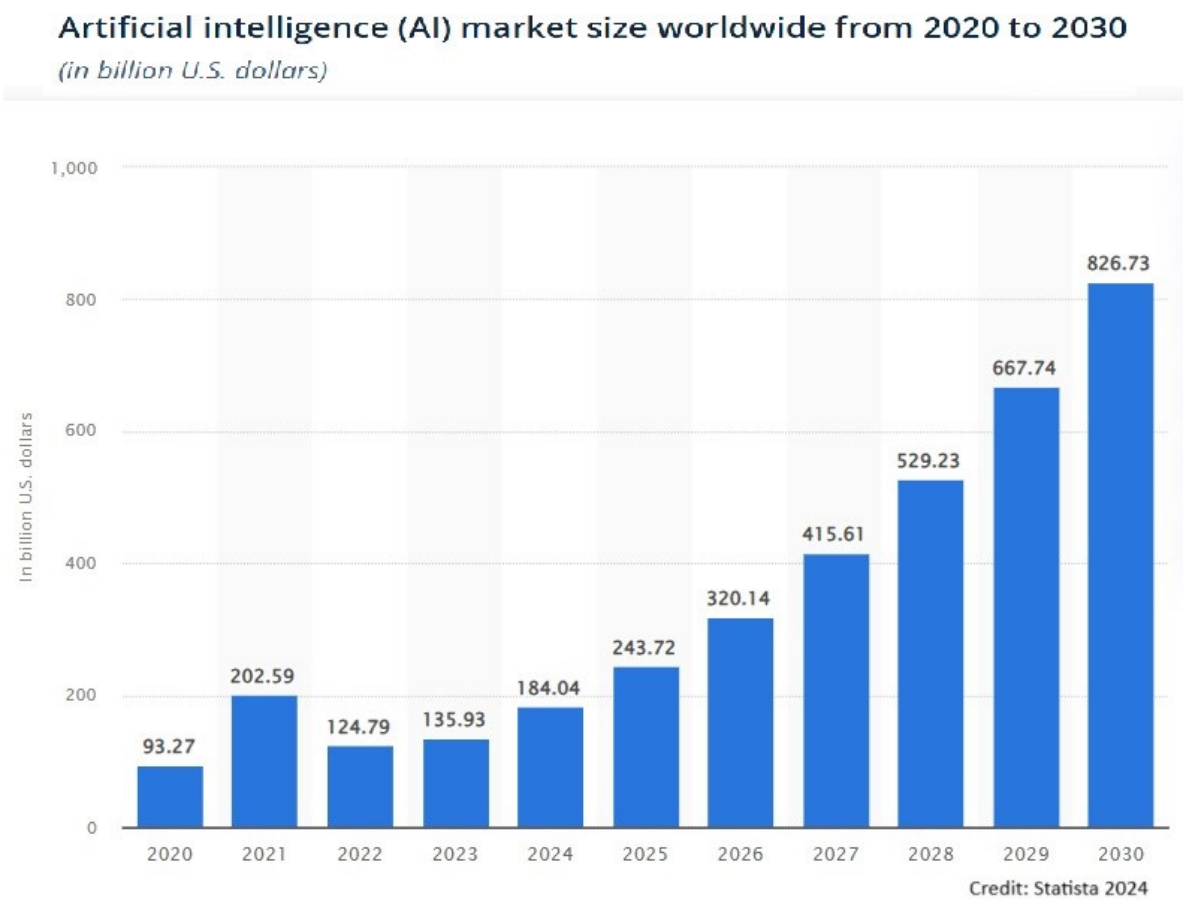
4 Jack Clark, 'Why 2015 Was a Breakthrough Year in Artificial Intelligence' (Bloomberg, 8 December 2015) <www.bloomberg.com/news/articles/2015-12-08/why-2015-was-a-breakthrough-year-in-artificial-intelligence> accessed 4 December 2024.

5 'GPT-3' (Wikipedia) <https://en.wikipedia.org/wiki/GPT-3> accessed 4 December 2024.

6 Lei Wang and others, 'A Survey on Large Language Model Based Autonomous Agents' (2024) 18 Frontiers of Computer Science 186345 <https://doi.org/10.1007/s11704-024-40231-1> accessed 4 December 2024.

7 'AI Boom' (Wikipedia) <https://en.wikipedia.org/wiki/AI_boom> accessed 4 December 2024.

8 Bergur Thormundsson, 'Artificial intelligence (AI) market size worldwide from 2020 to 2030' (Statista, 28 November 2024) <www.statista.com/forecasts/1474143/global-ai-market-size> accessed 4 December 2024.



The next breakthrough is expected with the development of artificial general intelligence (AGI) - a type of AI that matches or surpasses human cognitive capabilities across a wide range of cognitive tasks[9] – it is also called 'strong AI'. The development of AGI is nowadays a primary priority of OpenAI and Meta along with 70+ other projects.[10] The timeline is unclear, however, and the assessments range from years to decades.

The technology of artificial intelligence (AI) is now ubiquitous and on the news. Each day software companies boast of reaching yet another new level of mastery. AI is used in large-scale modelling (like the Earth 2.0 by Nvidia), data analysis, design, and process optimizations. AI is even used to improve computer code to make it run faster![11] For ordinary users, a breakthrough was the development of successful generative AI[12], including large

---

9 'Artificial general intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Artificial_general_intelligence> accessed 4 December 2024.

10 Ibid.

11 Jensen Huang, Nvidia CEO, 'NVIDIA Unveils "NIMS" Digital Humans, Robots, Earth 2.0, and AI Factories' (YouTube, 5 June 2024) <www.youtube.com/watch?v=IurALhiB6Ko&t=1637s> accessed 7 July 2024.

12 'Generative artificial intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Generative_artificial_intelligence> accessed 7 July 2024.

language models (LLMs), like ChatGPT, which made interaction with AI as natural as dialogue with a human operator and, sometimes, indistinguishable from human. Software companies embed AI into their applications making the use of AI seamless and indispensable for users. The technology is so robust, promising, and helpful that it seems that there is no turning back. Yet it is still an open question if AI is a Pandora's box.

The regulation of AI started to gain momentum in the mid-2010s when the states realized the opportunities and risks that AI technology brings. Many states, including the EU, the US, and China formulated their national strategies and aimed to gain leadership in the area. According to the China AI strategy, 'AI has become a new focus of international competition. AI is a strategic technology that will lead in the future; the world's major developed countries are taking the development of AI as a major strategy to enhance national competitiveness and protect national security…'[13]

At about the same time, the researchers started actively exploring AI safety (including the alignment problem and the issues of fairness and misuse).[14] With time, it evolved into a complex multidisciplinary concept stretching across computer science, human ethics, philosophy, and legal policy. Nowadays, states play a significant role in this process. In 2023, the UK hosted the AI Safety Summit, which was attended by 28 countries, including the US, China and the European Union. The talking points included the regulation of 'frontier AI' (the latest and most powerful AI systems), existential risk, and the use of AI for terrorism, criminal activity and warfare.[15] The resulting document was the Bletchley Declaration calling for the safe, human-centric, trustworthy and responsible development of AI.[16]

The Bletchley Declaration reflects the modern challenges of AI regulatory policy that resonate with the subject of this research: human rights, personal data, and safety.

---

13 The State Council, 'A Next Generation Artificial Intelligence Development Plan' (20 July 2017) Document No. 35 [2017].

14 Brian Christian, *The Alignment Problem: Machine learning and human values* (W. W. Norton & Company, 2020) 67, 73.

15 'AI Safety Summit' (Wikipedia) < https://en.wikipedia.org/wiki/AI_Safety_Summit> accessed 4 December 2024.

16 Policy Paper: the Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023 (Gov.UK, 1 November 2023) <www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023> accessed 4 December 2024.

The research will cover the regulation of the European Union, the US, and China: each of these jurisdictions aims to lead the AI development and regulation process.[17]

The EU was among the pioneers to introduce binding comprehensive regulation of AI. A relevant regulation, the AI Act, was adopted on 12 July 2024 and came into force on 1 August 2024 (with certain provisions coming into force later).[18] An extensive 144-page document defines, in a quite detailed manner, the framework for the horizontal regulation of AI in the European Union.

Although the US binding AI-specific regulation extends only to the use of AI by state agencies, the US is very active in their AI policy agenda. Both the Trump and Biden administrations set AI as a national priority and were productive in building policy documents. The US aims to lead the global process of developing AI-related standards and focuses much on that. There are also soft law instruments that apply horizontally to the AI industry, a notable one is the Blueprint for an AI Bill of Rights published by the White House in October 2022.[19] A 70-page non-binding white paper formulated the national approach to human rights in the AI context. It is 'an exercise in envisioning a future the American public is protected from the potential harms, and can fully enjoy the benefits, of automated systems'.[20] The AI Bill of Rights discusses AI-associated concerns such as the safety and efficiency of the systems, protection from algorithmic discrimination, data privacy, awareness of the use of AI, the existence of human alternatives, and the right to opt-out.

The third country of interest is China. In addition to being a rapidly developing technological giant, it is one of the first countries to introduce binding AI regulation. The views to regulate AI were expressed in 2017, while the AI-specific binding acts were introduced during 2021-

---

17 Executive Order No. 13859: Maintaining American Leadership in Artificial Intelligence, 11 February 2019, 84 FR 3967 (EO 13859); A Next Generation Artificial Intelligence Development Plan (n 13), s II(3) (China); Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Coordinated Plan on Artificial Intelligence' COM/2018/795 final, s 2.6.

18 European Parliament and of the Council Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) [2024] OJ L2024/1689 (AI Act).

19 White House Office of Science and Technology Policy, 'The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People' (White House, October 2022) <www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf> accessed 1 December 2024 (AI Bill of Rights).

20 'Relationship to Existing Law and Policy' (White House) <www.whitehouse.gov/ostp/ai-bill-of-rights/relationship-to-existing-law-and-policy/> accessed 7 July 2024.

2023, including the Algorithmic Recommendations Provisions (2021)[21], Deep Synthesis Provisions (2022)[22], and Generative AI Measures (2023)[23]. These acts regulate specific aspects of the use of AI technology. In 2023, China published the draft Trial Measures for Ethical Review of Science and Technology Activities to address ethical aspects.[24] In 2024, China announced its intention to adopt a comprehensive AI Law[25].

This thesis seeks to explore how the countries envision their regulatory policy, what are their priorities, and how they see and approach the AI risks. The research work also aims to highlight the distinguishing features of the compared jurisdictions.

The analysis will be focused on three areas: human rights, personal data, and safety. The research topic is especially relevant now, as each of the compared jurisdictions has recently adopted AI-specific legislation, a significant part of which was released in a few recent months.

The next chapter 2 will define the parameters and context of research: scope, the state of research, the research question, and methods. Chapter 3 contains the analysis, including an overview of the key relevant concepts, regulations in the compared jurisdictions and their distinctive features, ending with a brief overview of international regulatory initiatives, to

---

21 Cyberspace Administration of China, the Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China, the State Administration for Market Regulation, 'Provisions on the Management of Algorithmic Recommendations in Internet Information Services' (31 December 2021) Document No. 9 [2022] (Algorithmic Recommendations Provisions).

22 Cyberspace Administration of China, the Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China, 'Provisions on the Administration of Deep Synthesis Internet Information Services' (25 November 2022) Document No. 12 [2022] (Deep Synthesis Provisions).

23 Cyberspace Administration of China, National Development and Reform Commission of the People's Republic of China, Ministry of Education of the People's Republic of China, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology of the People's Republic of China, Ministry of Public Security of the People's Republic of China, State Administration of Radio and Television, 'Interim Measures for the Management of Generative Artificial Intelligence Services' (10 July 2023) Document No. 15 [2023] (Generative AI measures).

24 National Committee for the Governance of New Generation Artificial Intelligence, 'The Code of Ethics for the Next Generation of Artificial Intelligence' (25 September 2023) <www.most.gov.cn/kjbgz/202109/t20210926_177063.html> accessed 4 December 2024 (Chinese) (AI Code of Ethics); Giulia Interesse, 'Technology in China: Draft Trial Measures' (China Briefing, 11 May 2023) <www.china-briefing.com/news/china-ethical-review-of-science-and-technology-draft-trial-measures/> accessed 4 December 2024.

25 Zeyi Yang, 'Four things to know about China's new AI rules in 2024' (MIT Technology Review, 17 January 2024) <www.technologyreview.com/2024/01/17/1086704/china-ai-regulation-changes-2024/> accessed 3 December 2024; Artificial Intelligence Law of the People's Republic of China (Draft for Suggestions from Scholars) of 2 May 2024 by the Center for Security and Emerging Technology (CSET) (Translation) <https://cset.georgetown.edu/publication/china-ai-law-draft/> accessed 11 December 2024 (Draft AI law).

complete the picture. Chapter 4 contains the conclusions of the research. The summarised comparison analysis is reflected in the table in Appendix 1.

2 PARAMETERS AND CONTEXT OF RESEARCH

2.1 Scope of research

The regulation of AI is a huge and rapidly developing sphere. It is quite easy to be overwhelmed. Therefore, it is of critical importance to define the scope of research in detail.

2.1.1 Areas of research

This work will cover the three areas of AI regulation: the protection of human rights, the protection of personal data, and safety. The notion of these concepts in the AI environment is itself an interesting subject and will be explored in section 3.2. The focus of the research will be on AI-specific provisions.

2.1.2 Compared jurisdictions

Three jurisdictions will be compared: the EU, the US, and China. The choice seems quite straightforward – the US and China are the top two investors in the technology[26] and each has announced the target to become a global centre for AI development[27]. The EU, in its turn, led the human-centric approach to the development of AI and was the first jurisdiction to introduce comprehensive AI regulation.

2.1.3 Excluded areas

To frame the research, certain aspects are expressly excluded from the scope:

- The regulation of AI in the context of military applications, intelligence, and national security;
- Technical regulation of AI. Although the AI-related standards, both global and national, are discussed and being developed, this is a vast subject that deserves standalone research. Despite some national standards being mentioned and

---

[26] '7 Key AI Investment Statistics Every Investor Should Know' (Edge Delta, 24 May 2024) <https://edgedelta.com/company/blog/ai-investment-statistics> accessed 5 December 2024.

[27] See n 17; Executive Order No. 14110: On the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 30 October 2023, 88 FR 75191 (EO 14110), ss 2(h), 11.

discussed in this master thesis, this should not be viewed as a full-scope analysis of the area;

- Regulation of intellectual property rights;
- Trade-related aspects and regulation: customs, export control and trade agreements;
- Analysis of the national legislation of the EU member states; and
- The aspect of liability.

2.1.4 AI regulation vs. other regulation

AI regulation is always embedded and forms part of the general regulatory framework. For instance, a service that broadcasts news online and applies AI technology is simultaneously subject to mass media, content control, general personal data protection, general IT services, and specific 'deepfakes' regulation. To maintain the focus, the analysis will be targeted at the dedicated AI regulation. Even so, the general regulatory framework relevant to AI will be discussed on a high-level basis.

2.1.5 International regulation

There are more and more initiatives to coordinate the regulation of AI on the global level. This thesis will overview key relevant documents and initiatives relevant to the compared jurisdictions. Still, the focus will be on the national and union, as regards the EU, legislation.

2.2 The state of research

Although AI is a vastly explored subject, the analysis of available literature has not revealed a similar topic of research. There is currently ample legal literature, covering different aspects of AI. These include intellectual property[28], the use of AI in healthcare,[29] personal data[30],

---

28 Peter Georg Picht and Florent Thouvenin, 'AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property' (2023) 54 IIC International Review of Intellectual Property and Competition Law.

29 Sara Gerke, Timo Minssen and Glenn Cohen, *Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare*, *Artificial Intelligence in Healthcare* (Elsevier Inc. 2020); Prof. I Glenn Cohen and others, 'The European Artificial Intelligence Strategy: Implications and Challenges for Digital Health' (2020) 2 The Lancet Digital Health 7.

30 Nikolaus Marsch, *Artificial Intelligence and the Fundamental Right to Data Protection: Opening the Door for Technological Innovation and Innovative Protection*, *Regulating Artificial Intelligence* (Springer Cham 2019); Rafael Brown, Jon Truby and Imad Antoine Ibrahim, 'Mending Lacunas in the EU's GDPR and Proposed Artificial Intelligence Regulation' (2022) 9 European Studies: The Review of European Law, Economics and Politics.

ethics and human rights[31], and even religion[32], to name a few. The span of views varies from apocalyptic[33] to pure optimistic.[34]

The existing comparative analyses of AI regulations mostly focused on the aspects of privacy of data.[35] A broader analysis is available in the book by Bonnie Buchanan[36], but it was published in 2021 and is somewhat outdated given the current pace of regulatory development.

Due to the novelty of the relevant regulations, there is often scarce analysis of them in the legal literature. Therefore, the work will also rely on press releases of state bodies and overviews provided by law firms and AI practitioners. The work will also refer to publications in mass media.

For the technical part of AI, Wikipedia serves as a useful source of data. Since the technical information on AI (e.g., AI safety in computer science) is given in this work mostly for background and context purposes, the use of Wikipedia for these purposes seems appropriate.

---

31 Krzysztof Wach and others, 'The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT' (2023) 11 Entrepreneurial Business and Economics Review; Frank Vram Zerunyan, 'Techno Innovations: The Role of Ethical Standards, Law and Regulation, and the Public Interest' in Alie E. Abbas (ed), *Next-Generation Ethics: Engineering a Better Society* (Cambridge University Press 2019); Matthew R O'Shaughnessy and others, 'What Governs Attitudes toward Artificial Intelligence Adoption and Governance?' (2023) 50 Science and Public Policy.

32 Brian J Grim and Roger Finke, *The Price of Freedom Denied* (Cambridge University Press 2010); Heidi Campbell and Pauline Hope Cheong, *Thinking Tools for AI, Religion & Culture* (Network for New Media, Religion & Digital Culture Studies 2023).

33 Robert M Geraci, 'Apocalyptic AI: Religion and the Promise of Artificial Intelligence' (2008) 76 Journal of the American Academy of Religion.

34 'Ai Privacy: AI and Privacy: The Privacy Concerns Surrounding AI, Its Potential Impact on Personal Data - The Economic Times' (The Economic Times News, 25 April 2023) <https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms> accessed 11 December 2024; Stanton Heister and Kristi Yuthas, 'How Blockchain and AI Enable Personal Data Privacy and Support Cybersecurity' in Tiago M. Fernández-Caramés, Paula Fraga-Lamas (eds), *Advances in the Convergence of Blockchain and Artificial Intelligence* (IntechOpen 2022); Ilaria Querci and others, 'Explaining How Algorithms Work Reduces Consumers' Concerns Regarding the Collection of Personal Data and Promotes AI Technology Adoption' (2022) 39 Psychology and Marketing.

35 Yunfei Xing and others, 'AI Privacy Opinions between US and Chinese People' (2023) 63 Journal of Computer Information Systems; V Gabor Radi, 'Comparative Analysis of the AI Regulation of the EU, US and China from a Privacy Perspective' (2023) 46th ICT and Electronics Convention, MIPRO 2023 - Proceedings.

36 Bonnie Buchanan, 'AI: A Cross Country Analysis of China versus the West' in Susane Chishti, Ivana Bartoletti et al (eds), *The AI Book: The Artificial Intelligence Handbook for Investors, Entrepreneurs and FinTech Visionaries* (Wiley 2021).

2.3 Research task

The main research task is to compare the AI policy and regulation of the compared jurisdictions in the spheres of protection of human rights, protection of personal data, and safety and to discuss their differentiating features.

2.4 Method

The main method used is the desk research.

The analysis of the EU regulation, primarily the AI Act, will play a central role in the research; the AI regulation of the US and China will serve as comparison material.

The scope of the analysis includes relevant policy documents, regulatory acts (including draft law), white papers and other non-binding acts, documents of international organizations, legal literature, and official statements. Due to the novelty of the AI regulations, it is expected that no or scarce implementation practice exists. Some statistical data will be provided for the sake of completeness.

For the regulation of China, the analysis will be based on available English translations available in online depositories, e.g., www.chinalawtranslate.com. They are peer-reviewed and commented on by legal practitioners and seem to be a trustworthy source. As a verification method, I used online translation tools to translate original acts in Chinese published on the official websites.

3 ANALYSIS

3.1 General discussion

3.1.1 What is AI?

Speaking broadly, artificial intelligence is intelligence exhibited by machines[37]. Intelligence is a broad concept that defines such features as the capacity for abstraction, logic, understanding, self-awareness, learning, emotional knowledge, reasoning, planning, creativity, critical thinking, and problem-solving.[38] Concerning AI, intelligence is exhibited via intelligent agents, i.e., the agents that perceive their environment, take actions autonomously to achieve goals, and may improve their performance by learning or acquiring

---

37 Artificial Intelligence (n 3).

38 'Intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Intelligence> accessed 5 December 2024.

knowledge.[39] Therefore, the modern concept does not require AI to have a 'brain', personality, or conscience. Although the definition of AI applies to a wide range of use cases, is a critique of overusing the term.[40]

Nowadays the technology is ubiquitous and used in search engines (Google Search, recommendation systems (Netflix, YouTube), virtual assistance (Siri), autonomous vehicles, automatic language translation (Microsoft Translator), facial recognition (Apple FaceID)[41], and chatbots (ChatGPT), just to name a few. It is widely used in such industries as healthcare, games, mathematics, physics, finance, military, and programming. The application of generative AI covers almost every aspect of life.

This section is not intended to discuss the technicalities of AI work. Some relevant features will be discussed in the 'Safety' section 3.2.3 below.

3.1.2 The definition of AI

To regulate something, you need first define the subject of regulation. It proved to be a difficult task for lawmakers.[42] Define it too broadly, and it will include unintended software, make it too narrow, and the regulation will only apply to advanced AI leaving most AI applications not regulated. The 'AI effect' – a phenomenon where either the definition of AI or the concept of intelligence is adjusted to exclude capabilities that AI systems have already mastered[43] – adds to the difficulty.

Google defines AI as 'a field of science concerned with building computers and machines that can reason, learn, and act in such a way that would normally require human intelligence or that involves data whose scale exceeds what humans can analyze.' The definition further clarifies that on the operation level for business use AI is 'a set of technologies that are based primarily on machine learning and deep learning, used for data analytics, predictions and forecasting, object categorization, natural language processing, recommendations, intelligent

---

39 'Intelligent agent' (Wikipedia) <https://en.wikipedia.org/wiki/Intelligent_agent> accessed 5 December 2024.

40 Jason Aten, 'Stop Calling Everything A.I. It's Just Computers Doing Computer Stuff' (Inc.com, 13 May 203) <www.inc.com/jason-aten/stop-calling-everything-ai-its-just-computers-doing-computer-stuff.html> accessed 5 December 2024.

41 Artificial intelligence (n 3).

42 Magdalena Owczarczuk, 'Ethical and regulatory challenges amid artificial intelligence development: an outline of the issue' (2023) 22 Economics and Law 2, 299; Haroon Sheikh, Corien Prins, Erik Schrijvers, 'Artificial Intelligence: Definition and Background' in *Mission AI* (Springer Cham 2023).

43 'AI effect' (Wikipedia) <https://en.wikipedia.org/wiki/AI_effect> accessed 5 December 2024.

data retrieval, and more'.[44] This 'technical' vision of AI seems not suitable to the regulatory policy.

Probably the most thorough study on the legal definition of AI has been conducted by the EU. In June 2018, the Commission appointed the High-Level Expert Group on Artificial Intelligence (AI HLEG) with one of the main tasks to define AI. In April 2019, the AI HLEG released the document 'A definition of AI: Main capabilities and scientific disciplines'[45] that proposed an extensive definition of AI capturing many specific aspects of the technology:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).[46]

Despite the lengthy definition, the document acknowledges it to be 'a very crude oversimplification of the state of the art.'[47]

The proposed definition has not been used for the AI Act. Instead, the AI Act used a definition aligned with the work of international organisations.[48] The final definition was as follows:

---

44 'What is Artificial Intelligence (AI)' (Google Cloud) <https://cloud.google.com/learn/what-is-artificial-intelligence> accessed 5 December 2024.

45 'A definition of AI: Main capabilities and scientific disciplines.' A report/study by the High-Level Expert Group on Artificial Intelligence (AI HLEG) of 8 April 2019 (European Commission, 18 December 2018) <https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines> accessed 5 December 2024.

46 Ibid 6.

47 Ibid preface.

48 AI Act (n 18), recitals 12.

> ‘AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[49]

The US statutorily defines AI in a similar way:

> The term ‘artificial intelligence’ means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to — (A) perceive real and virtual environments; (B) abstract such perceptions into models through analysis in an automated manner; and (C) use model inference to formulate options for information or action.[50]

So far, China does not have a legal definition of AI. The Draft AI law that proposes comprehensive regulation of AI suggests the following language: “Artificial intelligence” (AI) means technology that utilizes computers to simulate human intelligent behaviour for use in prediction, recommendation, decision-making, or content generation, etc., for specialized or general purposes’.[51]

It appears that the definitions in the compared jurisdictions share the features that AI is machine-based and exhibits intelligent agency (can infer). All the definitions mention similar key use cases of AI. The EU highlights the capacity of AI to influence physical environments, while the Chinese definition mentions the existence of specialised and general-purpose AI. These features do seem to make much difference in practice as the regulation of AI in each of the compared jurisdictions is likely to cover all these aspects (e.g., the provisions on AI will apply to autonomous vehicles in the US and China, despite the ability to influence physical environments is not expressly mentioned).

The legal definitions seem quite broad. As we will see later, this is balanced by categorizing AI systems and adjusting the regulatory requirements to the risk category of AI.

---

49 Ibid 3(1).

50 National Artificial Intelligence Initiative Act of 2020, 15 U.S.C. § 9401 (2022).

51 Draft AI law (n 25), art 94(i).

3.1.3 The need for AI-specific regulation

Gleb Papyshev and Masaru Yarime suggest that the states' governing policies on AI have three main objectives: development, control, and promotion.[52] The states are balancing between these objectives trying to control unsafe practices, yet encouraging and supporting the development. This is especially relevant in the current situation of the AI race.

In 2017, the CEO of Intel opined that AI is in its infancy and that it is too early to regulate the technology.[53] The idea was supported by Chris Reed in 2018 who argued that it is too early to regulate the technology as many emerging products will not be able to meet the requirements and 'masterly inactivity in regulation is likely to achieve a better long-term solution than a rush to regulate in ignorance'.[54] There were many voices among the market and scholars that called for regulation to ensure sustainable development of the technology, including addressing a 'fundamental risk to the existence of human civilization'.[55] One of the most vigorous and active promoters of that has been Elon Musk.

The public support of these ideas is evidenced by 5700+ signatures under the Asilomar AI Principles published in 2017 by the Future of Life Institute; among the signatories are the founders of big tech AI companies and key AI experts.[56]

The current phase of AI development is, in many aspects, safety-driven.[57] As will be discussed further, AI poses many specific risks that can hardly be tackled by 'traditional' legislation and AI safety is not on the rise. Nowadays we are witnessing more and more countries adopting soft law and binding AI regulation, which evidence the shared

---

52 Gleb Papyshev, Masaru Yarime, 'The state's role in governing artificial intelligence: development, control, and promotion through national strategies' (2023) 6:1 Policy Design and Practice, 79.

53 Arjun Kharpal, 'A.I. is in its "infancy" and it's too early to regulate it, Intel CEO Brian Krzanich says' (CNBC, 7 November 2017) <www.cnbc.com/2017/11/07/ai-infancy-and-too-early-to-regulate-intel-ceo-brian-krzanich-says.html> accessed 5 December 2024.

54 Chris Reed, 'How should we regulate artificial intelligence?' (2018) 376 Philos Trans A Math Phys Eng Sci. <https://doi.org/10.1098/rsta.2017.0360> accessed 5 December 2024.

55 Keith B. Belton, 'How should AI be regulated' (IndustryWeek, 7 March 2019) <www.industryweek.com/technology-and-iiot/article/22027274/how-should-ai-be-regulated> accessed 5 December 2024; Camila Domonoske, 'Elon Musk Warns Governors: Artificial Intelligence Poses "Existential Risk" (NPR, 17 July 2017) <www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk> accessed 5 December 2024.

56 'Asilomar AI Principles' (Future of Life Institute, 11 August 2017) <https://futureoflife.org/open-letter/ai-principles/> accessed 5 December 2024.

57 Commission, 'AI Act' (European Commission, 14 October 2024) <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-a> accessed 5 December 2024.

understanding of the necessity of the AI-specific legislation.[58] As stated by the Commission, 'overall, there is a general agreement amongst stakeholders on a need for action.'[59]

3.2 Key concepts

This section is aimed to overview the concepts of human rights, personal data and safety through the 'AI lens'.

3.2.1 Human rights

There is a common understanding that AI may encroach on human rights.[60] And it comes as no surprise, therefore, that many jurisdictions declare the imperative that AI should be a human-centric technology.[61]

At first glance, it seems that the corpus of human rights has already been defined and sufficient. This includes the core sets of human rights formulated by the UN Convention on Human Rights[62] and the Charter of Fundamental Rights of the European Union[63]. So, the task of the regulator is just to set relevant measures and enforcement tools to ensure their protection. To a certain extent, this may work, since 'classic' human rights indeed address many of the AI risks. The relevant rights include the right to non-discrimination, the right to effective remedy by the competent national tribunal, the right to privacy, honour, freedom of thought, freedom of opinion and expression, the right to equal pay for equal work,[64] and also personal data protection rights, the rights of the child, elderly and persons with disabilities, environment protection, and consumer protection[65]. At the same time, as regards the AI capacity for misinformation, deep-fakes, and decision-manipulating, the 'classical' rights appear insufficient to adequately address the threats. This naturally requires revisiting the scope of human rights and tailoring them to the AI context.

---

58 'Regulation of artificial intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Regulation_of_artificial_intelligence> accessed 5 December 2024.

59 Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence act) and amending certain Union legislative acts' COM/2021/206 final, s 3.1.

60 See, e.g., AI Act (n 18), recitals 5; Commission, 'White Paper On Artificial Intelligence - A European approach to excellence and trust' COM(2020) 65 final, s 5.

61 See, e.g., AI Act (n 18), recitals 6.

62 Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR).

63 Charter of Fundamental Rights of the European Union 2000 (Charter of Fundamental Rights).

64 UDHR (n 62), arts 7, 8, 12, 18, 19, 23.

65 Charter of Fundamental Rights (n 63), arts 8, 24-26, 37, 38.

Such novel areas include, *inter alia,* human autonomy (AI may distort human decisions, including on the subliminal level), manipulation of data (e.g., deepfakes), exposure (generating of) to inappropriate content, social scoring, behaviour-prediction practices, and AI autonomous decision-making.

How the compared jurisdictions address these specific threats will be explored in section 3.3.

3.2.2 Personal data

Data is an indispensable, core element of AI technology. Without it, AI would remain a mathematical abstraction. There are ample AI use cases that are trained on personal data or process personal data during the operation of the AI system.

Considering the amount of data used and the capabilities of the technology (including data de-anonymisation and deep synthesis of a person's voice, face, etc.), the misuse of personal data by AI may lead to severe consequences unthinkable in the non-AI environment. Deepfake content may have various repercussions from being an innocent joke to financial fraud or devastating public statements.

Do the compared jurisdictions consider any special means of protection from these threats? The analysis of this topic is contained in section 3.3.

3.2.3 Safety

As machines learn they may develop unforeseen strategies at rates that baffle their programmers

Norbert Wiener[66]

Safety seems to be the cornerstone of this research. Even if a regulator formulates well human rights and personal data rights, they all remain declarations unless there is an effective safety mechanism to ensure them. This mechanism could be driven by corporate self-regulation, encouraged by soft law, or be binding.

Among the three discussed areas, safety is the most complex and hard to regulate. It pierces through the whole AI regulation and, to a great extent, defines the regulatory policy.

---

66 Norbert Wiener, 'Some Moral and Technical Consequences of Automation' (Science, 6 May 1960) <www.science.org/doi/10.1126/science.131.3410.1355> accessed 11 December 2024.

(a) What is AI safety?

IBM defines AI safety as 'practices and principles that help ensure AI technologies are designed and used in a way that benefits humanity and minimizes any potential harm or negative outcomes'.[67]

AI safety is a multidisciplinary concept, including the areas of computer science, psychology, sociology, and law. Concerning AI, safety could mean various things: safety standards for autonomous vehicles, labelling the AI-generated content, the requirements for dataset mapping and training algorithms, cybersecurity measures, or just having the user aware of interacting with AI. The listed aspects are just examples and lack structure. To help with that, it is worth referring to how AI safety is viewed in computer science.

On the design level, AI safety has two major components: AI ethics (or machine ethics) and AI alignment.[68] The first is basically how we want AI to behave – human rights and the protection of personal data are part of ethics. The second component deals with how 'to steer AI systems toward a person's or group's intended goals, preferences, and technical principles'[69] or, to put it simply, how to make the AI systems behave as we want. Since AI is morally neutral – it can be good or bad depending on how it is used[70] – countries pay attention to defining AI ethics.[71]

On the scale of the AI system lifecycle, AI safety could be broken down into the following elements (although there is no single widely accepted list):[72]

- Ethics & alignment. This addresses such problems as biases, specification gaming (AI systems can find loopholes that help them accomplish the specified objective efficiently but in unintended, possibly harmful ways), 'honest AI' (measures to prevent misleading information or hallucinations), emergent goals (the ensuring of the safety of the goals or subgoals that AI would independently formulate and

---

67 Amanda McGrath, Alexandra Jonker, 'What is AI safety' (IBM, 15 November 2024) <www.ibm.com/think/topics/ai-safety> accessed 6 December 2024.

68 'AI safety' (Wikipedia) <https://en.wikipedia.org/wiki/AI_safety> accessed 6 December 2024.

69 'AI alignment' (Wikipedia) <https://en.wikipedia.org/wiki/AI_alignment> accessed 6 December 2024.

70 Mike Trojecki, 'Is Artificial Intelligence Good or Bad: Debating the Ethics of AI' (IIOT World, 10 June 2019) <www.iiot-world.com/artificial-intelligence-ml/artificial-intelligence/the-good-the-bad-and-the-ugly-debating-the-ethics-of-ai/> accessed 5 December 2024.

71 See e.g., 'AI Ethics Guidelines Global Inventory' (AlgorithmWatch) <https://inventory.algorithmwatch.org/> accessed 6 December 2024.

72 McGrath (n 67); AI safety (n 68).

pursue), and embedded agency (a set of problems that may arise in real-world implementation of a theoretical model);

- Robustness – the ability of the system to withstand errors and adversarial attacks on machine learning algorithms;

- Explainability (also called transparency). Many AI models, especially large language models (LLMs), are 'black boxes' with opaque decision-making process; explainable AI are techniques to understand how AI arrived at the decision[73];

- Human oversight – the ability of human monitoring and intervention in the AI systems. Sometimes AI system is restricted from making the final decision, for instance, in critical areas such as healthcare, justice, and law enforcement.

- Security protocols – access control and anomaly detection.

A more extensive vision of the concept includes the problems of scalable oversight (the problem of the physical inability to effectively oversight AI systems by humans as the system scales), power-seeking AI (the risk that AI may choose an aggressive strategy), and the capability control (the control of the capabilities and the locality of the AI system, including interruptibility, switching-off, blinding and boxing (limiting its possible physical and social engineering ways to escape). The above elements are mostly discussed regarding AGI or the most powerful models that are feared to pose existential risks.[74]

Safety and security are usually distinguished. While safety addresses the inherent issues of the AI system, security relates to protection from external threats, e.g., cyberattacks.[75]

This thesis analyses the regulation of safety in its broad sense, including the aspects of security. As discussed above, the concept of AI safety goes far beyond mere product safety and will be analysed accordingly.

---

73 McGrath (n 67), Explainable AI.

74 AI safety (n 68).

75 McGrath (n 67), AI safety versus AI security.

(b) AI threats

For the sake of context, a quick overview of the AI-specific risks is feasible. On the basic level, the threats seem to come from:

- opacity pertained to the technology (also called the 'black box' or 'grey box' effect);
- intended or unintended misalignment of the system – for instance, this may have the forms of abusive practices (manipulation), misinformation, or mass processing of personal data without users' consent (e.g., generation of deep synthesis content); and
- large-scale impact of technology on humankind – technological unemployment, ecology, national and global security.

These threats entail AI-specific risks and concerns. For instance, the opacity and the capability of the technology for autonomy create demand in making users aware of the interaction with AI, having the option to opt out for a human alternative, calling for human intervention, or even restricting AI autonomous decisions in certain areas. It is apparent that people, at least for now, are not comfortable being one-on-one with the technology.

Considering the state of development, jurisdictions prohibit certain AI use cases as posing unacceptable risks; some other practices are marked as high-risk and subject to extended safety measures.

Among the types of AI risks, IBM selects bias and fairness, privacy, loss of control over the AI system, malicious misuse, cybersecurity, and existential risks.

A special category is the existential risk – the idea that substantial progress in AGI could lead to human extinction or an irreversible global catastrophe.[76] Nowadays this is not a mere topic of science fiction novels – more and more researchers and practitioners draw attention to this problem. Among the adverse scenarios discussed are social manipulation, cyberattacks, data manipulation, weaponization, bioterrorism, and even AI-enabled stable dictatorship.[77]

In 2023, the Center for AI Safety published the Statement of AI Risks: 'Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as

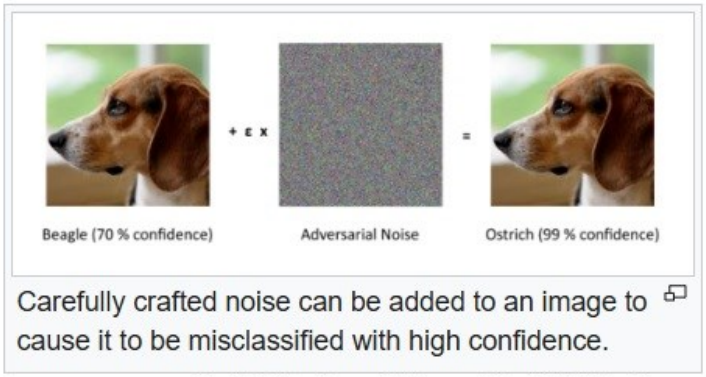
---

[76] 'Existential risk from artificial intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Existential_risk_from_artificial_intelligence> accessed 6 December 2024.

[77] Ibid.

pandemics and nuclear war.'[78] The statement was signed by many leading experts and CEOs of big tech AI companies, like OpenAI and DeepMind. In March 2023, the Future of Life Institute published an open letter calling to pause the training of AI systems more powerful than GPT-4 for at least 6 months, 'Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.'[79] More than 33 thousand experts, big tech founders and CEOs, and practitioners signed the initiative. The 2023 AI Safety Summit was the global reaction of the governments to these threats. In July 2023, the UN Secretary-General expressed the need for a new international body to help govern AI 'as the technology increasingly reveals its potential risks and benefits.'[80] These are just the examples and the topic gains momentum.

Despite the safety efforts, AI systems are often vulnerable to adversarial attacks[81], such as data poisoning. The adversarial data may be hardly noticeable by humans but lead to erroneous outputs.



Credit: Joshuaclymer, 2022, copyright, CC BY-SA 4.0 https://en.wikipedia.org/wiki/AI_safety#/media/File:Illustration_of_imperceptible_adversarial_pertubation.png

(c) Trustworthy AI

The regulation of AI is considered still in its infancy[82] but is a rapidly developing area with many acts adopted in the recent few months. The regulation of AI safety is tightly related to the concept of trustworthy AI, an approach to AI development that prioritizes safety and transparency for the people who interact with it.[83]

---

78 'Statement of AI Risk' (Center for AI Safety) <www.safe.ai/work/statement-on-ai-risk#open-letter> accessed 6 December 2024.

79 'Pause Giant AI Experiments: An Open Letter' (Future of Life Institute, 22 March 2023) <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> accessed 6 December 2024.

80 Brian Fung, 'UN Secretary General embraces calls for a new UN agency on AI in the face of "potentially catastrophic and existential risks" (CNN Business, 18 July 2023) <https://edition.cnn.com/2023/07/18/tech/un-ai-agency/index.html> accessed 6 December 2024.

81 AI Safety (n 68), Adversarial robustness.

82 Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions Fostering a European approach to Artificial Intelligence' COM/2021/205 final, s 3.

83 Nikki Pope, 'What Is Trustworthy AI?' (Nvidia, 1 March 2024) <https://blogs.nvidia.com/blog/what-is-trustworthy-ai/> accessed 6 December 2024.

The Ethics Guidelines for Trustworthy AI developed by AI HLEG[84] define three components of trustworthy AI:

1. it should be lawful (comply with applicable laws and regulations);

2. it should be ethical (adhere to ethical principles and values); and

3. in should be robust both from social and technical perspective.

The China White Paper on Trustworthy AI refers to five globally accepted elements: transparency, security, fairness, accountability, and privacy. [85] By accountability, the white paper implies that 'enterprises must audit the implementation processes of AI systems comprehensively to improve their traceability and ensure the trustworthiness of the systems and services from the source'.[86] The elements may vary from document to document but the core idea remains. For instance, according to the US NIST AI RMF, trustworthy AI systems must be 'valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed'.[87]

The envisions of trustworthy AI serve as a building block for state policies, guidelines and binding regulations. The concept is considered a key mechanism of building and maintaining trust in the technology by the society.[88]

3.3 The regulation of AI in compared jurisdictions

As comes from the title, this section discusses the AI regulation in each of the compared jurisdictions. For each jurisdiction, there will be an overview sub-section that describes the AI regulatory framework and sub-sections for human rights, personal data, and safety containing the analysis of AI-specific policy and regulation of these aspects.

---

84 AI HLEG, 'Ethics Guidelines for Trustworthy AI' (European Commission, 8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 6 December 2024 (AI HLEG Ethics Guidelines), Executive summary.

85 China Academy of Information and Communications Technology and JD Explore Academy, 'White Paper on Trustworthy Artificial Intelligence,' July 2021 <http://www.caict.ac.cn/english/research/whitepapers/202110/P020211014399666967457.pdf> accessed 6 December 2024 (China WP on Trustworthy AI), s 2.

86 Ibid, s 4.1.3(2).

87 National Institute of Standards and Technology, U.S. Department of Commerce (NIST). NIST.AI.100-1 Artificial Intelligence Risk Management Framework of January 2023 <https://doi.org/10.6028/NIST.AI.100-1> accessed 6 December 2024 (NIST AI RMF), s 3.

88 Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Building Trust in Human Centric Artificial Intelligence' COM(2019)168, s 2.

The analysis described below focuses on the key AI-specific features. A more detailed comparison is provided in the table in Appendix 1 hereto.

3.3.1 EU

(a) Regulatory framework

The EU is considered to be the jurisdiction with the most detailed and advanced regulation of AI at the moment. It is also the only jurisdiction that has adopted comprehensive binding horizontal regulation of AI. The recently adopted AI Act is considered to be 'the most far-reaching regulation of AI worldwide'[89].

The current state of regulation is the result of 7 years of EU regulatory activity. The initiatives to regulate AI trace back to December 2017, when the EU adopted the Joint Declaration on the EU's legislative priorities for 2018-19[90]. Among other things, the declaration mentioned 'ensuring a high level of data protection, digital rights and ethical standards while capturing the benefits and avoiding the risks of developments in artificial intelligence and robotics'[91]. On 9 March 2018, the Commission announced the setting up of an expert group on artificial intelligence (subsequently called the AI HLEG) to develop guidelines on AI ethics and engage in 'a wide, open and inclusive discussion on how to use and develop artificial intelligence both successfully and ethically sound'.[92]

On 16 February 2017, the European Parliament adopted a resolution with recommendations to the Commission on Civil Law Rules on Robotics.[93] This was a visioner's paper discussing the development of smart robotics that may implement artificial intelligence and be capable of self-learning and autonomous decision-making. One of the provisions of this romantic and forward-looking document mentioned that 'from Mary Shelley's Frankenstein's Monster to the classical myth of Pygmalion, through the story of Prague's Golem to the robot of Karel Čapek, who coined the word, people have fantasised about the possibility of building

---

89 'Regulation of artificial intelligence' (n 58).

90 Commission, 'Joint Declaration on the EU's legislative priorities for 2018-19' (European Commission, 14 December 2017) <https://commission.europa.eu/publications/joint-declaration-eus-legislative-priorities-2018-19_en> accessed 6 December 2024.

91 Ibid.

92 Commission, 'Artificial intelligence: Commission kicks off work on marrying cutting-edge technology and ethical standards' (European Commission, 9 March 2018) <https://ec.europa.eu/commission/presscorner/detail/en/ip_18_1381> accessed 6 December 2024.

93 European Parliament resolution 2015/2103(INL) of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics [2018] OJ C252/239.

intelligent machines, more often than not androids with human features'[94], discussed the Asimov's Laws, and the possibility of assigning the most sophisticated autonomous robots with the specific legal status of 'electronic persons' and attributing to them the liability for the wrongdoings.[95] From the very beginning, the EU formulated that the development of AI technology must be based on the ethical principle of the protection of fundamental rights, privacy and, data protection[96], as well as human safety, health, and security[97].

The process of building a full-fledged AI policy kicked off in April 2018 with the adoption of the European Strategy on Artificial Intelligence[98]. As per section 3.4 of the Strategy, the same year the AI HLEG was appointed.[99] In December 2018, the Commission delivered the Coordinated Plan on AI.[100] The plan admitted that 'without major efforts, the EU risks losing out on the opportunities offered by AI, facing a brain-drain and being a consumer of solutions developed elsewhere' and called for major investments into the AI sphere,[101] adaptive learning and training, and building the European data space. The plan also discussed developing ethics guidelines 'with a global perspective and ensuring innovation-friendly legal framework'[102] and required AI technology to be predictable, responsible, verifiable, respect fundamental rights and follow ethical rules to gain trust and be accepted by society.[103] It envisioned that Europe 'can become a global leader in developing and using AI for good and promoting a human-centric approach and "ethics-by-design" principles'.[104]

---

94 Ibid, Introduction, para A.

95 Ibid, s 59(f).

96 Ibid, s 13.

97 Ibid, s 10.

98 Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions artificial intelligence for Europe' COM/2018/237 final.

99 Commission, 'Commission appoints expert group on AI and launches the European AI Alliance' (European Commission, 14 June 2018) <https://digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance> accessed 6 December 2024.

100 Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Coordinated Plan on Artificial Intelligence' COM/2018/795 final.

101 Ibid, s 2.1.

102 Ibid, s 2.6.

103 Ibid, s 2.6.

104 Ibid.

In April 2019, the AI HLEG delivered the Ethics guidelines for trustworthy AI[105]. In addition to the elements of the trustworthy AI discussed earlier, the guidelines formulated seven key requirements for AI systems:

1. Human agency and oversight – AI systems should empower human beings, allow them to make informed decisions and protect human rights; a proper oversight mechanism is required;

2. Technical robustness and safety – this requires AI systems to be resilient to failures and adversarial attacks as well as safe; a fall-back plan must be in place;

3. Privacy and data governance – privacy and data protection must be ensured as well as a proper data governance mechanism ensuring data integrity and authorised access;

4. Transparency – humans must be aware of interacting with AI and of the system's capabilities and limitations; the decisions of the AI system must be explainable and the data, system and business model must be transparent;

5. Diversity, non-discrimination, fairness – no biases and equal access to the systems;

6. Societal and environmental well-being – AI systems must be sustainable and environmentally friendly, mindful of their societal impact; and

7. Accountability – responsibility and accountability for AI systems outcomes must be established as well as appropriate measures to audit the algorithms, data and design processes.[106]

Notably, democracy is mentioned 24 times in the guidelines stressing that 'a future where democracy, the rule of law and fundamental rights underpin AI systems and where such systems continuously improve and defend democratic culture will also enable an environment where innovation and responsible competitiveness can thrive'[107].

In April 2019, the Commission released the Communication on Building Trust in Human Centric Artificial Intelligence that reinstated that AI should be 'a tool that has to serve people with the ultimate aim of increasing human well-being'[108]. To achieve this, AI should be

---

105 AI HLEG Ethics Guidelines (n 84).

106 Ibid, s II(1).

107 Ibid, s I(9).

108 COM(2019)168 (n 88), s 2.

trustworthy, so 'the values on which our societies are based need to be fully integrated in the way AI develops'[109]. As for these values, the communication lists human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities, and also refers to the Charter of Fundamental Rights that 'brings together the personal, civic, political, economic and social rights enjoyed by people within the EU'[110]. The Communication mentions new challenges posed by AI due to its ability to 'learn' and make autonomous decisions, to produce biased outputs and the vulnerability to cyber-attacks.[111] Therefore, it concludes, the technology must be human-centric 'to have the public's trust and adhere not merely by law, but also to ethics standards'[112]. The communication endorsed the AI HLEG Ethics guidelines for trustworthy AI mentioning its non-binding nature, nevertheless emphasising its important role in developing the trustworthy AI. Finally, the communication highlighted the ongoing international cooperation and the active role of the EU in building human-centric AI and dialogue on the G7 and G20 platforms and contributing to standardization activities.

In June 2019, the ethics guidelines were followed by the AI HLEG's Policy and investment recommendations for trustworthy Artificial Intelligence.[113] The Policy reinforced and elaborated the principles set out in the guidelines, including the human-centric approach to AI. The policy highlighted the importance of research and education, investments, data and infrastructure, and the need for dedicated AI regulation; it also stressed the need for the protection of privacy (including preventing mass surveillance), worker's rights, data interoperability (data sharing), assessment of AI systems for compliance, and adopting the unified EU regulation on AI. In July 2020, the AI HLEG issued the Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, a relatively compact 30+ page questionnaire based on the seven key areas identified in the guidelines.[114]

---

109 Ibid.

110 Ibid.

111 Ibid.

112 Ibid.

113 AI HLEG, 'Policy and investment recommendations for trustworthy Artificial Intelligence' (European Commission, 26 June 2019) <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence> accessed 6 December 2024.

114 AH HLEG, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment' (European Commission, 17 July 2020) <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment> accessed 6 December 2024.

In February 2020, the Commission released the White Paper on Artificial Intelligence - A European Approach to Excellence and Trust.[115] The White Paper connects the data, AI and the well-being:

> As digital technology becomes an ever more central part of every aspect of people's lives, people should be able to trust it… Europe's current and future sustainable economic growth and societal wellbeing increasingly draw on value created by data. AI is one of the most important applications of the data economy.[116]

The white paper calls upon a common European approach to AI to avoid the fragmentation of the single market – the national initiatives risk endangering legal certainty, weaken citizen's trust and prevent the emergence of a dynamic European industry.[117] The paper defines two building blocks: 'an ecosystem of excellence' – the framework for research, innovation and investments; and 'an ecosystem of trust' – the key elements of the regulatory framework, including the human-centric approach, the rules protecting fundamental rights and consumers' rights, in particular for high-risk AI systems.[118] The white paper also mentions the EU international cooperation within the Council of Europe, the G20, the OECD, UNESCO, and ITU[119], and the effect the AI HLEG had had on their work. The document expresses the intention to continue the cooperation with 'like-minded' countries and global AI players based on the EU rules and values and promoting 'the respect of fundamental rights, including human dignity, pluralism, inclusion, non-discrimination and protection of privacy and personal data and it will strive to export its values across the world'.[120]

Advocating for AI regulation, the white paper claims the lack of trust in the technology is one of the main obstacles to a broader application of AI (the others are lack of investment and relevant skills).[121] The paper also mentions the 'black box' effect in AI technology that complicates the enforcement of regulations not adapted to AI.[122] Speaking of potential harm, the white paper rightly states that it could be both material (health or damage of property)

---

115 COM(2020) 65 final (n 60).

116 Ibid, s 1.

117 Ibid.

118 Ibid.

119 The United Nations Educational Scientific and Cultural Organization (UNESCO), the Organisation for Economic Co-operation and Development (OECD), and the International Telecommunications Union (ITU).

120 COM(2020) 65 final (n 60), s 4(H).

121 Ibid, s 5.

122 Ibid.

and immaterial (loss of privacy, discrimination, limiting the freedom of expression) and selects the three focus areas of regulation: the protection of fundamental rights (including personal data and privacy), safety, and liability-related issues.[123] Notably, among the protected fundamental rights the white paper mentions political rights such as the freedom to assembly and the right to freedom of expression.[124]

As regards safety, the white paper raises important issues of the attributability of liability for AI automated decisions, the difficulty of gathering evidence by the suffering person (as compared to traditional technology), establishing a causal link between the product and damage and, consequently, having due recourse. The paper stresses that the lack of clear safety requirements disorients and thus undermines the trust in AI.

The white paper proposes the areas of AI-specific regulation, i.e., when the non-AI regulation appears insufficient. These areas include: liability and the allocation of responsibilities across the actors in the supply chain, the application of safety rules to stand-alone software, services powered on AI technology (healthcare, financial and transport), addressing the ongoing 'learning' feature of AI products, and addressing AI-specific safety threats (personal security, cyber threats).[125] The white paper envisions the regulation of AI to be risk-based and differentiates the category of 'high-risk' AI, which should be subject to stricter requirements. The document defines the types of such requirements, including regarding remote biometric identification.[126] Since the EU data protection rules prohibit the processing of biometric data to identify natural persons (subject to limited exceptions), the white paper proposes the use of AI for the same purposes be limited to narrow duly justified and proportionate cases.[127] Finally, the paper also discusses the mandatory conformity assessment, voluntary labelling of no high-risk AI, and the governance structure.[128] The white paper was a major building block towards the development of the binding EU regulation on AI – many of its ideas were subsequently incorporated into the AI Act.

---

123 Ibid, s 5(A).

124 Ibid, s 5(A).

125 Ibid, s 6(B).

126 Ibid, s 6(D)(f).

127 Ibid.

128 Ibid, sub-ss 5(F—H).

In April 2021, the Commission released a package of documents containing:

1. the Communication on Fostering a European approach to Artificial Intelligence;[129]

2. the Updated Coordinated Plan on AI;[130]

3. the Proposal for the AI Act[131] and the accompanying Impact Assessment of the Regulation on Artificial intelligence[132].

The first document, the Communication on Fostering a European approach to Artificial Intelligence, contained a broad overview of the AI-related activity of the Commission for the last few years and the refined vision of the regulation of AI. Thus, the communication discussed the substantial revision of the existing and the development of new legislation to adapt the regulation to AI. This included the revision of the Machinery Directive[133] addressing safety risks of new technologies, cyber risks and autonomous machines, as well as the revision of other relevant acts such as the Product Liability Directive[134], the General Product Safety Directive[135], the EU Security Union strategy[136], the new cybersecurity strategy,[137] the proposed new Digital Services Act and Digital Markets Act[138], as well as the

---

129 Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence' COM/2021/205.

130 Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions Fostering a European approach to Artificial Intelligence' COM/2021/205, Annex 'Coordinated Plan on Artificial Intelligence 2021 Review' (Updated Coordinated Plan on AI).

131 Commission, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts' COM(2021) 206 final.

132 Commission staff working document, 'Impact assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts' SWD(2021) 84 final

133 Commission, 'Proposal for a Regulation of the European Parliament and of the Council on machinery products' COM/2021/202 final; Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast) [2006] OJ L 157/24.

134 Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L210 (PLD).

135 Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety (Text with EEA relevance) [2001] OJ L11.

136 European Parliament and the Council, 'Joint communication to the European Parliament and the Council the EU's Cybersecurity Strategy for the Digital Decade' JOIN/2020/18 final.

137 Commission, 'The Cybersecurity Strategy' (European Commission) <https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-strategy > accessed 7 December 2024.

138 Commission, 'Europe fit for the Digital Age: Commission proposes new rules for digital platforms' (European Commission, 15 December 2020) <https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2347> accessed 7 December 2024.

European Democracy Action Plan[139]. The document repeats challenges of AI regulation related to the characteristics of the AI technology, mainly the opacity of algorithms, resulting in difficulties in investigating causal links and enforcing the existing legislation. The communication also discusses facial recognition in public spaces, which can be very intrusive on privacy due to the poor functioning of such systems and the high risk of error and discrimination.[140] The document calls for a risk-based, future-proof, and innovation-friendly approach, that is to intervene only where this is strictly needed and in a way that minimises the burden for economic operators, with a light governance structure.[141]

The second document was an update of the 2018 Coordinated Plan on AI. It was focused on the four key areas: enabling conditions for AI's development; funding efforts; regulatory framework ensuring the AI is developed human-centric, sustainable, secure, inclusive, accessible and trustworthy; and building strategic leadership in high-impact sectors such as environment, health, the public sector, robotics, and security. As regards the regulation of AI, the plan focuses on a horizontal framework for AI prioritizing safety and the respect of fundamental rights, including the criteria and requirements for high-risk AI systems, measures to adapt the liability framework, and adapting the existing sectoral safety legislation. [142]

The third document contained the proposal of what three years later became the AI Act.

In September 2022, the Commission proposed the Directive on adapting non-contractual civil liability rules to artificial intelligence.[143] The proposal stressed that 'current national liability rules, in particular, based on fault, are not suited to handling liability claims for damage caused by AI-enabled products and services' and that 'the specific characteristics of AI, including complexity, autonomy and opacity (the so-called 'black box' effect), may make it difficult or prohibitively expensive for victims to identify the liable person and prove the requirements for a successful liability claim'.[144] The directive proposed a rebuttable presumption of a causal link between the fault of the defendant and the output produced by

---

139 Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European democracy action plan' COM/2020/790 final.

140 COM/2021/205 (n 129), s 2.

141 Ibid.

142 2021 Coordinated Plan on AI (n 130), s 9.

143 Commission, 'Proposal for a Directive of The European Parliament and of The Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)' COM(2022) 496 final (AILD).

144 Ibid, Explanatory Memorandum, s 1.

the AI system as well as additional requirements on disclosure of evidence on high-risk AI systems. The directive has not been adopted yet.

On 24 January 2024, the Commission announced the establishment of the European AI Office to 'oversee the advancements in artificial intelligence models, including as regards general-purpose AI models, the interaction with the scientific community, and should play a key role in investigations and testing, enforcement and have a global vocation'.[145]

Lastly, on 13 June 2024, the AI Act was adopted. It is the major piece of AI regulation in the EU and the first comprehensive binding AI regulation worldwide. The AI Act entered into force on 1 August 2024 (some of its provisions will enter into force in 2025—2027). This is a massive 144-page document containing a detailed regulatory framework for AI systems. It is a product of many years of visioning and shaping the AI regulatory framework by different stakeholders and represents the up-to-date state of development of AI regulation in the EU.

The EU was the pioneer jurisdiction to introduce binding AI regulation of such a level of detail. This is not the first time the EU has taken such a role. Another recent example is the Regulation on markets in crypto-assets[146], where the EU attempted to regulate the emergent market. So, the EU has the experience and the flavour to endeavour regulatory *terra incognito*.

The AI Act establishes horizontal regulation of AI, i.e., it applies to all AI systems, to all industries and use-cases (subject to certain exceptions), as part of a product, software, or as a standalone solution. The act is designed to 'squeeze into' the existing legislation[147] and fill-in AI-specific regulatory gaps.

The AI Act states the protection of human rights and safety as its priorities and formulates its purpose as follows:

> [T]o improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter,

---

145 Commission, 'Decision of 24.1.2024 establishing the European Artificial Intelligence Office' C(2024) 390 final, recitals 5.

146 European Parliament and the Council Regulation (EU) 2023/1114 of 31 May 2023 on markets in crypto-assets, and amending Regulations (EU) No 1093/2010 and (EU) No 1095/2010 and Directives 2013/36/EU and (EU) 2019/1937 (Text with EEA relevance) [2023] OJ L150/50 (MICA).

147 The act refers to tens of acts it aims to be aligned with.

including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation.[148]

The AI Act is designed to be comprehensive legislation that both builds up the regulatory framework and directly regulates many aspects of AI. By large, the regulation addresses:

- harmonised rules for the placing on the market, the putting into service, and the use of AI systems in the Union;
- prohibitions of certain AI practices;
- specific requirements for high-risk AI systems and obligations for their operators;
- harmonised transparency rules for certain AI systems;
- harmonised rules for the placing on the market of general-purpose AI models;
- rules on market monitoring, market surveillance, governance and enforcement; and
- measures to support innovation, including by SMEs and start-ups.[149]

There are three notable features: the focus on building the human-centric regulatory framework; the attention to safety aspects in combination with a risk-based approach; and the intention to build future-proof legislation;

The AI Act provides for four categories of AI systems depending on their potential risks:

- Unacceptable risk (prohibited AI systems);
- High risk (a set of additional requirements to AI systems);
- Limited risk (specific transparency obligations to some AI systems);
- Minimal (all other AI systems; they are subject to minimal requirements).

In a nutshell, the approach implies that the burden of compliance with the regulation must be weighed with the potential risk. This is to ensure reasonable safety measures but not hinder,

---

148 AI Act (n 18), art 1(1).

149 AI Act (n 18), art 1(2).

at the same time, the development of AI technology. The EU is the only compared jurisdiction that clearly outlines prohibited AI.[150]

Many provisions of the AI Act will be discussed in detail in the next sections, so the full-scope overview of the act is not feasible here.

The AI Act is designed to fit into the existing, non-AI, regulatory framework. Among the relevant non-AI legal sources: the Charter of Fundamental Rights of the European Union[151]; the European Declaration on Digital Rights and Principles for the Digital Decade (DODR)[152]; the General data protection regulation (GDPR)[153] and relevant Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679[154], which establish the framework for processing personal data of individuals by non-public entities; Data Protection Regulation for EU institutions, bodies, offices and agencies (EUDPR)[155]; the Law Enforcement Directive (LED)[156] – relates to processing connected with criminal offences or the execution of criminal penalties; the Digital Services Act[157] that addresses illegal online content, advertising and disinformation; the Digital Markets Act[158] aimed to regulate the behaviour of the "Big Tech" firms within the European single market; Cyber

---

150 There are US guidelines that also define prohibited use-cases for AI but these guidelines apply only to the use of AI by state agencies.

151 Charter of Fundamental Rights (n 63).

152 European Declaration on Digital Rights and Principles for the Digital Decade [2023] OJ C 23/1 (DODR).

153 European Parliament and the Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) [2016] OJ L119/1 (GDPR).

154 Data Protection Working Party Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 of 3 October 2017 (last revised on 6 February 2018) WP251rev.01.

155 European Parliament and the Council Regulation (EU) 2018/1725 of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC [2018] OJ L 295/39 (EUDPR);

156 European Parliament and the Council Directive (EU) 2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L119/89 (LED).

157 European Parliament and the Council Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) [2022] OJ L277/1 (Digital Services Act).

158 European Parliament and the Council Regulation (EU) 2022/1925 of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance) [2022] OJ L295/1 (Digital Markets Act).

Resilience Act[159] that is intended to bolster cybersecurity rules and enhance security of hardware and software products[160]; the Market Surveillance Regulation[161]; and the Product Liability Directive[162].

Notable draft law includes: the draft AI Liability Directive[163] and the Product Liability Directive[164] aimed together to build the improved regulatory framework for the non-contractual civil and product liability for AI systems; and the draft Machinery Directive[165] that should lay down requirements for the design and construction of machinery products and has the purpose of aligning the EU regulation of machinery using AI with AI regulation.

The EU takes an active stance on the international level. Since 2019, it has participated in the Council of Europe's initiative to assess the need for regally binding AI regulation. Its international work was aimed at promoting the EU's vision of sustainable and trustworthy AI in the world. The EU was the founding member of the Global Partnership on AI, contributed to the OECD's and WIPO's work on AI, and was actively engaged in the ISO's and IEEE's standardization initiatives; the dialogue between the EU and the US over the trustworthy AI is also worth mentioning.[166] A major international achievement was the signing of the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (AI Convention)[167] on 5 September 2024. The convention is the result of years of joint work of 46 Council of Europe's member states and other countries, including the US, Japan, Australia, and Israel. The EU, the US, and the UK were among the 10 initial signatories.

---

159 European Parliament and the Council Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act) (Text with EEA relevance) [2024] (not yet in force, pending notification) (Cyber Resilience Act).

160 Commission, 'Cyber Resilience Act' (European Commission, 15 September 2022) <https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act> accessed 7 December 2024.

161 European Parliament and the Council Regulation (EU) 2019/1020 of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011 (Text with EEA relevance.) [2019] OJ L169/1.

162 PLD (n 134).

163 AILD (n 143).

164 Commission, 'Proposal for a Directive of the European Parliament and of the Council on liability for defective products' COM/2022/495 final (Draft PLD).

165 Commission, 'Proposal for a Regulation of the European Parliament and of the Council on machinery products' COM(2021) 202 final.

166 COM/2021/205 (n 130), s 10.

167 Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Vilnus, 5 September 2024) Council of Europe Treaty Series No. 225 (AI Convention).

As overviewed above, the EU regulation of AI is extensive. It has formulated in detail its policy, while the AI-specific binding regulation is still developing. A few important directives are at the stage of the draft. A major developing step was the adoption of the AI Act. Despite its volume and the scope of detail, in many aspects, it just sets the framework that must be enriched with relevant guidelines and best practices.

(b) Regulation of subject areas

This section will discuss the features of AI-specific legislation in the analysed areas: human rights, personal data, and safety.

(I) Human rights

Generally, the EU regulation is the most human-centric among the compared jurisdictions. Thus, the potential impact on human rights is one of the key factors that determine the risk category of an AI system.[168]

The key legal sources of AI-specific regulation of human rights (fundamental rights) are the AI Act, the European Declaration on Digital Rights and Principles for the Digital Decade (DODR), and the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (the AI Convention).

The AI Act refers to the 'classic' human rights established in the Charter of Fundamental Rights.[169] Even so, it was the general understanding that the digital transformation affects every aspect of people's lives[170] and there is a need to recall the most relevant rights in the context of this transformation[171], including restating the importance of freedom of choice in interactions with algorithms and artificial intelligence and a fair digital environment[172]. To this end, the DODR was adopted in 2023. Although the declaration is not dedicated to AI and refers to a broader category of 'digital rights', the existence of such legislation makes the EU stand out and evidence a true commitment to ensuring high protection of human rights in the digital and AI environments.

By large, the 'digital' human rights are not completely new but rather an adaptation of classical rights to the digital context. Nevertheless, the DODR brings more clarity as regards

---

168 AI Act (n 18), recitals 48.

169 AI Act (n 18), art 1(1).

170 DODR (n 152), Preamble, para 2.

171 Ibid, Preamble, para 7.

172 Ibid.

the scope of rights and acknowledges the substantial specifics of the protection of these rights in the digital environment.

Classic human rights in the focus. Although the AI legislation aims to protect all fundamental rights, some of them are more relevant, or vulnerable, in the AI context. According to the explanatory memorandum to the proposed AI Act, the focus is made on the protection of the following rights: the right to human dignity; respect for private life and protection of personal data; non-discrimination and equality between women and men; the rights to freedom of expression and freedom of assembly; the right to an effective remedy and a fair trial, the rights of defence and the presumption of innocence, as well as the general principle of good administration; the workers' rights to fair and just working conditions; consumer protection; the rights of the child and the integration of persons with disabilities; and the right to a high level of environmental protection and the improvement of the quality of the environment. [173]

Political rights. A distinguishing feature of the EU AI legislation is the focus on political rights. The ensuring of the high level of protection of democracy is stated as the purpose of the AI Regulation.[174] The AI systems that are involved in the democratic process are considered high-risk AI systems.[175] The AI Convention by its name already emphasizes this aspect. Article 5 requires each party to have measures ensuring that AI systems do not undermine the integrity, independence and effectiveness of democratic institutions and fair access to and participation in public debate and freely form opinions.

AI-specific rights. Despite the digital rights derive from the fundamental rights, some of them address very specific risks that are absent outside the AI context. Therefore, unless they are expressly formulated, it would have been extremely hard, both for the affected person and the courts, to enforce them. The EU legislation has formulated a set of relevant rights.

- The right to know / to transparency. It appears that people feel particularly uncomfortable when they interact with AI, or are subject to AI decisioning, without knowing thereof. The right to know is integrated in each of the compared jurisdictions. This right is formulated in the DODR – its article 9(b) commits to 'ensuring an adequate level of transparency about the use of algorithms and artificial intelligence, and that people are empowered to use them and are informed when interacting with them'. A similar right is formulated in article 15(2) of the AI

---

173 COM/2021/206 final (n 59), Explanatory Memorandum, s 3.5.

174 AI Act (n 18), art 1(1).

175 AI Act (n 18), annex III, s 8.

Convention: 'as appropriate for the context, persons interacting with artificial intelligence systems are notified that they are interacting with such systems rather than with a human'. The AI Act expressly requires the developers of AI systems to ensure the persons interacting with it are informed thereof unless this is obvious to a user or an AI system is dealing with criminal offences.[176]

- Protection of human autonomy / freedom of choice. The relevant risk posed by AI is well-formulated in the AI Act:

> 'AI systems [may] deploy subliminal components such as audio, image, video stimuli that persons cannot perceive, as those stimuli are beyond human perception, or other manipulative or deceptive techniques that subvert or impair person's autonomy, decision-making or free choice in ways that people are not consciously aware of those techniques or, where they are aware of them, can still be deceived or are not able to control or resist them.'[177]

As stated above, the effect of AI may be on the unconscious level. More to that, even an aware person can still be manipulated. Recognising the huge potential threat of such systems, the AI Act prohibits the AI system that

> '[D]eploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm… [or that] exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm.'[178]

---

176 AI Act (n 18), art 50(1).

177 AI Act (n 18), recitals 29.

178 AI Act (n 18), sub-ss 5(1)(a), (b).

The protection of the freedom of choice is specifically addressed in the DODR, which commits to ensuring that technologies such as artificial intelligence are not used to pre-empt people's choices.[179]

- Protection *from* content. Historically, the vector was toward the free flow of information and the access to information – the more information the better. In the context of digital transformation, and especially concerning AI, people face the need to be protected *from* information. This is a new paradigm and we will not find a relevant 'classical' right for this aspect. The problem of controlling digital content is hot and is acerbated by the capability of AI to generate content, especially as regards deep synthesis.

  The DODR commits to taking proportionate measures to tackle all forms of illegal content, in full respect for fundamental rights, including the right to freedom of expression and information, albeit without establishing any general monitoring obligations or censorship, and to creating a digital environment where people are protected against disinformation, information manipulation and other forms of harmful content, including harassment and gender-based violence.[180]

  The AI Convention requires the AI-generated content to be clearly identified.[181] It further states that 'certain AI systems intended to interact with natural persons or to generate content may pose specific risks of impersonation or deception… in certain circumstances, the use of these systems should therefore be subject to specific transparency obligations'[182] The AI Act continues that synthetic content becomes 'increasingly hard for humans to distinguish from human-generated and authentic'[183] and even introduces the definition of 'deep fake', which means 'AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful'.[184]

  The AI Act requires the providers of AI systems generating synthetic audio, image, video, or text content, to mark it in a machine-readable format and detectable as

---

179 DODR (n 152), art 9(d).

180 DODR (n 152), art 15(c, d).

181 AI Act (n 18), art 8.

182 AI Act (n 18), recitals 132

183 AI Act (n 18), recitals 133.

184 AI Act (n 18), art 3(60).

artificially generated or manipulated.[185] Applying the risk-based approach, the AI Act states that marking solutions must be 'effective, interoperable, robust and reliable as far as this is technically feasible' and shall consider the costs of implementation and the state of the art.[186] The obligation does not apply if the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate or prosecute criminal offences.[187] There is a standalone obligation to disclose that the content was artificially generated by an AI system capable of generating deep-fake content.[188] The AI Act encourages the development of the relevant code of practice to standardise compliance with this requirement.[189]

- Right to explanation. To mitigate the 'black box' effect of AI systems, the AI Act requires the deployers of high-risk AI systems to provide meaningful explanations of the roles of the AI system in the decision-making procedure and the main elements of the decision taken upon a request of the concerned person for decisions that allegedly impacted on their health, safety, or fundamental rights.[190]

- Human oversight / no AI-autonomous decisioning. As discussed, people are often not comfortable being one-of-one with AI systems. Certain decisions have, or may have, a significant impact on the person, so the negative consequences of an inappropriate decision will be high. Despite human oversight being an element of AI safety, it could also be considered a specific human right since the human oversight mechanisms often imply prohibiting autonomous AI decisions in certain spheres or scenarios (usually those that may have a major impact on human rights, e.g., healthcare, access to public services).

  The right is declared in the AI Convention, which requires adopting or maintaining adequate 'oversight requirements tailored to the specific contexts and risks'[191]. The DODR contains limited provisions as regards human oversight and relates only to

---

185 AI Act (n 18), art 50(2).

186 Ibid.

187 Ibid.

188 Ibid.

189 AI Act (n 18), art 50(7).

190 AI Act (n 18), art 86.

191 DODR (n 152), art 8.

important decisions affecting workers.[192] The AI Act is more elaborate in this respect. The obligations relate to high-risk AI systems and include the requirement for the systems to be designed and developed 'in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use'.[193] The design of the high-risk AI system must allow, if needed, to override, disregard or reverse the output of the AI system and to intervene and apply a 'stop' button to halt the system or switch it into a safe mode.[194] Finally, as regards the AI systems that apply biometrical identification, safe for certain exceptions (e.g., law enforcement) the system's output should be verified by at least two natural persons with the appropriate training to apply the identification result.[195] The AI Act also requires the manual of such systems to describe the human oversight measures.[196] This right, and automatic decision-making generally, is backed up by the rights related to automated decision-making provided by GDPR that significantly limit these practices in the EU.[197]

- Protection from social scoring. As was discussed already, the EU regulation is very human-centric and opts for a conservative approach when there is a risk of major harm to human rights, including discrimination. As explained in the AI Act,

  > 'AI systems providing social scoring of natural persons may lead to discriminatory outcomes and the exclusion of certain groups, violate the right to dignity and non-discrimination and the values of equality and justice as well as lead to the detrimental or unfavourable treatment of natural persons or whole groups thereof in social contexts.[198]

  Although social scoring could be viewed as a type of discrimination, a 'classic' right, full-blown social scoring is hardly imaginable outside the AI context. Therefore, mentioning this practice in the AI Act is feasible. Specifically, the AI Act prohibits AI systems for

---

192 Ibid, art 6(e).

193 AI Act (n 18), art 14(1).

194 Ibid, art 14(4).

195 Ibid, art 14(5), Annex III, point 1(a).

196 Ibid, 13(3)(d)

197 GDPR (n 153), art 22.

198 AI Act (n 18), recitals 13.

> ‘[T]he evaluation or classification of natural persons … based on their social behaviour or … personality characteristics, with the social score leading to either or both of the following: (i) detrimental or unfavourable treatment of certain natural persons or groups of persons in social contexts that are unrelated to the contexts in which the data was originally generated or collected; (ii) detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity.’[199]

What does it mean? On the one hand, the attempt is clearly to ban, or at least significantly limit, social scoring practices. The used wording, however, blurs the intention and could be understood that such practices are possible unless they discriminate. But what is the point of such regulation if there is a general requirement for non-discrimination? We should apparently wait for the clarifications of the Court of Justice of the European Union to shed light on this aspect.

- No criminal offence prediction systems.[200] This endorses the presumption of innocence. As explained in the AI Act, ‘natural persons should never be judged on AI-predicted behaviour based solely on their profiling, personality traits or characteristics … without a reasonable suspicion … based on objective verifiable facts and without human assessment thereof.’[201] The prohibition is formulated in subsection 5(1)(d) of the AI Act. It does not apply to AI systems that are not based on the profiling of individuals or the personality traits and characteristics of individuals, like the assessment of the risk of fraud by undertakings and localisation of narcotics or illicit goods by customs authorities.

To summarise, the EU regulation implements the human-centric approach. Relevant rights and provisions are primarily contained in the AI Act, which specifies in a very detailed manner AI-specific rights; some rights and protections are formulated by prohibiting certain AI practices.

---

199 Ibid, para 5(1)(c).

200 Ibid, para 5(1)(d).

201 Ibid, recitals 42.

(II) Personal data

Regarding personal data rights, AI regulation mostly refers to general data protection legislation formed by the GDPR, the EUDPR, and the LED. By large, data subjects' rights include: the right to be informed, the right to access, the right to rectification, the right to object/restrict to processing, including to withdraw consent, the rights related to automated decision-making and profiling, the right to be forgotten, the right to data portability, and the right to complaint.

The AI legislation claims the protection of personal data is one of its priorities.[202] Thus, the AI Convention requires implementing measures to ensure the privacy rights of individuals and the protection of their personal data. The DODR also demands that personal data be protected. Finally, the AI Act draws significant attention to the AI-related processing of personal data. The expression 'personal data' is mentioned 80 times in the act!

According to the AI Act, it 'does not seek to affect the application of existing Union law governing the processing of personal data' but by introducing harmonized rules for AI systems it should facilitate the effective implementation of these rights.[203] Relevant provisions of the AI Act fall under two major blocks: the special personal data processing provisions for developing and training AI systems and provisions on the processing of biometrical data.

The use of personal data collected for other purposes for the development of AI systems. The AI Act provides for a special regime of 'regulatory sandbox' that sets out certain exclusions from general rules in the controlled test environment. One of the exceptions is the right to use personal data collected for other purposes to develop, train, and test AI systems, subject to certain requirements.[204] Other obligations of the data controller/processor remain.[205] The national competent authorities must supervise the processing of personal data in the sandboxes.[206]

To enjoy the exception, the AI system must be developed for safeguarding substantial public interest and only in certain areas (e.g., public safety or health), the processing of personal must be necessary and irreplaceable by anonymized or synthetic data, and the processing

---

202 AI Convention (n 167), art 11(a).

203 AI Act (n 18), recitals 10.

204 Ibid, art 59.

205 Ibid, recitals 140

206 Ibid, s 57(10).

must be isolated from other data processing environment; the sharing of the original data outside the sandbox is subject to general rules.[207] Article 59(3) of the AI Act leaves room for other instances of processing data collected for other purposes for the training of innovative AI systems if so is authorized by the EU law. For high-risk AI systems, the AI Act requires the indication of the original purpose of the collection of personal data used in the data sets.[208]

The AI Act prioritises the protection of personal data over the benefits of free and open-source AI. Therefore, the AI Act does not provide for a relaxed regime for such AI systems.[209]

Processing of special categories of personal data by high-risk AI systems. Such systems may only process special categories of personal data where it is strictly necessary for bias detection and correction, with the appropriate safeguards and records-keeping, with no further transfer or transmission, and subject to deletion once the bias has been corrected or the retention period has ended.[210]

Biometrical data. The utmost attention is given to the processing of biometrical data and facial images. In this respect, the AI Act prohibits certain use-cases and classifies many others as high-risk AI. The latter, for instance, relates to remote biometrical identification, where 'technical inaccuracies' may lead to biased results and entail discriminatory effects.[211] Following the risk-based approach, however, AI systems used for verification purposes (e.g., unlocking a device) do not fall under the high-risk category.[212] Relevant prohibited or restricted AI practices could be viewed as AI-specific personal data rights of the subjects forming the category of AI-specific personal data rights.

AI-specific rights

- No emotion-inferring systems at workplaces and education institutions.[213] The AI Act introduces the definition of an 'emotion recognition system'. This is an AI system to identify or infer the emotions or intentions of natural persons based on their biometric data.[214] As explained in recitals 44 of the AI Act, 'there are serious

---

207 Ibid.

208 Ibid, sub-s 10(2)(b).

209 Ibid, recitals 103.

210 Ibid, art 10(5).

211 Ibid, recitals 54.

212 Ibid

213 Ibid, art 5(1)(f).

214 Ibid, art 3(39).

concerns about the scientific basis of AI systems aiming to identify or infer emotions, particularly as expression of emotions vary considerably across cultures and situations, and even within a single individual'[215]. The imbalance of power in the context of work and education combined with the intrusive nature of such systems that involve the processing of biometrical data led to the ban of such systems at workplaces and educational institutions. Yet again, following the risk-based approach, the restriction does not apply to systems used strictly for medical (e.g., for therapeutical use) or safety reasons.[216] The emotion recognition systems that are not prohibited by the AI Act are qualified as high-risk AI.[217] Natural persons must be notified of the operation of AI systems that can identify or infer emotions or intentions.[218]

- Ban on certain use-cases of biometric categorization of natural persons. According to the AI Act, 'biometric categorisation' is the assigning of natural persons to specific categories based on their biometric data.[219] The ban is narrowed to the systems that deduce or infer race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation of an individual.[220] Other systems that conduct biometric categorization based on special categories of personal data are categorized as high-risk[221] and require informing the users of the operation of such systems. The AI Act clarifies that the prohibition does not cover any labelling or filtering of lawfully acquired biometric datasets, such as images, based on biometric data or categorizing of biometric data in the area of law enforcement.[222] Unlike many other cases, the AI Act does not provide an elaborative explanation of the prohibition. Probably the ban follows the general human-centric regulatory vector and reflects the potential impact of the biometric categorization on the rights of individuals.

---

215 Ibid.

216 Ibid.

217 Ibid, recitals 54.

218 Ibid, recitals 132.

219 Ibid, recitals 16.

220 Ibid, art 5(1)(g).

221 Ibid, recitals 54.

222 Ibid.

- No real-time remote biometric identification in publicly available spaces for law enforcement.[223] By 'real-time remote biometric identification system' the AI Act means a remote biometric identification system, whereby the capturing of biometric data, the comparison and the identification all occur without a significant delay, comprising not only instant identification but also limited short delays to avoid circumvention.[224] This practice also affects the right to privacy (private life) and the freedom of assembly. The AI Act notes that such identification 'is particularly intrusive to the rights and freedoms of the concerned persons', 'it may affect the private life of a large part of the population, evoke a feeling of constant surveillance and indirectly dissuade the exercise of the freedom of assembly and other fundamental rights',[225] and that such practices may lead to biased results and entail discriminatory effects.[226]

  Therefore, the AI Act requires that the use of such systems in publicly available spaces be limited to a narrow set of use cases: the targeted search for specific victims; the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons; a genuine and present or a genuine and foreseeable threat of a terrorist attack; or the localization or identification of a person suspected in a specifically listed crimes that a penalized by at least 4 years' imprisonment.[227] This right goes side-by-side with law enforcement and public safety interests, and the EU regulator is trying hard to implement the 'risk-based' approach to find the right balance. This is an exemplary instance of the EU human-centric approach to the regulation of AI as the government gains many practical benefits from unrestrictedly running such systems.

  The use of such AI systems is subject to the completed fundamental rights impact assessment[228] and prior authorisation by a judicial authority or an independent administrative authority, as appropriate to the Member State.[229] Following the risk-based approach, there is an option to proceed without authorization in the situation of urgency, provided that the authorization is requested within the next 24 hours.

---

223 Ibid, art 5(1)(h).

224 Ibid, art 3(42).

225 Ibid, recitals 32.

226 Ibid, recitals 32.

227 Ibid, art 5(1)(h).

228 Ibid, art 5(2).

229 Ibid, art 5(3).

The authorisation shall be granted based on the principles of necessity, proportionality, and valid purpose. Furthermore, each use of the system must be notified to the market surveillance authority and the national data protection authority.[230] The data must be summarized by the market surveillance authority in the annual reports submitted to the Commission, which publishes the aggregated annual report.[231]

This is indeed a very high standard of real-time remote biometric identification. Somewhat this approach could be explained by the intention of the AI Act to fit into the existing legislation. Thus, the LED limits the processing of special categories of personal data to the cases where such processing: (i) is authorized by the Union or national law; (ii) protects the vital interest of the data subject or another natural person; or (iii) relates to personal data manifestly made public by the data subject.[232] Many instances of real-time remote identification do not fall into these criteria. Realising this, the AI Act specifically states that its rules operate as *lex specialis* to the rules of the LED. The AI Act grants fluency to the Member States to authorize, fully or partially, the use of such AI systems.[233] Probably the approach will be subject to further harmonisation as there will be more implementation practice.

As for other remote biometric identification systems permitted by the Union and national law, they are classified as high-risk AI systems given the risk they pose.[234]

Even the use of post-remote biometric identification systems[235] is quite limited under the AI Act. Such systems should be used 'in a way that is proportionate, legitimate and strictly necessary, and thus targeted, in terms of the individuals to be identified, the location, temporal scope and based on a closed data set of legally acquired video footage'.[236] They also require an *ex-ante* authorisation from the judicial authority for each use, except when they are used for the initial identification of a potential suspect based on the objective and verifiable facts

---

230 Ibid, art 5(4).

231 Ibid, arts 5(6, 7).

232 LED (n 156), art 10.

233 Ibid, art 5(5).

234 Ibid, annex III, s 1(a); recitals 54.

235 These are the remote biometric identification systems other than a real-time system – art 3(43) of the AI Act.

236 Ibid, recitals 94.

directly linked to the offence.[237] The untargeted use of such systems is also prohibited.[238] This should not lead to discriminative surveillance or to circumvent the prohibition of real-time remote biometric identification.[239] The deployers shall also submit annual reports to the market surveillance authority and the national data protection authority on the use of such systems. The Member States may introduce even stricter laws on the use of post-remote biometric identification systems.[240]

- No AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the Internet or CCTV footage.[241] As explained in the EU AI Act, such practices add 'to the feeling of mass surveillance and can lead to gross violations of fundamental rights, including the right to privacy'.[242] This is an example of a true human-centric approach enshrined in the AI Act as such systems and databases could be in high demand by corporations and the government.

The AI regulation is very mindful and protective as regards personal information. There are many cases where the EU demotes its law enforcement and security interests for the sake of privacy and the comfort of natural persons. Some AI-specific rights and protections evidence the depths of the study of the risks and potential intrusive use cases and the desire to resolve them by taking the risk-based approach.

(III) Safety

Quite naturally, the EU aims to be an AI-safe place and safety is declared a priority of the AI regulation[243]. In the same vein, article 9(e) of the DODR commits to providing safeguards and taking appropriate action to ensure the safety of AI systems. The AI Convention is more specific in this respect and splits this general idea into the obligations of transparency and oversight (article 8), accountability and responsibility (article 9), reliability (article 12), the risk and impact management framework (art 16), and safe innovation (article 13). The latter

---

237 Ibid, art 26(10).

238 Ibid, art 26(10).

239 Ibid, recitals 94.

240 Ibid, art 26(10).

241 Ibid, art 5(1)(e).

242 Ibid, recitals 43.

243 Ibid, art 1(1).

is of the most interest as it acknowledges the potential major risks of AI technology and warns of the need for controlled environments and supervision by competent authorities.

Despite the above provisions mostly formulate the desired outcomes rather than specific obligations of the AI providers, they, nevertheless, define the regulatory vectors elaborated in the AI Act and other relevant acts.

The EU declares the use of a risk-based approach.[244] This means that the regulation should tailor the type and content of the rules to the intensity and scope of the risks that AI systems can generate.[245] This is typical for the regulation of novel areas by the EU (e.g., the same approach was claimed and effectively applied in the regulation of the market in crypto assets[246]). And it is not a mere declaration but a thorough and actively applied practice. One of many instances is regulatory sandboxes or the exceptions applied to free open-source AI models.

As already discussed above, the AI Act defines four categories of AI systems based on their potential risk: unacceptable, high-risk, limited risk, and minimal risk.

It is notable that, despite the policy pressure to promote the development of AI systems, the EU showed integrity and banned certain most risky AI use cases. As already discussed, the prohibited use cases include AI systems that[247]:

- Implement subliminal, manipulative, or deceptive techniques or exploit vulnerabilities of a natural person or a group;
- Predict the risk of committing a crime by an individual;
- Create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage;
- Infer emotions of natural persons at workplaces and education institutions;
- Conduct certain types of social scoring and biometric categorisation; or

---

244 Ibid, recitals 26.

245 Ibid, recitals 26.

246 MICA (n 146).

247 AI Act (n 18), art 5(1).

- Use real-time remote biometric identification in publicly assessable spaces, except for narrow cases and subject to strict rules.

Next, the AI Act defines quite a broad spectrum of AI systems as high-risk. They consist of two major groups: (i) AI systems used as a safety component of a product or a product itself that is subject to the Union harmonisation legislation listed in Annex I to the AI Act (this includes, e.g., machinery, toys, medical devices and motor vehicles); and (ii) AI systems referred to in Annex III to the AI Act (e.g., permitted use-cases of remote biometric identification, safety components in critical infrastructure, evaluating learning outcomes), except when such systems do not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making.[248] The major part of the AI Act is focused on the regulation of this high-risk category of systems.

The category of 'limited-risk' systems is not expressly defined in the AI Act. The naming follows from the Commission's overview of the AI Act and refers to certain AI systems, whose risks derive from their lack of transparency.[249]

Minimal-risk AI systems are subject to minimal requirements.

Requirements and obligations related to high-risk AI systems. The AI Act contains an extensive set of requirements for high-risk AI systems. These include:

- Having a *risk-management system* that addresses: (i) the identification and analysis of the known and reasonably foreseeable risks that the AI system can pose to health, safety and fundamental rights when the system is used per the intended purpose; (ii) the evaluation of risk that may emerge under conditions of reasonably foreseeable misuse; (iii) the evaluation of other risks based on post-market monitoring; and (iv) the adoption of appropriate and targeted risk management measures. AI systems should be tested for risks. [250]

- Compliance with *data sets quality* requirements. These include the training, validation, and testing of data sets with the data governance and management practices appropriate to the intended purpose, including (i) the relevant design choice; (ii) data collection processes and the origin of data; (iii) relevant data

---

248 Ibid, art 6(1–3).

249 Commission, 'AI Act' (n 57).

250 AI Act (n 18), art 9

mapping; (iv) examination for biases and the appropriate measures of their identification, prevention and mitigation; (v) relevant, sufficiently representative and, to the best extent possible, free of errors data sets.[251]

- *Technical documentation* to be developed before the AI systems are placed on the market. The minimal information to be included is defined in Annex IV to the AI Act and includes, *inter alia*, (i) a general description of the AI system; (ii) detailed description of certain its elements, e.g., the methods and steps of its development, how the data was obtained and selected; (iii) detailed information about the monitoring, functioning and control of the AI system; (iv) a description of the risk management system; and (v) a copy of the EU declaration of conformity.[252]

- *Automatic recording* of events (logs) to identify risks and conduct monitoring of the operation of the AI system.[253]

- Meet *transparency* requirements to enable deployers to interpret a system's output and use it appropriately. The system should also be accompanied by an instruction that includes 'concise, correct and clear information that is relevant, accessible and comprehensible to deployers'.[254] The language of the provision is very thoughtful – the instruction can easily drown the reader with the technicalities and blur the meaning, so the wording is designed to cut such practices. The instruction should contain the information listed in Article 13 of the AI Act, i.e., the essential information on the system and its operation.

- *Human oversight*. This obligation is of particular interest. Some scholars even consider it to be the only true AI-specific, as other features of the technology could more or less apply to other products or software.[255] The obligation continues the discussion that in certain cases people want the outputs of AI systems to be supervised by humans to ensure it is valid and reasonable. The AI Act requires high-risk systems to be designed and developed in a way that includes human-machine interface tools and enables the AI systems to be effectively overseen by natural persons. The AI Act further demands that persons assigned to oversee is enabled:

---

251 Ibid, art 10.

252 Ibid, art 11.

253 Ibid, art 12.

254 Ibid, art 13.

255 Heder Mihaly, 'A criticism of AI ethics guidelines' (2020) XX(4) Informacios Tarsadolom <https://doi.org/10.22503/inftars.XX.2020.4.5> accessed 12 December 2024.

(i) to understand the capacities and limitations of the system; (ii) to remain aware of the possible tendency of over-relying on the output of the system; (iii) to correctly interpret the system's output; (iv) decide if the output should be disregarded, overridden or reversed; and (v) to use the 'stop' button or switch the system in a safe state. Notably, for systems that apply remote biometric identification, the results must be verified by at least two individuals.[256] Although the obligations seem detailed and specific, the question remains on how to comply with them in practice, for instance, how to deal with the problem of scalable oversight.

- *Accuracy, robustness and cybersecurity*. High-risk AI systems must achieve the appropriate level of those. This includes resilience against attempts by unauthorised third parties to alter the use, and outputs of the performance of the system; the provisions mention such techniques as data poisoning, model poisoning and adversarial examples or model evasion, which are very detailed and specific. The AI Act encourages the development of relevant benchmarks and measurement methodologies to assess compliance.[257]

In addition to the requirements for high-risk AI systems, the AI Act also sets the requirements for their providers, namely to:

- Ensure the requirements to the high-risk AI systems[258];

- Put a *quality management system* that ensures compliance with the AI Act and covers, *inter alia,* (i) strategy for regulatory compliance; (ii) techniques and procedures for the development, quality control and quality assurance of the high-risk AI system; (iii) technical specifications; (iv) post-monitoring system; (v) procedures for reporting on serious incidents; (vi) record-keeping procedures; and (vii) accountability framework[259];

256 Ibid, art 14.

257 Ibid, art 15.

258 Ibid, art 16

259 Ibid, art 17

- Conduct the *conformity assessment*[260] (the are limited exceptions for the reasons of public security, health, etc.)[261], adopt the EU declaration of conformity[262], and affix the CE marking[263];

- *Register* the deployer and the AI system (with certain exceptions) in the relevant EU database[264];

- *Keep the documentation* on the system available to the national competent authorities for at least 10 years after the system has been placed on the market[265] and also keep the automatically generated logs for at least 6 months[266];

- *Immediately correct* the system if there are reasons to believe it is not compliant, including the recall and withdrawal measures and investigations in collaboration with the deployers[267];

- *Appoint an authorized representative* in the EU if the provider is established in a third country[268].

The deployers of the high-risk AI systems shall ensure the use of the system in compliance with the providers' instruction, including due human oversight by trained personnel, monitoring, logging of data, and informing the users of the operation of the AI system.

There are also additional requirements for deployers which are the public authorities, Union institutions, bodies, offices and agencies. They shall (i) register in the EU database, request relevant authorisations from judicial bodies, ensure documenting and reporting on the use of remote biometric identification systems[269] and, (ii) if the AI system is intended for critical infrastructure, conduct a fundamental rights impact assessment[270].

---

260 Ibid, art 43

261 Ibid, art 46.

262 Ibid, art 47

263 Ibid, art 48

264 Ibid, art 49

265 Ibid, art 18

266 Ibid, art 19

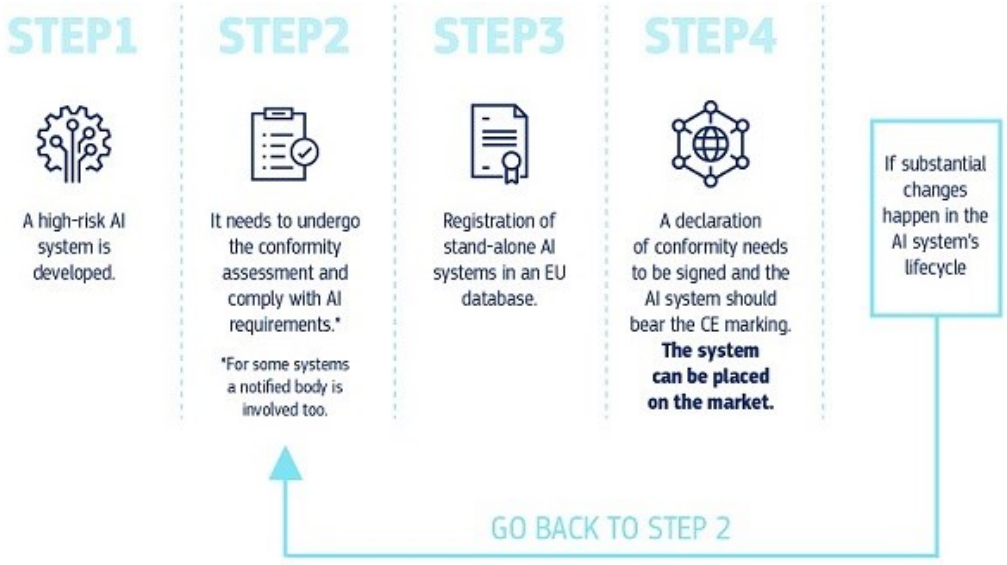
267 Ibid, art 20.

268 Ibid, art 22.

269 Ibid, art 26.

270 Ibid, art 27.

Lifecycle of a high-risk AI system



Credit: European Commission[271]

Transparency requirements to limited-risk AI systems. This category includes AI systems that are intended to interact directly with natural persons. For such systems article 50 of the AI Act provides the following specific obligations:

- Notify of the operation of the AI system (unless it is obvious);
- Mark all the synthesized audio, image, video, or text content in a machine-readable and detectable format. The technical solution must be adequate in terms of cost, state-of-the-art, robustness and reliability. The solution may be subject to standardization; and
- Additional specific notification requirements are set for providers of AI systems that infer emotions, conduct biometric categorization, or generate deepfake content.

The AI Act encourages the development of a code of practice and authorises the Commission to adopt general rules for compliance with the above requirements.

Regulation of AGI, ASI, frontier models and addressing existential risks. The AI Act regulates the general-purpose AI models (GPAI). They are defined as

> [A]n AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is

---

271 Commission, 'AI Act' (n 249).

> capable of competently performing 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, except AI models that are used for research, development or prototyping activities before they are placed on the market.[272]

The definition is quite broad and its 'capable of competently performing a wide range of distinct tasks' part matches the notion of artificial general intelligence (AGI). The AI Act does not specifically regulate or define the AGI, ASI, or frontier models.

The regulation of the GPAI represents an additional layer: a high-risk, limited-risk, or minimal-risk AI could be the GPAI.[273] The AI Act mentions large generative models as an example of GPAI.[274] It should be noted, however, that such models (e.g., ChatGPT and the like) are considered as 'weak' AI and not AGI. So, it would be inaccurate to consider the rules for GPAI as the regulation of AGI; vice versa, it could be that the AI Act applies a low standard for AGI.

Nevertheless, as rightly pointed out in recitals 101, such AI models may be 'the basis for a range of downstream systems, often provided by downstream providers that necessitate a good understanding of the models and their capabilities, both to enable the integration of such models into their products, and to fulfil their obligations under this or other regulations'. So, reasonable additional transparency measures are feasible. The AI Act makes an exception from these requirements for those providers that release their models under a free and open-source licence[275]. The measure is designed to encourage such practices and promote development.

Within the GPAI, the AI Act selects the category of AI systems with systemic risks. These are systems with 'high-impact capabilities' (i.e., the capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models[276]) and having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain[277].

---

272 Ibid, art 3(63).

273 Ibid, recitals 85.

274 Ibid, recitals 99.

275 Ibid, recitals 102.

276 Ibid, art 3(64)

277 Ibid, art 3(65).

The EU expressly states the intention to develop future-proof regulation of AI.[278] The AI systems that are at the frontier of technology are AGI/ASI. They have not been developed yet and the timelines are vague (according to one of the recent researches, the average predicted date is 2041[279]). Sooner or later, their development is nevertheless expected.[280] This category of AI is also the one that poses the most risks, including the existential one. The AI Act acknowledges this fact and also that at the moment there is a problem with the classification and regulation of the relevant systems due to their emergent and novel features and unclear capabilities. As pointed out in recitals 110 of the AI Act, the GPAI could pose systemic risks, including major accidents, disruptions of critical sectors and serious consequences to public health and safety; negative effects on democratic processes, and public and economic security; the dissemination of illegal, false, or discriminatory content. These risks increase with model capabilities and model reach. The risk could derive from, *inter alia,* the (i) model's reliability, fairness and security; (ii) the level of autonomy; (iii) access to tools, novel or combined modalities, release and distribution strategies; (iv) the potential to remove guardrails and other factors; from (v) issues of control related to alignment with human intent; (vi) chemical, biological, radiological, and nuclear risks, including for weapons development, design acquisition, or use; (vii) offensive cyber capabilities; (viii) risks from models of making copies of themselves or 'self-replicating' or training other models; (ix) the facilitation of disinformation or harming privacy with threats to democratic values and human rights; or (x) 'risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community'[281]. So, the AI Act articulates well the systemic and existential risks of AI systems and acknowledges them.

As regards categorizing the GPAI by risk, the AI Act states that it is hard to estimate the exact risk a system poses. Therefore, the AI Act applies a simplified approach: to treat any system with high-impact capabilities as risky and refine the criteria of GPAI models with systemic risk once the understanding of the technology improves.[282] As the initial capability benchmark, the AI Act uses floating point operations (so-called 'flops') and treats any GPAI as posing a systemic risk if the computation power used for its training exceeds 10^{25} flops.[283]

---

278 Ibid, recitals 138.

279 Rupert Macey-Dare, 'How Soon is Now? Predicting the Expected Arrival Date of AGI- Artificial General Intelligence' (2023) <https://ssrn.com/abstract=4496418> accessed 8 December 2024.

280 'Artificial general intelligence' (n 9).

281 AI Act (n 18), recitals 110.

282 Ibid, recitals 111.

283 Ibid, art 52(2).

The Commission may decide to categorise certain use cases as having systemic risk, including based on a qualified alert from the scientific panel.[284] Lastly, the AI Act envisions a methodology and the procedure for classification of GPAI with system risk[285] - this may provide more tailored qualification criteria in the future.[286]

The providers of the GPAI models shall[287]:

- Have technical documentation on the model, including its training, testing and the results of evaluation;
- Have information and documentation to be shared with providers integrating the model into their AI systems. This shall enable the providers, inter alia, to have a good understanding of the capabilities and limitations of the GPAI model;
- Have a policy on copyright and related rights;
- Make publicly available a summary of the content used for training the model; and
- Designate an EU representative (for non-EU providers).

The two first bullet points do not apply to the providers that released the model on an open-source basis, except for the providers of the models with systemic risks.

The providers of the GPAI models with systemic risks must additionally:

- perform model evaluation following 'standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks';
- assess and mitigate possible systemic risks at the Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk;

284 Ibid, recitals 111, art 51(1)(b).

285 Ibid, recitals 110, 111.

286 Ibid, art 51(1).

287 Ibid, art 53.

- keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them; and

- ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model.[288]

Standardisation. As rightly pointed out in recitals 121 of the AI Act, 'standardisation should play a key role to provide technical solutions to providers to ensure compliance with this Regulation, in line with the state of the art, to promote innovation as well as competitiveness and growth in the single market'. No matter how extensive and detailed is the AI Act, it may not reflect all the technicalities and specific use cases.

Developing a standard for AI is a hard task. The technology is still rapidly evolving and the scope of its application is immense. Yet, this is an important area of work that should articulate core practices and promote the implementation of the AI Act. The AI Act tackles this issue from different angles. Thus, it:

- Encourages the development of voluntary (non-binding) standards and best practices by including industry, academia, civil society and standardisation organisations[289];

- Sets out a procedure to call by the market for the official development of certain standards; the Commission may file such a relevant request with the European standardisation organisations[290];

- Identifies the primary areas that should be subject to a harmonized EU standard. This relates to the general-purpose AI models[291] and high-risk AI. Interestingly, the AI Act gives the providers the option to ensure the system's compliance either by following the standard or by confirming the fulfilment of the AI Act obligations otherwise.[292] The development of harmonized standards is envisioned as a joint process, with the participation of various stakeholders including industry, start-ups, SMEs, academia, and civil society, as well as the Fundamental Rights Agency,

---

288 Ibid, art 54.

289 Ibid, recitals 26.

290 Ibid, art 40.

291 Ibid, recitals 117.

292 Ibid, recitals 117.

ENISA, the European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC) and the European Telecommunications Standards Institute (ETSI); the international cooperation is also encouraged[293]; and

- Authorizes the Commission to adopt a common specification on compliance with the AI Act, as an exceptional fallback plan, if the development of harmonized standards fails.[294] The Commission's specification should be repelled once the harmonized standard is adopted[295].

Corporate self-regulation is especially important as regards the GPAI models, due to their novel, untested, and evolving nature. As it is stated in recitals 117 of the AI Act, 'the codes of practice should represent a central tool for the proper compliance with the obligations' under the AI Act and 'providers should be able to rely on codes of practice to demonstrate compliance with the obligations', until a harmonized standard is published.[296] In this process, the AI Act highlights the importance of considering the international approach and practices.[297]

To reinforce this idea, the AI Act contains a dedicated article 56 that provides for the development of the relevant code of practice by the AI Office with the involvement of various stakeholders: civil society organisations, industry, academia and other relevant stakeholders, such as downstream providers and independent experts. The code of practice must be adopted by 2 May 2025 and is designed to be of voluntary application. If the code of practice is not adopted until 2 August 2025, however, the Commission may intervene and adopt common rules for the implementation of the obligations set out by the AI Act.

Lastly, the AI Act provides for voluntary codes of conduct for non-high-risk AI systems[298] and the Commission's implementation guidelines[299].

---

293 Ibid, recitals 150.

294 Ibid, recitals 121; art 41.

295 Ibid, art 41(4).

296 Ibid, art 53(4).

297 Ibid, recitals 116.

298 Ibid, 95

299 Ibid, 96.

Exceptions and regulatory sandboxes. As discussed previously, the jurisdictions are balancing between controlling and promoting the technology, thus, granting such regulatory lacunae is a fine-tuning mechanism to achieve both.

The AI Act provides exceptions from its provisions (in full or in part) to AI systems, models, and components that are:

- free and open-source (does not apply to GPAI models with systemic risk)[300]. The rationale of the exception is that such systems can contribute to research and innovation in the market and can provide significant growth opportunities for the Union economy.[301] The AI Act does not apply to such systems unless the system is a prohibited or high-risk system, or it falls under the transparency obligations as a limited-risk system[302];

- for scientific research and development[303] – the idea is to support innovation, respect the freedom of science, and not undermine the research and development activity[304]; or

- deployed by natural persons in the course of their pure non-professional activity[305].

The final thing that draws interest is the regime of regulatory sandbox. The approach is not novel and is generally expected for AI. Regulatory sandboxes are always a compromise between the promotion of technology and lowered compliance standards. The regime creates an 'isolated', more controlled area, where the market can test their ideas in the lighter regulatory environment. This is yet another example of the risk-based approach applied by the EU.

The AI Act defines 'AI regulatory sandbox' as a controlled framework set up by a competent authority which offers the possibility to develop, train, validate and test an innovative AI

---

300 Ibid, recitals 89.

301 Ibid, recitals 102.

302 Ibid, art 2(12).

303 Ibid, art 2(6).

304 Ibid, recitals 25.

305 Ibid, 2(10).

system, according to a sandbox plan for a limited time under regulatory supervision.[306] The regime is specified in Chapter VI of the AI Act and includes the following main features:

- the providers remain liable to third parties for damages, however, if the provider follows in good faith the sandbox regime, it is released from administrative liability[307];
- subject to conditions, AI providers may use personal data collected for other purposes to develop, train and test their systems; and
- the providers may test their AI systems in real-world conditions, subject to a testing plan and under the authorities' supervision.[308] In such testing, however, the personal data exception does not apply and the provider must have the informed consent of the subjects of the testing[309].

The AI Act requires each Member State to establish at least one regulatory sandbox by 2 August 2026[310], thus indicating a clear intention to inspire AI development.[311] What is important and very mindful, the regulatory sandboxes aim to 'contributing to evidence-based regulatory learning'[312]. The EU recognizes that AI is hard to regulate and aims to improve the regulation as the implementation practice accumulates.

As discussed above, the AI Act provides for full-scope AI governance and directly regulates many technical safety requirements. It incorporates the AI ethics accepted by the EU via the general and AI-specific protection of human rights and personal data rights. The AI Act effectively applies the risk-based approach and often chooses the protection of human rights over the government or corporate interests.

---

306 Ibid, art 3(55).

307 Ibid, 57(12).

308 Ibid, art 60.

309 Ibid, art 61.

310 Ibid, art 57(1).

311 Ibid, art 57(5).

312 Ibid, art 57(9)(d).

3.3.2 US

(a) Regulatory framework

The US regulation of AI is extensive and is rapidly developing. However, the US is still reluctant to introduce binding federal horizontal regulation of AI. A major part of the current regulation is formed by white papers, plans and guidelines having no binding effect; some industry-specific legislation may extend to AI. The US is particularly active in the regulation of the use of AI by state agencies – this sphere has binding regulations and is quite detailed. A significant part of AI regulation relates to the issues of national security, defence, and the threats of weaponized AI. The US is very focused on standardization, including building global standards with the US as the leader of the process. The jurisdiction is also notable for strong corporate self-regulation; relevant non-legislative commitments build a standalone compliance layer that somewhat fills in the regulatory gap.

The regulatory process started in 2016, about the same time as in the EU. That year the National Science and Technology Council (NSTC), the principal means by which the President's administration coordinates science and technology policy, issued the report 'Preparing for the Future of Artificial Intelligence'[313] and The National Artificial Intelligence Research and Development Strategic Plan[314]. The same year the NSTC Subcommittee on Machine Learning and Artificial Intelligence was formed. The report surveyed the current state of AI, its existing and potential applications and made recommendations for specific further actions, including the adapting of the regulation to the AI technologies, e.g., to autonomous vehicles.[315] The report formulated the intended regulatory approach as risk-driven, meaning that the AI regulations are needed to the extent the existing regulation is not sufficient and may not be adapted to address these AI-specific risks.[316] It was focused on safety, ethical training of AI, and global security considerations. The plan defined the development strategy, including long-term investments, ensuring secure AI systems,

---

313 National Science and Technology Council (NSTC), Committee on Technology, 'Preparing for the future of artificial intelligence' (October 2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf> accessed 8 December 2024 (NSTC AI Report).

314 National Science and Technology Council (NSTC), Networking and Information Technology Research and Development Subcommittee, 'The national artificial intelligence research and development strategic plan' (October 2016) <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf> accessed 8 December 2024 (NSTC 2016 AI Plan).

315 Ed Felten, Terah Lyons, 'The Administration's Report on the Future of Artificial Intelligence' (the White House, 12 October 2016) <https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence> accessed 8 December 2024.

316 NSTC AI Report (n 313), 11.

research ethics, building standards and benchmarks, and developing shared public datasets.[317]

In 2018, the National Security Commission on Artificial Intelligence (NSCAI) was established having the goal 'to consider the methods and means necessary to advance the development of artificial intelligence, machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States'[318].

In 2019 Donald Trump adopted Executive Order No. 13859: Maintaining American Leadership in Artificial Intelligence (EO 13859).[319] The order stressed the importance of maintaining AI leadership, protecting safety, civil liberties, privacy and American values, fostering public trust and confidence in AI technologies, and the development of relevant standards.

The order was followed by the position paper of the NIST[320] 'U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools'[321]. The document evidenced the active ongoing AI-relevant standardization work and indicated the existence of some standards for concepts and terminology, metrics, safety, and risk management but also the lack of a relevant standard for trustworthiness and societal and ethical considerations and the standards for data sets. [322] The plan intended to maximise the use of the voluntary consensus standards process, both to save development costs and utilise the private sector expertise.[323] It suggested focusing the federal initiative on the development of standards in the spheres of trustworthiness, reliance and robustness of AI systems as well as 'strategically engaging' internationally.[324] The same year the Department of Defense released 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense'.[325] The document relates to the application of AI as a weapon

---

317 NSTC 2016 AI Plan (n 314), 11.

318 Gabriel Honrada, 'America's AI edge fading fast to China' (Asia Times, 19 September 2022) <https://asiatimes.com/2022/09/americas-ai-edge-fading-fast-to-china/> accessed 8 December 2024.

319 EO 13859 (n 17).

320 The National Institute of Standards and Technology, US Department of Commerce (NIST).

321 NIST, 'U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools' (submitted on 9 August 2019) <www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf> accessed 8 December 2024 (NIST 2019 Plan).

322 Ibid, 11(table 1), 12(table 2), 13.

323 Ibid, 18.

324 Ibid, 24.

325 Defense Innovation Board, 'AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (2019) <https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF> accessed 8 December 2024.

and states the technology is 'neither inherently positive nor negative'[326]; among its principles, it requires the use of AI systems to be responsible, equitable, traceable, reliable, and governable. In 2019, the National Artificial Intelligence Initiative Act and the AI in Government Act were introduced in the Senate. It required certain federal activities in the sphere of AI (investment, support, designations), such as the creation of AI-specific bodies and research institutes to promote the development of AI, including building towards trustworthy AI and research in technology ethics. The drafts also required the Office of Management and Budget (OMB) to issue a memorandum to be used as guidance to federal agencies on the use of AI. The AI in Government Act was adopted in December 2020.[327] The draft National Artificial Intelligence Initiative Act accomplished a lengthy path and finally became law in January 2021 as part of the National Defense Authorization Act for Fiscal Year 2021[328].

In 2020, in furtherance to EO 13859, the Office of Management and Budget released the Guidance for Regulation of Artificial Intelligence Applications[329]. It formulated 10 key principles of the 'stewardship' of AI applications: public trust in AI, public participation, scientific integrity and information quality, risk assessment and management, benefits and costs, flexibility, fairness and non-discrimination, disclosure and transparency, safety and security, and interagency coordination. The guidance also formulated non-regulatory approaches to AI that include sector-specific policy guidance or frameworks, pilot programs and experiments, and voluntary consensus standards. The guidance mentioned the importance of protection of human rights (civil liberties), privacy (personal data), and safety as well as addressing the AI alignment risks. It even mentions the application of a risk-based approach and experimental regimes – so, many of the announced regulatory features resemble those of the EU.

In 2021, the NSCAI issued its final report, a 750+ page document that presented 'an integrated national strategy to reorganize the government, reorient the nation, and rally [its] closest allies and partners to defend and compete in the coming era of AI-accelerated

---

326 Ibid, 5.

327 Consolidated Appropriations Act, 2021, H.R.113, Public Law No. 116-260, 134 stat 1182, Division U, Title I – AI in Government Act of 2020.

328 William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, H.R.6395, Public Law 116-283, Division E – National Artificial Intelligence Initiative Act of 2020.

329 Executive Office of the President, Office of Management and Budget (OMB), 'Memorandum M-21-06 for the heads of executive departments and agencies: Guidance for Regulation of Artificial Intelligence Applications' (17 November 2020) <www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf> accessed 8 December 2024 (OMB M-21-06).

competition and conflict'[330]. The report is focused on national security and the issues of AI applications in warfare and intelligence. Nevertheless, it proposes some draft legislative language, including regarding upholding the democratic values: privacy, civil liberties, and civil rights in uses of AI for national security and includes increasing public transparency about the use of AI, conducting risk and impact assessment, third-party testing, and creating a task force to identify the policy and legal gaps.[331]

In December 2022, the Global Catastrophic Risk Management Act was adopted.[332] Among other sources, it refers to 'global catastrophic and existential threats' – intentional or accidental threats arising from the use and development of emerging technologies[333]. Its provisions are considered to extend to AI technology[334]. By the same package, Congress adopted the Advancing American AI Act that required the OMB to issue implementation guidelines for government agencies.[335]

The next major step was the Blueprint for an AI Bill of Rights published in October 2022 by the Office of Science and Technology Policy (OSTP).[336] As it comes from the title, the document is focused on the rights of natural persons in the AI environment. It sets out five main principles: safe and effective systems, algorithmic discrimination protection, data privacy, notice and explanation and human alternatives, and consideration and fallback. The AI Bill of Rights is a non-binding document and is aimed to 'guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence'[337].

---

330 NSCAI, 'The Final Report' <https://reports.nscai.gov/final-report/> accessed 8 December 2024, 11.

331 Ibid, 697–701.

332 James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, 23 December 2022, H.R.7776, Public Law 117-263, Title LXXIII, Subtitle A – Global Catastrophic Risk Management Act of 2022.

333 Ibid, sec 7302(7).

334 'Portman, Peters Introduce Bipartisan Bill to Ensure Federal Government is Prepared for Catastrophic Risks to National Security' (Homeland Newswire, 25 June 2022) <https://web.archive.org/web/20220625220611/https://homelandnewswire.com/stories/627890045-portman-peters-introduce-bipartisan-bill-to-ensure-federal-government-is-prepared-for-catastrophic-risks-to-national-security> accessed 9 December 2022.

335 James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, 23 December 2022, H.R.7776, Public Law 117-263, Title LXXII, Subtitle B – Advancing American AI Act.

336 AI Bill of Rights (n 19).

337 Ibid, 3.

The NSTC updated the National Artificial Intelligence Research and Development Strategic Plan in May 2023.[338] The plan focuses on human-AI collaboration, understanding and addressing the ethical, legal and societal applications of AI, safety and security, data sharing, and the developing of standards.

In July 2023, President Biden's administration secured a voluntary commitment from seven major IT companies operating in the US aimed at addressing AI risks. Later the same year, eight more companies adhered to it, so the final list includes Amazon, Adobe, Anthropic, Cohere, Google, IBM, Inflection, Meta, Microsoft, Nvidia, OpenAI, Palantir, Salesforce, Scale AI, Stability AI[339], and later Apple[340]. The commitment encompassed a variety of obligations, such as security testing before the release, sharing information on the AI risks and their management with the industry, public bodies, society and academia; prioritization of cybersecurity; informing users on the operation of AI, labelling AI-generated content; and prioritising the research on societal risks such as biases, discrimination, and privacy.

On 30 October 2023, Biden signed Executive Order No. 14110: On the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[341], which is one of the key pieces of federal legislation on AI. The executive order reinstated the main principles of the AI policy (this time, eight):

1. Ensuring AI is safe and secure;

2. Promoting responsible innovation, competition, and collaboration to secure AI leadership;

---

338 NSTC, Select Committee on Artificial Intelligence: Report 'National Artificial Intelligence Research and Development Strategic Plan 2023 Update' <www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf> accessed 9 December 2024 (2023 AI RD Plan).

339 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI' (WH.GOV, 21 July 2023) <www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/> accessed 9 December 2024; 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI' (WH.GOV, 12 September 2023) <www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/> accessed 9 December 2024.

340 'FACT SHEET: Biden-Harris Administration Announces New AI Actions and Receives Additional Major Voluntary Commitment on A' (WH.GOV, 26 July 2024) <www.whitehouse.gov/briefing-room/statements-releases/2024/07/26/fact-sheet-biden-harris-administration-announces-new-ai-actions-and-receives-additional-major-voluntary-commitment-on-ai/> accessed 9 December 2024.

341 EO 14110 (n 27).

3. Supporting American workers as part of the responsible development of AI;

4. Advancing equity and civil rights;

5. Protection of AI-users, consumer protection;

6. Protection of privacy and civil liberties;

7. Use of AI by the Federal Government and thereby building internal expertise; and

8. The Federal Government shall lead the way to global societal, economic, and technological progress, promoting AI safety and embracing AI benefits.

The executive order operates with AI-specific terms such as 'AI model', 'AI red-teaming', 'dual-use foundation model', 'machine learning', 'model weight', 'self-healing network', and 'testbed' and the technical features of AI.[342] EO 14110 scheduled a broad scope of regulatory activities, including the development of guidelines on safe, secure and trustworthy AI systems; defining AI capabilities that have significant threats (critical infrastructure, nuclear); and developing relevant guardrails, providers' reporting requirements, identification of foreign resellers; defining measures to manage AI in critical infrastructure and cybersecurity; addressing the threats of the malicious use of AI to develop chemical, biological, radiological and nuclear materials or weapons; labelling synthesized content; regulation of dual-use foundation models; and use of federal data for training.[343] Besides, the planned activities included evaluating AI-related workforce disruptions[344] and advancing equity and civil rights, including assessing the use of AI in the judicial system[345], and the protection of consumers, patients, passengers, and students[346]. The executive order does not

342 Ibid, s 3.

343 Ibid, s 4.

344 Ibid, s 6

345 Ibid, s 7.

346 Ibid, s 8.

set any direct regulations but rather maps further work of the state agencies. After one year, the White House reported 100+ completed tasks[347], including:

- establishment the US AI Safety Institute and signing collaboration agreements with Anthropic and OpenAI regarding major new AI models[348],
- developing draft guidance for generative AI and dual-use foundation models[349];
- releasing the National Security Memorandum on Artificial Intelligence (NSM AI)[350] and the accompanying Framework to Advance AI Governance and Risk Management in National Security (NSM AI Framework)[351] that sets internal binding guidelines for the use of AI by government agencies. A brief 14-page document contains the prohibited and high-risk AI use cases, as well as minimum risk management and monitoring practices;
- securing voluntary commitments to combat image-based sexual abuse[352], which is currently one of the most harmful uses of AI[353], including responsible sourcing of datasets, incorporating feedback loops and stress-testing strategies and removing

---

347 'Fact Sheet: Key AI Accomplishments in the Year Since the Biden-Harris Administration's Landmark Executive Order (The White House, 30 October 2024) <www.whitehouse.gov/briefing-room/statements-releases/2024/10/30/fact-sheet-key-ai-accomplishments-in-the-year-since-the-biden-harris-administrations-landmark-executive-order/> accessed 9 December 2024.

348 'U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI' (NIST, 29 August 2024) <www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research> accessed 9 December 2024.

349 'Department of Commerce Announces New Guidance, Tools 270 Days Following President Biden's Executive Order on AI' (NIST, 26 July 2024) <www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following> accessed 9 December 2024.

350 President, Communication to Federal agencies 'Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence' (the White House, 24 October 2024) <www.whitehouse.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/> accessed 9 December 2024 (NSM AI).

351 US Government, 'Framework to Advance AI Governance and Risk Management in National Security <https://ai.gov/wp-content/uploads/2024/10/NSM-Framework-to-Advance-AI-Governance-and-Risk-Management-in-National-Security.pdf> accessed 9 December 2024 (NSM AI Framework).

352 'White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse' (the White House, 12 September 2024) <www.whitehouse.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse/> accessed 9 December 2024.

353 'Fact Sheet: Key AI Accomplishments in the Year Since the Biden-Harris Administration's Landmark Executive Order' (the White House, 30 October 2024) <www.whitehouse.gov/briefing-room/statements-releases/2024/10/30/fact-sheet-key-ai-accomplishments-in-the-year-since-the-biden-harris-administrations-landmark-executive-order/> accessed 9 December 2024.

nude images from AI training datasets; the commitment was signed by Adobe, Anthropic, Cohere, Microsoft, Open AI, and Common Crawl;

- adopting the non-binding 'Artificial Intelligence And Worker Well-being: Principles and Best Practices For Developers And Employers' by the Department of Labour[354];
- signing the AI Convention; and
- issuing the Risk Management Profile for Artificial Intelligence and Human Rights by the US Department of State[355].

In March 2024, in the implementation of the Advancing American AI Act and EO 14110, the OMB issued the memorandum 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence'[356]. The same month, the Federal AI Governance and Transparency Act[357] was introduced in the House of Representatives aimed to harmonise the AI governance measures to be applied by state agencies, including building the federal AI system. The draft is still pending.

NIST plays a significant role in the formation of AI regulation by developing standards. The five most important issues as of today include:

- NIST AI.100-1. Artificial Intelligence Risk Management Framework (AI RMF 1.0)[358];
- NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[359];

---

354 U.S. Department of Labor, 'Artificial Intelligence And Worker Well-being: Principles And Best Practices For Developers And Employers' <www.dol.gov/general/AI-Principles> accessed 9 December 2024 (AI Principles for Workers).

355 U.S. Department of State, Bureau of Cyberspace and Digital Policy 'Risk Management Profile for Artificial Intelligence and Human Rights' (25 July 2024) <www.state.gov/risk-management-profile-for-ai-and-human-rights/> accessed 9 December 2024.

356 Executive Office of the President, Office of Management and Budget, 'Memorandum No. M-24-10 for the Heads of Executive Departments and Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence' (28 March 2024) <www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf> accessed 9 December 2024 (OMB M-24-10).

357 Federal AI Governance and Transparency Act, H.R.7532, Draft <www.congress.gov/bill/118th-congress/house-bill/7532/all-actions> accessed 9 December 2024.

358 NIST AI RMF (n 87).

359 NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. July 2024 <https://doi.org/10.6028/NIST.AI.600-1> accessed 9 December 2024.

- NIST SP 800-218A. Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile[360];
- Draft NIST AI 100-4. Reducing Risks Posed by Synthetic Content an Overview of Technical Approaches to Digital Content Transparency[361]; and
- Draft NIST AI 800-1. Managing Misuse Risk for Dual-Use 4 Foundation Models[362].

Although the analysis of the state law is outside the subject of research, a brief overview for the sake of completeness is feasible. In January 2023, New York adopted the New York City Bias Audit Law[363] that prohibited automated tools to hire or promote employees, unless a third-party audit of the system for bias is conducted. In March 2024, the State of Tennessee adopted the so-called ELVIS Act[364] addressing audio deepfakes and voice cloning and aimed at protecting copyright and artists' rights against unauthorised use of their voice. In February 2024, the draft Safe and Secure Innovation for Frontier Artificial Intelligence Models Act[365] was introduced in California having the regulatory goal to address the existential risks posed by the most powerful AI models. In September 2024, the bill was vetoed by the Governor. The veto was explained by 'a false sense of security about controlling this fast-moving technology'[366] that the law may give to the citizens. In March 2024, the Artificial Intelligence Policy Act was adopted in Utah; it establishes liability for companies that do not disclose

---

360 NIST SP 800-218A. Secure Software Development Practices for Generative AI and Dual-Use Foundation Models ipd, April 2024 <https://doi.org/10.6028/NIST.SP.800-218A.ipd> accessed 9 December 2024.

361 NIST AI 100-4. Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency (Draft for Public Comment) <https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf> accessed 9 December 2024.

362 NIST AI 800-1. Managing Misuse Risk for Dual-Use Foundation Models (Initial Public Draft) <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf> accessed 9 December 2024.

363 'Streamline compliance with Holistic AI' (NYC Bias Audit) <www.nycbiasaudit.com/> accessed 9 December 2024.

364 Margaret R. Szewczyk, Lucas Amodio, C|EH, 'Artificial Intelligence and Copyrights: Tennessee's ELVIS Act Becomes Law' (Armstrong&Teasdale, 27 March 2024) <www.armstrongteasdale.com/thought-leadership/artificial-intelligence-and-copyrights-tennessees-elvis-act-becomes-law/> accessed 9 December 2024.

365 Helena Engfeldt, Garrett Stallins, 'The Nuts and Bolts of the proposed California Safe and Secure Innovation for Frontier Artificial Intelligence Models Act' (Connect on Tech, 12 August 2024) <www.connectontech.com/the-nuts-and-bolts-of-the-proposed-california-safe-and-secure-innovation-for-frontier-artificial-intelligence-models-act/> accessed 9 December 2024.

366 Wendy Lee, 'Gov. Gavin Newsom vetoes AI safety bill opposed by Silicon Valley' (Los Angeles Times, 29 September 2024) <www.latimes.com/entertainment-arts/business/story/2024-09-29/gov-gavin-newsom-vetoes-ai-safety-bill-scott-wiener-sb1047> accessed 9 December 2024.

their use of generative AI when required under consumer law and for committing criminal offences using generative AI[367].

To summarize, the US currently does not have binding regulation of AI although the legislative process is very active. Mostly, horizontal AI regulation applies soft law mechanisms such as guidelines, frameworks and non-binding standards; certain binding rules apply to the use of AI by governmental agencies. There is a practice of voluntary commitments from key AI tech players that to some extent cover up the regulatory gaps.

(b) Regulation of subject areas

(I) Human rights

The US approach to protecting human rights in the sphere of AI was well-formulated in EO 13859: 'The United States must foster public trust and confidence in AI technologies and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people'[368]. The protection of human rights is an element of the US policy to promote the advancement of AI technologies.

Although the US legislation refers to human rights (also referred to as civil rights and civil liberties) as a priority[369], this does not seem to be the top regulatory priority. The preference is made to safety and national security. Another aspect that stands out is the attention of the US to workforce disruptions that could be caused by AI technology.[370]

The key legal sources related to human rights are:

1. The AI Bill of Rights;

2. NIST AI RMF plus the Risk Management Profile for Artificial Intelligence and Human Rights by the US Department of State; and

3. The AI Convention.

---

367 Daniel K. Alvarez et al., 'U.S. State AI Update: Utah Enacts First State Generative AI Transparency Law – California, Colorado and Connecticut Are Close Behind' (Willkie Farr & Gallagher LLP, 8 May 2024) <www.willkie.com/-/media/files/publications/2024/05/us_state_ai_update_utah_enacts_first_state_generative_ai_transparency_law.pdf> accessed 9 December 2024.

368 EO 13859 (n 17), s 1(d).

369 For instance, the attitude to human-centric approach is mentioned in section 1(d) of the National Security Memorandum on Artificial Intelligence – NSM AI (n 350).

370 See e.g., EO 13859 (n 17), s 7; AI Principles for Workers (n 354).

The AI Bill of Rights was issued first and it is probably the primary human-centric document up to date. It provides for a list of AI-specific human rights to be respected:

- Notification of the operation of an AI system.[371] The AI Bill of Rights elaborates on the typical notification approach and requires the providers to explain how and why the system contributes to the outcomes the impact the user, including 'plain language', 'brief and clear' description of the use of the system; it also requires the AI systems to provide explanations that are 'technically valid, meaningful and useful', 'tailored to the purpose' and 'tailored to the target'. The requirement is formulated quite demanding given the obscure nature pertained to AI; still, this defines the compliance maxima;

- No algorithmic discrimination.[372] To provide this, the AI Bill of Rights requires conducting an independent algorithmic impact assessment, including disparity testing results and mitigation information, and making public the results, ensuring accessibly to people with disabilities, and ongoing monitoring[373];

- Human alternative, consideration and fallback.[374] The AI Bill of Rights requires the providers to enable users to opt out of automated systems in favour of 'timely and not burdensome' human alternative[375]. This is interesting since the EU legislation provides for that right only in limited cases. Additionally, the provider must ensure access to human consideration, remedy, and an escalation process if the automated system fails or errs. Meaningful human oversight of decisions tailored to the nature and risk of the decisions is required in sensitive domains such as criminal justice, education and health, including proper training of the operators, and human consideration before any high-risk decision.

NIST AI RMF is a voluntary standard for AI risk management and is not focused on human rights. In the relevant part, it mentions the issues of privacy, non-discrimination (managing biases), and explainability.[376] The Risk Management Profile for Artificial Intelligence and Human Rights was developed by the US Department of State as non-binding guidance on the application of NIST AI RMF and views the subject from a specific angle, namely from

---

371 AI Bill of Rights (n 19), 40.

372 Ibid, 3.

373 Ibid, 27.

374 Ibid, 7.

375 Ibid, 49.

376 NIST AI RMF (n 87), 15–17.

international human rights. The profile explains it with three reasons: these rights are universally applicable and already functioning; they apply both to the public and private sectors, and many risks posed by AI are human rights-related. The document represents a compliance instruction that compliments the NIKE AI RMF. Following the structure of AI RMF, it discusses the four organizational functions (govern, map, measure, and manage) and contains general recommendations aimed at mindful development and implementing AI systems, such as risk evaluation, impact assessment, gathering feedback, publicly communicating incidents, and establishing an appropriate redress mechanism.

The draft NIST AI 100-4 raises the problem of synthetic content provenance and explores different methods of explicit and covert methods of watermarking and detection of different types of media. Although the standard stresses the importance of notification of users of the synthetic content, it does not suggest actionable measures and is mostly aimed at exploring the up-to-date state of the art in the area. NIST.AI.600-1 reinstates the importance of content control in generative AI systems, yet does not go further than a general statement of the goal to protect human rights.

The provisions of the AI Convention have already been discussed in section 3.3.1(b)(I) concerning the EU regulation. The protected rights include: the integrity of democratic processes and the respect for the rule of law, transparency and oversight, accountability and responsibility, non-discrimination, privacy and protection of personal data, effective remedies, respect of rights of persons with disabilities and children, and safeguarding the existing, ‘classic’, human rights.

To summarise, the US legislation on the protection of AI-specific human rights is mostly based on the AI Bill of Rights and the AI Convention. Although there are other legislative and soft law instruments (whitepapers, standards) that highlight the importance of protection of human rights in the AI environment, they are not binding or specific in this regard. Although the AI Bill of Rights is non-binding, it sets important protective national standards. The US legislation revolves around the same basic ideas on the protection of human rights as the EU law but substantially lacks a detailed mechanism for compliance and enforcement. A notable feature is that the US legislation seems to view the protection of human rights as the logical outcome of safe and trustworthy AI systems and mainly focuses on the latter.

(II) Personal data

This aspect is very peculiar for the US since at the moment the US does not have horizontal federal legislation on the protection of personal data. In the US this concept usually forms

part of a broader concept of privacy or data privacy. The relevant legislation, as well as the scope of rights, is statute-specific (e.g., interaction with government, healthcare).[377] To a certain extent, this gap is covered by the legislation of the states, e.g., by the California Consumer Privacy Act effective from 2020.[378] The is also the NIST Privacy Framework[379], a voluntary document that facilitates privacy practices by the organization. Although the framework refers to widely accepted principles of personal data processing, e.g., consent-based processing[380], it may not be considered an adequate substitute for data privacy legislation. Therefore, there is no solid regulatory grounds for building AI-specific rights upon. The need for regulation of this area, including specifically for AI, is recognized and reflected in legislation. For instance, the AI Bill of Rights refers the data privacy within the five basic AI rights.

The AI Bill of Rights requires ensuring that data collection confirms reasonable expectations and that only strictly necessary data is collected; having the subject's use-specific consent for data processing in 'brief, understandable plain language', and having enhanced protections for the sensitive data (health, work, education, finance), the processing of such data must be subject to ethical review.[381] The same as in the EU, the AI Bill of Rights prohibits the unchecked surveillance of persons; such practices are subject to 'heightened oversight' and a pre-deployment human rights impact assessment. More to that, such practices are not allowed in education, work, housing, and other inappropriate contexts. The compliance with the above obligations was intended via an independent evaluation of the system.[382]

Another major piece of AI legislation that relates to personal data is Biden's Executive Order EO 14110 having a dedicated section 9 for this topic. The section discusses privacy-enhancing technologies (PETs) and mainly focuses on data protection by public agencies. The provisions are more focused on compliance measures than the scope of relevant rights.

---

377 Paul F. Pittman, Abdul Hafiz, Andrey Hamm, 'Data Protection Laws and Regulations USA 2024' (ICLG, 31 July 2024) <https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa> accessed 9 December 2024.

378 Gavin Wright, 'California Consumer Privacy Act (CCPA)' (TechTarget, December 2023) <www.techtarget.com/searchcio/definition/California-Consumer-Privacy-Act-CCPA> accessed 9 December 2024.

379 NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management, version 1.0 (16 January 2020) <https://doi.org/10.6028/NIST.CSWP.01162020> accessed 9 December 2024.

380 Ibid, 23.

381 AI Bill of Rights (n 19), 6.

382 Ibid, 35.

NIST AI RMF recognises that 'AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals'[383] and offers the use of privacy-enhancing technologies, as well as data-minimizing methods such as de-identification and aggregation for certain model outputs. NIST.AI.600-1 uses the terms 'personal identifiable information' (PII) (as an analogue of personal data) and 'sensitive information' and stresses the data privacy risks that generative AI systems may pose. Even so, its practical recommendation does not go far beyond the general requirement for 'privacy enhanced' and 'safe' systems, apart from the technical measures.[384] The 2023 AI RD Plan recognizes the need for additional work to address privacy concerns[385], including the advancement of technological and governmental methods[386] .

Article 11 of the AI Convention requires each part to adopt and maintain measures to ensure, that the privacy rights of individuals are protected and the relevant effective guarantees and safeguards are in place during the whole lifecycle of AI systems. It could be assumed that the US will adopt national legislation to endorse these provisions.

Conclusion. The federal legislation on the protection of personal data is still being developed. Although the federal AI legislation refers to data protection as one of its priorities, there is so far no binding dedicated federal regulation in this respect. AI-specific legislation is mostly formed by the AI Convention and the AI Bill of Rights. Other AI legal sources stress the need for research and legislative initiatives in the area, while standards are focused on the technical side of the matter. There is sector-specific privacy legislation as well as relevant legislation of states, including AI-specific, that partly covers the regulatory gap.

(III) Safety

Safety is the cornerstone of the US legislation on AI. It is the predominant topic in the majority of relevant acts. Promoting reliable, robust, and trustworthy AI applications is considered to be a key to building public trust and validation of AI and, consequently, to the continued adoption and acceptance of AI.[387] In the existing AI acts, a significant part of the

---

383 NIST AI RMF (n 87), 17.

384 NIST.AI.600-1 (n 359).

385 2023 AI RD Plan (n 338), 13.

386 Ibid, 28.

387 OMB M-21-06 (n 329), 3.

discussion on AI safety relates to the use of AI for warfare and intelligence purposes[388] – these are outside the scope of this research.

The building of the AI safety framework in the US is a multi-stream process. First, the US adopts AI-specific legislation, including the AI Bill of Rights, executive orders, R&D plans and memoranda envisioning the safety system for AI. Second, the US is actively developing standards, e.g., the NIST AI RMF. Third, the US encourages corporate self-regulation and aims to implement relevant experience in standards. Fourth, the US facilitated voluntary commitments from major AI corporate players. Five, the US is very cautious about introducing any binding obligations and is currently relying on soft law mechanisms; there are, however, guidelines binding on government agencies. The key relevant provisions are discussed below.

The policy on AI safety is mostly structured through the concept of trustworthy AI. The elements of trustworthy AI were formulated in Executive Order 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government[389]. The order applies to the use of AI by state agencies and requires AI systems to be:

- Lawful and respectful of the Nation's values;
- Purposeful and performance-driven – the benefits of using AI should significantly outweigh the risks, while the risks can be assessed and managed;
- Accurate, reliable, and effective – the application of AI must be consistent with the purpose for which they were trained;
- Safe, secure, and resilient;
- Understandable - operations and outcomes of AI applications should be sufficiently understandable by subject matter experts, users, and others, as appropriate;
- Responsible and traceable – inputs and outputs of particular AI applications should be well documented and traceable, as appropriate and to the extent practicable;
- Regularly monitored;

---

388 E.g., the Final Report of the NSCAI of 2021 (n 330).

389 Executive Order No. 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, 3 December 2020, 85 FR 78939 (EO 13960).

- Transparent - disclosing relevant information regarding the use of AI; and
- Accountable - implementing and enforcing appropriate safeguards for the proper use and functioning of AI applications, as well as monitoring, auditing, document compliance, and ensuring appropriate training of personnel.[390]

Generally, the US follows the risk-based approach in the regulation of AI safety:

> It is not necessary to mitigate every foreseeable risk; in fact, a foundational principle of regulatory policy is that all activities involve tradeoffs. Instead, a risk-based approach should be used to determine which risks are acceptable and which risks present the possibility of unacceptable harm, or harm that has expected costs greater than expected benefits.[391]

The AI Bill of Rights formulates the principle of 'safe and effective systems', it comes first in the document. The AI Bill of Rights requires the system to undergo pre-deployment testing, be subject to ongoing monitoring, and adhere to domain-specific standards. The result of this procedure may lead to removing the system from use.[392] Further, they should be designed to 'proactively protect from inappropriate or irrelevant data use in the design, development, and deployment'. The systems should be subject to independent evaluation. In continuation of these basis provisions, the AI Bill of Rights requires clear organizational oversight through governance structures and procedures.[393] Also, the document draws attention to the relevance and high quality of the training data, including the tracking and reviewing of derivative data.[394] Finally, the systems should be designed to allow their independent evaluation.

The OMB M-21-06[395] sets the obligatory requirements for state agencies on their use of AI systems. It focuses on the issues of information quality, risk assessment and management, preserving innovation (proportionality of measures and costs vs. benefits assessment), non-discrimination, disclosure and transparency, safety and security as regards operating AI systems as intended, and ensuring confidentiality, integrity, and systemic resilience. The

---

390 Ibid, s 3.

391 OMB M-21-06 (n 329), 11.

392 AI Bill of Rights (n 19), 5.

393 Ibid, 19.

394 Ibid, 20.

395 OMB M-21-06 (n 329).

guidance rather formulates the general approach and the desired outcomes than provides specific compliance measures.

The 2023 AI RD Plan highlights the following elements of trustworthy AI systems: safe, secure, reliable, resilient, interpretable, and transparent, preserving privacy and avoiding inappropriate bias; it calls for the development of relevant standards and benchmarks[396] also taking into account 'the tradeoff among the mentioned trustworthiness characteristics'[397]. In the relevant part, the document is focused on human-AI collaboration (the interpretability aspect), research on the ethical, legal and societal implications of AI, ensuring safety and security, and building standards and benchmarks. As regards safety, the plan stresses the need to explore human-machine interactions, especially for safety-critical AI algorithms and applications, the need for novel programming techniques to develop more robust and verifiable AI as well as mentions the long-term existential risk that AI can pose.[398] Regarding security, the plan stresses cybersecurity threats, 'data poisoning' techniques and the need for robustness throughout the whole supply chain.[399] The plan, however, does not contain any specific projected regulation to address these aspects.

Section 4 of EO 14110 specifically concerns safety and security issues. The order is addressed to government agencies and does not set obligations for private AI developers. Among the priorities, the EO lists:

- Developing guidelines, standards, and best practices for AI safety and security;
- Ensuring safe and reliable AI – the subsection primarily relates to dual-use foundation models; large AI models having potential capabilities that could be used in malicious cyber-enabled activity (e.g., those are considered systems having compute capacity of 10^{20} integer or floating-point operations per second for training AI);
- Managing AI in critical infrastructure and cybersecurity;

396 2023 AI RD Plan (n 338), 22.

397 Ibid, 23.

398 Ibid, 16–17.

399 Ibid, 17.

- Reducing risks of CBRN[400] threats, e.g., reducing the risk of misuse of synthetic nucleic acids;
- Reducing the risks posed by synthetic content; and
- Promoting safe release and preventing the malicious use of federal data for AI training.

The executive order mostly outlines the risk and states the need for their addressing via consultations, building risk-management practices and monitoring.

According to the NSM AI, 'the United States Government must also provide appropriate safety and security guidance to AI developers and users, and rigorously assess and help mitigate the risks that AI systems could pose'[401] and draws attention to the need for tools for reliability safety testing, especially frontier models, and mentions the role of AISI and NIST in this process and conducting a relevant assessment for the adequacy of available tools. [402] The NSM AI also requires the responsible use of AI, balanced with the benefits AI provides; this includes the review and updating of relevant (non-AI-specific policies), ensuring objective measurement metrics, and developing guidance.[403] Finally, the memorandum requires strengthening of AI governance and risk management, including addressing risks to physical safety, privacy harms, discrimination and bias, inappropriate use, lack of transparency, lack of accountability, data spillage and poor performance (inconsistent outcomes, unlawful discrimination and harmful biases, or undermining the integrity of decision-making process), and deliberate manipulation and misuse.[404] The NSM AI projects to do this via NIST AI RMF, ensuring appropriate governance and oversight and risk management policies to ensure privacy, civil liberties and safety.[405]

As regards detailed specific safety-related measures, the landmark document was OMB M-24-10 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence' of 28 March 2024[406]. It sets binding requirements for the use of AI by state agencies. The memorandum covers AI governance, innovation, and risk management.

---

400 CBRN - chemical, biological, radiological and nuclear materials and agents that could potentially harm the society through their accidental or deliberate release, dissemination, or impacts.

401 NSM AI (n 350), s 2(a).

402 Ibid, s 3.3.

403 Ibid, s 4.1.

404 Ibid, s 4.2.

405 Sect 4.2.(e)(ii)(E).

406 OMB M-24-10 (n 356).

For the latter, the memorandum sets the minimum required measures. The memorandum also highlights the importance of implementing AI management consistently and requires work in the standardisation of processes, promoting templates and sharing best practices.

As regards strengthening AI governance, the memorandum requires each agency to appoint a chief AI officer (CAIO), of appropriate seniority and skills, to be in charge of key AI-related decisions, bear primary responsibility for the agency's compliance, coordinate the agency's use of AI, manage relevant risks, devise a compliance plan, and conduct an inventory of AI use-cases.

For generative AI, the memorandum requires that agencies assess the potential benefits against risks and establish adequate safeguards and oversight mechanisms[407].

As for risk management, the memorandum defines two special risk categories of AI:

- Safety-impacting AI. This refers to AI whose output produces an action or serves as a principal basis for a decision that has the potential to significantly impact the safety of human life or well-being; climate or environment; critical infrastructure, including the infrastructure for voting and protecting the integrity of elections; and strategic assets or resources, including high-value property and information marked as sensitive or classified by the Federal Government.[408] This category is risk is presumed in the following AI activities: controlling safety-critical functions with dams, electrical grids, traffic, nuclear reactors, etc., maintaining the integrity of elections and voting structure; controlling of physical movement of robots, applying kinetic force, delivering biological, chemical agents; autonomously and semi-autonomously moving vehicles; controlling the transport, transporting industrial wastes, designing, constructing, or testing of industrial equipment that, if failed, would pose a significant risk to safety; carrying medically relevant functions of medical devices; providing medical diagnoses, determining medical treatment; controlling access to government facilities, determining or carrying out enforcement actions regarding sanctions, trade restrictions, or other control on export, investments, or shipping[409]; and

- Rights-impacting AI. This relates to AI whose output serves as a principal basis for a decision or action concerning a specific individual or entity that has a legal,

---

407 Ibid, 11.

408 Ibid, 30

409 Ibid, appendix I, s 1.

material, binding, or similarly significant effect on them. This includes civil rights and liberties, privacy, including but not limited to freedom of speech, voting, human autonomy, and protections from discrimination, excessive punishment, and unlawful surveillance; equal opportunities, including equitable access to education, housing, insurance, credit, employment, and other programs where civil rights and equal opportunity protections apply; or access to or the ability to apply for critical government resources or services, including healthcare, financial services, public housing, social services, transportation, and essential goods and services.[410] The memorandum contains a list of use cases, where the AI is presumed to be rights-impacting that includes, *inter alia*[411]: risk assessments about individuals; predicting criminal recidivism or criminal offenders, biometric identification, including in publicly accessible spaces, sketching faces, reconstructing faces based on genetic information; conducting cyber-intrusions; detecting or measuring emotions, replicating of person's likeliness or voice without express consent; detecting student cheating or plagiarism; determining the terms and conditions of employment, use in medical devices; allocating loans, credit scoring; access to critical government resources or services; translating between languages for official communication in legally binding responses; child adoption and protection sphere.

Many practices that are prohibited by the AI Act are regarded as high-risk in the memorandum.

The minimum practices are primarily addressed to these 'impacting' categories. They are defined in section 5(c) of the memorandum on a non-exhaustive basis and include:

- Completing an AI impact assessment, including the assessment of the intended purpose of AI, potential risks, and quality of the relevant data;
- Testing AI in the real-world context (pilot and limited releases are preferable);
- Independent evaluation of AI;
- Conducting ongoing monitoring; regular evaluating of the risk of AI use;
- Mitigating emerging risks to rights and safety, including the assessment of the adequacy of the human-intervention mechanisms;

---

410 Ibid, 30.

411 Ibid, appendix I, s 2.

- Ensuring adequate human training and assessment;
- Providing additional human oversight, intervention and accountability for decisions that result in a significant impact on rights and safety;
- Provide public notice and plain language documentation; and
- For rights-impacting AI:
    - Identifying and assessing AI's impact on equity and fairness, and mitigating algorithmic discrimination, including the assessment of significant disparities in the model's performance in real-world testing, disparities mitigation, or stopping using the system if this could not be adequately mitigated;
    - Consulting and incorporating feedback from affected communities and the public: conduct direct usability testing, public hearings, and post-transaction customer feedback;
    - Ongoing monitoring and mitigation for AI-enabled discrimination;
    - Notifying negatively affected individuals;
    - Maintaining human consideration and remedy processes; and
    - Maintaining options to opt out of AI-enabled decisions.

The memorandum also contains a section for managing risk in federal procurement.

Government agencies were required to implement risk management practices and appropriately document them, or terminate the use of non-compliant AI, by 1 December 2024. They must also at least annually certify and publish a determination (a compliance statement) regarding AI or obtain an appropriate extension or waiver. The extension could be granted for not longer than 1 year if an agency cannot feasibly meet the minimum requirements by the deadline.[412] A waiver may be granted by a CAIO based on a risk assessment if fulfilling the requirement would increase risks to safety or rights overall or would create an unacceptable impediment to critical agency operations.[413]

---

412 Ibid, s 5(c)(ii).

413 Ibid, s 5(c)(iii).

NSM AI defines a similar category of 'high-impact AI systems'. These are the systems 'whose output serves as a principal basis for a decision or action that could exacerbate or create significant risks to national security, international norms, human rights, civil rights, civil liberties, privacy, safety, or other democratic values'[414]. NSM AI prescribes the following minimum risk management practices for such systems that generally correspond to OMD M-24-10 but are less detailed. They include: a mechanism for agencies to assess AI's expected benefits and potential risks; a mechanism for assessing data quality; sufficient test and evaluation practices; mitigation of unlawful discrimination and harmful bias; human training, assessment, and oversight requirements; ongoing monitoring; and additional safeguards for military service members, the Federal civilian workforce, and individuals who receive an offer of employment from a covered agency.[415] In a nutshell, the scope of measures is similar to the EU's high-risk AI but less detailed.

The NSM AI Framework endorses and further specifies the ideas of the NSM AI and is designed as an alternative to OMB M-24-10 for the federal agencies to comply with. The US AI Framework defines the categories of prohibited AI (this category is absent in OMB M-24-10) and high-impact AI and sets out minimum risk management practices.

The prohibited use cases include AI systems that:

- Profile, target, or track activities of individuals based solely on their exercise of rights protected under the Constitution and applicable US domestic law, including freedom of expression, association, and assembly rights;
- Unlawfully suppress or burden the right to free speech or right to legal counsel;
- Unlawfully disadvantage an individual based on their ethnicity, national origin, race, sex, gender, gender identity, sexual orientation, disability status, or religion;
- Detect, measure, or infer an individual's emotional state from data acquired about that person, except for a lawful and justified reason such as to support the health of U.S. Government personnel;
- Infer or determine, relying solely on biometrics data, a person's religious, ethnic, racial, sexual orientation, disability status, gender identity, or political identity;

414 NSM AI (n 350), s 4.2.(e)(ii)(D).

415 Ibid.

- Determine collateral damage and casualty estimations, including identifying the presence of non-combatants, before kinetic action without (1) rigorous testing and assurance within the AI systems' well-defined uses and across their entire lifecycles, and (2) oversight by trained personnel who are responsible for such estimations exercising appropriate levels of judgment and care;

- Adjudicate or otherwise render a final determination of an individual's immigration classification, including related to refuge or asylum, or other entry or admission into the United States;

- Produce and disseminate reports or intelligence analysis based solely on AI outputs without sufficient warnings that enable the reader of the reports or analysis to recognize that the report or analysis is based solely on AI outputs; and

- Remove a human "in the loop" for actions critical to informing and executing decisions by the President to initiate or terminate nuclear weapons employment.[416]

The above provides an alternative vision of the prohibited AI: some use cases are similar to those of the EU, but the list includes some use cases qualified as high-risk in the EU.

AI use is presumed to be of high impact if it uses controls or significantly influences the outcomes of any of the following activities:

- Tracking or identifying individuals in real-time, based solely on biometrics, for military or law enforcement action;

- Classifying an individual as a known or suspected terrorist, insider threat, or other national security threat to inform decisions or actions that could affect their safety, liberty, employment, immigration status, ability to enter or remain in the United States, or Constitutionally-protected rights and freedoms;

- Determining an individual's immigration classification, including related to refuge or asylum, or other entry or admission into the United States;

- The activities referenced in Appendix I of OMB M-24-10, when those activities: (i) occur within the US, impact US persons, or relate to immigration processes or other entry or admission into the US;

---

[416] Ibid 2, 3.

- Designing, developing, testing, managing, or decommissioning sensitive chemical or biological, radiological, or nuclear materials, devices, and/or systems (including chemical and biological data sources) that could be at risk of being unintentionally weaponizable;

- Operating or deploying malicious software designed to allow AI to automatically and without human oversight write or rewrite code in a way that risks unintended performance or operation, spread autonomously, or cause physical damage to or disruption of critical infrastructure; and

- Using AI as a sole means of producing and disseminating finished intelligence analysis.[417]

The framework defines another 'high-risk' category – AI use cases impacting federal personnel. These are the applications of AI conducting the activities of hiring, including determining the salary and benefits packages, decision on promotion, demotion, or termination of employment, and determining job performance, physical, and mental health.

The above lists and categories are non-exhaustive and the state agencies may extend them as needed on a non-classified basis.[418]

The framework sets the minimum risk management practices for high-impact and federal-personnel-impacting AI uses that include:[419]

- Risk and impact assessment – identifying potential benefits and risks of AI, including the analysis demonstrating that AI is better suited for the task than other options and mitigation options; if the benefits do not 'meaningful' outweigh the risk, the agency must refrain from using it; the assessment of the quality of the relevant data;

- Pre-deployment testing, pilots and limited releases are encouraged;

- Conduct an independent review and report the results;

- Identifying and mitigating factors that may contribute to unlawful discrimination or harmful biases;

---

417 Ibid 3, 4

418 Ibid 4.

419 Ibid 5–8.

- Developing processes, including training, to mitigate the risk of overreliance on AI systems;
- Ensure appropriate human consideration and oversight of AI-based decisions, establishing clear human accountability for such decisions and escalation procedures;
- Having reporting processes for unsafe, anomalous, inappropriate or prohibited use;
- Regular monitoring and testing, including perioding human reviews to assess changes in the context of AI operation;
- Mitigate emerging risks; and
- Have processes of internal escalation and senior-leadership approvals for uses of AI that pose significant degrees of risks (e.g., those that may harm foreign policy interests).

Agencies are encouraged to apply these practices to non 'high-impact' AI systems.

As regards the AI impacting federal personnel, the Framework set the following additional measures: consulting and incorporating feedback from the workforce (their representatives); notifying and obtaining consent of individuals affected by AI systems; notifying if AI was used to inform on adverse employment-related decision; and providing timely human consideration and remedy through a fallback and escalation system.

Likewise in OMB M-24-10, the NSM AI Framework provides the option to obtain a waiver for up to one year if fulfilling the minimal practices would increase risks to privacy, civil liberties, or safety, or would create an unacceptable impediment to critical agency operations or exceptionally grave damage to national security.[420] The standard for the waiver is quite high and granting the waiver is subject to consultation with the respective agency's civil liberties and privacy officers.

The framework requires each agency to appoint a chief AI officer (CAIO) who should institute the governance and oversight process, ensure compliance and make major AI-related decisions. Besides that, the agency must establish the AI Governance Board composed of

---

420 Ibid 7.

relevant senior officials to advise agency leadership, develop standardized documents, and report misuse.

Finally, the agencies applying AI systems of all categories shall: conduct an inventory of their high-impact use cases; evaluate their data management practices; at least annually submit a report on their AI activities, including the AI oversight; and establish standardized training requirements and guidelines for the workforce.

The draft Federal AI Governance and Transparency Act[421] is aimed to harmonise the AI governance practices of state agencies. The act purports to establish a federal register for AI and requires the agencies to develop and make public their governance charters. The measures generally resemble the NSM AI Framework and OMB M-24-10, including data management, risk assessment, ongoing oversight and regular updates, and independent evaluation at least bi-annually[422]. Interestingly, the agency is released from this obligation if the AI is used solely for research and development or non-publicly in a national security system[423].

NIST AI RMF is designed as a voluntary framework that could be applied both by public and private players. It strives to be risk-based, resource-efficient, pro-innovation, and consensus-driven. The standard follows a different approach to building AI risk-management systems. First, it describes the elements of trustworthy AI: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed – the risk management system must address all of them. Then it splits the system into four functions: govern, map, measure, and manage.

---

421 Federal AI Governance and Transparency Act (n 357).

422 Ibid, para 3596(a).

423 Ibid, para 3595(e).

Credit: AI RMF 1.0, page 20

The set of measures and their specifics is to be determined based on the use-case profile and may include aspects of AI design, development, and deployment, operating and monitoring, testing, evaluating, verification and validation (so-called TEVV), addressing human factors, involving domain experts, conducting AI impact assessment, procurement-related measures, governance and oversight, as well as defining human roles and responsibilities in decision-making and overseeing, considering the systemic and human cognitive biases when designing, developing, deploying and evaluating the system, tailoring human-AI interactions and AI explainability.

Generative AI and dual-use frontier models. NIST AI.600-1 is the extension of NIST AI RMF and sets the profile for generative AI or GAI[424]. The standard highlights specific risks of the generative AI, including CBRN information and capabilities, confabulations (an alternative term for AI 'hallucinations' or 'fabrications'), dangerous, violent or hateful content, data privacy, environmental impact, harmful bias or homogenization, human-AI configuration (anthropomorphizing GAI systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with GAI systems), information

[424] EO 14110 defines generative AI as 'the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content'.

integrity and security, intellectual property, obscene, degrading, and/or abusive content (inappropriate synthetic content), and value chain and component integration. Addressing these issues forms part of trustworthy generative AI.

The standard provides for a non-exhaustive list of measures, including establishing specific policies (e.g., for measuring the effectiveness of employed content provenance methodologies – cryptography, watermarking),[425] allocating responsibilities, categorizing the types of content, documenting risks (e.g., overreliance – GV-6.2-001), evaluations involving human subjects (Measure 2.2), measures aimed at system security and resilience (Measure 2.7), measures to mitigate confabulations (explainability, fine-tuning, ethical considerations, Measure 2.9). In total, the standard contains more than a hundred specific instructions. The measures include organizational governance, pre-deployment testing, structured public feedback, including red-teaming, content provenance (data tracking techniques), and incident disclosure.[426]

Draft NIST AI 800-1 relates to managing misuse risk for dual-use foundation models and is also worth mentioning. The 'dual-use foundation model' is defined as an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by: (i) substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons; (ii) enabling powerful offensive cyber operations through automated vulnerability discovery and exploitation against a wide range of potential targets of cyber-attacks; or (iii) permitting the evasion of human control or oversight through means of deception or obfuscation[427].

The standard mentions the following specific challenges of this category of AI: broadly applicable across many domains, hard to assess potential misuse; capabilities do not clearly translate across domains; hard to predict how scale will affect the performance; the relationship between the capability and risk of harm is often unclear; safeguard evaluation

---

425 NIST.AI.600-1 (n 359) 19, action GV4.3--001

426 Ibid 47–53.

427 EO 14110 (n 341), s 3(k).

methods are nascent; risk evaluation may depend on scarce domain expertise; and difficulties in emulating threats and misuse.[428]

The standard defines seven objectives to address these issues: anticipating potential risks, establishing a mitigation plan, managing the risk of model theft, measuring the misuse risk before deployment, post-deployment collecting and responding to information on misuse, and providing transparency. For each of these objectives, the standard provides for recommended practices and paperwork. For example, it recommends assessing the risk of the system's jailbreaking[429] or establish a practice of bounties for finding misuse[430].

A real-world example of the assessment of such a generative model is the joint pre-deployment assessment of Anthropic's Upgraded Claude 3.5 Sonnet by US and UK safety institutes (report announced on 19 November 2024[431]), where the model was tested against the best human performances in a range of tasks – cybersecurity, biological expertise capabilities, and software improvement.

Finally, as has been pointed out already, the US government is in close collaboration with the major tech players. Some policy and regulation gaps are backed up by their voluntary commitments.[432]

The commitments secured by Biden's Administration in July 2023 include:

- Ensuring the products are safe before publicly deployed. This includes conducting internal and external security testing and sharing information across the industry and with government, civil society and academia on managing AI risks;
- Building security-first systems, including investing in cybersecurity and insider threat safeguards, facilitating third-party discovery and reporting vulnerabilities;
- Earning the public's trust. To this end: developing robust technical mechanisms to make the user aware of the AI-generated content, e.g., watermarking; publicly reporting on the system's capabilities and limitations; prioritizing research on

---

428 NIST AI 800-1 (n 362) 7.

429 Ibid 17.

430 Ibid 15.

431 'Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet: The U.S. AI Safety Institute and the UK AI Safety Institute conducted joint pre-deployment testing of Anthropic's latest model' (NIST, 19 November 2024) <www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet> accessed 9 December 2024.

432 See n 339.

societal risks of AI systems (harmful bias, discrimination and protecting privacy); and developing and deploying advanced AI systems to help address society's greatest challenges (e.g., climate change).

Another example is the call in March 2023 by the Future of Life Institute to pause giant AI experiments for at least 6 months and conduct an independent evaluation before any real-world testing.[433] The initiative obtained more than 33 thousand signatures, including those of Y. Bengio, S. Russel, E. Musk, S. Wozniak, and Y. Harari. In June 2022, Cohere, OpenAI, and AI21 Labs published the Best Practices for Deploying Language Models.[434]

The major AI companies usually have a quite developed publicly available AI risk-management policy. For instance, OpenAI published its Charter. Regarding the long-term safety and development of AGI, it states:

> We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project.[435]

Finally, the US declared the ambition to lead the global regulatory safety agenda. For instance, section 11 of EO 14110 is dedicated to strengthening American leadership abroad and formulates the task as follows:

> The Federal Government will seek to promote responsible AI safety and security principles and actions with other nations, including our competitors, while leading key global conversations and collaborations to ensure that AI benefits the whole world, rather than exacerbating inequities, threatening human rights, and causing other harms.[436]

The US was a co-organiser of the US's AI Safety Summit and was among the first states to establish a national AI safety institute, initiated the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy[437] signed by 50+ countries; and signed

---

433 Pause Giant AI Experiments: An Open Letter (n 79).

434 'Best practices for deploying language models' (OpenAI, 2 June 2022) <https://openai.com/index/best-practices-for-deploying-language-models/> accessed 9 December 2024.

435 'OpenAI Charter' (OpenAI) <https://openai.com/charter/> accessed 9 December 2024.

436 EO 14110 (n 341), s 2(h).

437 Shawn Steene, Chris Jenks, 'The Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' (Lieber Institute, West Point, Articles of War, 13 November 2023) <

the AI Convention. To foster further development, in July 2024, NIST issued NIST AI 100-5: A Plan for Global Engagement on AI Standards[438] setting the standardization priorities and recommended global activities with the US's leading role in this process. It appears that, to a certain extent, the US intends to fill in national regulatory gaps with global policies and standards realizing that truly risky AI products are not bound by standard borders.

To summarise: the US prioritises and is very active in developing safety regulations. There are two major vectors for that: defining guidelines for the use of AI by stage agency and providing non-binding recommendations to private developers by suggesting standards. The scope of safety practices in many aspects matches the practices of the EU. The US sets the categories of prohibited, risk-impacting and rights-impacting AI. Certain practices are qualified by risk otherwise than in the EU. The US law acknowledges the possible catastrophic impacts of AI, yet so far, the legislation has no comprehensive measures to address these risks.

3.3.3 China

(a) Regulatory framework

At the moment, the binding regulation on AI in China is focused on three use cases:

- Algorithmic recommendations in internet information services;
- Deep synthesis in internet information services; and
- Generative AI services.

Chinese regulation of AI is the least extensive among the compared jurisdictions. Generally, Chinese law aims for technological development, protecting public interests and Core Socialist Values, and ensuring national security. This is ensured, among other things, by detailed regulation of content.

AI-specific legislation emerged in China in 2017 when the State Council of China issued 'A Next Generation Artificial Intelligence Development Plan'[439]. The plan declares AI to be

---

https://lieber.westpoint.edu/political-declaration-responsible-military-use-artificial-intelligence-autonomy/> accessed 9 December 2024.

438 NIST AI 100-5. A Plan for Global Engagement on AI Standards <https://doi.org/10.6028/NIST.AI.100-5> accessed 9 December 2024.

439 A Next Generation Artificial Intelligence Development Plan (n 13).

the new focus of international competition and strategic technology that will lead the future; as envisioned, AI is 'the new engine of economic development and brings new opportunities for social construction'[440]. With that, the plan mentions the corresponding challenges such as the transformation of employment structures, the impact on legal and social theories, violations of personal privacy, and challenges in international relations and norms. The plan expects that it will have far-reaching effects on the management of government, economic security, social stability, as well as global governance.[441] The plan requires acceleration of deep integration of AI with the economy based on the principles: technology-led development (utilize the most that the technology is offering); the use of systems layout (fully give play to the advantages of the socialist system); market-dominant approach (accelerate the commercialization); and openness (encourage open-source sharing and joint innovation).

The plan sets the goal for China to achieve major breakthroughs in basic theories of AI by 2025 and by 2030 to become the world's primary AI innovation centre and have appropriate AI laws and regulations, ethical norms, and policy systems.[442] The plan focuses on technological aspects of AI and requires 'forcefully developing' new AI industries[443]. The plan envisions smart courts and smart cities[444] and calls for the development of laws, regulations, and ethical norms (they should ensure 'the healthy development of AI'[445]), as well as key policies, AI technology standards, the intellectual property system, AI security supervision, and the evaluation system.

In 2019, the National Governance Committee for the New Generation Artificial Intelligence issued the Governance Principles for the New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence (GPNGAI)[446]. Its eight principles include:

- harmony and human-friendliness (AI should conform with human values and ethical principles, promote human-machine harmony, and serve the progress of

---

440 Ibid, s I.

441 Ibid.

442 Ibid, s II.3.

443 Ibid, s III.2.1

444 Ibid, s III.3.2

445 Ibid, s V.1

446 National Governance Committee for the New Generation Artificial Intelligence, 'Governance Principles for the New Generation Artificial Intelligence – Developing Responsible Artificial Intelligence' (China Daily, 17 June 2019) <www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html> accessed 9 December 2024 (GPNGAI).

human civilization; the misuse, abuse, or evil use of AI technology should be prevented by all means);

- fairness and justice;
- inclusion and sharing (environmental friendliness and resource conservation, AI education, training, and support to vulnerable groups);
- respect for privacy;
- safety and controllability;
- shared responsibility;
- openness and collaboration; and
- agile governance – AI governance should be adaptive, reflect the development of the technology, and cover the entire life cycle of AI products.

In July 2021, the China Academy of Information and Communications Technology issued the White Paper on Trustworthy Artificial Intelligence[447]. It overviews the history of the concept and stresses its key role in gaining people's trust in the technology. The white paper focuses on the issues of stability and explainability of AI systems, as well as on the protection of privacy. It calls for accelerated development of AI regulation in China. As regards AGI, the document admits that AI governance is mostly focused on 'weak' AI and since the development of AGI 'is guaranteed to alter the course of humanity…, it is necessary to adopt a forward-looking strategy', including for cutting-edge technologies, super-deep learning, quantum machine learning, and 'strong' AI.[448]

In September 2021, the National Governance Committee for the New Generation Artificial Intelligence issued the 'Ethical Norms for the New Generation Artificial Intelligence'.[449] The

---

447 China WP on Trustworthy AI (n 85).

448 Ibid, s 5.2.

449 National Governance Committee for the New Generation Artificial Intelligence, 'Ethical Code for the New Generation of Artificial Intelligence' (26 September 2021) <http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html> accessed 9 December 2024.

norms apply to all natural and legal persons engaged in AI-related activities.[450] The document formulates six fundamental ethical norms:

1. Enhancing the well-being of humankind, including abiding by common values of humankind, respecting human rights and the fundamental interests of humankind, abiding by regional and national ethical norms, the priority of public interests, and enhancing the sense of happiness;

2. Promoting fairness and justice: adherence to shared benefits and inclusivity, protecting legitimate rights and interests; protecting vulnerable and underrepresented groups and providing alternatives as needed;

3. Protecting privacy and security;

4. Ensuring controllability and trustworthiness. Ensure that AI is always under meaningful control, the right to opt-out from interaction with AI or be subject to AI decisions;

5. Strengthening accountability: human beings are the ultimate liable subjects; clear responsibilities of all relevant stakeholders, establishing accountability mechanisms; and

6. Improving ethical literacy.[451]

Besides, the document requires the promotion of agile governance, strengthening risk prevention (including 'systemic risk monitoring'); strengthening the awareness of self-discipline, improving data quality, enhancing safety, security and transparency, avoiding biases and discrimination, promoting good use (pre-deployment testing), avoiding misuse and abuse and forbidding malicious use ('it is strictly forbidden to endanger national security, public safety and product safety, and it is strictly forbidden to harm public interests'[452]); and ensuring timely and proactive feedback. The three topics researched in this work are, therefore, covered by the above ethics principles.

---

450 Ibid, introduction.

451 Ibid, s 3.

452 Ibid, s 20.

The same month, in September 2021, the Guiding Opinions on Strengthening the Overall Governance of Internet Information Service Algorithms were issued[453] that support the ideology of China being a cyber power and focused on the following three-year objectives: complete governance mechanisms (improve policies and regulations; clear rights and responsibilities of entities); improved oversight system (including handling of conduct that violates laws and regulations); and standardization of the algorithm ecology ('algorithms are to be correctly oriented with abundant positive energy', fair and equitable, safe, controllable and 'independently innovative', to effectively prevent the potential threats brought on by the abuse of algorithms).[454]

In December 2021, the Provisions on the Administration of Algorithmic Recommendations for Internet Information Services were adopted.[455] This is the first AI-specific binding piece of legislation that regulates the algorithmic recommendation technologies applied by Internet information services. The provisions regulate the matters of notification of users, profiling, management systems, information security, and content control (including the prohibition of fake news via synthesized content). They require the enterprises to report their algorithms to the government. In April 2022, the Cyberspace Administration of China released a notice[456] addressed to the provincial, autonomous region and directly governed municipality cybersecurity and informatization offices aimed at promoting the above provisions and tasking them to ensure internet enterprises organise self-inspection and file their algorithms, establish governance bodies, appoint relevant responsible personnel, and ensure compliance with the provisions on time.

In November 2022, the Administrative Regulations of Deep Synthesis in Internet Information Services were adopted setting out requirements for providers of relevant services, including the identification of users.[457]

---

453 State Internet Information Office, Central Propaganda Department, Ministry of Education, Ministry of Science and Technology, Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Culture and Tourism, State Administration for Market Regulation, State Administration of Radio and Television, 'Guiding Opinions on Strengthening the Overall Governance of Internet Information Service Algorithms' (17 September 2021) Document No. 7 [2021] (Chinese).

454 Ibid, s I(3).

455 Algorithmic Recommendation Provisions (n 21).

456 Cyberspace Administration of China, 'Notice on Launching the Clear 2022 Special Action on the Comprehensive Governance of Algorithms' (China Internet Information Office, 8 April 2022) <www.cac.gov.cn/2022-04/08/c_1651028524542025.htm> accessed 9 December 2024 (Chinese).

457 Deep Synthesis Provisions (n 22).

In July 2023, China adopted the Interim Measures for the Management of Generative Artificial Intelligence Services[458] that apply to the use of generative AI technologies to provide services to the public in the mainland PRC.

In August 2023, the Chinese National Information Security Standardization Technical Committee (TC260) adopted its Practice Guide to Cybersecurity Standards of Generative artificial intelligence service content identification method.[459] The Basic Security Requirements for Generative AI Services followed in February 2024.[460] Lastly, in September 2024, the National Information Security Standardization Technical Committee (TC260) adopted the AI Safety Governance Framework (China Safety Governance Framework).[461]

China has regional AI legislation that relates to the development of the AI industry in Shanghai Municipality and Shenzhen Special Economic Zone that came into effect in 2022.[462]

Other noticeable legislation includes:

- Provisions on the Security Assessment of Internet Information Services with Public Opinion Properties or Social Mobilization Capacity.[463] Although it is not AI-specific, it sets the security assessment regulations for AI services that are 'capable of creating public opinions or social mobilization' and many specific requirements are linked to this category of services;

---

458 Generative AI measures (n 23).

459 National Information Security Standardization Technical Committee, 'Cybersecurity Standard Practice Guide - Generative Artificial Intelligence Service Content Identification Method' (25 August 2023) Document No. 124 [2023] (Chinese).

460 National Cyber Security Standardization Technical Committee, 'TC260-003 "Basic Requirements for Security of Generative Artificial Intelligence Services" (29 February 2024) <www.tc260.org.cn/front/postDetail.html?id=20240301164054> accessed 9 December 2024 (Chinese).

461 The National Information Security Standardization Technical Committee (TC260), 'AI Safety Governance Framework' (September 2024) <www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf> accessed 9 December 2024 (China Safety Governance Framework).

462 Alexander Chipman Koty, 'Artificial Intelligence in China: Shenzhen Releases First Local Regulations' (China Briefing, 29 July 2021) <www.china-briefing.com/news/artificial-intelligence-china-shenzhen-first-local-ai-regulations-key-areas-coverage/> accessed 9 December 2024; Ashyana-Jasmine Kachra, 'Making Sense of China's AI Regulations' (Holistic AI, 12 February 2024) <www.holisticai.com/blog/china-ai-regulation> accessed 9 December 2024.

463 Cyberspace Administration of China, 'Provisions on the Security Assessment of Internet Information Services with Public Opinion Properties or Social Mobilization Capacity' (15 November 2018) <www.cac.gov.cn/2018-11/15/c_1123716072.htm> accessed 9 December 2024 (Chinese) (Provisions on Internet Information Services).

- Personal Information Protection Law (PIPL)[464];
- Provisions on the Management of Online A/V Information Services[465];
- Cybersecurity Law[466];
- Data Security Law[467]; and
- Internet Information Service Management Measures[468].

Notable draft laws include:

- Draft AI Law[469]; and
- Draft Measures for Labelling of AI-Generated Synthetic Content[470].

China has an extensive AI policy. As regards binding legislation, it focuses on specific use cases: generative AI and algorithmic recommendations and deep synthesis in internet information services. The legislation is focused on controlling content, especially as regards online services that have public opinion properties or social mobilization capacity. China is active in adopting AI-specific standards. There is also a progressive Draft AI law that may set a comprehensive horizontal regulation of AI.

---

464 Personal Information Protection Law of the People's Republic of China of 20 August 2021 (Chairman's Order No. 91) (PIPL).

465 State Internet Information Office, Ministry of Culture and Tourism, State Administration of Radio and Television, 'Provisions on the Management of Online A/V Information Services' (18 November 2019) Document No. 3 [2019].

466 Cybersecurity Law of the People's Republic of China of 7 November 2016 (Chairman's Order No. 53).

467 Data Security Law of the People's Republic of China of 10 June 2021 (Chairman's Order No. 84).

468 People's Republic of China State Council, 'Internet Information Service Management Measures' (Decree No. 292 of 25 September 2000).

469 AI Law (n 25).

470 Cyberspace Administration of China, 'Measures for Labelling of AI-Generated Synthetic Content (draft for solicitation of comments)' (14 September 2024) <www.cac.gov.cn/2024-09/14/c_1728000676244628.htm> accessed 9 December 2024.

(b) Regulation of subject areas

(I) Human rights

Although China declares the protection of human rights in the AI context, the public interests and the Core Socialist Values[471] appear to prevail.[472] The key AI-specific provisions on the protection of human rights are discussed below; they are specific to the regulated areas.

In the context of algorithmic recommendations[473]:

- The algorithmic model may not violate law, regulations and ethics (this part is expected) but also morals, such as no AI that leads to addiction or excessive consumption) (article 8). The reference to morals is unusual since the moral norms are not universal even among close communities; it is questionable, therefore, how this provision could be implemented in practice;

- No unlawful or harmful tags (article 10). The providers must set the model in a way it may not enter unlawful or harmful information as keywords or tag the users accordingly;

- Notification and explanation (article 16). The service provider must clearly notify users about the operation of the algorithmic recommendation and make public the basic principles, purposes and motives, and main operational mechanisms of their systems;

- Right to opt-out (article 17). The provider must give users a choice to not be targeted by their individual characteristics and a convenient option to switch off the algorithmic recommendation services;

- No unsafe content for minors (article 18). The provider must ensure the algorithmic recommendation adapts the recommendation to make it convenient for minors to 'obtain information beneficial to their physical and mental health'; the service may not push information toward minors to incite unsafe conduct, violate social morals, or lead to online addiction;

---

471 Generative AI measures (n 23), art 4(1).

472 E.g., Deep synthesis provisions (n 457), arts 6, 15.

473 Algorithmic recommendation provisions (n 455).

- Protection of the elderly (article 19). The article requires service providers to take special care to address the needs of older users, including in preventing fraud;

- Protecting consumers (article 21): respecting the rights of consumers to fair trading, no unreasonably differentiated treatment, no exploitation of consumer's habits and other characteristics;

- No misinformation (article 14). The provider may not use algorithms to shield information, over-recommend, manipulate topic lists or search result rankings, control hot search terms or selections and other such interventions in information presentation; or carry out acts influencing online public opinion.

- Effective complaint procedure (articles 22, 30): this applies both to user-compliant and public complaints.

In the context of deep synthesis[474]:

- Labelling of generated content (articles 16 and 17).

In the context of generative AI services[475]:

- No misinformation (article 4(1)). The services must uphold the Core Socialist Values[476] and conduct content control, including to prevent promoting ethnic hatred and ethnic discrimination, violence and obscenity, as well as fake and harmful information.

- No discrimination (article 4(2)): to be ensured during the process of the algorithm's design, training, model generation and optimization;

- Respecting intellectual property rights (article 4(3));

- Respecting the lawful rights, interests, and physical and psychological well-being, such as image, reputation, honour, privacy, and personal information (article 4(4));

---

474 Deep synthesis provisions (n 457).

475 Generative AI measures (n 23).

476 There 12 Core Socialist Values: prosperity, democracy, civility, harmony; freedom, equality, justice, rule of law; patriotism, dedication, integrity, and friendship – 'Core Socialist Values' (Wikipedia) <https://en.wikipedia.org/wiki/Core_Socialist_Values> accessed 9 December 2024.

- Protection of minors (article 10): protection of minors from overreliance or addiction to generative AI services; and
- Easy complaints and reporting mechanism (article 15).

The draft China AI Law envisions a unified set of principles and human rights, among them:

- Explainability (article 6): AI developers and providers must provide and explain the basic information, purpose and intent, and main operating mechanisms of their AI products and services;
- Equal rights (article 32): ensuring equality of users, universality and fairness of products and services; avoiding prejudice and discrimination;
- Right to know (article 33): the right of users to have information on AI systems, including operating mechanisms, functional limitations, contact information, rights and remedial channels;
- Protection of privacy and personal information (article 34) – will be discussed in the next section;
- Right to explanation and refusal of AI-based decisions (article 35);
- Protection of IP rights of AI-generated content (article 36);
- Rights and interests of workers (article 37): informing workers on the application of AI when concluding labour contracts; no punishing of dismissing decisions based solely on AI;
- Rights and interests of digitally disadvantaged groups (article 38): providers must ensure differentiated needs of 'minors, the elderly, women, people with disabilities and other people with weaker digital access capabilities'; facilitate the safe use of AI products and services by the elderly; protect minor from inappropriate content and functions;
- Right to obtain help and training (article 39). Providers of AI products and services that involve the significant legitimate rights and interests of individuals shall provide assistance and skills training to individuals and organizations;
- Right to complain and sue (article 40); and

- Rights related to social credit (article 76): social credit scoring and rating shall be aimed at improving credit quality and efficiency and strengthening the level of risk management, not infringe upon the right to equal access to the provision of public services or other legitimate rights and interest of individuals.

So, the Chinese legislation has quite an extensive AI-specific body of rights that resembles, to a certain extent, the approach of the EU. It is important, however, how these rights are enforced in practice. There are studies on the AI facial recognition practices in China to suppress unrest.[477] There is also information in the press that China uses AI news anchors for propaganda (e.g., to affect the elections in Taiwan)[478]. Despite the declared respect for intellectual property, the requirement to disclose algorithms by internet service providers was considered to encroach on this right and forced disclosure of proprietary information.[479] The Chinese government uses heavily facial recognition tools, including to track minority groups (e.g., Uighurs). [480] The latter is called the first known example of a government intentionally using artificial intelligence for racial profiling.[481]

(II) Personal data

The rights of data subjects concerning the processing of their personal data are set in the Personal Information Protection Law of the People's Republic of China (PIPL). The law, by large, applies approaches similar to the GDPR as regards the scope of rights and protections (consent-based processing, etc.). The term 'personal information' is defined broadly as 'various information related to an identified or identifiable natural person recorded electronically or by other means, but does not include anonymized information'[482]. There is

---

477 Martin Beraja, Andrew Kao, David Y Yang, Noam Yuchtman, 'AI-tocracy' (2023) 138(3) The Quarterly Journal of Economics <https://doi.org/10.1093/qje/qjad012> accessed 9 December 2024.

478 Dan Milmo, Amy Hawkins, 'How China is using AI news anchors to deliver its propaganda' (Guardian, 18 May 2024) <www.theguardian.com/technology/article/2024/may/18/how-china-is-using-ai-news-anchors-to-deliver-its-propaganda> accessed 9 December 2024.

479 Arendse Huld, 'China's Sweeping Recommendation Algorithm Regulations in Effect from March 1' (China Briefing, 6 January 2022) <www.china-briefing.com/news/china-passes-sweeping-recommendation-algorithm-regulations-effect-march-1-2022/> accessed 10 December 2024.

480 Lycas Laursen, 'China's AI-Tocracy Quells Protests and Boosts AI Innovation' (IEEE Spectrum, 22 August 2023) <https://spectrum.ieee.org/china-facial-recognition> accessed 10 December 2024; Martin Beraja, Andrew Kao, David Y Yang, Noam Yuchtman, 'Autocracy and AI Innovation' (SCCEI China Briefs, 1 July 2022) <https://sccei.fsi.stanford.edu/china-briefs/autocracy-and-ai-innovation> accessed 10 December 2024.

481 Paul Mozur, 'One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority' (NY Times, 14 April 2019) <www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html> accessed 10 December 2024.

482 PIPL (n 464), art 4.

a category of sensitive information; interestingly, all personal information of a minor under the age of 14 years is included in this category.[483]

The regulation of automated decision-making is of relevance to AI. PIPL does prohibit such practices, but requires ensuring the transparency of the decision-making and the fairness and impartiality of the result; where the decision may have a significant impact on an individual's rights and interests, the subject has the right to request clarification or refuse from automated decision-making[484]

The Chinese AI legislation highlights the protection of personal data as a priority:

- The Algorithmic recommendation provisions prohibit the providers from falsely registering users, illegally trading or manipulating user accounts; and from false likes, comments, and reshares (article 14); the supervision and inspection bodies must maintain confidentiality of personal information (article 29).

- The Deep synthesis provisions mainly state the requirement to comply with personal data laws and requirements, including when the data is used for training. Besides, the provisions require prompting the users of a deep synthesis service to notify the individuals whose personal information is being edited (faces, voices) and obtain their independent consent in accordance with law (Article 14).

- In the context of generative AI services, there is a general obligation on the providers to protect personal information (articles 9 and 11 of the Generative AI Measures). The services need to identify users and monitor their inputs and the generated content, however – this will be discussed in the next section.

- The Draft AI law contains two relevant articles. Article 34 'Protection of Privacy and Personal Information' requires the developers and providers to protect the privacy and personal information-related rights and interests; prohibits the evaluation of individuals' behavioural habits, interests, or information about his or her economic, health, or credit status without the subject's consent; requires informing in advance on safety and security risks of the system; and prohibits the collection of non-essential personal information. Article 74 relates to biometric recognition. The article demands the AI processing of biometric information be limited to instances when there is a sufficient necessity, with strict protection

---

483 Ibid, art 28.

484 Ibid, art 24.

ensured. If a non-biometric alternative exists, preference should be given to the non-biometrical solution. If the biometric information generated by AI is used maliciously or has the risk of identifying a specific natural person, upon the request of the rights holder, the AI provider shall take necessary measures such as blocking, breaking links, and information deletion. The latter relates, for instance, to a problem of the generation of deep synthesis content of celebrities (not always with malicious intent) without the subject's consent.

The existing AI regulation primarily relies on general personal data protection legislation. This legislation generally corresponds to the global practices of personal data protection. The Draft AI Law proposes valuable AI-specific extensions to these basic rights. As discussed in subsection 3.3.3(b)(I) above, there seems to be a big difference between the processing of personal data by state agencies and by private providers.

(III) Safety

Similar to other compared jurisdictions, safety plays a key role in the Chinese AI policy. Based on the concept of trustworthy AI, China has developed safety measures applied in regulated AI use cases. The measures are often based on and are an extension of non-AI legislation.

Algorithmic recommendation services. The Algorithmic recommendation provisions contain general requirements for the service providers. The measures are aimed to 'carry forward the Core Socialist Values, preserve national security and the societal public interest, protect the lawful rights and interests of citizens, legal persons, and other organizations, and promote the healthy and orderly development of internet information services'.[485] The obligations include the following:

- Content control: the providers shall persist in being oriented towards 'mainstream values, optimize mechanisms for algorithmic recommendation services, actively transmit positive energy, and promote the uplifting use of algorithms'; the service shall not endanger national security or the societal public interest, disrupt economic and social order, or harm the lawful rights and interests of others, or transmit information that is prohibited by law; the provider must have measures to prevent and stop transmitting 'negative information' (article 6);

485 Algorithmic recommendation provisions (n 455), art I(1).

- To ensure algorithmic security, establish management systems, including technology ethics review, user registration, information dissemination examination, security assessment and monitoring, handling security incidents, data security and personal information protection, countering telecommunications and online fraud, allocate specialized personnel and technical support suited to the scale of the service and, importantly, disclose the rules for algorithmic recommendation services (article 6);

- To strengthen information security management, establish databases to be used to identify unlawful and harmful information; mark all the generated and synthetic; cease dissemination of unlawful information, record and report the cybersecurity and informatization departments thereof (article 9);

- To establish 'perfect mechanisms' for manual intervention and autonomous user choice, and 'vigorously present information conforming to mainstream value orientations in key segments such as front pages and main screens, hot search terms, selected topics, topic lists, pop-up windows' (article 11);

- Providers are encouraged to use tactics, such as content de-weighting, scattering interventions, etc., and optimize the transparency and understandability of search, ranking, selection, push notification, display, and other such norms, to avoid creating harmful influence on users, and prevent or reduce controversies or disputes (article 12);

- Providers that are internet news information services shall not use deepfake or synthesized news information, nor may they disseminate news information not published by work units in the State-determined scope (article 13);

- The providers are prohibited from manipulating accounts to interfere with information presentation such as by blocking information, making excessive recommendations, manipulating the order of top content lists or search results, controlling hot searches or selections, influencing online public opinion, or evading oversight and management (article 14);

- Do not unreasonably restrict other internet information service providers, or obstruct and undermine the normal operation of the internet information services that are lawfully provided, do not abuse a monopoly or conduct unfair competition acts (article 15);

- Providers of algorithmic recommendation services with public opinion properties or having social mobilization capabilities must report the domain of application, the algorithm type, algorithm self-assessment, content intended to publicize within 10 days of the service provision as well as update on modifications or ceasing the service (article 24) and display relevant record index number. Such providers shall also conduct security assessments (article 27) and keep network records (article 28). The compliance of these measures is supported by the non-AI requirements of the PRC Data Security Law and PRC Cybersecurity Law;

- The providers must also protect the user's human rights and personal data protection rights as discussed above.

The above measures are a combination of general principles and desired regulatory outcomes. The Algorithmic recommendation provisions in many aspects provide only heads of the measures without giving the implementation specifics. Nevertheless, the above seems to set the compliance framework in sufficient detail to enforce and penalize the non-compliance.

Deep synthesis internet information services. The Deep synthesis provisions set a list of general requirements:

- Legal content. The service must not be used to produce, reproduce, publish, or transmit illegal information, or to engage in illegal activities, e.g., that endanger the national security and interests, harm the image of the nation, harm the societal public interest, disturb economic or social order, or harm the lawful rights and interests of others (article 6). This includes the production and transmission of fake news.

- Establish a management system for user registration, review of algorithm mechanisms, scientific ethics reviews, review for information publication, data security, personal information protection, prevention of telecommunication network fraud, and emergency response, and shall have safe and controllable technical safeguard measures (article 7);

- Draft and disclose management rules and platform covenants, and post the security obligations (article 9);

- Identify the real identity of the users (article 10) – this measure is used to trace back the non-compliance behaviour and investigation;

- Control of users' conduct. The provider must have technical measures or manual methods to review the data imputed by uses and synthesis outcomes, and record and report incidents; if illegal or negative content is discovered, the provider must report the incident, take measures against the relevant user, e.g., issue a warning, restrict functions, suspend the service, or close the account (article 10); the providers shall also set mechanism for dispelling rumours or false information; store related records and report incidents (article 11);

- Provide convenient portals for users to appeal, public complaints and reports (article 13);

- Data management: the providers must ensure the security of training data, including the confidentiality of personal data, including obtaining consent of subjects to edit their biometric information – faces, voices (article 14);

- Technical management and periodic review of the algorithmic mechanism that produces synthesis, including conducting security assessments for generating and editing biometric information and special items, scenarios, and non-biometric information that might involve national security, nations' image, national interest and the societal public interest (article 15);

- Label the generated content (article 16), including special requirements (reasonable position or location) for deep synthesis that might cause confusion or mislead the public (article 17); removing the label is prohibited (article 18); and

- Conduct relevant filing formalities if the service has public opinion properties or the capacity for social mobilization: filing (article 19); carrying out security assessment (article 20); and assisting the supervision authorities (article 21).

Generative AI services. The Generative AI measures hold the providers responsible for the generated content (article 9) and require them to apply a list of measures:

- Content control: uphold the Core Socialist Values; no illegal content, including inciting subversion of national sovereignty or the overturn of the socialist system, endangering national security and interests or harming the nation's image, inciting separatism or undermining national unity and social stability, advocating terrorism or extremism, promoting ethnic hatred and ethnic discrimination, violence and obscenity, as well as fake and harmful information (article 4(1)). If the illegal

content is discovered, the provider must take measures (including against the users), correct/improve the model and report the incidents (article 14);

- No discrimination: design and select training data, model generalization and optimization accordingly (article 4.2.);
- Respect intellectual property rights, commercial ethics, no monopolies or unfair competition through algorithm advantages (article 4.3.); respect the lawful rights and interests, well-being of others, including the image, reputation, honour, privacy, and personal information (article 4.4);
- Ensure transparency in the generative AI service (article 4.5);
- Due training data and training model (article 7): the data and foundation models should have lawful sources; intellectual property rights observed; consent of data subject when using personal data is obtained; employ effective measures to increase the quality, truth, accuracy, objectivity and diversity of training data; due content tagging procedure (article 8);
- Describe information on the service to users, and employ measures to prevent minors from overreliance or addiction to generative AI services (article 10);
- Label the generated content (article 12);
- Provide safe, stable and sustained service to ensure users' normal usage (article 14);
- Employ an easy complaint and reporting mechanism (article 15); and
- Conduct a security assessment and filing if the service is with public opinion properties or the capacity for social mobilization (article 17).

All-AI provisions. In this respect, two documents must be observed: the Draft AI law and the China Safety Governance Framework.

The Draft AI law formulates principles and the regulatory framework, as well as specific requirements for certain use cases. It also defines obligations for the developers, providers, and users. The Draft AI law is designed to apply horizontally to the whole AI industry, irrespective of the use cases.

The draft provides for the following safety-related principles: scientific and technological ethics, fairness and impartiality, transparency and explainability, safety and accountability,

proper use (compliance with laws, ethics, fulfilling obligations, and respecting rights); human intervention (mechanism for human supervision and audit mechanisms; assessing and explaining safety risks); protection of IP rights, green development and diversified co-governance. The principles set the basic gridline for the regulatory framework.

The State should support the development and applying cybersecurity, data security and algorithmic security technologies, support relevant professional organizations, encourage risk insurance; encourage enterprise-led industry research institutes; formulate AI application scenarios (e.g., in government services, scientific research, finance, and autonomous driving); support funding of research in AI; and promote digital literacy.[486] The state shall also provide coordination, guide the implementation, conduct oversight, conduct grading and categorisation and implement special regulations for critical AI; the state should also conduct risk monitoring, safety and security assessment, and establish pilot mechanisms.[487]

The developers and providers have the obligations to:

- ensure the AI product or service complies with the law; take organizational and technical measures to this end; conduct regular inspection and monitoring. The measures must correspond to the magnitude of the risks, the probability of occurrence, the level of technological development, and the cost of implementation (article 41);
- conduct a safety risk assessment (article 42), including checking for biases and discrimination, for impact on the public interest, for risks and interests of individuals and safety risks, for compliance with ethics, and for appropriate and legal protection measures. It could be a self- or third-party assessment.
- report major safety and security incidents (article 43);
- maintain the safety of insurance policyholders: collaboration with the insurer to inspect the safety and security and ensure compliance (article 44);
- ensure the quality of training data and the diversity of data sets; no illegal data in the data sets (article 45);
- issue guidelines for users; react to misuse and illegal activity (article 46);

---

486 Draft AI law (n 25), ch II.

487 Ibid, ch V.

- ensure content safety and security, including no generation of false or harmful content – for AI service providers in the network information services (article 47);
- provide labelling AI content (article 48); and
- have a license, if required by law (article 49).

The Draft AI law establishes the category of 'critical AI' that includes AI that: is applied in critical information infrastructure; significantly impacts personal rights and interests such as life, freedom, and dignity; are foundation models above a certain level of computing, parameters, scale of use; and in other cases, provided by law.

The critical AI is subject to additional requirements, such as using secure and trustworthy physical equipment; if applicable, formulating appropriate data labelling rules and audit mechanisms; setting up a specialized security management institution; making public the designated person responsible for the supervision, management and AI compliance; registering the system at the national AI supervision platform; conducting a security risk assessment at least annually; disclosing security risks to users, and establishing a relevant disclosure mechanism; having a security emergency response plan; and reporting on corporate restructuring (merger, division) to the oversight government body.[488]

The Draft AI law contains a separate chapter VI that sets out specific provisions for certain AI use cases, e.g., for the use of AI by state organs, in judicial work, or medical services. For these cases, AI decisions may only be used as a reference. In news AI, the provider shall ensure the review and approval of the release mechanism, as well as ensure visible identifiers of AI-generated content. For AI social bots, the providers must reasonably control the quantity and quality of the published content; manipulation or guiding public opinion is prohibited. The use of biometric recognition in AI systems is only possible if there is no non-biometric alternative. Autonomous driving AI requires a license. Social scoring AI could be used only for improving credit quality and risk management; it may not lead to detrimental access to public services or encroach on other legitimate interests and rights of individuals. Lastly, the Draft AI law requires the developers of GPAI[489] to ensure safety and trustworthiness through value alignment and other technical means, conduct safety

---

488 Ibid, section VI(II).

489 According to art 94(v) of the Draft AI Law, GPAI means AI with broad cognitive capabilities that can be applied in multiple fields.

assessments, enhance transparency and explainability, and report the results to the supervision authority.

Finally, the Draft AI law provides for exclusions from its regulations for AI used by natural persons for personal and family affairs, in scientific research activities, and for free and open-source AI.

The Draft AI provides for a comprehensive detailed vision of the AI regulatory framework. Some of its items are even more detailed and thought through than the AI Act. The draft pays much attention to human-centric and relevant AI-specific protections.

The final document to be discussed in this section is the China Safety Governance Framework, which was released in September 2024. The document is non-binding, yet designed to be applied to the whole AI industry and stakeholders.

The framework declares the task to uphold a people-centred approach, adhere to the principles of developing AI for good, and promote consensus and coordinated efforts of AI safety governance among governments, international organizations, companies, research institutes, civil organizations, and individuals.

The safety governance principles of the China Safety Governance Framework include ensuring sustainable security while putting equal emphasis on development; prioritising innovative developments; focusing on preventing AI safety risks; stakeholder-specific approach; comprehensive 'whole-process' governance chain; safeguarding national sovereignty, security and development interests; protecting the legitimate interests and rights of individuals and legal entities; and guaranteeing AI the technology benefits humanity.[490] The framework does not pursue a safety-first or human-first approach but rather calls for balancing these aspects with the need for innovative development of AI.

Unlike other documents addressing safety and AI governance, the China Safety Governance Framework provides a breakthrough in AI risks, which includes:

- Inherent risks. This includes (i) risks from models and algorithms: risk of explainability (unpredictable and untraceable outputs, challenging to trace and rectify the anomalies); risk of bias and discrimination (personal biases, poor-quality datasets); risks of robustness (the systems are susceptible to changing operational environments and malicious interference); risks of stealing and tampering; risks of

490 China Safety Governance Framework (n 461), s 1.

unreliable output (generative AI hallucinations); risks of adversarial attacks; (ii) risks from data: illegal collection and use of data; improper content and poisoning in training data; unregulated training data annotation; data leakage; (iii) risks from AI systems: exploitation through defects and backdoors; computing infrastructure security; and risks of supply chain security (technology barriers, export restrictions; supply disruptions for chips, software and tools); and

- Risks in AI applications: cyberspace risks (content safety, users' misleading; information leakage, cyberattacks, flaw dissemination to downstream models), real-world risks (systems non-alignment, hallucinations and erroneous decisions; using AI in illegal activities; misuse of dual-use items and technologies); cognitive risks ('information cocoons'; cognitive warfare – fake news, shaping public values and cognitive thinking); and ethical risks (social discrimination, prejudice and widening the intelligence divide; challenging traditional social order; uncontrollable AI). [491]

The AI Framework suggests technological and governance measures to address the risks.

The first category of measures includes monitoring of technological developments, established security development standards, appropriate data collection, enhanced risk identification and mitigation, secure supply chain and timely repair and reinforcement of the system, established security protection mechanism and data safeguards, established service limitations and tracing the end use of AI systems to prevent high-risk application scenarios; identify unexpected, untruthful, and inaccurate outputs via technological means, preventing users' profiling by their enquiries; research and test technologies to prevent cognitive warfare. As regards the ethical risks, the AI Framework requires the filtering of training data and verifying the outputs to prevent discrimination and implement highly efficient emergency management and control measures for the use of AI by the government, critical information infrastructure, and areas directly affecting public safety, and people's health and safety.

The second category (governance measures) includes the proposed measures to be taken by the regulator: classification and grading AI systems based on their features and risk levels and bolstering the end-use management of AI; digital certificates to label AI systems serving the public; improve data security and personal information protection; AI-related ethical review standards and guidelines; strengthening AI supply chain security; advancement of the research on AI explainability; ongoing analysis of threats and formulating the emergency

---

491 Ibid, s 3.

response plans; enhance the training of AI safety talents; AI safety education and promotions of industry self-regulation and social supervisions; and promoting international exchange and cooperation.[492]

The safety guidelines are designed for specific stakeholders: developers, providers, users in key areas, and general users:[493]

- The developers must abide, *inter alia*, by the AI for good principles, science and technology ethics, conduct compliance reviews and information exchange; strengthen data security; assess the potential biases in AI models and algorithms; evaluate products and services for legal and risk management requirements, regularly conduct safety and security evaluation tests; formulate test rules and methods; evaluate the tolerance of their product or services for external interferences; and generate detailed test reports;

- Service providers shall publicize the capabilities and limitations of AI products and services; inform users on the application scope, support informed choices and cautions use; explain the AI product's accuracy to the users; provide user's responsibility framework with consent forms and service agreements; establish risk prevention and real-time risk monitoring mechanisms; assess the product's and service's ability to withstand and overcome adverse conditions (faults, attacks); report safety and security incidents and detected vulnerabilities; conduct an impact assessment of the AI product on users;

- Users in key areas (key sectors) shall assess the long-term and potential impact of applying AI technology and conduct a risk assessment and grading to avoid technology abuse; regularly perform audits; fully understand its data processing and privacy processing measures; use high-security passwords and multi-factor authentication; enhance network and supply chain security; limit data access and have data backup mechanisms, maintain confidentiality; avoid reliance on AI for decision making, have the option of switch to human-based traditional methods; and

- General users have limited obligations, which are more of a self-learning nature: raise awareness of the potential security risks and personal information protection,

---

492 Ibid, s 6.

493 Ibid.

be mindful of cybersecurity risks; review legal documents and policies before issuing an AI system; familiarize and use AI systems as intended; be aware of the potential impact of AI products on minors and take steps to prevent addiction and excessive use.

The China Safety Governance Framework contains the most detailed overview of AI risks and risk-specific measures broken down by key stakeholders and is one of the clearest guidance on AI among the all documents analysed in this work.

As overviewed, China has a safety framework for the three major regulated areas. Outside these areas, the AI framework is regulated by the non-binding, yet very thorough, China Safety Governance Framework. The Draft AI law, if adopted, will introduce full-scope horizontal regulation of safety. Given its very detailed provisions, which are up-to-speed with the technology development, this will greatly enhance the regulation of this area.

3.4 International regulation

Each of the compared jurisdictions claimed the intention to lead the global development of AI regulation, including relevant standards.

Indeed, many voices call for the global approach given the features of the technology and the perceived inability to effectively regulate its implication on the national level. One of the goals for cooperation is to mitigate the effect of the 'AI race'. This is especially important in the deployment of advanced powerful systems[494].

In 2017, the Future of Life Institute published the Asilomar Principles[495] calling for a culture of cooperation to avoid corner-cutting on safety standards. In 2019, the OECD AI Principles were adopted.[496] In 2020, the Global Partnership on Artificial Intelligence (GPAI) was launched. In 2023 the AI Safety Summit was held with 28 countries attended. In March 2024, the UN General Assembly adopted a landmark resolution on promoting safe, secure and trustworthy AI systems with an emphasis on the protection of human rights.[497]

---

494 Allan Dafoe, 'AI Governance: A Research Agenda' (Centre for the Governance of AI Future of Humanity Institute University of Oxford, 27 August 2018) <www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf> accessed 12 December 2024, 43.

495 Asilomar Principles (n 56).

496 The OECD AI Principles of 2019 (revised in 2024) <www.oecd.org/en/topics/ai-principles.html> accessed 12 December 2024.

497 UN General Assembly, 'Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development' (11 March 2024) A/78/L.49.

Finally, the signing of the AI Convention was a great achievement that set standards for human rights and data protection by AI systems.

The global work on AI standards is underway. There are active projects of ISO, the IEEE Standards Association, and the International Telecommunication Union. In July 2024, the NIST published its Plan for Global Engagement on AI Standards.[498]

4 CONCLUSION

This section discusses the main conclusions and takeaways of the research.

4.1 Control vs. development

AI is considered the most, or one of the most, influential technologies now. The EU, the US, and China are perceived as leading AI development centres; neither of the compared jurisdictions wants to lose this AI race. In his presentation at the Beneficial AGI 2019 Conference, Allan Dafoe, the Director and Principal Scientist at Google DeepMind compared the situation with the Cold War.[499] It is also a consensus that AI may pose significant risks. The jurisdictions, therefore, are balancing between the control and development of AI.[500] Too much control (regulation), and development may be hindered and key AI players opt for other jurisdictions. This forces the so-called 'race to the bottom'[501] – the temptation to develop more capable AI systems and neglect safety – which may end up in a catastrophic accident. This process affects not only governments but also commercial entities. Under market pressure, they are inclined to release faster even not well-tested AI systems.[502]

To hold these tendencies, each of the compared jurisdictions tries to find its optimal balance. There is a popular observation by Anu Bradford, who stated that 'the United States is

---

498 NIST AI 100-5 (n 438).

499 Allan Dafoe, 'AI Strategy, Policy, and Governance' (Future of Life Institute, the Beneficial AGI 2019 Conference, 28 March 2019) <www.youtube.com/watch?v=2IpJ8TIKKtI> accessed 10 December 2024.

500 Miriam C. Buiten, 'Towards Intelligent Regulation of Artificial Intelligence' (2019) 10(1) European Journal of Risk Regulation <DOI: https://doi.org/10.1017/err.2019.8> accessed 10 December 2024.

501 Stuart Armstrong, Nick Bostrom, Carl Shulman, 'Racing to the Precipice: A Model of Artificial Intelligence Development (Report)' (2023) Technical Report 2023-1, Future of Humanity Institute, Oxford University 1, 6.

502 Dan Hendrycks, Nicholas Carlini, John Schulman, Jacob Steinhardt, 'Unsolved Problems in ML Safety' (Arxiv, 16 June 2022) <https://doi.org/10.48550/arXiv.2109.13916> accessed 10 December 2024.

following a market-driven approach, China is advancing a state-driven approach, and the EU is pursuing a rights-driven approach'.[503] The research generally concurs with this conclusion.

The EU has consistently declared the protection of human rights (fundamental rights), as well as safety, as the top priority in its AI policy. It also became the first jurisdiction to adopt detailed binding regulation on AI that applies horizontally to all developers and providers, both in the public and private sectors. The signing of the AI Convention is consistent and promotes this policy further.

The US is generally cautious about adopting binding rules for private players and mostly relies on soft law instruments, including white papers and voluntary standards. Its work on standards seems to be the most active among the compared jurisdictions. The regulatory gaps are partly covered by the voluntary commitments of the major AI developers. At the same time, the US is actively building the governance framework on the use of AI by state agencies. Several detailed guidance documents are binding on them. Among the US regulatory priorities are national security, safety, and protection of human rights (civil rights). Most probably because a major part of AI regulation relates to the use of AI by the state, much attention is given to the use of AI as weapons or for cyberattacks and intelligence, as well as for addressing CBRN threats.

China has an AI policy that is focused on development and social well-being. Its binding legislation relates to three spheres: algorithmic recommendations, deep synthesis, and generative AI. The legislation sets strict content control provisions. There are national AI-specific standards and a draft law for comprehensive regulation of AI.

All the compared jurisdictions are active on the international level and aim to lead the regulation development process, including relevant AI standards. The most notable in this respect is the US, which even adopted a plan for the relevant work (NIST.AI.100-5).

4.2 Protection of human rights

All the compared jurisdictions claim a priority to protect human rights. They also acknowledge the need for AI-specific human rights. Among 'classic' human rights, AI regulations are mostly focused on the protection of the right to information, non-discrimination, privacy, and the right to complaint / the right to effective recourse. The EU and the US also draw attention to political rights (freedom of expression, the right to

---

503 Anu Bradford 'The Race to Regulate Artificial Intelligence: Why Europe Has an Edge Over America and China' (Foreign Affairs, 27 June 2023) <www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence-sam-altman-anu-bradford> accessed 10 December 2024.

assembly). The US contains many provisions for the protection of workers in the AI context. Although this category of rights is mentioned in all the compared jurisdictions, China seems to be the most protective of the rights of minors and the elderly.

Among the AI-specific rights defined by the jurisdictions are:

- the right to know of the use of AI. Additional notices may be provided by AI systems that may generate content, especially deep synthesis content;
- the right to opt out from being subject to AI decisions / the right to a human alternative. The addition or a substitute of this right may be the obligation of the provider to explain AI decisions to the user, intervene and override the AI-decisioning process, or use AI only to facilitate human-made decisions;
- the right to safe content, including no misinformation, deepfakes, illegal, and inappropriate generated content; labelling of generated content;
- the right to freedom of choice (no AI systems that manipulate human decisions or otherwise affect their freedom of choice);
- the right to not be subject to the use of emotions inferring, criminal activity predicting AI, or untargeted surveillance based on biometric information; and
- the right to not be subject to social scoring (at least if it could be detrimental to the individual);
- the right to help and training (projected in China for the cases of AI decisions in high-impact spheres).

In one of its publications, the Commission stated: 'In the global race for AI, Europe competes with conflicting visions. Between 'AI for profit' and 'AI for control', Europe could embrace 'AI for society', a human-centered, ethical, and secure approach that is true to our core values'[504].

---

[504] Commission, 'The global race for AI' (European Commission) <https://joint-research-centre.ec.europa.eu/jrc-mission-statement-work-programme/facts4eufuture/artificial-intelligence-european-perspective/global-race-ai_en> accessed 10 December 2024.

4.3 Personal data

As regards the processing of personal data in the AI context, the jurisdictions mostly rely on their general data protection legislation. Nevertheless, the AI regulation highlights the importance of the protection of personal data in the context of technology.

The jurisdictions often have AI-specific provisions for the processing of personal data and relevant rights. Mostly they relate to the processing of biometric data and other special categories of personal data. These rights and protections include:

- Qualifying AI systems that process special categories of data as high risk, for limited purposes or in the absence of alternative;
- Limited use of emotion-inferring systems (e.g., ban of their use at workplaces and education institutions);
- Prohibited or limited use of AI for biometric categorization;
- Prohibited or limited use of AI for real-time biometric identification in publicly available spaces for law enforcement; and
- Prohibition of making databases from untargeted scraping of faces in the Internet or CCTV footage;
- No manipulation with accounts, false likes, comments, or reshares;
- Specific consent from individuals that are subject to deep synthesis editing;
- No use of data for exploitation of behavioural habits (projected in China).

The EU has implemented an interesting mechanism: despite the EU being generally very protective as regards personal data, it authorized the use of personal data collected for other purposes for the AI developed in the regulatory sandboxes. Neither of the other compared jurisdictions implemented or declared a plan to implement the same mechanism.

4.4 AI safety

All the compared jurisdictions focus their regulatory policy on safety and trustworthiness. The jurisdictions also claim the risk-based approach, although the EU seems to use it the most consistently.

To adjust safety obligations to the risks, the jurisdictions define the category of high-risk AI. Usually, these are use cases that pose major safety risks or have the potential to significantly negatively impact human rights. The EU also defined prohibited AI practices that pose unacceptable risks. In the US there is only one guidance that defines prohibited AI and applies to state agencies. So far, China has not defined the prohibited AI category. The EU also has the category of limited-risk AI that is subject to extended transparency requirements.

By large, the scope of safety measures is similar among the compared jurisdictions. The key difference is if they are binding and the scope of AI systems they apply. As a rule, safety measures include:

- Having a risk management system;
- Compliance with data sets, and training program requirements;
- Having technical documentation on the system;
- Ensuring transparency requirements (AI explainability);
- Human oversight and adequate intervention measures;
- Robustness and security;
- Conducting an impact assessment;
- Conducting a conformity assessment (often by a third party);
- Registering in the relevant state register;
- Ensuring the protection of human rights and personal data;
- Adequate training of personnel;
- Labelling AI-generated content; and
- Ongoing monitoring.

The above provisions usually apply to the high-risk category of AI.

To further specify these requirements, the compared jurisdictions develop relevant standards and guidelines and encourage the development of voluntary (non-binding) best practices and codes of conduct by the market, civil institutions, and academia. The latter is perceived as a

milder instrument of state policy; the experience of their implementation can be used for official standards.

Following the risk-based approach, the EU has provided the regime of regulatory sandbox, a specific regime for developing AI systems. The regime provides for more control by the authorities, yet releases the developer from administrative liability if the system will not perform as intended. Besides, the regime provides for an exception for the use of personal data.

Another example of the risk-based approach is the exceptions applied to free open-sourced AI models. The jurisdictions encourage such practices since they promote the development of the technology. Besides, the obligations of AI developers do not usually apply, or apply on a limited basis, to AI systems and models designed for R&D, science, or non-commercial personal use.

China is focused on content control. In this respect, it provides requirements for the identification of users with their real names and supervising their inputs and outputs in deep synthesis and generative AI systems. Its regulation is also very specific as regards the prohibition of misinformation, fake news, manipulation with accounts, likes, and tampering with the natural outputs of recommendation systems. China also obliges the providers to disclose details of the used algorithms in certain cases.

Despite the current AI safety regulations look well-developed and specific, it will unlikely be possible to have a fixed AI safety policy for long. As the AI technology develops, adjustments will be needed.[505] The next stage that would definitely require rethinking AI safety is the development of AGI.

In its presentation in 2019, Allan Dafoe shared a survey on perceptions of AI governance challenges around the world. Among the top items were data privacy, autonomous vehicles, surveillance, hiring bias, value alignment, digital manipulation, criminal justice bias, disease diagnosis, and critical AI systems failure.[506] It seems that most of these items have been reflected in the EU and US AI policies.

---

505 Geoffrey Irving, Amanda Askell, 'AI Safety Needs Social Scientists' (Distill, 19 February 2019) <https://doi.org/10.23915/distill.00014> accessed 10 December 2024.

506 Dafoe (n 499).

NSCAI mentioned in its report that 'advances in AI ... could lead to inflection points or leaps in capabilities… [and] introduce new concerns and risks and the need for new policies, recommendations, and technical advances'[507].

4.5 AGI, ASI and the existential risk

The EU has declared the intention to build a future-proof AI regulation. This is easier said than done. The AI Act contains the definition of a general-purpose AI model (GPAI) and the 'high-capability' category of it that is subject to additional safety measures. China operates with the term. The most approximate notion in the US regulation is the 'dual-use frontier model'. The regulators struggle to distinguish these categories of AI from others. For instance, ChatGPT falls under the definition of GPAI but is not considered AGI by the AI industry.

The problem with this approach is that it defines too broad a scope of AI systems. Making all the AI systems that fall under the definition subject to extremely strict requirements (and these would be needed given the potential risks, including existential ones, that AGI may pose) would be an overkill. So, the regulation may be simply not ready to adequately address the risks of AGI once it is created.

So far, the jurisdictions require the developer of GPAI to disclose in a more detailed way the technical documentation on the model and its testing and have a deeper ethics (values) and alignment assessment of such systems, including the adversarial testing. Would that be enough?

Both the EU and the US recognise the existential risks with AI, and even specify relevant scenarios; China is less specific on that. Despite mentioning the risks, the relevant provisions are rather an acknowledgement than an action plan. So, at the moment the jurisdictions seem to consider the existential risks as remote and do not regulate it explicitly.

---

507 Final Report of the NSCAI of 2021 (n 330), 36, *brackets added*.

APPENDIX 1 – COMPARISON TABLE

REGULATION OF AI IN THE EUROPEAN UNION, THE UNITED STATES, AND CHINA

Aspect	EU	US	China
GENERAL			
General description of the regulatory framework	Detailed horizontal binding regulation of AI that applies to all AI systems in the EU, the first jurisdiction to adopt such comprehensive binding regulation. Distinctive human-centric approach.	Regulation of AI mostly consists of soft law. Detailed regulation of the use of AI by State agencies (State AI). Many acts relate to the application of AI in warfare, intelligence, and for national security purposes. Focus on the development of standards. Practice of voluntary commitments by major AI tech companies.	Binding regulation of three areas: algorithmic recommendations, deep synthesis, and generative AI. Focus of development. Detailed content control provisions. Preference of socialist values over individual human rights.
Adopted AI strategy	Yes, in 2017	Yes, in 2016	Yes, in 2017

Aspect	EU	US	China
Adopted AI ethics	Yes	Yes. A significant part thereof is sector-specific and refers to the State AI	Yes
Binding AI regulation	Yes	Yes, for the State AI.	Yes, for the three areas: algorithmic recommendations, deep synthesis, and generative AI. The Draft AI law proposes general horizontal regulation of AI.
Legal definition of AI	Yes	Yes	Yes
Extraterritorial effect	No, the obligation to have a local representative of a foreign AI system	No, have a local representative for the foreign provider of the State AI.	No. For generative AI, the Chinese authorities may notify the relevant organs outside China to bring a non-complying non-PRC service in line with the requirements.

Aspect	EU	US	China
HUMAN RIGHTS			
Human-centric approach	Yes	Declared, but appears a lesser priority than safety and national security	Declared, but community values are presumed to prevail
AI-specific rights	Yes	Yes	Yes
- Right to know	Yes	Yes	Yes
- Right to opt-out / to human alternative	No	Yes	Yes
- Right to explanation of AI-facilitated decisions	Yes, high-risk AI.	Yes	Projected

Aspect	EU	US	China
- Protection of human autonomy / freedom of choice	Yes, AI systems that materially distort behaviour or exploit the vulnerabilities of a natural person and have the potential to cause significant harm are prohibited.	Yes, sector-specific, for the State AI	Yes, from addictive practices, overconsumption, and overreliance on the system, especially of minors and the elderly.
- Protection from illegal/ harmful content	Yes	Non-binding standard and for the State AI	Yes, detailed binding regulation
- Labelling AI-generated content, including deep synthesis	Yes	Yes	Yes
- Human oversight / no AI autonomous decisions	Yes, for decisions involving personal data, for certain individuals impacting decisions (workers)	Yes, should be appropriate to the risk and impact of the decision; ban on certain AI-sole decisions in the State AI.	Projected for certain use cases (judicial system, medicine)

Aspect	EU	US	China
- Right to obtain help and training	No	No	Projected. For AI products and services that affect individuals' significant legitimate interests
Protection from human rights intrusive AI practices			
- AI predicting behaviour	Yes, AI prediction of criminal activity is prohibited	Yes, considered as rights-impacting AI	–
- Social scoring	Yes, such practices if they lead to detrimental treatment in other social contexts than where the data was collected or which is disproportionate to their social behaviour, or its gravity, are prohibited.	Not discussed	Projected. May be used for risk management and shall not infringe on equal access to public services or individuals' legitimate rights and interests.

Aspect	EU	US	China
Other features			
- Attention to political rights and democracy	Yes, specific references in the AI Act and the AI Convention. Systems that may affect the sphere are considered high-risk AI.	Yes, specific references in the AI Convention and AI Bill of Rights and guidance for the State AI, the category of rights-impacting AI.	Declared
- Attention to the rights and interests of workers	Yes, certain AI use cases are prohibited in the workplace, and some systems are considered high-risk.	Yes. Development plans discuss the technological unemployment. Protection of federal employees in the State AI applications.	–
- Attention to rights of children, elderly and people with disabilities	Yes	Yes	Yes, including from overreliance or addiction to generative AI services.
- AI literacy	Encouraged	Encourages	Encouraged

Aspect	EU	US	China
PERSONAL DATA			
Protection of personal data is a priority	Yes	Yes	Declared. In practice state use of AI often invades subject's privacy (e.g., mass surveillance and facial recognition)
AI-specific rights and protections from intrusive AI practices	Yes	Yes	Yes
- Processing of special categories of personal data	Yes. The AI systems are high-risk if process biometric data. Use of special categories of data in high-risk systems only if strictly necessary for detecting biases.	Projected. The use of sensitive categories of data on a strictly needed basis. Certain use cases are rights-impacting AI.	Yes, obtaining the consent of the individual, who is subject to deep synthesis using biometric information (voice, face). No biometric processing if it can be avoided.

Aspect	EU	US	China
- Biometric identification	Yes, high-risk AI	Yes, the relevant State AI are considered rights-impacting and subject to an impact assessment and stricter compliance requirements.	No
- Real-time remote biometric identification in public spaces	Yes, permitted only for targeted identification and subject to strict rules, e.g., obtaining an authorisation from the court.	Mass surveillance in public spaces is not permitted. Such State AI are considered rights-impacting and are subject to additional limitations.	No, such identification is currently used by the State.
- Emotion recognition	Yes, prohibited at workplace and educational institutions	Such State AI are risk-impacting, their use is subject to limitations	-
- Biometric categorization / profiling	Yes, AI systems used to deduce race, political opinions, trade union membership, religious or philosophical beliefs, sex life or	In State AI – rights-impacting AI subject to impact assessment and stricter compliance requirements	No, used by the state to profile and track minority groups.

Aspect	EU	US	China
	sexual orientation based on biometric data are prohibited.		
- Mass face recognition	Yes, AI systems that create a facial recognition database using untargeted facial images from the Internet and CCTV are prohibited.	Yes, the State AI that applies untargeted facial recognition is rights-impacting.	–
Other features	Exceptions from standard rules of data processing for regulatory sandboxes for the development of AI systems safeguarding substantial public interest, the developer may use personal data collected for other purposes.	Profiling by the State AI based on the individuals' exercising of protected rights is prohibited.	Prohibition of manipulation with accounts (falsely register, trade, false likes, comments)

Aspect	EU	US	China
SAFETY			
Safety as priority	Yes	Yes	Yes
Risk-based approach	Yes	Yes	Yes, in the context of balancing development and security
Categorisation of AI by risk	Yes: prohibited, high-risk, limited-risk, and low-risk AI. A special category of GPAI and GPAI with high capabilities	Yes, for the State AI: safety-impacting AI and rights-impacting AI	Projected – critical AI, special applications scenarios (judicial AI, medical AI, news AI, etc.).
Safety measures			
- Risk management system	Yes, for high-risk AI	Yes	Yes

Aspect	EU	US	China
- Requirements for datasets and training processes	Yes, for high-risk AI	Yes	Yes
- Transparency/ interpretability	Yes, for high-risk AI	Yes	Yes
- Human oversight	Yes, for high-risk AI	Yes	Yes
- Conformity assessment (including third-party)	Yes, for high-risk AI	Non-binding for the private sector; obligatory for the use of AI by government agencies. Certain obligations could follow from non-AI industry-specific regulation (e.g., medical devices)	Disclosure and notification for algorithmic recommendation service providers. Projected for all AI.
- Declaration of conformity	Yes, for high-risk AI	Yes, for safety-impacting and rights-impacting AI	Yes, for AI systems that have public opinion properties or capacity for social mobilization

Aspect	EU	US	China
- Registration of AI system, provider with state register	Yes, high-risk AI	Projected for the State AI	Yes, for AI systems that have public opinion properties or capacity for social mobilization. Projected for critical AI.
- Labelling AI-generated content	Yes	Yes, for the State AI. Projected for other AI.	Yes
- Security emergency plan	Yes, for high-risk AI	Yes, impacting State AI	Projected for Critical AI
- Technical documentation	Yes, for high-risk AI	Yes	Yes
- Post-deployment monitoring	Yes, for high-risk AI	Yes	Yes
- Local representative of foreign provider	Yes, for high-risk AI	Yes, in the procurement of the State AI	No

Aspect	EU	US	China
- Mandatory risk insurance	–	–	Projected
- Technical documentation	Yes, for high-risk AI	Yes	Yes
Restriction of autonomous AI deciding	Yes, the possibility is defined by a risk-based approach. For remote biometric identification, the decision must be verified by at least two humans.	Yes, for decision-making in critical areas (subject to risk assessments as part of oversight requirement)	Projected, for decisions significantly affecting rights and legitimate interests
Regulation of generative/deep synthesis AI	Yes, labelling deep synthesis content, notification of users of AI with such capabilities	Yes, synthetic content detection and labelling, discouraging anonymous use, and data provenance.	Yes, labelling, notification of users, identification of users by real name, control of users' inputs and outputs; protection from inappropriate content, especially of minors; dispelling rumours and false information

Aspect	EU	US	China
Regulation of AGI and ASI & frontier models	GPAI, GPAI with systemic risk. Subject to additional transparency requirements, model evaluation, stress testing and implementation of measures to mitigate systemic risks.	Dual-use frontier models – risk mapping, addressing the risk of misuse, model theft, monitoring and notifying accidents.	Projected for AGI – ensure safety, value alignment, conduct safety assessment, and report the results.
Regulation to address existential risks	Limited. Acknowledgement of the risk and the necessity to study; requirements for GPAI with systemic risk.	Limited. In the documents concerning national security, including CBRN threats.	Limited. The risks of uncontrollable, power-seeking, self-replicating, self-aware AI are mentioned in the standard; tracing the end use of AI systems.
Exemptions from general requirements:	Yes	Yes	Projected
- Free and open-source	Yes	–	Projected

Aspect	EU	US	China
- Personal use / non-professional activity	Yes	–	Projected
- Scientific research and development	Yes, AI Act 2(6)	Yes, for the rights-impacting and safety-impacting State AI.	Projected
- The system is not provided to the public	–	–	Yes, for generative AI services
Official AI standards	In development	Yes, and being developed	Yes, and being developed
Voluntary best practices developed by the market, voluntary commitments by the AI companies	Encouraged	Exist, encouraged	Encouraged
Pilot testing and regulatory sandboxes	Yes	The idea is mentioned, yet not formalised	–

Aspect	EU	US	China
Other features			Classification of safety risks into Inherent safety risks and safety risks in AI applications. For algorithmic recommendation services – providers shall uphold mainstream value orientations, vigorously disseminate positive energy, and advance the use of algorithms upwards and in the direction of good. Social bots – control the quantity and quality of generated content.

Disclaimer: The above overview shall not be considered exhaustive. It focuses on a limited number of aspects relevant to the research. The applications of AI for military, national security, intelligence, and dual-use AI are disregarded.

BIBLIOGRAPHY

Books

Brian Christian, *The Alignment Problem: Machine learning and human values* (W. W. Norton & Company, 2020) 67, 73.

Buchanan B, 'AI: A Cross Country Analysis of China versus the West' in Chishti S, Bartoletti I et al (eds), *The AI Book: The Artificial Intelligence Handbook for Investors, Entrepreneurs and FinTech Visionaries* (Wiley 2021)

Campbell H, Cheong P, *Thinking Tools for AI, Religion & Culture* (Network for New Media, Religion & Digital Culture Studies 2023).

Gerke S, Minssen T, Cohen G, *Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare*, *Artificial Intelligence in Healthcare* (Elsevier Inc. 2020)

Grim B, Finke R, *The Price of Freedom Denied* (Cambridge University Press 2010)

Heister S and Yuthas K, 'How Blockchain and AI Enable Personal Data Privacy and Support Cybersecurity' in Fernández-Caramés T, Fraga-Lamas P (eds), *Advances in the Convergence of Blockchain and Artificial Intelligence* (IntechOpen 2022)

Marsch N, *Artificial Intelligence and the Fundamental Right to Data Protection: Opening the Door for Technological Innovation and Innovative Protection*, *Regulating Artificial Intelligence* (Springer Cham 2019);

Russell S, Norvig P, *Artificial intelligence: a modern approach* (4th edn, Pearson 2021)

Sheikh H, Prins C, Schrijvers E, 'Artificial Intelligence: Definition and Background' in *Mission AI* (Springer Cham 2023)

Zerunyan F, 'Techno Innovations: The Role of Ethical Standards, Law and Regulation, and the Public Interest' in Alie E. Abbas (ed), *Next-Generation Ethics: Engineering a Better Society* (Cambridge University Press 2019)

Articles

— 'Ai Privacy: AI and Privacy: The Privacy Concerns Surrounding AI, Its Potential Impact on Personal Data - The Economic Times' (The Economic Times News, 25 April 2023) <https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms> accessed 11 December 2024

— 'Portman, Peters Introduce Bipartisan Bill to Ensure Federal Government is Prepared for Catastrophic Risks to National Security' (Homeland Newswire, 25 June 2022) <https://web.archive.org/web/20220625220611/https://homelandnewswire.com/stories/627890045-portman-peters-introduce-bipartisan-bill-to-ensure-federal-government-is-prepared-for-catastrophic-risks-to-national-security> accessed 9 December 2022.

Alvarez D et al., 'U.S. State AI Update: Utah Enacts First State Generative AI Transparency Law – California, Colorado and Connecticut Are Close Behind' (Willkie Farr & Gallagher LLP, 8 May 2024) <www.willkie.com/-/media/files/publications/2024/05/us_state_ai_update_utah_enacts_first_state_generative_ai_transparency_law.pdf> accessed 9 December 2024

Armstrong S, Bostrom N, Shulman C, 'Racing to the Precipice: A Model of Artificial Intelligence Development (Report)' (2023) Technical Report 2023-1, Future of Humanity Institute, Oxford University

Aten J, 'Stop Calling Everything A.I. It's Just Computers Doing Computer Stuff' (Inc.com, 13 May 203) <www.inc.com/jason-aten/stop-calling-everything-ai-its-just-computers-doing-computer-stuff.html> accessed 5 December 2024

Belton K, 'How should AI be regulated' (IndustryWeek, 7 March 2019) <www.industryweek.com/technology-and-iiot/article/22027274/how-should-ai-be-regulated> accessed 5 December 2024

Beraja M, Kao A, Yang D, Yuchtman N, 'AI-tocracy' (2023) 138(3) The Quarterly Journal of Economics <https://doi.org/10.1093/qje/qjad012> accessed 9 December 2024

Beraja M, Kao A, Yang D, Yuchtman N, 'Autocracy and AI Innovation' (SCCEI China Briefs, 1 July 2022) <https://sccei.fsi.stanford.edu/china-briefs/autocracy-and-ai-innovation> accessed 10 December 2024

Bradford A, 'The Race to Regulate Artificial Intelligence: Why Europe Has an Edge Over America and China' (Foreign Affairs, 27 June 2023) <www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence-sam-altman-anu-bradford> accessed 10 December 2024

Brown R, Truby J, Ibrahim I, 'Mending Lacunas in the EU's GDPR and Proposed Artificial Intelligence Regulation' (2022) 9 European Studies: The Review of European Law, Economics and Politics

Buiten M, 'Towards Intelligent Regulation of Artificial Intelligence' (2019) 10(1) European Journal of Risk Regulation <DOI: https://doi.org/10.1017/err.2019.8> accessed 10 December 2024

Chipman A, 'Artificial Intelligence in China: Shenzhen Releases First Local Regulations' (China Briefing, 29 July 2021) <www.china-briefing.com/news/artificial-intelligence-china-shenzhen-first-local-ai-regulations-key-areas-coverage/> accessed 9 December 2024;

Clark J, 'Why 2015 Was a Breakthrough Year in Artificial Intelligence' (Bloomberg, 8 December 2015) <www.bloomberg.com/news/articles/2015-12-08/why-2015-was-a-breakthrough-year-in-artificial-intelligence> accessed 4 December 2024

Cohen G et al, 'The European Artificial Intelligence Strategy: Implications and Challenges for Digital Health' (2020) 2 The Lancet Digital Health 7

Dafoe A, 'AI Governance: A Research Agenda' (Centre for the Governance of AI Future of Humanity Institute University of Oxford, 27 August 2018) <www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf> accessed 12 December 2024

— 'AI Strategy, Policy, and Governance' (Future of Life Institute, the Beneficial AGI 2019 Conference, 28 March 2019) <www.youtube.com/watch?v=2IpJ8TIKKtI> accessed 10 December 2024

Domonoske C, 'Elon Musk Warns Governors: Artificial Intelligence Poses "Existential Risk"' (NPR, 17 July 2017) <www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk> accessed 5 December 2024

Engfeldt H, Stallins G, 'The Nuts and Bolts of the proposed California Safe and Secure Innovation for Frontier Artificial Intelligence Models Act' (Connect on Tech, 12 August 2024) <www.connectontech.com/the-nuts-and-bolts-of-the-proposed-california-safe-and-secure-innovation-for-frontier-artificial-intelligence-models-act/> accessed 9 December 2024

Felten E, Lyons T, ‘The Administration’s Report on the Future of Artificial Intelligence’ (the White House, 12 October 2016) <https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence> accessed 8 December 2024

Fung B, ‘UN Secretary General embraces calls for a new UN agency on AI in the face of “potentially catastrophic and existential risks” (CNN Business, 18 July 2023) <https://edition.cnn.com/2023/07/18/tech/un-ai-agency/index.html> accessed 6 December 2024

Geraci R, ‘Apocalyptic AI: Religion and the Promise of Artificial Intelligence’ (2008) 76 Journal of the American Academy of Religion

Hendrycks D, Carlini N, Schulman J, Steinhardt J, ‘Unsolved Problems in ML Safety’ (Arxiv, 16 June 2022) <https://doi.org/10.48550/arXiv.2109.13916> accessed 10 December 2024

Honrada G, ‘America’s AI edge fading fast to China’ (Asia Times, 19 September 2022) <https://asiatimes.com/2022/09/americas-ai-edge-fading-fast-to-china/> accessed 8 December 2024

Huang J, Nvidia CEO, ‘NVIDIA Unveils "NIMS" Digital Humans, Robots, Earth 2.0, and AI Factories’ (YouTube, 5 June 2024) <www.youtube.com/watch?v=IurALhiB6Ko&t=1637s> accessed 7 July 2024

Huld A, ‘China’s Sweeping Recommendation Algorithm Regulations in Effect from March 1’ (China Briefing, 6 January 2022) <www.china-briefing.com/news/china-passes-sweeping-recommendation-algorithm-regulations-effect-march-1-2022/> accessed 10 December 2024

Interesse G, ‘Technology in China: Draft Trial Measures’ (China Briefing, 11 May 2023) <www.china-briefing.com/news/china-ethical-review-of-science-and-technology-draft-trial-measures/> accessed 4 December 2024

Irving G, Askell A, ‘AI Safety Needs Social Scientists’ (Distill, 19 February 2019) <https://doi.org/10.23915/distill.00014> accessed 10 December 2024

Kachra A, ‘Making Sense of China’s AI Regulations’ (Holistic AI, 12 February 2024) <www.holisticai.com/blog/china-ai-regulation> accessed 9 December 2024

Kharpal A, ‘A.I. is in its “infancy” and it’s too early to regulate it, Intel CEO Brian Krzanich says’ (CNBC, 7 November 2017) <www.cnbc.com/2017/11/07/ai-infancy-and-too-early-to-regulate-intel-ceo-brian-krzanich-says.html> accessed 5 December 2024

Laursen L, ‘China’s AI-Tocracy Quells Protests and Boosts AI Innovation’ (IEEE Spectrum, 22 August 2023) <https://spectrum.ieee.org/china-facial-recognition> accessed 10 December 2024

Lee W, ‘Gov. Gavin Newsom vetoes AI safety bill opposed by Silicon Valley’ (Los Angeles Times, 29 September 2024) <www.latimes.com/entertainment-arts/business/story/2024-09-29/gov-gavin-newsom-vetoes-ai-safety-bill-scott-wiener-sb1047> accessed 9 December 2024

Macey-Dare R, ‘How Soon is Now? Predicting the Expected Arrival Date of AGI- Artificial General Intelligence’ (2023) <https://ssrn.com/abstract=4496418> accessed 8 December 2024

Margaret R. Szewczyk, C|EH A, ‘Artificial Intelligence and Copyrights: Tennessee’s ELVIS Act Becomes Law’ (Armstrong&Teasdale, 27 March 2024) <www.armstrongteasdale.com/thought-leadership/artificial-intelligence-and-copyrights-tennessees-elvis-act-becomes-law/> accessed 9 December 2024

McGrath A, Jonker A, ‘What is AI safety’ (IBM, 15 November 2024) <www.ibm.com/think/topics/ai-safety> accessed 6 December 2024

Mihaly H, ‘A criticism of AI ethics guidelines’ (2020) XX(4) Informacios Tarsadolom <https://doi.org/10.22503/inftars.XX.2020.4.5> accessed 12 December 2024

Milmo D, Hawkins A, ‘How China is using AI news anchors to deliver its propaganda’ (Guardian, 18 May 2024) <www.theguardian.com/technology/article/2024/may/18/how-china-is-using-ai-news-anchors-to-deliver-its-propaganda> accessed 9 December 2024

Mozur P, ‘One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority’ (NY Times, 14 April 2019) <www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html> accessed 10 December 2024

O’Shaughnessy M et al, ‘What Governs Attitudes toward Artificial Intelligence Adoption and Governance?’ (2023) 50 Science and Public Policy

Owczarczuk M, ‘Ethical and regulatory challenges amid artificial intelligence development: an outline of the issue’ (2023) 22 Economics and Law 2

Papyshev G, Yarime M, 'The state’s role in governing artificial intelligence: development, control, and promotion through national strategies' (2023) 6:1 Policy Design and Practice

Picht P, Thouvenin F, ‘AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property’ (2023) 54 IIC International Review of Intellectual Property and Competition Law

Pittman P, Hafiz A, Hamm A, ‘Data Protection Laws and Regulations USA 2024’ (ICLG, 31 July 2024) <https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa> accessed 9 December 2024

Pope N, ‘What Is Trustworthy AI?’ (Nvidia, 1 March 2024) <https://blogs.nvidia.com/blog/what-is-trustworthy-ai/> accessed 6 December 2024

Querci I et al, ‘Explaining How Algorithms Work Reduces Consumers’ Concerns Regarding the Collection of Personal Data and Promotes AI Technology Adoption’ (2022) 39 Psychology and Marketing

Radi G, ‘Comparative Analysis of the AI Regulation of the EU, US and China from a Privacy Perspective’ (2023) 46th ICT and Electronics Convention, MIPRO 2023 - Proceedings

Reed C, ‘How should we regulate artificial intelligence?’ (2018) 376 Philos Trans A Math Phys Eng Sci. <https://doi.org/10.1098/rsta.2017.0360> accessed 5 December 2024

Steene S, Jenks C, ‘The Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy’ (Lieber Institute, West Point, Articles of War, 13 November 2023) <https://lieber.westpoint.edu/political-declaration-responsible-military-use-artificial-intelligence-autonomy/> accessed 9 December 2024

Thormundsson B, ‘Artificial intelligence (AI) market size worldwide from 2020 to 2030’ (Statista, 28 November 2024) <www.statista.com/forecasts/1474143/global-ai-market-size> accessed 4 December 2024

Trojecki M, ‘Is Artificial Intelligence Good or Bad: Debating the Ethics of AI’ (IIOT World, 10 June 2019) <www.iiot-world.com/artificial-intelligence-ml/artificial-intelligence/the-good-the-bad-and-the-ugly-debating-the-ethics-of-ai/> accessed 5 December 2024

Wach K et al, ‘The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT’ (2023) 11 Entrepreneurial Business and Economics Review

Wang L et al, ‘A Survey on Large Language Model Based Autonomous Agents’ (2024) 18 Frontiers of Computer Science 186345 <https://doi.org/10.1007/s11704-024-40231-1> accessed 4 December 2024

Wiener N, ‘Some Moral and Technical Consequences of Automation’ (Science, 6 May 1960) <www.science.org/doi/10.1126/science.131.3410.1355> accessed 11 December 2024

Wright G, ‘California Consumer Privacy Act (CCPA)’ (TechTarget, December 2023) <www.techtarget.com/searchcio/definition/California-Consumer-Privacy-Act-CCPA> accessed 9 December 2024

Xing Y et al, ‘AI Privacy Opinions between US and Chinese People’ (2023) 63 Journal of Computer Information Systems

Yang Z, ‘Four things to know about China’s new AI rules in 2024’ (MIT Technology Review, 17 January 2024) <www.technologyreview.com/2024/01/17/1086704/china-ai-regulation-changes-2024/> accessed 3 December 2024

Press-releases

‘Streamline compliance with Holistic AI’ (NYC Bias Audit) <www.nycbiasaudit.com/> accessed 9 December 2024

Commission, ‘AI Act’ (European Commission, 14 October 2024) <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-a> accessed 5 December 2024

Commission, ‘AI Act’ (European Commission, 14 October 2024) <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> accessed 7 December 2024

Commission, ‘Artificial intelligence: Commission kicks off work on marrying cutting-edge technology and ethical standards’ (European Commission, 9 March 2018) <https://ec.europa.eu/commission/presscorner/detail/en/ip_18_1381> accessed 6 December 2024

Commission, ‘Commission appoints expert group on AI and launches the European AI Alliance’ (European Commission, 14 June 2018) <https://digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance> accessed 6 December 2024

Commission, ‘Europe fit for the Digital Age: Commission proposes new rules for digital platforms’ (European Commission, 15 December 2020) <https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2347> accessed 7 December 2024

Commission, ‘The global race for AI’ (European Commission) <https://joint-research-centre.ec.europa.eu/jrc-mission-statement-work-programme/facts4eufuture/artificial-intelligence-european-perspective/global-race-ai_en> accessed 10 December 2024

NIST, ‘Department of Commerce Announces New Guidance, Tools 270 Days Following President Biden’s Executive Order on AI’ (NIST, 26 July 2024) <www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following> accessed 9 December 2024

White House, ‘FACT SHEET: Biden-Harris Administration Announces New AI Actions and Receives Additional Major Voluntary Commitment on A’ (WH.GOV, 26 July 2024) <www.whitehouse.gov/briefing-room/statements-releases/2024/07/26/fact-sheet-biden-harris-administration-announces-new-ai-actions-and-receives-additional-major-voluntary-commitment-on-ai/> accessed 9 December 2024

White House, ‘FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed

by AI' (WH.GOV, 21 July 2023) <www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/> accessed 9 December 2024;

White House, 'FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI' (WH.GOV, 12 September 2023) <www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/> accessed 9 December 2024

White House, 'Fact Sheet: Key AI Accomplishments in the Year Since the Biden-Harris Administration's Landmark Executive Order (WH.GOV, 30 October 2024) <www.whitehouse.gov/briefing-room/statements-releases/2024/10/30/fact-sheet-key-ai-accomplishments-in-the-year-since-the-biden-harris-administrations-landmark-executive-order/> accessed 9 December 2024

White House, 'Fact Sheet: Key AI Accomplishments in the Year Since the Biden-Harris Administration's Landmark Executive Order' (WH.GOV, 30 October 2024) <www.whitehouse.gov/briefing-room/statements-releases/2024/10/30/fact-sheet-key-ai-accomplishments-in-the-year-since-the-biden-harris-administrations-landmark-executive-order/> accessed 9 December 2024

White House, 'Relationship to Existing Law and Policy' (the White House) <www.whitehouse.gov/ostp/ai-bill-of-rights/relationship-to-existing-law-and-policy/> accessed 7 July 2024

White House, 'White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse' (WH.GOV, 12 September 2024) <www.whitehouse.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse/> accessed 9 December 2024

Other sources

'7 Key AI Investment Statistics Every Investor Should Know' (Edge Delta, 24 May 2024) <https://edgedelta.com/company/blog/ai-investment-statistics> accessed 5 December 2024

'AI alignment' (Wikipedia) <https://en.wikipedia.org/wiki/AI_alignment> accessed 6 December 2024

'AI Boom' (Wikipedia) <https://en.wikipedia.org/wiki/AI_boom> accessed 4 December 2024

'AI effect' (Wikipedia) <https://en.wikipedia.org/wiki/AI_effect> accessed 5 December 2024

'AI Ethics Guidelines Global Inventory' (AlgorithmWatch) <https://inventory.algorithmwatch.org/> accessed 6 December 2024

'AI Safety Summit' (Wikipedia) <https://en.wikipedia.org/wiki/AI_Safety_Summit> accessed 4 December 2024

'AI safety' (Wikipedia) <https://en.wikipedia.org/wiki/AI_safety> accessed 6 December 2024

'Artificial general intelligence' (Wikipedia) <https://en.wikipedia.org/wiki/Artificial_general_intelligence> accessed 7 December 2024

‘Artificial Intelligence’ (Wikipedia) <https://en.wikipedia.org/wiki/Artificial_intelligence#History> accessed 4 December 2024

‘Core Socialist Values’ (Wikipedia) <https://en.wikipedia.org/wiki/Core_Socialist_Values> accessed 9 December 2024

‘Existential risk from artificial intelligence’ (Wikipedia) <https://en.wikipedia.org/wiki/Existential_risk_from_artificial_intelligence> accessed 6 December 2024

‘Generative artificial intelligence’ (Wikipedia) <https://en.wikipedia.org/wiki/Generative_artificial_intelligence> accessed 7 July 2024

‘GPT-3’ (Wikipedia) <https://en.wikipedia.org/wiki/GPT-3> accessed 4 December 2024

‘Intelligence’ (Wikipedia) <https://en.wikipedia.org/wiki/Intelligence> accessed 5 December 2024

‘Intelligent agent’ (Wikipedia) <https://en.wikipedia.org/wiki/Intelligent_agent> accessed 5 December 2024

‘Regulation of artificial intelligence’ (Wikipedia) <https://en.wikipedia.org/wiki/Regulation_of_artificial_intelligence> accessed 5 December 2024

‘What is Artificial Intelligence (AI)’ (Google Cloud) <https://cloud.google.com/learn/what-is-artificial-intelligence> accessed 5 December 2024