Building a better Artificial Intelligence (AI) future – Agile regulations for risk mitigation

Jana Sukkarieh, PhD, CTO at Noha Z (www.nohaz.co).

AI is not just another technology. It is a foundational shift because of key technological advancements that have made it a “general-purpose” technology[1] which means today’s AI is a transformative force with the potential to reshape entire industries and society.

The main question facing us then is how will future generations look back on this decade? Will they see it as the time we wisely guided this AI power to solve humanity’s great challenges and ensure prosperity for all, or as the time we became slaves to systems that influence our every thought and action? Or, even worse, will they see this decade as the time we lost control through our own short-sightedness, or we widened injustice and inequity worldwide?

Integrating AI successfully into our society is not predetermined. It is a goal we must actively choose by implementing agile, risk-based regulations that encourage innovation while protecting our core values. The alternative is a path of reactive, fear-based, or overly rigid rules. And it poses just as great a threat as no regulation at all.

Although AI algorithms[2] themselves are not inherently harmful the potential for negative outcomes generally depends on how an AI system or product is designed, developed or deployed. In many cases, negative outcomes could be avoided by educating users and managers, especially in high stake industries such as health, legal, or financial about AI algorithms and their potentially harmful influences or outcomes.

Remember, in some cases, even top AI experts are uncertain about the potential harm or outcomes of some AI models, and as AI is getting more sophisticated and pervasive into society, the risk of harm increases.

Principles for guiding AI regulation in the right direction.

What follows are guidelines to be observed while navigating general regulations and action steps we can take. How can we move forward?

- **Focus on the AI application rather than the algorithm**: For example, a medical diagnostic AI requires different policy rules than a generative art tool or autonomous weapons.
- **Use agile, outcome-based standards**: Agile regulation works much better than rigid, static laws. So, make sure you support creating adaptable frameworks designed for resilience, capable of integrating new developments and evolving insights. Regulators must be empowered to foster innovation and this calls for organizations to maintain open communication throughout the process. The use of regulatory sandboxes[3] is essential for testing and adapting to new developments.
- **Mandatory transparency and audits**: For high-risk AI, regulators should require a list of “regulatory labels” like “nutrition labels,”[4] independent bias audits,[5] and clear explanations for significant decisions. Build trust through verification, not obscurity.
- **Global cooperation, not fragmentation**: We need to work toward international standards and norms to prevent chaotic patchwork of conflicting laws that could stifle global innovation and security.
- **Fostering open innovation and competition**: Ensure regulations do not accidentally cement the power of tech giants by creating rules only they can afford to follow. Regulators need to support open-source models and a diverse market – a market that benefits everyone.

Although we support regulations and the establishment of clear policies, it is important to emphasize that good intentions do not always make for good rules.

How well-intentioned regulation can fail – We believe the missteps include the following:

- **The precautionary paralysis misstep**: strangling nascent innovation with overly burdensome, pre-emptive rules based on hypothetical long-term risks, causing a brain drain and ceding leadership to less scrupulous nations.
- **The "one-size-fits-all" misstep**: applying the same blunt rules to a university researcher, a startup and a big tech company, crushing competition and smothering creativity.
- **The theatre over substance misstep**: mandating symbolic steps on AI ethics without real power or enforcing hollow transparency that does not lead to accountability or change. More concerning is if the well-intentioned laws exist but only enforced on "our enemies" but not "our friends".
- **The reactive whack-a-mole misstep**: legislating after each scandal, for example, a deep fake election incident, instead of building a proactive, resilient framework capable of handling new challenges.

As non-legal experts, many of us struggle to navigate legal guidelines.

In general, to navigate AI regulations, legal frameworks or guidelines and better understand your obligations, consider questions that fall under the following categories.

General AI regulation

Asking about the scope or definition of AI in the regulation or framework is paramount. What is being regulated? Because this has major implications on who is liable when something goes wrong, are those being regulated the data scientists, developers, researchers, deployers, providers and/or even users?

Think about the consequences for your hiring process. If you ask a developer to sign a contract making him or her responsible for everything that might go wrong, you will be left with no applicants. No one would agree to take on that level of risk, and you would not be able to hire anyone.

In the EU AI Act[6], everyone, in one capacity or another involved in limited or high-risk systems is mentioned whereas in the US Senate Bill 1047 in California (https://legiscan.com/CA/text/SB1047/id/2919384), developers particularly seem to be most responsible for what the bill calls Frontier AI.

Another question that fits within the general AI regulation is what are the criteria or how do you determine risk level or harm? If we were to discuss any AI application – such as facial recognition, document discovery or contract generation – we might have some disagreements on what constitutes a risk level.

To illustrate the ambiguity of risk assessment, consider the weather prediction tools that are widespread and used by most people worldwide, common sense implies that risk level is minimal due to the unpredictability of the weather. In the UK, it is hard to imagine someone suing the meteorology department when the four seasons could materialize, unexpectedly, on the same day although some global lawsuits filed against some meteorology departments were successful. In such cases, the judge might have considered the harm of the weather prediction tool as limited or high-risk and not minimal, especially if it causes someone's death.

A third question that falls under general AI regulation is about the role humans must play. Do they oversee everything at every stage? What is their required AI literacy?

Data governance

Questions under data governance could include how to ensure my data is relevant and representative and not erroneous? Note that in some cases, the EU AI act mentions data completeness under the general AI regulation. What is data completeness?

Another question under data governance that we should ask is how to strike a balance between the need to train AI and privacy compliance concerns? What about cross border data flows?

Ethical considerations

This focuses on the ethical principles and guidelines that must be followed. Or how should we address potential biases, discrimination or societal impacts of AI?

Enforcement and compliance

Here we should ask about what specific regulations apply to my AI system. How can we ensure compliance? And of course, who or which roles are liable?

So again, the worldwide need for well-designed and effective regulations, legal frameworks or guidelines is essential.

Global laboratory for AI governance

Many concerns existed even before the rise of Large Language Model (LLM)-based Generative AI. These include the individual, or union, and the enterprise concerns.

Individual or unions concerns include the following:

- **Job Substitution:** AI and automation may replace jobs, especially in repetitive or predictable tasks, leading to economic displacement and the need for workforce retraining.

- **Misinformation:** AI can be used to generate convincing but false information, such as deep fakes or fake news, which can spread rapidly and undermine trust in media and institutions.

- **Biases:** AI systems can learn, amplify, and perpetuate existing human biases in their training data, leading to unfair or discriminatory outcomes in areas like hiring, lending, and criminal justice.

- **Lack of privacy:** The data required to train and operate many AI systems can be collected and used to compromise personal privacy, often without an individual's full awareness or consent.

- **Surveillance:** AI-powered surveillance technologies, such as facial recognition, employee-monitoring and wiretapping, raise concerns about the loss of civil liberties and the potential for a society under continuous observation.

- **Impersonation:** AI can be used to create realistic impersonations of individuals through voice cloning or deep fakes, posing risks for fraud, social engineering, and identity theft.

- **Unreliable profiling:** AI systems may create profiles of individuals based on incomplete or inaccurate data, leading to flawed decisions or unfair targeting in marketing, law enforcement, or other applications.

- **AI applications learning curve:** The complexity of many AI tools or agents means that users must undergo a significant learning process to use them effectively, which can be a source of frustration and a barrier to adoption.

- **Ethical implications:** This surrounds the moral dilemmas posed by AI, such as accountability. For example, when an AI makes a harmful decision, it could affect the ethical use of autonomous weapons.

- **Energy consumption and the effects on the environment:** Training and running large-scale AI models require significant computational power, consuming vast amounts of electricity and water and contributing to carbon emissions. Big Tech companies like Microsoft, Google, and Amazon are facing criticism for building data centers in drought-prone areas and for demonstrating the lack of transparency in their water usage. The high energy consumption of AI data centers directly impacts the environment, potentially creating climate change by straining water and power grids.

- **Larger inequity in our societies:** AI development and access are not evenly distributed. A few large tech companies and wealthy nations dominate power and resources, which could exacerbate existing social and economic inequalities globally.

- **Ransomware:** AI could enhance ransomware by automating and improving every step of an attack, from initial reconnaissance to random negotiation. This might create a digital “arms race” between attackers and defenders.

- **Weaponry and wars:** AI is being integrated into military and defense systems to supposedly enhance speed, precision, and efficiency, leading to the development of highly advanced, and controversial, weaponry systems. This might create an “arms race” between attackers and defenders and the indiscriminate killing of many civilians.

The enterprises’ concerns include:

- **Return on investment (ROI):** The significant upfront costs and resources required for developing and deploying AI systems don't always translate into a clear or timely financial return, making it a risky investment for businesses.

- **Misinformation:** Again, AI can be used to generate highly convincing but false information, such as deep fakes and fake news, which can spread rapidly and erode public trust.

- **Bad behavior/fraud/misuse:** Malicious actors can use AI to automate and scale fraudulent activities, like producing phishing attacks, or identity theft, or creating malicious software, posing a serious threat to individuals and organizations.

- **Intellectual property (IP) or copyright infringements:** AI models trained on vast amounts of data can inadvertently reproduce or closely mimic copyrighted material, raising complex legal questions about ownership and intellectual property rights.

- **Difficulty of integration:** Integrating new AI systems with a company’s existing legacy software and workflows can be complex and time-consuming, creating operational friction and unexpected costs.

- **Vulnerability to cyber threats:** AI systems can be targets of cyberattacks, and can be exploited to launch more sophisticated attacks, such as adversarial attacks that manipulate AI models into making incorrect decisions.

- **Poor data quality:** The performance of an AI model directly depends on the quality of its training data. Inaccurate, incomplete, or biased data can lead to flawed and unreliable outcomes.

- **Lawsuits:** Companies deploying AI systems face legal risks, including lawsuits related to a lack of explainability, discrimination, or privacy violations, which can result in significant financial penalties and reputational damage.

- **Market competitiveness:** The rapid pace of AI innovation means that businesses failing to adopt and adapt to new AI technologies risk falling behind their competitors.

- **Reputation:** A malfunctioning, biased, or misused AI system can lead to public backlash and severely damage a company's brand and reputation.

- **Fake reviews:** AI can be used to generate large numbers of inauthentic reviews and ratings, manipulating consumer decisions and undermining the integrity of online platforms and enterprises.

- **Another AI winter** refers to a period of reduced funding and interest in AI research and development, like historical periods when the hype around AI failed to meet expectations. The fear is that inflated promises and high-profile failures could lead to another such downturn.

In a high-stake or regulated domain, all of the above are even more consequential. One would like to assume that governments or enterprises are moral enough to not unleash products that exacerbate or fuel these concerns. But, unfortunately, this level of protection is not happening! And that’s why the need for regulations and frameworks that address these concerns are stronger than ever.

We all agree that AI communities and users, in general, have been talking, for years, about protective principles and measures, even without specific legal frameworks, guidelines or regulations. However, it is of utmost importance that some careful decisions are moved up to legislative levels to ensure that these measures address and even go beyond current human concerns about AI.

The world has become a laboratory for AI regulations and principles. What follows reveals how different parts of the world are addressing some of these concerns.

The European Union's (EU's) comprehensive rulebook described as a rights-based, centralized, and legally binding framework

The EU AI act is a substantially comprehensive, legally binding piece of legislation. It focuses on regulating AI development and applications based on four levels of risk in descending order of "harm level." The four harm levels are:

1. unacceptable
2. high-risk
3. limited, and
4. minimal.

The unacceptable level represents prohibited AI practices or AI systems with unacceptable risk such as systems that target the elderly or children as illustrated by the following two "quasi-clauses" in article 5 (*https://artificialintelligenceact.eu/article/5/).*

Quasi-clause*: AI systems that pose a clear and present risk to the public, such as those designed to manipulate individuals or groups, or to exploit vulnerable groups are prohibited.*

Quasi-clause*: The following artificial intelligence practices shall be prohibited... the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm;*

The above shows the EU's willingness to draw bright red lines based on fundamental rights.

At the core of the risk-based approach is the Classification of High-Risk AI systems (https://artificialintelligenceact.eu/article/6/):

Quasi-clause*: AI systems shall be classified as high-risk if they are intended to be used as a safety component of a product... or are themselves a product covered by the Union harmonization legislation listed in Annex I (https://artificialintelligenceact.eu/annex/1/), and the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment.*

The above ties AI regulation to existing product safety legislation (e.g., for medical devices, cars, toys). This demonstrates a pragmatic strategy of building on previous regulatory frameworks rather than starting entirely from scratch.

In general, the EU's regulation of AI systems is a tiered, risk-based approach. While minimal-risk systems, such as an AI-powered email spam filter, are often unregulated, systems with limited or high-risk face specific oversight.

The requirements for high-risk systems are stricter and more comprehensive versions of those for limited-risk systems. It is as if the set of measures for limited risk is a subset of the measures for high-risk. These measures, which are essential for quality and safety, include:

- **Transparency:** ensuring it is clear how the AI works.
- **Traceability and Explainability:** the ability to follow the AI's processes and understand its decisions.
- **Accuracy and Robustness:** guaranteeing reliable and consistent performance.
- **Security:** protecting the system from threats.
- **Accessibility:** making sure the system is usable for all.
- **Privacy:** safeguarding people's personal information.

Beyond these operational requirements, a crucial measure is **conformity assessment**, which ensures the system meets standardized desiderata agreed upon by independent bodies. We will now consider a couple of these obligations, in a bit more detail.

Transparency (traceability and explainability) obligation

Quasi-clause*: Developers, providers, operators and deployers of high-risk AI systems shall develop, operate and use high-risk AI systems in a manner that ensures transparency. Including: Appropriate information to users that the system is AI and its intended purpose and, in some cases, including information about the data used and necessitates that AI systems are developed and used with the capacity for appropriate traceability and explainability, making their decision-making understandable to users.* See article titled, *Transparency and Provision of Information to Deployers* (*https://artificialintelligenceact.eu/article/13/*)

The term explainability is not formally defined in the EU AI Act's list of definitions (Article 3). The implicit requirement is that AI systems should make their decision-making understandable to users. This ambiguity is a double-edged sword.

On the one hand, it offers organizations the flexibility to innovate on how they achieve compliance, recognizing the significant technical, practical, and legal challenges associated with this requirement. On the other hand, failure to properly address explainability - especially in regulated domains or for life-altering decisions like a medical diagnosis or university admission - could lead to serious legal repercussions and lawsuits.

Articles 52, 53, and 55 deal with general-purpose AI, which implicitly, includes Generative AI. They include a statement that the output of AI systems must be clearly marked as artificially generated.

Quasi-clause: *'Providers of foundation models and of AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video (deep fakes) shall ensure that the outputs are marked in a machine-readable format and detectable as artificially generated or manipulated.'*

The above is a direct, technologically specific response to a known threat, i.e. disinformation. It shows the regulation is attempting to keep pace with innovation by mandating a technical solution, such as watermarking.

Accuracy, robustness and cybersecurity obligation

The EU AI Act requires high-risk AI systems to be consistently accurate, robust, and cybersecure throughout their use, with declared performance metrics and resilience against errors, manipulation, and cybersecurity threats, including biased learning loops and specific AI vulnerabilities. See article titled *Accuracy, Robustness and Cybersecurity* (https://artificialintelligenceact.eu/article/15/).

Much is condensed in article 15 and we do not claim to be able to discuss everything. But focus on some observations or questions regarding the concept of robustness to help you understand this requirement. Again, if you look at the defined terms in the EU AI act (see Article 3: Definitions (https://artificialintelligenceact.eu/article/3/), you would not find a definition for a robust system or robust data. It might be the case that regulators are assuming some knowledge on behalf of the reader, or it is just that it is too complex to know what constitutes robustness. The act states how it can achieve robustness "The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans". The question is what other solutions "ensure" robustness in high-risk AI? Continuous monitoring and validation, adversarial training, stress testing, explainability, human-in-the-loop, formal verification, or regular audits? What else?

The above paragraph speaks to the point mentioned at the beginning of this document about the nonexistence of a one-size-fits all but again, not getting this requirement right has severe repercussions from erosion of public trust in AI systems to potential cyberwarfare (one of the conspiracy theories justifying the shutdown of the Spanish and Portuguese energy infrastructure in 2025). An adversary could exploit these weaknesses to disrupt or disable a nation's vital services, with potentially devastating results.

Protection of privacy obligation

The right to privacy encompasses several clauses, such as the report titled *Recital 69* (https://artificialintelligenceact.eu/recital/69/), which requires that AI systems must guarantee privacy and data protection throughout their lifecycle using data minimization and privacy protection by design and default. *Article 78: Confidentiality* (https://artificialintelligenceact.eu/article/78/) elaborates on the principles of confidentiality, protection and privacy for data in general, not just personal data. Also, the broader Article 10 *Data and Data Governance* (https://artificialintelligenceact.eu/article/10/) strongly establishes the privacy obligation via several principles, namely, data minimization and purpose limitation, data governance and management practices, data quality and bias detection, strict conditions for processing special categories of personal data, integration with GDPR, transparency and documentation, and emphasis on fundamental rights.

Consider the expression "privacy protection by design and default. How easy could you achieve this requirement? Maybe you could but you would likely need to create a new fundamental shift in your mindset. Instead of treating privacy as just a compliance checklist, it must become a core engineering principle alongside functionality and security. This means you would need a proactive approach to prioritize data minimization to collect only what is absolutely necessary and build in new privacy-enhancing technologies from the ground up.

Achieving this core principle remains a challenge. Not a single, universally agreed-upon methodology for implementing it exists, and it can be difficult to scale up to large, complex networked systems. For a great discussion on "privacy by design" and jurisdiction-specific definitions of personal data, check out Jaap-Henk Hoepman (2021).[7]

Data governance obligation

Consider this quasi-clause:

Developers and operators shall develop, operate and use High-risk AI systems in a manner that ensures that the data used to train and develop them is of high quality, relevant, and sufficiently representative. See Article 10: Data and Data Governance (https://artificialintelligenceact.eu/article/10/). Here again, AI researchers have been exploring the term "sufficiently representative" for decades. And yet, no single, bullet-proof testing approach to guarantee that a training dataset for an AI system can sufficiently represent "sufficiently representative." It is so complex and depends heavily on the AI system's specific application and context.

Data protection and fairness obligation

What about this quasi-clause?

Developers and operators shall develop, operate and use high-risk AI systems in a manner that ensures compliance with applicable Union data protection law, including Regulation (EU) 2016/679 and Regulation (EU) 2018/1725. See *Article 10: Data and Data Governance* (https://artificialintelligenceact.eu/article/10/)

Again, this clause is another example of how we build on previous regulations, and we do not start from scratch.

Human oversight obligation

Quasi-clause: *Developers and operators shall develop, operate and use High-risk AI systems in a manner that ensures appropriate human oversight throughout their lifecycle.* See *Article 14: Human Oversight* (https://artificialintelligenceact.eu/article/14/)

Again, is this an open-ended obligation?

We have highlighted specific requirements to show that while the EU AI Act mandates requirements that AI researchers have been discussing for decades, its high-risk systems framework raises questions about the practical enforceability of these standards. The ambiguity surrounding compliance, as we illustrate above, suggests the Act, though an excellent start, may not be adequately addressing the problems it was intended to solve. Furthermore, it is notable and surprising that the Act explicitly excludes military applications, leaving AI systems used in conflicts unregulated.

America's Principles-Based Patchwork blends non-binding federal principles (AI Bill of Rights) and a growing mosaic of state laws

In the US, certain rules or acts about banning fake reviews and testimonials exist.   Two examples:

- The Federal Trade Commission's article titled, Federal Trade Commission Announces Final Rule Banning Fake Reviews and Testimonials (https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials).

- The defiance act punishing the creation of non-consensual deep fake porn (S.3696 - DEFIANCE Act of 2024 (https://www.congress.gov/bill/118th-congress/senate-bill/3696/text).

The US AI bill of rights is a more principle-based framework which focuses on protecting individual rights in the context of AI. The measures suggested to protect individual rights include equity, safety, efficacy and plain language information about AI systems and products which can be seen as part of a transparency requirement. Similarly, to the EU AI act, explainability, accessibility, and several layers of evaluation and protection of privacy are essential. Furthermore, there is a need for fallbacks or alternative plans, the case with any robust technology.

The AI Bill of Rights is a set of principles, not binding law. Its "clauses" are aspirations.

Safe and effective systems

Quasi-clause*: You should be protected from unsafe or ineffective systems. Automated systems should be developed in consultation with diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system.*

Contrast this with the EU's legally binding Article 5. The US principle uses non-binding language like "should be" and focuses on process or consultation rather than a specific, enforceable outcome ("shall be prohibited"). This highlights the US's flexible, guidance-oriented approach.

Algorithmic discrimination protections

Quasi-clause*: You should not face discrimination by algorithms and systems should be used and designed in an equitable way. Designers, developers, and deployers of automated systems should take proactive and continuous* *measures to protect individuals and communities from algorithmic discrimination.*

Again, note the language: “should be used and designed”, “should take... measures.” It sets a national expectation but lacks the legal force of a specific requirement for pre-market conformity assessments or audits, which are in the EU Act.

China’s approach: targeted and authoritarian governance ensuring AI serves state security and socialist values

China has moved quickly to implement a series of targeted regulations that align with its national and political interests. This is a crucial counterpoint to the Western models. The key regulations included:

- Algorithmic Recommendation Regulations (2022) *China’s policy on Algorithms* (https://digit-research.org/insights/chinas-regulations-on-algorithms/. The focus is on controlling the spread of information, requiring transparency, and giving users the right to opt out of algorithmic recommendation services. While the regulation's text is general, its purpose and enforcement were a direct response to the market dominance and perceived social harms of the country's largest tech conglomerates like ByteDance, Tencent and Alibaba.

Article 8 mentions ethics and addiction. See article titled, *Provisions on the Management of Algorithmic Recommendations in Internet Information Services*
BY CHINA LAW TRANSLATE ON 2022/01/04
https://www.chinalawtranslate.com/en/algorithms/):

The providers of algorithmic recommendation services shall periodically check, assess, and verify algorithm mechanisms, models, data, and outcomes, and must not set up algorithmic models that violate laws and regulations, or go against ethics and morals, such as by inducing users to become addicted or spend too much (time or money).

This vague and sweeping clause gives regulators broad discretion to interpret what violates “social ethics”, allowing for flexible enforcement based on the state’s priorities.

- Deep Synthesis Regulations (2023) (https://www.loc.gov/item/global-legal-monitor/2023-04-25/china-provisions-on-deep-synthesis-technology-enter-into-effect/) Article titled *China: Provisions on Deep Synthesis Technology Enter into Effect.* This targets deep fakes and AI-generated content by setting out responsibilities "concerning data security and personal information protection, transparency, content management and labeling, technical security, etc.". It mandates clear labeling of synthetic media and requires user consent for their creation. Generative AI Measures (2023) (https://www.loc.gov/item/global-legal-monitor/2023-07-18/china-generative-ai-measures-finalized/ Article titled *China: Generative AI Measures Finalized*

Regulate services like those offered by Baidu or Alibaba. Emphasize that generated content must align with core socialist values, must not contain subversion of state power, and must respect intellectual property. They also require security assessments for public-facing tools.

Article 4 establishes the general requirements for content created by generative AI. The provision states that https://www.chinalawtranslate.com/en/overview-of-draft-measures-on-generative-ai/):

4(1) Content generated using generative AI shall embody the Core Socialist Values and must not incite subversion of national sovereignty or the overturn of the socialist system, incite separatism, undermine national unity, advocate terrorism or extremism, propagate ethnic hatred and ethnic discrimination, or have information that is violent, obscene, or fake, as well as content that might disrupt the economic or social order.

This is perhaps the most powerful clause to contrast with Western models. It explicitly mandates that AI output must align with a specific political ideology. This is not just about safety or rights; it is about ensuring AI serves as a tool for state control and stability.

China represents a model of state-controlled innovation. The goal is not just to mitigate risk but to actively direct AI to serve state security and social stability objectives. This is a fundamentally different vision than the EU's rights-based or the US's, aspirational, market-and-rights-based approach.

China's regulations are focused on control and alignment with state goals.

The UK's Experiment: A Principles-Based, Pro-Innovation Framework

The UK's experiment is about delegating authority to existing sectoral regulators with a light-touch, principles-based mandate.

The UK is attempting to position itself as an AI governance hub with a distinctly light-touch, sector-specific approach. In February 2024, the UK government published its response to the AI regulation white paper (https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper), outlining its "pro-innovation" approach. The core tenets of the response can be summarized as 1) there is no central AI regulator . Instead, it empowers existing regulators (e.g., for health, finance, competition) to create context-specific rules based on a set of cross-sectoral principles such as safety, security, transparency, and fairness, and 2) the response is characterized by its heavy reliance on voluntary measures and partnerships with the private sector, notably through the establishment of new organizations such as the AI Safety Institute, named now the AI Security Institute (https://www.aisi.gov.uk/).

The UK model is the polar opposite of the EU's centralized one. It bets on agility and expert regulators but risks creating a confusing patchwork within the UK itself and lacking the teeth to enforce its principles. It is a fascinating "wait-and-see" experiment.

<u>Context-Specific Guidance</u>

Quasi-clause: Regulators will require that AI is used safely within their respective domains (e.g., healthcare, finance).

The UK does not have a single text like the EU Act. Its core philosophy is decentralization. The requirement is for regulators to act, not for a central body to create one-size-fits-all rules. This highlights a completely different governance model.

For global businesses and citizens alike, navigating this cacophony of rules is becoming AI’s next great challenge. The pressing question is no longer if we should regulate AI, but how can we possibly reconcile these competing visions into a coherent global framework.

Some other national efforts

- **Canada:** The Artificial Intelligence and Data Act (AIDA) is part of a Bill (https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document). It shares similarities with the EU AI Act (e.g., a focus on high-impact systems and requirements for risk mitigation) but is generally seen as less prescriptive. It is still working its way through Parliament.

Definition of Harm

Definition*: "Harm" means (a) physical or psychological harm to an individual; (b) damage to an individual's property; or (c) economic loss to an individual.*

You might notice that this definition is notably broad, especially including “economic loss.” This could theoretically cover a wider range of harms than other frameworks and shows, again, how different jurisdictions are scoping the very problem regulation is meant to address.

- **Brazil:** Has a draft AI Bill (PL 2338/2023) (https://artificialintelligenceact.com/brazil-ai-act/) Brazil AI Act that is also inspired by the EU’s risk-based approach but places a stronger emphasis on fundamental rights and establishing a dedicated regulatory authority.

- **Japan:** Leaning towards a more flexible, business-friendly approach to avoid stifling innovation, while still developing guidelines around transparency and ethics. Japan passed recently a bill that passed through the House “Act on the Promotion of Research and Development and the Utilization of AI-related Technologies” https://www.japantimes.co.jp/news/2025/05/28/japan/japan-ai-law/).

Japan enacts bill to promote AI development and address its risks. It is also closely watching and engaging with the G7 process.

- **Singapore:** A leader in the "soft law" approach. Its "AI Verify" foundation provides a toolkit for companies to demonstrate responsible AI through voluntary self-testing, rather than imposing strict government mandates. Singapore "…expanded it [AI verify] for generative AI systems (Project Moonshot) and is working hard to ensure its AI governance frameworks are interoperable/aligned with the international community's". (https://practiceguides.chambers.com/practice-guides/artificial-intelligence-2025/singapore/trends-and-developments) Trends and Developments

In February 2025, Singapore announced new AI safety initiatives, including the Global AI Assurance Pilot (https://digitalpolicyalert.org/event/27039-singapore-announced-global-ai-assurance-pilot) and the Joint Testing Report with Japan, to assess LLMs across different languages (https://digitalpolicyalert.org/event/27042-singapore-and-japan-announced-a-joint-testing-report-for-large-language-model-safety). Singapore keeps on publishing specific AI policy documents such as the Agentic AI Primer in 2025 (https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/agentic-ai-primer.html) to promote transparency, fairness, and accountability in AI development and deployment. Agentic AI Primer

International and Multilateral Initiatives

These are crucial for addressing the global nature of AI and preventing fragmentation.

- G7 Hiroshima AI Process: (https://www.soumu.go.jp/hiroshimaaiprocess/en/index.html) resulted in the creation of a voluntary International Code of Conduct for Advanced AI Systems. This is a significant effort by the world's largest advanced economies to create a common baseline on issues like security, safety, and trust.

- High-Level Advisory Body on AI: The UN established a High-Level Advisory Body on AI (https://www.un.org/digital-emergingtechnologies/ai-advisory-body) to explore global governance options. This is a move toward a more inclusive, global conversation, though concrete binding outcomes are a long way off. For us, the key is their understanding of the value that interdisciplinary teams bring.

- Global Partnership on AI: OECD AI Principles (https://www.oecd.org/en/topics/sub-issues/ai-principles.htm One of the first and most influential sets of principles, agreed upon by over 50 countries. They are high-level and voluntary but have set the normative foundation for many national policies. Also, by the OECD, a Global Partnership on AI (GPAI) (https://www.oecd.org/en/about/programmes/global-partnership-on-artificial-intelligence.html) has been established, which is a multi-stake holder initiative which brings together experts from science, industry, civil society, and governments to collaborate on responsible AI development.

- Global treaty on AI by the Council of Europe: A global AI treaty has been also opened for signature. As the Council of Europe announced, this is the “first-ever international legally binding treaty aimed at ensuring that the use of AI systems is fully consistent with human rights, democracy, and the rule of law.” (https://www.coe.int/en/web/portal/-/council-of-europe-opens-first-ever-global-treaty-on-ai-for-signature).

These regulations are of utmost urgency not only to ensure that humans are in the driver’s seat but also as some concerns might threaten the idea of democracy itself. However, this fragmentation presents a new complex geopolitical and practical realities risk. The main questions are: 1) do we have a “spaghetti bowl” of conflicting regulations that could paralyze global developers and create a race to the bottom? and 2) how do we achieve a better global path? Therefore, finding interoperability between these regulatory models will be a central challenge.

We support Agile methodologies and believe these regulations, by design, are agile themselves. This will allow them to evolve with the accelerating AI landscape and the success of future sandboxes. When deploying an AI system, consider how you can proactively anticipate these changes to stay ahead of the curve. It is important to stress that the goals of these frameworks do not stifle innovation. The intention is to harness the power of AI while mitigating risks or minimizing AI potential harm.

Furthermore, it is worth saying that while AI legal frameworks, principles, and guidelines are a crucial step forward, there is a pressing need for clearer liability frameworks. The question of ‘who is liable?’, I think, remains unresolved. Is it the developer, researcher, provider, or the company that adopted the system? The EU AI Act and US SB 1047 offer some guidance, but the issue is complex and likely case-by-case. As researchers and developers, we are concerned about the uncertainty surrounding legal compliance. It is unclear when our work might cross a legal boundary. The prospect of needing legal counsel before every project is daunting. At the end of the day, regulations are nothing but contracts that one has no option but to sign. One needs to understand them carefully. Interesting enough already several researchers in the UK and US are using AI to help people navigate all these regulations about AI (Kelts and Sharma, 2024) and (Marino et al., 2024).[8]

Finally, for any new AI application, a critical question is whether you can achieve compliance with the EU AI Act or any other act, before set deadlines. This requires a pragmatic assessment. If you are in the early stages of a project, you should weigh the benefits and drawbacks of compliance by design.

Conclusion: call for action to forge a successful path

In conclusion, the future of AI is not yet written. We must offer concrete, actionable recommendations to ensure it is one we want. Actionable steps include:

- **Legislators** should pass flexible laws, based on core tenets, that empower expert agencies, like the US Food and Drug Administration for Health, or the Bureau of the Federal Trade Commission for consumer protection, to update standards as technology evolves.

- **The Industry** must move beyond ethical principles and invest in concrete implementation through safety research, rigorous internal auditing, and transparent engagement with regulators.
- **The Public and Academia** should demand accountability, participate in the process and support research on AI safety and ethics to inform evidence-based policy.

Achieving the masterpiece of AI-driven prosperity requires us to be deliberate architects, not passive passengers. The conversation is not “regulation versus ethical innovation”, it is “what smart regulation will foster the innovation we want and prevent the future we fear?” Let us choose a path defined by science, wisdom, foresight, and success.

Finally, it is essential for international bodies to regulate AI used in conflicts to ensure that human judgment and accountability remain at the core of military decision-making, safeguarding against a future where technology, not humanity, dictates the course of wars.

Think of this as an introduction to a complex and ever-changing subject, not the final word. We have not examined every single detail of the acts and frameworks discussed, so we urge you to use this as a guide to begin.

This piece is intended for a general audience and is not a legal opinion. Legal professionals are encouraged to provide their constructive feedback.

END NOTES

1. A **general-purpose technology**, a term coined by economists, is a major innovation that fundamentally transforms economies and societies. It is not a single invention, but a platform for a cascade of future innovations that are applicable across a variety of sectors and improve over time, such as electricity.
2. An **AI algorithm** is a set of step-by-step instructions designed to enable a machine to perform a task that typically requires human intelligence such as learning, problem solving, pattern-recognition and decision-making.

3. A **regulatory sandbox** is a closed testing environment where regulators and companies can safely experiment with new technologies and business models without the immediate burden of all existing regulations.
4. In their simplest form, we envision a list of **regulatory labels** as a list of pairs of <regulatory_requirement, compliance_status>.
5. **Independent bias audit** is a third-party evaluation of a technology to identify and address biased or discriminatory outcomes affecting individuals.
6. The EU Artificial Intelligence Act (EU AI Act)
7. Jaap-Henk Hoepman. “Privacy is hard: and seven other myths”, 2021. MIT Press.
8. Steven A. Kelts and Chinmayi Sharma. “Leveling Up to Responsible AI Through Simulations”, March, 2025. See also https://www.techpolicy.press/leveling-up-to-responsible-ai-through-simulations/ Bill Marino, Yaqub Chaudhary, Yulu Pi, Rui-Jie Yew, Preslav Aleksandrov, Carwyn Rahman, William F. Shen, Isaac Robinson, and Nicholas D. Lane. “Compliance Cards: Computational Artifacts for Automated AI Regulation Compliance”, September 2024. See also https://arxiv.org/abs/2406.14758

ABOUT THE AUTHOR

AI-driven innovator with a proven track record in transforming research into real-world solutions. From academia to entrepreneurship, Jana has driven groundbreaking projects in cybersecurity, education, and law. As Founder & CTO of Noha-Z, she is building a platform that transforms knowledge management through "computable" documents, focusing currently on contracts. Backed by a PhD in AI from Cambridge University and experience at various tech companies, Jana is passionate about leveraging language, mathematics and artificial intelligence to create positive change.

ABOUT Noha Z, LLC(https://www.nohaz.co/), We harness a hybrid of human and artificial intelligence to drive innovation. We partner with companies to help them ask the right questions, build custom software, and automate complex processes. In parallel, we are developing our own product, a content discovery ecosystem focused on mining and creating "smarter" documents, for better knowledge management.