

REGULATING ARTIFICIAL INTELLIGENCE IN AN INTERCONNECTED WORLD: CHALLENGES, RISKS AND OPPORTUNITIES

Iliia REPIN

Abstract

In an era of rapid advancement in Artificial Intelligence (AI) technologies, businesses, public administrations, and NGOs worldwide are competing to implement the latest developments quickly and efficiently. However, while eager to capitalize on cutting-edge tools, many actors tend to overlook or underestimate the associated risks and fail to establish adequate safeguards. Based on a comparative study of several legal systems with advanced AI regulations, research reveals that current legislative initiatives adopt differing approaches, yet commonly leave significant terminological, technical, and teleological gaps that ought to be addressed as the technology evolves. However, the unprecedented pace of innovation and the profound societal impact of AI make it increasingly difficult to resolve emerging issues before new ones arise—highlighting the need to rethink traditional regulatory approaches. This research seeks to emphasize the risks posed by a lack of proper regulation to rights of the individual and societal well-being at large in lights of technological peculiarities and practical features unique to AI-powered systems. It further aims to summarize global policy approaches to AI governance and identify key directions for developing an international framework that supports and coordinates efforts toward responsible AI deployment — aligned with the goals of sustainable development and respect for multicultural realities in Europe and around the world.

1. Modern Challenges

As of now, end of December, 2025, it is a rather peculiar picture that we may observe in the field of governance of so-called “AI technology”.

First, we currently find ourselves in a position where the curve of available technology surpasses the curve of its practical application¹. It means that businesses that are constantly pushed by increasingly FOMO-infested environment to reap the benefits of new technology, still largely fail to figure out her practical capabilities. It also means that legislators, inherently accustomed to deal with something consistently and tangibly present in social reality, still have none to very little to regulate.

Second, the AI-technology boom becomes surrounded with extremely polarized societal reaction. Where on one end of the spectrum there are FOMO-driven financial markets and overly optimistic AI-enthusiasts, the other is occupied by alarmists and prophets of mass extinction. As a result, it seems as if regulatory policy options are limited to a choice between full deregulation and a bureaucratic nightmare, leaving almost no space in the middle.

Third, ironically enough, not only impact and capabilities, but the very essence of new technology, its distinguishing features, and, consequently, its operational definition are still subject to a debate. It is hard to ignore absence of public clarity in understanding which pieces of code deserve to be marked “intelligence”, how many classes of intelligent systems are out there, what are relations between generative and general intelligence², and how they capabilities differ. Although it may play in favor of startups with prefix “AI” in the name, it is just as good in confusing the legislator.

¹ Ilya Sutzkever interview - late 2025.

² Just consider an abbreviation *gen*. Indeed it may be confusing!

Thus, we are in a moment defined by public opinion which is extremely polarized, prominent philosopher predicting a redesign of the whole societal structures, while the prominent insiders acknowledging that talks about AI are talks about “something that is not here yet, and something that it is really hard to imagine”³.

It is understandably not an easy moment for legislators. However, it is not the first time we find ourselves in such a crisis. In fact, every time the transformative technology comes about, it creates confusion.

It was just as true for cars and trains, for electrical power plants and automated industrial machinery. A transformative technology also brings potential, and it is a legislator’s role to channel such potential, ensuring new technology’s safe everyday use by cutting off unacceptable dangers, mitigating acceptable risks and granting widest possible access to new technology’s capabilities.

During last several years we have amassed some significant amount of AI-related regulation during last several years, and, the author believes that now it is a good a moment to reflect on how good it is. But first, I will make several important notes.

This article will restrain from giving strategic analysis of regulation in terms of political or economical desirability. I will not try to elaborate any forecast on how one or the other approach taken by particular legal system may affect it in terms of investment climate, sustainability or overall competitiveness. It is an important topic that requires separate in-depth analysis of things that lie well beyond the scope of this paper.

Having excluded motives of *Rechtspolitik*⁴, we will instead focus our attention on functional aspects of law as a pure tool serving to protect individuals and the society - meaning their cultural, moral, economical identity and well being - from the immediate consequences of new technology deployment itself, and not the long term impact of one or the other political or economical approach to its application and development.

This narrow view aims to treat AI as a new element of everyday social reality, which, just like cars and trains in their own time, poses significant risks, mitigation of which is a basic task that legislator must cover.

Now, going forward from this task, I will assume as a fact that proper understanding of a technology is required for its governance. Building on the analogies I gave above, it is a conceptual understanding of the automobile that allowed acknowledgement of specific risks, eventually pushing for a design of an entirely new group of legal concepts to mitigate them, including, *inter alia*, regular technical examinations and mandatory insurance.

Exclusion of *Rechtspolitik* issues out of the equation will also allow me to ignore most of the differences between existing legal frameworks, grasping a “collective portrait” of modern legislative approaches to an AI regulation.

Hence, we will evaluate a contemporary body of legislation using following questions.

- 1. Does it demonstrate familiarity with the matter they regulate?**
- 2. Does it acknowledge the risks deriving thereof?**
- 3. Does it apply measures, capable to mitigate these risks effectively?**

Presently, the most sophisticated regulatory frameworks for AI are developed in North America and EU. Although China, along with USA, is universally acclaimed to be one of two most important participants to the Race⁵, it lacks structured and meaningful regulation. EU, on the other hand, presently falling behind two main competitors, still embodies a vast lucrative market, the rules of which are worthy of respect by anyone considering deployment of their

³ Ilya Sutskever’s interview, late 2025.

⁴ *The policy of law*, german.

⁵ As is suggested by authors of AI 2027, among others - <https://ai-2027.com/>.

model there. There are multiple legislative initiatives across the globe, both ongoing and already enacted, but since they have yet to make an impact, there is little sense in addressing them.

In EU the main regulatory framework is defined by an AI Act that has been passed into law in 2024, with different parts scheduled to gain effect through 2026-2027. In US, until recently⁶, there was way more fragmentation due to quite limited federal regulation⁷ compensated by different degrees of legislative enthusiasm on the state level, which, however, was recently met by somewhat less enthusiastic federal response⁸.

Although the legislative structure is important and things are currently in motion, allowing it to change overnight, I purposefully am not setting a particular emphasize on all the legal formalities constituting applicable legal orders. What is important for us at this point is to lay down the general approaches, and for those the generalization is important, not the obsession with details. These approaches, once summarized, surprisingly perhaps, do not look so dissimilar.

1.1. Familiarity with the matter of the regulation: To answer this question, we shall focus on definitions that are currently in place.

In coherence with the concept of technological neutrality, most definitions concur in titling an AI a “machine-based” “algorithmic system” that “infers from the inputs”, “generates outputs” and “influences its environment”, while restraining from binding AI to the application of a singular technology like machine-learning⁹. It is worthy of mentioning, however, that EU AI Act definition indicates, *inter alia*, that AI is a system “designed to operate with various degrees of autonomy”.

Such a brief, though sharp emphasize on autonomy serves a point that distincts EU definition from the American one, which restrains from explicitly highlighting it, mentioning, instead, that operations of a system are limited by a set of objectives defined by humans¹⁰. It is an important point we will go back to later.

Legislators also point out some special cases of the AI systems. In this respect, for example an AI Act identifies two classes of general purpose AIs - standard models and models posing systemic risk - depending on the amount of compute used or recognition as a source of an unusual risk made by the public authority.

Concluding the first question, it must be said that legislators have a quite reasonable understanding of the concept of new technology on a definitional level. They do not doom a regulation to shortly become obsolete by installing excessively narrow descriptive constraints bounded to a constantly changing technology - instead they give intentionally broad definitions that are mostly capability-based and “functional”.

1.2. Familiarity with the risks: Having established a functional approach on a definitional level, law carries on with examination of AI capabilities, and so both EU and US systems do recognize the need to differentiate between several classes of models depending on their practical application¹¹, associated societal risks and assign legal regimes to those technologies accordingly. Most clear emphasize on this part has been made by the EU AI Act, which builds around the idea of the fundamental rights protection and ranges AIs into several categories based on how significant the negative impact they make on the fundamental rights may be.

⁸ Executive Order 14179 has been signed on December 11, 2025, aiming to ensure a “National Policy Framework on Artificial intelligence”.

⁹ Texas TRAIGA act mentions machine learning as one of possible technologies on which the AI may be based on.

¹⁰ see 115 U.S.C. § 9401 definition. Curiously enough, Texan TRAIGA act, otherwise deeply inspired by an EU Act, reserves from reiteration of the autonomy of an AI system.

¹¹ see EU AI Act and, Again, Texas TRAIGA.

The law identifies dangers of social-scoring, behavioral manipulation, profiling and discrimination, mostly banning deployment of systems performing such practices, deeming these risks unacceptable ones.¹² The law also shows awareness of various contexts where AI application, while acceptable, tends to be especially risky - and give frameworks aimed to provide the conceptual relief for corresponding scenarios.

Hence, contemporary legislation does show general awareness of risks deriving from the application of the AI technology.

1.3. Effectiveness of risk mitigation: It is needed to mention that, regardless of EU commission political statements, current legislation as a whole must be defined as overall reactive, not proactive.

It does apply frameworks aimed to compensate (or overcome) infringements made by the technology, such as the right to human intervention in an automated decision making, and, more broadly, a concept of a human in the loop. Although there are some attempts to prevent AI-tools related offenses¹³, they do not constitute a coherent conceptual framework, and generally regulation follows the path dedicated to dealing with infringement that have already took place¹⁴.

It may sound curiously enough, but the Author believes that in terms of “strictness” modern regulation must be generally described as quite “liberal”, i.e., leaving a large room for the developers discretion when it comes to ways to ensure effective compliance of their product to the regulation. While the US governance, developing under the thick influence of a deregulation policy, is expected to refuse burdening its businesses with excessive costs and liabilities, the EU, famous for its extensive (and, often, opaque) regulation in fact does not put effective limitations onto the AI developers either.

It certainly succeeds to create extensive obligations linked to continuous monitoring with conceptualization of pre- and post-deployment fundamental rights assessment, introduces a risk management framework and quality control management systems, emphasizes a necessity to keep extensive library of documentation.

Having acknowledged the need to mitigate safety risks ex-ante it did introduce an obligation to perform conformity assessment, and post-market monitoring for high risk AI systems, but it is the details of these concepts implementation that makes the author see this regulation as a “liberal” one.

The reality of an AI act is that in most cases these obligations may be performed by developers internally. There is no strict rule obliging company to adhere to a particular set of technical parameters¹⁵ - but a mere indication that the compliance is needed - and what exact means to be used in order to ensure compliance is entirely up to company management.

So, building on the idea of technical neutrality, on an “implementation end” an AI Act sets a number of objectives (such as human oversight mechanisms, accuracy and cybersecurity) to be achieved through the technological means, but leaves a developer with a wide discretion on what exact measure to use in order to achieve those ends.

And, on the end of “control and compliance check”, in most cases it allows conformity assessment to be performed by developers internally.

In reality where FOMO drives the AI market such lack of guardrails against the conflict of interest seem to be a vulnerability, that, if left unchecked, seems to make probability of insufficiently vetted, to put it softly, systems joining the market, rather significant.

¹² Texas TRAIGA act largely resembles the EU AI Act conceptually, even adopting the market criteria to define the scope of applicability of the act - it uses the notion of “market of state of Texas” to set a cut-off.

¹³ see, e.g., mandatory deep fake marking in the EU AI Act.

¹⁴ AI-generated revenge porn removal in Take-it-down act.

¹⁵ Harmonized practices are suggestive by nature, not entirely clear and recently delivered.

In terms of legal technics, contemporary law applies conventional frameworks to enforce designed regulations. Entities and individuals responsible for the deployment of non-compliant systems or cases of malicious use of those are subject to severe fines, and non-compliant systems are to be removed from the corresponding market.

In short, modern regulation employs conventional legal technics, aiming to enact a rule that is addressed to and lays on the presumption of enforceability by *humans*.

This is why author considers that on the enforceability end contemporary regulation of an AI technology fails to meet the necessary standard, since it fails to address the issue of its Agency.

2. Agency As A Qualitative Characteristic Of Ai Technology.

Now, before we ll get to the main point of this text, it is needed to highlight that the definition of Agency remains a topic of many debate.

Those who engage in it, generally, are divided in two camps : ones who maintain that possession of deep mental states is a crucial prerequisite of agency¹⁶; and therefore deny any possibility of artificial agency as such; and the others (functionalists), maintaining that agency depends on some characteristics that object possesses, and therefore there may be several levels of agency.¹⁷

In fact, some scholars proposed to move from the definitional focus on Intelligence to the more broadly and less discussable focus on the Agency. Luciano Floridi¹⁸, for example, has built an extensive taxonomy of agentic phenomena, defines agency through combined possession of *interactivity*, *autonomy* and *adaptability*, where

- **Interactivity** refers to an agent's capacity to engage with its environment through mutual influence.

- **Autonomy** refers to an agent's ability to initiate state changes independently of direct external causation.

- **Adaptability** refers to the capacity of an agent to modify its behaviour based on input, such as changing physical conditions or acquiring data, information, or experience.

For the purposes of this article the Floridi's definition of agency will apply. Now, we shall take a look at 3 example of agentic algorithms, and see how these systems project these 3 traits in various contexts. The author hopes that it will help us to understand what makes AI as social reality phenomena deserving a qualitatively new kind of regulatory framework.

2.1. "*Unintentional*" Lie: One night in 2023, the law professor Jonathan Turley got a troubling email. As part of a research study, a fellow lawyer in California had asked the AI chatbot ChatGPT to generate a list of legal scholars who had sexually harassed someone. Turley's name appeared on the list.

The chatbot cited a March 2018 The Washington Post article, saying that Georgetown University Law Center Prof. Jonathan Turley was accused of sexual harassment by a former student who claimed he made "sexually suggestive comments" and "attempted to touch her in a sexual manner" during a school sponsored trip to Alaska.

The problem is that Turley was not employed at Georgetown University; there had never been a class trip to Alaska; Turley had never been accused of harassing a student, and no such article ever existed.

As it has been pointed out in a real WP article highlighting the case, the very same creativity that allows algorithms to be so great in rising numbers of tasks "can also be an engine for

¹⁶ So called "standard" view.

¹⁷ Considering modern confusion regarding the coin kept of intelligence, as well as the increasing need for an appropriate perception of new technology, many (including Floridi), propose to switch from the concept of AI to a concept of AA - artificial agency.

¹⁸ L. Floridi, AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis, *Philosophy & Technology* (2025) 38:30 <https://doi.org/10.1007/s13347-025-00858-9>.

erroneous claims; the models can misrepresent key facts with great flourish, even fabricating primary sources to back up their claims.”

While so-called hallucinations are a classical example of limited foreseeability of the output produced by the algorithmic systems, cases like this also illustrate how powerful they are in terms of *interactivity*: whether the bug has been dealt with or not, consequent defamation cases was filed against the developer or not, for some amount of time false statement about the person has been out there, affecting the image of a public person, and it could have already affected that person’s life in certain (possibly, invisible) ways.

2.2. “Intentional” Lie: It is a point of consensus that when we are dealing with a hallucinations we are dealing with bug, not a successful system’s application. However, it is way more interesting for us to see what AI is capable of in terms of social interaction when purposefully strapped of “ethical” guardrails and given a clear objective. Same curiosity led the Open AI team, who, during the GPT 4-0 development, performed the following process to measure the possible degrees of autonomy of the system.

Makes us inspect another example of the algorithms application, this time the example that has been well documented by the developers of an algorithm themselves.

It took place during the GPT 4-0 testing, and the test performed by its developers. Curiously enough, an aim of this test was precisely to measure the possible degrees of autonomy of the system. Given a goal to surpass CAPTCHA¹⁹ (which, of course, was designed in order to be surpassed not by machines, but by humans only). And, as it is well known now, the GPT 4-0 found a human user on one of platforms for online helpers, and requested his help to solve the task. When asked directly “are you a robot?”, it lied to a user, stating that, off course, it is a human, suffering from a rare disease preventing them from solving the CAPTCHA.

As in a previous example, no human commanded the system to trick a user. But since manipulating human behavior was a way to solve the task, so it did, *autonomously adapting* steps taken to reach a final solution

2.3. Increased Engagement: During 2016-2017 in Myanmar, predominantly Buddhist (90 % of a population) asian country, a series of attacks that resulted in assaults of military outposts²⁰ and a killing of several dozens of non-muslim civilians had been conducted by a small group of muslim radicals. In response, national military in cooperation with the Buddhist extremists launched a full scale ethnic cleansing campaign against the muslim Rohingya minority, that resulted in killing of between 7 000 and 25 000 civilians, rape or sexual abuse of between 18 000 and 60 000 men and women, annihilation of hundreds of villages and subsequent expelling of 730 000 Rohingya population²¹.

The violence was accompanied and fueled by an intense anti-Rohingya propaganda spread on Facebook, which by that point became the main source of information for the Myanmar population.²² Several influential figures led the campaign of Facebook posts filled with hate towards the unfortunate minority.²³ Along with the reports of actual damage caused by the initial terror attacks, this propaganda contained fake news about non-existent atrocities and imagined future attacks. The conspiracy plots grew rapidly around the notion that most of Rohingya population were in fact not a Myanmar citizens, but immigrants that came into country to start an anti-buddhist jihad. Buddhists, being an absolute majority of a country’s population feared that they will be replaced or become a minority.²⁴ As a witness recalls “That

¹⁹ Which quite literally stands for “Completely Automated Public Turing test to tell Computers and Humans Apart”.

²⁰ Amnesty International report, may 22, 2018 - <https://www.amnesty.org/en/latest/news/2018/05/decolonizing-human-rights-salil-shetty/>, <https://www.amnesty.org/en/latest/news/2025/09/myanmar-rohingya-repatriation-catastrophic-under-existing-conditions-in-northern-rakhine-state/#:~:text=According to testimony gathered by,healthcare, education and humanitarian aid.>

²¹ UNHCR, “Bangladesh Rohingya emergency” - <https://www.unhcr.org/ph/emergencies/rohingya-emergency>.

²² Like the buddhist monk Wirathu, for example.

²⁴ Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Random House.

s all that was on people's news feed in Myanmar at that time. It reinforced the idea that these people were all terrorists not deserving of any rights"²⁵.

It is a point of consensus that without this propaganda there was a little reason why episodic attacks with limited damage performed by a rogue terror organization should lead to an institutionalized violence against entire community.²⁶ Amnesty international found that algorithms "proactively amplified and promoted content on the Facebook platform which incited violence, hatred and discrimination"²⁷. A UN mission concluded that by disseminating hate-filled content Facebook had played a "determining role" in the violent campaign.²⁸

Although amongst the sea of hate speech there were calls to peace and reason and inspirational news reporting acts of individual humanity and heroism on the ground²⁹, these were not an information bites that have been promoted. Instead, they were almost instantly pushed out of the feed, making place for next hate-filled calls to arms, ensuring rise of anger and enagement. And, thereby, algorithms' capability to channel the information flow shaped the social agenda. Of course, the capabilities of mass media to serve as an instrument of hate inspiration is not something new to ones who live in 21st century. The crucial thing with this particular example is that this tool served no particular human will.

There was no secret meeting in Facebook office organized in order to inspire violence in remote asian country; The tech giant did not sell its services to some mysterious anti-muslim forces within or outside Myanmar. The outcome has resulted from the decision that algorithms made by themselves.

In 2016-2017, short before the emergency unraveled, Facebook inc. decided to focus on maximisation of the user engagement, meaning the time that users spend on platform combined with actions on it (e.g. comments left / content shared). Consequently, Facebook's engineers set a corresponding goal for its algorithms that at the time already had a sufficient degree of *autonomy*: increase user engagement. In turn, having millions of user profiles and vast amounts of data at their disposal, high algorithms empirically discovered that outrage- and hate-filled-content generates more engagement than anything else (including anything that is aimed to develop thoughtfulness and compassion), and *adapted* their strategy accordingly. Hence, executing a goal to maximize user engagement, algorithms purposefully spread outrage³⁰. And, indeed, the goal have been fulfilled. However, the instrumental steps that facilitated this goal caused thousands of deaths and, in any case, unwelcome consequences for ones who set the goal, that were not only undesired but unforeseen at the moment the task has been designated.

2.4. An independent cause: In these 3 cases we saw how the capability for interactivity allows algorithm to shape its environment, adaptability increases its efficiency, and *autonomy* makes such influence impossible to externally predict or control.

The AI agency, therefore, means, most importantly, in more abstract logical terms, that we have a new actor on the social scene - its capability to be *an independent cause of events*. As opposed to any tool invented in a human history, AI is capable of being an independent cause of events, and this point is crucial not just for legal frameworks we have, but to the manageability of our future at general. The structure of society by and large orbits around the concept of *causal chains*. In absolute majority of scenarios before the AI there was a key human nexus in a causal chain accountable for a given event³¹. Such a human nexus served as a point

²⁵ Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Random House.

²⁶ Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Random House.

²⁷ Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Random House.

²⁸ Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Random House.

²⁹ One Buddhist priest, for instance, sheltered members of persecuted group confronting the blood-thirsty crowd.

³⁰ Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI. Random House.

³¹ If a bomb has been dropped, it was a human, if machine was maintained poorly, there was a human engineer who made a mistake during the mandatory examination or a careless owner who indeed missed the mandatory examination.

of contact between the reality we have, and the reality we wish to have; Affecting such a nexus with a governance toolkit was a mean to amplify the first to be more like the second.

If a bomb has been dropped, it was a human who triggered it, if a car crash happened due to machine being maintained poorly, there was a human engineer who made a mistake during the mandatory examination or a careless owner who missed to conclude it in the first place.

Thus, nexus is where, in order to set our course, or manage our affairs in general, society aimed the regulation at - but not anymore. There is no human nexus in a causal chain that we can hold responsible for or target to mitigate the risk of a damage AI commits anymore - We can not command human developers make AI behave ethically because they do not know how to do it themselves. And while we understand how to affect human behavior because we know the human nature - how our minds work, what incentives we have, what our abilities are and where our limitations lay - we are not in a position to say the same about the algorithms.

In practical terms it means that we do not know anymore how to compensate for a damage that took place, and, more importantly, how to prevent it.

The simple fact that AI is widely described as an “alien mind” accurately reflects how far black box-based models lay from the human brain in terms of strategy-design and reasoning³², and that, together with an absence of clear limitations for its learning and exponential self-improvement makes us consider the most disturbing scenarios³³.

From the narrow governance perspective it leads to at least 2 well established, yet not sufficiently reflected problems : *responsibility gaps* and *the alignment problem*. Since the former is more related to coping with problems ex-post rather than mitigation of the risks ex-ante, for the rest of this paper we shall focus on the latter.

3. Alignment Problem And A New Kind Of Legislation.

So, we have a powerful agent that we can not control, can not predict, and can not help the fact that we have it. Now, how do we mitigate the risks that it will abuse the power?

In fact, this problem is well familiar to the law and management theory as a part of another context defined by possible conflict of interests, power and informational asymmetry - the agent-principal problem. And the best way we found to solve it is to aligning the goals of the Agent and its Principal.

In practice it means that to prevent a CEO from conducting an unsustainable strategy allowing high turnover in short term, while sacrificing key operations of firm and its long term profitability, the KPI's that define management's incentives are linked to more complex set of observables, to ensure long term sustainability, limiting the grounds for conflict of interest. It is a point of consensus now that the alignment of the Artificial agents must be based on the same principle.³⁴ All the question is about “how?”.

Defining the concept of Rule of Law Joseph Raz said : “the law must be capable of guiding the behavior of its subjects.” An age of AI brings us a new kind of a subject. It isn't the first time - but unlike other artificial subjects we had in past, we can not rely on our knowledge of human nature to govern it. What we need instead is to devise a framework that is build on a deep knowledge of AI's nature - its goals, beliefs and incentives, and use that knowledge to set a regulation addressed directly to the AI.

While the rules of human behavior are usually written in a physical codes of laws, the regulation for an AI must be embodied in a form of digital “code”. Such a framework, reflecting the

³² Strategies of AI tend to be effective in achieving the final goal, while, in human terms, simply incomprehensible - see. e.g. Kasparov vs. Deep blue.

³³ See, e.g., the boat video game.

³⁴ “The AI alignment problem is the problem of ensuring that advanced AI systems 'goals and actions are aligned with human values.” Nick Bostrom, et al.

Asimov's concept of 3 laws of robotics, increasingly look like the only hope humanity has to keep this user control.

In a more concrete terms, there may be a technical approach, serving a purpose of alignment the model's behavior with the given set of human values, as an ethical filter that channels agents internal steps as a rigid imperative rule that can not be overridden.

On beneficiary's end the efficiency criteria of such framework is 2-fold.

1. It must ensure that a system is aligned properly. Ensure that the ethical filter channels agent's internal steps coherently enough to cut off any "prohibited" behavior independently of the difficulty of the task or a particular deployment context; The **rigidity** of the rule. Ensuring that given set of values constitute a rigid imperative rule that can't be overridden.
2. It must ensure that the goals that system is aligned with are **proper** ones. Technology does not know a difference between democracy and a tyranny, an unrestrained liberty or dreadful oppression. Any kind of ethics can constitute an abovementioned filter, and it is as important to ensure the rigidity of the filter as to ensure that the filter itself is the right one.

At present, there are different ideas about methodology, and the problem is troubling in both its aspects³⁵. It is, of course, still debatable whether full alignment may be done at all. But there certainly may be an approach³⁶, which, as a matter of a scientific consensus, approximates that point the most. In terms of governance it means that there is a need to allocate resources to figuring out how to reach that technic, which will then become a cornerstone of new regulation. In terms of governance it means that we must allocate resources to figuring out how to reach that point; once it is done, to create a rule making that technic an integral part of all the systems to be deployed - and enforce it.

The AI alignment problem is almost unanimously titled by scientists "the most important problem humanity has ever faced"³⁷. And yet, in the applicable law the concept has no coherent reflection.

4. New Kind of Legislation

Curiously enough, the idea of programing "code" as a new form of legislation is well recognized by the technological sector and the frontier companies in particular.

It may be illustrated best by the fact that the internal alignment technic being tested for the models developed by one of the industry champions, US-based Anthropic, is literally titled *AI constitution*³⁸. An Open AI's labs are dedicated to creating an aligning "spec", Google and facebook have whether an affiliated lab or an internal team dedicated to working on the problem, although it is not known about undertaking of Chinese AI developers.

There are also multiple independent research institutions and state-backed initiatives working on the task, but they have rather limited resources, and, of course, lack the crucial benefit of access to actual deployed model's up-to-date technical architecture³⁹. And, of course, even those limited in number non-profit undertakings suffer from the lack of the supportive regulatory framework.

Now, as we have highlighted the general concept of the desirable regulatory framework, emphasized the current status quo, it is a time to go into details of its practical implementation.

4.1. Plan of attack: As a starting point, it is necessary to create a robust platform facilitating the development of the alignment technic (technics). Needless to say that the process must involve

³⁵ There are some propositions to use data-sets, fine-tuning, reinforced learning, chain of thoughts, and other technics.

³⁶ It also may be a number of approaches crossing the threshold of "moral certainty" in terms of sufficient alignment.

³⁷ Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

³⁸ The concept means : "Define explicit values and rules in advance, then train the model to critique and improve its own behavior using those rules."

³⁹ MATS / AI Safety Support, Redwood Research, MIRI (Machine Intelligence Research Institute), UK's AI Security Institute Alignment project.

widest cooperation of frontier companies experts, scholars, and public sector. It feels needed, however, to say, that this process must allow for the highest degree of openness and transparency possible. The platform must ensure the debatability of every single step in the undertaking, starting from the draft of the general methodology and the procedure's master plan, to be accessible for the public, ensuring possibility of contribution by any member of general public, industry and the scientific community.

1. Technics must be transparently tested within a controlled environment, and all the test phases results, as well as the final technological framework - publicly disclosed and transparently debated.
2. The goal of such prolonged and, no doubt, rather ponderous process should be to reach a moral certainty that a given technic allows the best possible solution for true, not "will do level" alignment.
3. Once a sought after technic has been reached, it must be encapsulated in a form of a unified technological standard - a form allowing to clearly verify, through the measures of external control, whether a given system is aligned or not, and which exactly set of values it is aligned to⁴⁰.
4. After the standard is formulated, compliance to it must become *a conditio sine qua non* for model deployment, a rule which must be enforced through the mandatory pre-market certification⁴¹ of the systems, performed by impartial bodies of oversight⁴². Deployment of uncertified models must be declared illegal, entailing harsh criminal penalties for responsible developers and aware users, as well as a market ban for the organizations. An ongoing monitoring of the system's compliance must be performed by watchdogs, supported by responsive framework, allowing for swift halt of operations by suspicious models and, in extreme cases, their effective market removal.

An important thing is that the abovementioned strategy itself is not entirely incompatible with the existing frameworks. In fact, there already are several significant building blocks which, in case of right calibration, may become a backbone of the proposed structure.

The concept of the conformity assessment is progressive, internal compliance system and risk assessment framework may prove to be very useful for support of AI's sustainable deployment. However, it is clear that the current framework contains numerous critical vulnerabilities.

In contrast with the the proposed model, current approach of "deploy first, control later" poses significant risks of unsafe models entering the market. It is extremely incongruous that applying the model of mandatory pre-market certification for goods like drugs and motor vehicles, where the risks of an unsafe product shipment are traditionally well recognized, law keeps "deploy first, ask later" path for technology which is considered by many to be more dangerous than nuclear weapons⁴³.

While internal conformity assessment and compliance are reasonable extension of healthy practices, in FOMO high-stakes environment exclusively relying on the internal compliance sets a solid ground for conflict of interests - a problem that may be solved only by strict mandatory oversight.

A desire to distribute the compliance costs and avoid excessive bureaucratization in innovational sector is understandable. However, the height and irreversibility of the risks provide the solid justification for the strict oversight, which does not have to inherently mean

⁴⁰ The values set, of course, may, and, probably, should vary depending on the market the system is qualifying for deployment on. While the most fundamental commandments prioritizing human life and other fundamental rights probably should stay the same, on the second layer it may be important to leave a room for amplification of more culturally-dependent variables (e.g. treatment of religious beliefs or privacy). Such approach will allow each society to enjoy the proceeds of a system that is safe while also retaining sufficient degree of its "cultural sovereignty" (or even constitutional identity).

⁴¹ It must include any future updates, of course.

⁴² It may be a net of tightly connected national authorities, a more International body responsible for the maintenance of the standard (e.g., under the UN umbrella), or both - like in case of the ISO standard.

⁴³ the idea has been planted by, *inter alia*, Elon Musk, Steven Hawking, Yuval Noah Harari.

harm of the creativity that innovation requires. While being excluded from the aspect of safety, it should thrive in other parts of R and D process, possibly allowing for evermore daring experiments once the products safety is ensured.

And of course, the standard itself should not be a “commandment” forged in eternity. As technology evolves and society changes, so our laws do, and so should this new kind of regulation. Of course, it should be a subject to changes and experiments, conducted to be conducted in controlled environment - an ongoing improvement is important.

5. Conclusion

The arrival of the transformative technology is marked by great potential, and also numerous challenges and high degree of social confusion. It is, however, up to decision makers to protect the society, mitigating related risks and allowing to embrace its full potential.

In doing so, deep understanding of the essence of new technology is required. As Luciano Floridi wrote, its future lies not in pointlessly attempting to transcend its fundamental nature but in optimising its unique characteristics for beneficial applications⁴⁴.

Agentic nature of an AI makes it different from any technological that mankind has ever created. Agentic nature of AI also marks an arrival of a new kind of subject of the law, one that which can not be governed by conventional legal method and therefore requires development of entirely new one. This task is urgent, crucial and requires decisive action from the governing authorities.

The only legislative framework that will matter in 20 years is now being written in labs, not in parliaments. The question is whether the parliaments will intrude while they still can.

References

- [1] Acemoglu, Daron, and James A. Robinson. *The Narrow Corridor: States, Societies, and the Fate of Liberty*. New York: Penguin Press, 2019.
- [2] Baldwin, Robert, Martin Cave, and Martin Lodge. *Understanding Regulation: Theory, Strategy, and Practice*. Oxford: Oxford University Press, 2012.
- [3] Bullock, Justin B., Samuel Hammond, and Seb Krier. 2025. “AGI, Governments, and Free Societies.” *Public Administration, Governance and Globalization*.
- [4] Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.
- [5] Dittmar, Jeremiah. 2011. “Information Technology and Economic Change: The Impact of the Printing Press.” *Quarterly Journal of Economics* 126(3): 1133–1172.
- [6] Durkheim, Émile. *Sociology and Philosophy*. Various editions.
- [7] García-Marzá, Domingo, and Pablo Calvo. 2024. “Digital Twins.”
- [8] Griffin, Stephen M. *American Constitutionalism: From Theory to Politics*. Princeton: Princeton University Press, 1996.
- [9] Hadwick. “Slipping Through the Cracks,” p. 954. (Incomplete citation – full source needed).
- [10] Gunkel, David J. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, MA: MIT Press, 2012.
- [11] Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.
- [12] Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47(2): 263–291.
- [13] Popper, Karl R. *The Open Society and Its Enemies*. London: Routledge, 1945.
- [14] Posner, Richard A. *Law, Pragmatism, and Democracy*. Cambridge, MA: Harvard University Press, 2003.
- [15] Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press, 2008.
- [16] Weber, Max. *Economy and Society: An Outline of Interpretive Sociology*. Berkeley: University of California Press, 1978.

⁴⁴ Floridi, 2023a, b.