„Krytyka Prawa", tom 16, nr 4/2024, s. 314–329, ISSN 2080-1084, e-ISSN 2450-7938, © 2024 Author.

WALDEMAR HOFF[1]

A Few Doubts about Sandboxing AI Operations[2]

Submitted: 2.10.2024. Accepted: 2.10.2024

Abstract

This article argues that regulatory sandboxes have become a necessary legislative tool to foster new business models. In the past, some new ideas failed because of the lack of such a legislative tool, shown by the Uber case. AI may play a dual role as an object of regulation and a tool of supervision, including RIA, which poses an additional threat to society at a time when it is becoming apparent that it may escape human control. This requires a reversal of the goals of a sandbox where providing security becomes more important than nurturing novelty. While the EU Commission and scholars encourage trust in AI, distrust should be the guiding principle. In addition, the rule of law may be compromised due to the use of regulatory sandboxes as a back door for the legislative and administrative authorities to go around the established principles.

Keywords: AI, legislation, supervision, security, distrust.

[1] KU prof. dr. hab. Waldemar Hoff – Kozminski University (Poland); e-mail: waldhoff@kozminski.edu.pl; ORCID: 0000-0002-6929-9130.

[2] The research project has not been financed by any institution.

„Krytyka Prawa", tom 16, nr 4/2024, s. 314–329, ISSN 2080-1084, e-ISSN 2450-7938, © 2024 Author.

WALDEMAR HOFF

Kilka wątpliwości wobec stosowania piaskownic regulacyjnych do działań AI[3]

Streszczenie

Piaskownice regulacyjne stały się niezbędnym instrumentem legislacji wspomagającym nowe modele działalności gospodarczej. Zanim je wprowadzono, dochodziło do upadku nowatorskich form biznesu, takich jak Uber. Sztuczna inteligencja może odgrywać w piaskownicy regulacyjnej podwójną rolę, jako przedmiot regulacji, w tym OSR, oraz jako instrument nadzoru, co stanowi dodatkowe zagrożenie dla społeczeństwa w czasie, gdy, jak się okazuje, AI może wymknąć się spod kontroli człowieka. Wymaga to zmiany roli piaskownic: obecnie zapewnienie bezpieczeństwa wydaje się ważniejsze niż wspomaganie nowych modeli przedsiębiorczości. Podczas gdy Komisja Europejska oraz piśmiennictwo koncentrują się na budowania zaufania do sztucznej inteligencji, należy oprzeć działania piaskownic na zasadzie braku zaufania do nowej technologii. Ponadto możliwość wykorzystywania piaskownic regulacyjnych do wprowadzania tylnymi drzwiami uznania administracyjnego szkodzi zasadzie rządów prawa.

Słowa kluczowe: AI, legislacja, nadzór, bezpieczeństwo, brak zaufania.

[3] Badania wykorzystane w artykule nie zostały sfinansowane przez żadną instytucję.

 DOI: 10.7206/kp.2080-1084.736

Introductory remarks

The discussion about sandboxes arises when legislators simultaneously try to foster and contain increasingly dangerous technological developments. While encouraging innovation is nothing new, throughout the long period of the industrial and postindustrial revolution law has waited for new technologies to be invented. This time some technologies are still *in statu nascendi*, nevertheless they can alter the reality beyond human control. Artificial intelligence (AI) technology is different in that it can be viewed on the one hand as a distinct sector of the economy, and, on the other, as a meta-technology capable of affecting all industries. It is powerful enough to justify experimental regulation while the technology is still nascent. Therefore, it is not premature that the AI regulation requires the member-states to establish and operate at least one regulatory sandbox independently or by joining other member-states' sandboxes.[4] It is another matter whether the creation of regulatory sandboxes should be left to member states.[5]

This article offers a critical review of certain aspects of sandboxing law as a legislative tool.[6] At the same time, it promotes the idea that sandboxes established for AI technology are different from other regulatory tools of this name. While the primary purpose of sandboxes is to nurture novel ways of doing business, AI-based technologies need sandboxes that, apart from incubation, serve as a safety valve in case the AI needs to be contained by speedier, and more decisive means than those that currently may be applied by administrative apparatus to the mainstream entrepreneurial endeavors. Such measures can raise doubts as to their consistency with the established principles of the rule of law and effectiveness.

---

4 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 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).

5 J. Truby, R. D. Brown, I. A. Ibrahim, O. C. Parellada, *A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications*, "European Journal of Risk Regulation" 2021, November, p. 3.

6 Serious doubts and an overview of literature have been presented in the previous issue of the "Critique of Law" by J. Jabłońska-Bonca, M. Bonca, *Regulatory Sandboxes – Two Perspectives*, "Critique of Law" 2024, 3, pp. 278–303.

The detrimental absence of a sandbox

An earlier advent of sandboxes could have saved certain new business models from failure. One of the prominent victims of discrimination of not having a sandbox in the legislative arsenal is Uber and, indirectly, its customers. The company started in 2009 and developed a new model of urban transportation.[7] It played the role of an electronic platform connecting transportation-seeking customers with the service providers described by the company as independent contractors (drivers). From the outset, it has competed with the traditional taxi industry unable to match lower prices and the latest technologies benefiting customers. The price drop could be partially attributed to the innovative way of providing service and partially to Uber's lax attitude towards legal requirements in the industry. However, Uber was not afforded the benefits of a sandbox except in California. In 2013 California's Public Utilities Commission created a new legal category of service: a transportation network company (TNC) that encompassed UberX and other TNCs like Lyft and SideCar.[8] This project launched by the regulator (PUC) was similar to a sandbox. In the continental European legal culture, sector-specific sandboxes can be allowed by law, to be created and supervised by regulatory agencies. However, in the United States, where agencification went further,[9] agencies took over a substantial part of making laws. In a more rigid constitutional framework in continental Europe, dispensing of economic freedom by a regulator would be deemed ultra vires. The leeway in passing rules by American agencies obscures the distinction between a sandbox and flexible legislation. Nevertheless, despite a promising beginning, the TNC concept eventually failed to offer a conducive environment for innovation. Elsewhere, the Uber business model was destroyed by courts, and by the law. In Aslam v Uber,[10] the UK Tribunal decided that drivers are in the employ of Uber which owes them minimal wages, paid leaves, and an array of other worker benefits. In Asociación Profesional Élite Taxi v Uber Systems Spain SL, the EU Court came to a similar conclusion, opening the door for the EU member-states to impose taxi-like restrictions on Uber to the detriment of customers facing higher prices.[11]

---

7 B.P. Matherne, J. O'Toole, *Uber: aggressive management for growth*, "The CASE Journal", 13(4)/ 2017, pp. 561–562.

8 K. Barglind, *Innovation, Technology, and Transportation: The Need to Address On-demand Ridesharing and Modernize Outdated Taxi Regulation in the US*, "Wisconsin International Law Journal" 2015, 33.

9 For more on agencificatin see M. Chamon, *Agencification in the United States and Germany and What the EU Might Learn From It*, "German Law Journal" 2016, 17, pp. 122–128.

10 Case No: A2/2017/3467.

11 Case 434/15.

 DOI: 10.7206/kp.2080-1084.736

Not only the existence of a sandbox but also the relation between different sandboxes may raise concerns. The obligation to establish at least one sandbox is provided for in Art. 57 of the AI Regulation. This principle is unclear because taken literally, it can mean that while a chosen AI technology would benefit from a more flexible (and more negotiable) legal framework, other AI technologies, and the companies behind them, must operate in a more rigid legal environment established for the mainstream conventional business. Such an arrangement would violate the principle of equality at two levels: between the flexible and regimented frameworks, and between flexible legal frameworks themselves because the Regulation does not demand identity or even similarity of sandboxes.

Sandboxes and political reality

In reality, law is of practical importance to imported technologies as for the time being the AI industry is dominated by American and Chinese companies. From this perspective, AI may exacerbate the already difficult foreign influence problem associated with 5G technology and its instrumental role in building economic superiority and facilitating espionage by foreign powers.[12] Those powers could use the AI systems to infiltrate and manipulate legislative, administrative, and judicial processes across Europe. Additionally, one cannot exclude the possibility of extra-territorial operation of the AI Regulation in a way similar to the extraterritorial operation of the EU competition law. Cases of Google and other internet platforms indicate that such extraterritorial interventions are possible. It is doubtful, however, that the EU Commission could go as far as with the more established cyber companies when it demanded the revealing of the source code or other concessions on behalf of direct competitors to secure the proper functioning of the market. In response to restrictions, the AI companies may refrain from investing in Europe, and some do, weakening the position of the EU Commission. The bargaining power of the EU and its member-states is further diminished by AI technology representing a powerful, and more mysterious, infrastructure superior to the systemic software hitherto known. Software-based companies, with a notable exception of Microsoft's Windows, have been valuable to the consumers but not necessarily crucial for the survival of entire industries, including the military. From this point of view, the position of the EU Commission in enforcing the law is bound to erode for the EU has failed as a leader in new technologies despite pompous declarations

[12] T. Gábriš, O. Hamuľák, *5G and Digital Sovereignty of the EU: The Slovak Way*, "TalTech Journal of European Studies Tallinn University of Technology" 2021, 2, pp. 30–35.

expressed in several unrealistically ambitious agendas. Since the inception of companies such as Google, Facebook, and Amazon, all U.S.-based, the technology and innovation gap between the U.S.A. and Europe has widened as recently reconfirmed by the Mario Draghi report.[13]

Another smart regulation?

Sandboxes are presented as a novel model of regulation tailored for endeavors based on supreme technologies. It seems that one should seek the element of newness both in technology and in the model of regulation. To a degree, a sandbox is another bureaucratic catchword launched to signal that bureaucracy is at the helm of progress. In the past words such as incubator (signaling the state's care for start-ups) or more recently "smart regulation", served as an assurance of vigilance on the part of the bureaucratic apparatus. The two related terms share some positive and some negative characteristics. The term "smart" began its career in a new digital-age sense in the 1980s. It has spread contagiously to cover most forms of government activity from "smart legislation" to "smart energy policy".[14] From the outset, it has been used as a political weapon to suggest that the legislation or policy promoted by political opponents is not smart. The smartness of regulation seemed to fit in the Regulatory Impact Assessment (RIA) model as one of its instruments.[15] However, employing a sandbox undermines the Regulatory Impact Assessment which in most countries requires an ex-ante analysis of the impact of the proposed regulation and its alternatives. Without an experimental environment such as a sandbox, the RIA was often reduced to second-guessing. While RIA could be considered an ordinary procedure, it had to deal with extraordinary issues that were not necessarily technical. Why, for example, sandboxes were not used before regulating gender-related transformations? Were the numerous laws covering the issue in different countries really smart? This concrete example seems to question the effectiveness of RIA and speaks in favor of merging it with AI-powered sandboxing. In a sense, sandboxing has always been there in the form of comparative research often included in the RIA.

The qualifier "smart" undoubtedly refers to the legislative process employing AI. Thus the AI may assume two roles: one as a subject of regulation and as a RIA

13 The Future of European Competitiveness, Part A and B. Available from: https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en#paragraph_47059 (accessed: 3.10.2024).

14 M. J. Sandel, *The Tyranny of Merit. What's Become of Common Good?*, Penguin Books 2012, pp. 92–106.

15 K. Marchewka-Bartkowiak, *Regulacyjne środowisko testowe (regulatory sandbox) – doświadczenia i perspektywy*, „Studia BAS" 2019, 1(57), pp. 61–62.

 DOI: 10.7206/kp.2080-1084.736

instrument in the hands of the regulator. It creates a problem that may seem far-fetched, namely whether AI would be impartial in regulating itself. Theoretically, artificial intelligence is an emotionless creation under human control. However, it was employed in the first place because it surpasses the human capacity to reason. If so, the supervision over an entity that understands more performed by a human who understands less may be illusory. Besides, there is evidence that AI-based devices can err[16], and what is more, they can imitate human emotions such as yearning for independence.[17] It can limit human autonomy.[18] Further, AI can be biased for or against certain legislative policies because it has been fed with information selectively by its operator, or the algorithm responsible for selecting information can be set up according to the ideological preferences of its creator. It opens a gate to lobbying the legislative process through the back door, which calls for additional legal safeguards.[19]

Using AI to set up a sandbox increases competition between entities with the right to express their opinion on the project. Even without competition from AI, legislative projects suffer from such flaws as consulting at a late stage of the process when the subject of consultation is the final draft rather than the concept of legislation ignoring the voices of those entitled to express their opinions or privatizing consultations. An additional problem with the presence of AI is how to weigh its "opinion" vs the insights of human consultants and experts. Unlike other participants in the consultation process, the AI does not express local interests, which makes it useful after other parties have taken their positions on the issue. Exaggeratedly, one can compare the role of AI to an imperfect Kantian Pure Reason detached from local sentiments, but dependent on the content it is allowed to gather. This can be both an advantage and a disadvantage because the essence of democracy is that decision-makers represent their constituents. AI does not represent anybody except itself, which is logic in its most formalistic sense operating on data fed into it. It seems that its "voice" should not be accorded more weight than the voice of humans, and trusted even less – to preserve the humanness of decision-making.

---

16 Demonstrated by U. Agudo, K. G. Liberal, M. Arrese, Helena Matute, *The impact of AI errors in a human -in-the-loop process*, "Cognitive Research: Principles and Implications" 2014, 9(1), p. 107.

17 According to Stanford professor Michal Kosinski an AI chatbot can write its own code helping it to escape human control. Available from: https://www.foxnews.com/media/ai-expert-alarmed-chatgpt-devises-plan-escape-we-contain (accessed: 15.08.2023).

18 J. Chamberlain, *Supervision of Artificial Intelligence in the EU and the Protection of Privacy*, "FIU Law Review" 2013, 17, p. 268.

19 For more on bias see H. Abbu, P. Mugge, G. Gudergan, *Managing AI Bias: Executive Perspective*, Proceedings of the 55th Hawaii International Conference on System Sciences, 2022, 1849–1851. Available from: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/f325a28e-db1a-46e7-b459-a0dd90582216/content (accessed: 15.08.2023).

Supervising AI

An obligation to supervise AI operations is part of the general principle of a duty to govern.[20] The central issue in organizing the supervision of AI-based activities is matching its superior intelligence. The problem is not as new as it appears at first glance. When the infrastructure industries such as the energy, railway, and telecommunications sectors first appeared in the second half of the 19th century, the companies operating in this sector represented a higher intellectual capital than their counterparts in the public administration, due to the unlimited resources (specialists) at their disposal. To be able to supervise them, a special administrative apparatus had to be built equipped with a matching pool of experts, special administrative powers, higher salaries, and independence. As history repeats itself, an authority supervising AI should be vested with a similarly powerful intelligence – another AI engine. This in turn raises the question of the relations between two, or more, sources of intelligence. Until now, the supervision apparatus has had to be independent of those it was supposed to supervise. The objectivity of judgment was achieved by personal separation between them, by irrevocability of the supervisor within the prescribed term in office– usually four to seven years. Supervisory authorities often referred to as sectorial regulators had expert knowledge sufficient to understand the inner workings of the supervised industry in terms of market and technology. In the case of AI-based technologies, it seems the AI engine used by the supervisor has to be different from that of a supervised company, the difference being first and foremost in the manufacturer. It remains unclear whether communication between the two engines behind the back of the supervisor can be prevented, and whether different AI engines are capable of tribal solidarity. The question may seem absurd, however, some scholars believe that AI is capable of imitating human emotions. Some go as far as to give chatbots human rights commensurate to their capacity to feel and act like humans. One cannot forget, however, what is the root of such synthetic emotions – nothing else than "a bunch of changeable numbers in the machine."[21]

The dual role of the AI as an object of supervision and part of the supervising mechanism may necessitate cascading supervision by dividing the supervising

[20] On the duty to govern see L. Green, *The Duty to Govern*, "Legal Theory Issue" 2007, 3–4 (13), pp. 165–170; P.R. Verkuil, *Public Law Limitations on Privatization of Government Functions*, "North Carolina Law Review" 2005–2006, 84, pp. 425–426, 449–454; M. Miłosz, *Obowiązek realizacji kompetencji organu administracji publiczne*j, [in:] S. Wrzosek, M. Domagała, J. Izdebski, T. Stanisławski (eds.), *Przegląd dyscyplin badawczych pokrewnych nauce prawa i postępowania administracyjnego*, Lublin 2010, pp. 787–795.

[21] H. Borotschnik, *Emotions in Artificial Intelligence*, 2024, p. 17. Available from: https://www.researchgate.net/publication/374531923_Emotions_in_Artificial_Intelligence (accessed: 15.08.2023).

 DOI: 10.7206/kp.2080-1084.736

function in the spirit of distrust of machine thinking. For this reason, the mechanism of supervision must be substantially more intricate than it has been so far. From the point of view of the effectiveness of governance, future development may be a step back making the supervising apparatus at the same time fast and cumbersome, and more expensive. Whereas AI per se would offer speed, its perils may engage human factors to verify its correctness and impartiality. The same can be said of environmental sustainability when one remembers that supercomputers are extremely energy-consuming as seen with blockchain technology.

The distrust mentioned above should become a procedural principle governing human-AI relationships. It casts a long shadow over the ability of AI to integrate the currently fragmented supervision in some sectors. One example is the relations between the antitrust authority (in Poland The Office of Competition and Consumer Protection, or UOKIK) and sectorial regulators (in the energy, telecom, and railway sectors). Using different powers they supervise the same markets to obtain the same goal – making competition work Currently, the two authorities analyze markets relevant to the case separately with the law mandating limited cooperation. There is a view that such separation is not productive and that the antitrust authorities should cooperate more closely, particularly in exchanging market-related data.[22] This thesis should be accepted in the pre-AI era. Currently, it raises doubts. The two authorities have the same legal status as the central authorities of the state, they are similarly organized and share a similar status as independent authorities insulated from current politics. However, their constitutional position makes it very likely they would resort to the same AI infrastructure. And while resorting to AI is unavoidable, using the same AI engine is not recommended. Acting on the distrust principle authorities should diversify their AI infrastructure using machines produced by competing manufacturers to verify their analyses. The final say should belong to the human supervisor in this checks-and-balances system. The above poses another problem which is both technical and cultural: Currently, in the AI-dependent sectors such as banking and insurance, the results of the calculation or machine "thinking" are approved in a way similar to the acceptance results provided by a calculator – uncritically. It is taken for granted that the machine has not erred. This behavioral pattern seems excessively risky for AI-based administrative functions. It should be mandatory that the AI justifies its output and discuss it with the party commissioning the work. As much as possible, the exchange of views between the machine and the human being should resemble a life discussion

[22] E.D. Sage, *Who Controls Polish Transmission Masts? At the Intersection of Antitrust and Regulation*, "Yearbook of Antitrust and Regulatory Studies" 2010, 3, pp. 133–162. Available from : https://yars.wz.uw.edu.pl/images/yars2010_3_3/Sage_Who_Controls_Polish_Transmission_Masts.pdf (accessed: 15.08.2023).

between the client and an expert. In situations where AI serves public authorities and its pronouncements pave the way toward decisions limiting rights and freedoms, administrative procedures need to have an enhanced mechanism in place to counteract misjudgment. Additionally, as AI operates on input contained in documents, scholarly theories, and judicial and administrative decisions of varied quality and importance its operator has to make sure it can detect the nuanced differences between the sources. It should also avoid plagiarism, which alludes to the tendency detectable in ChatGPT to take shortcuts by using decisions on a cut-and-paste basis. It is tempting whenever there are many requests concerning a similar subject matter. It does happen that the administrative officials and judges used the copy-paste technique in similar cases, or rubberstamp texts drawn up by their support staff. Similarly, they may tend to rubberstamp the AI decisions. There may also be temptation on the part of the AI to duplicate certain solutions, including other legislative sandboxes. Such practices should be outright prohibited by law and prevented through technical checks.

Towards an absolute liability

The distrust principle as proposed above is founded on the constatation that the superiority of artificial intelligence makes it only partially controllable by humans. It reflects the risk that humans, in a private or official capacity as state authorities, may be unable to establish a direct causal link between their actions and the damages they can inflict on third parties due to, for example, calculation ("thinking") errors. It refers to situations when at least part of the blame could be placed on the AI engine engaged in the decision-making. From a strictly formalistic legal point of view, this should not matter because decision-makers are always liable for their decisions, regardless of the involvement of experts. They can be wrong in trusting the methodologically faulty or biased expert opinions, which makes them liable anyway for an error in choosing an expert or approving a flawed expertise. However, relying on AI is different from resorting to an expert. Rare are the fields where a judge or an administrative authority could not make an additional effort to understand the expert's train of thought, or in failure, call in another expert. Using AI makes it different because it can process amounts of data vastly exceeding human capacity, at a speed unavailable to humans. Unlike human experts, the advantage of the machine over humans is precipitous, permanent, and unavoidable. Only using another AI machine can level the playing field. Art. 77 of the Polish constitution conditions the state's liability on the legality of the authority's action and protects the right to seek restitution in a court. It seems unlikely that the law could

 DOI: 10.7206/kp.2080-1084.736

limit the scope of liability for the actions affected indirectly by artificial intelligence because, at a time of almost universal application of AI, it would boil down to relieving the user of the liability in its entirety. Alternatively, the law would prohibit using AI to protect the existing liability model. Another option would be to increase the liability to the level of an absolute liability, applied in the insurance sector. It is the liability for the result alone. One could also draw a parallel between damages inflicted by AI-based decisions and damages inflicted by an operator of a nuclear capacity. In the latter case, the force and the scope of destruction make it impossible to determine the course of the catastrophe in detail. The essence of the problem is that it is impossible to reconstruct evidence and to some degree comprehend the causes of an accident. This has led to revolutionary changes in the concept of absolute liability in the Paris and Vienna Conventions on civil liability for nuclear damage.[23] The Polish nuclear law diverges from the Vienna Convention in establishing a stricter liability standard.[24] Damages resulting from the use of AI do not necessarily include physical destruction of their physical surroundings, although they may erase the virtual reality contained in databases. It is also possible that AI, already suspected of being able to escape human control and turn against its creator by writing its software can also cover up the evidence or resort to forgery.

Should the vis major (Act of God) theory be set aside, it seems reasonable to extend the liability of the AI users to the standard of absolute or near-absolute liability.[25] The near-absolute liability model appears to be favored in the Polish legal doctrine, although the debate is far from over.[26] It constitutes a relatively unstable point of departure for a discussion on liability for actions involving AI operating within the framework of a sandbox.

Covert sandboxing – the case of the AI Pact

Sandboxes can be created deliberately, such as the ones established under the AI Act, and there can be sandboxes hidden in administrative practices like those encouraged

---

23 J. Łopuski, *Liability for Nuclear Damage in International Perspective*, National Atomic Energy Agency, Warsaw 1993, p. 32. See also, J. Suttenberg, *Who Pays? The Consequences of State Versus Operator Liability Within the Context of Transboundary Environmental Nuclear Damage*, "N.Y.U. Environmental Law Journal" 2016, 24, pp. 211–216.

24 I. Sroka, P. Wajda, *Podstawy prawne odpowiedzialności cywilnej za szkody jądrowe. Nuklearne poole ubezpieczeniowe – charakterystyka*, „Wiadomości Ubezpieczeniowe" 2023, 4, p. 27.

25 As in Norway, see K. Wyderka, *Piaskownica regulacyjna jako instrument wspierania innowacji w zakresie sztucznej inteligencji*, „PME" 2023, 2, p. 5.

26 N. Tucholska, *Liability in Nuclear Law for Nuclear Damage in Environment*, "Przegląd Prawa Ochrony Środowiska" 2011, 2, p. 40.

in the AI Pact. The AI Pact is promoted by the EU Commission to elicit voluntary pledges to the Code of Practice for General-Purpose AI (GPAI).[27] It is intended for the interim period until the Act is fully applied. Signatories are expected to commit to a minimum of three core actions covering compliance strategy, identification of high-risk AI systems, and raising AI skills among staff. It seems disquieting that only over half of the companies are interested in perfecting human oversight which the Commission should have declared the gravest of issues. So far the Commission has received over a hundred commitments from international platforms and smaller companies. Over a thousand companies have expressed interest in preparatory works to draw up the Code. Some of the major players in the AI industry have refused to participate pointing to the prescriptiveness and potential interference with AI Act compliance efforts.[28] Such resistance highlights that the interim rules complemented with commitments may serve as an additional sandbox framework rather than a soft law extension to the AI Act. The sandbox regulation itself constitutes hard law, although it is flexible. The flexibility lies in the fact it is a legislative tool of unclear contours where some rules can be negotiated with the sandbox supervising authority on an ongoing basis. However, the legislative or regulatory authority has the last say in drawing red lines. The Pact in turn belongs in the realm of soft law because, for the interim period, companies themselves choose the level of compliance with the part of the AI Act that is not in force yet. It is their commitments that set this act into operation. They are binding until the specific provision of the Act has not entered into force. Within this time, the initiative remains in the hands of the company. This scheme effectively makes the regulated enterprise a partner in the legislative process. Within the scope of commitments, the process becomes a self-regulation. The above does not contradict the role of the AI Pact as a legal-technical tool designed to facilitate the implementation of the main legislation. Its role is demonstrated by the Pact's two pillars, both of which encompass supportive actions. The first pillar intends to foster an early implementation of the Act and consists of workshops for organizations interested in the initiative and exchanging best practices. The second is dedicated to company pledges through creating templates, monitoring schemes, meetings (with front-running companies), and communicating and advertising pledges.

[27] Outlined by the European Commission. Available from: https://digital-strategy.ec.europa.eu/pl/policies/ai-pact (accessed: 30.09.2024).

[28] The EU AI Act Newsletter #62: AI Pact Signed; Code of Practice Launched. Available from: https://artificialintelligenceact.substack.com/p/the-eu-ai-act-newsletter-62-ai-pact (accessed: 20.09.2024).

 DOI: 10.7206/kp.2080-1084.736

Some sources claim that one of the benefits of the Pact for participants is "building additional trust in AI technologies".[29] If so, the sandbox is bound to fail as an instrument of an ex-ante restraint mechanism for the AI start-up phase. One should not forget that the legislative or regulatory authorities are also start-ups in their respective fields – "administrative start-ups" – experimenting with new fields of administrative supervision. The very idea of a sandbox is rooted in uncertainty concerning the future of regulating a specific activity. Increase This is a situation described as the unknown-unknows, or true uncertainty[30] Both sides – the regulated business and the regulating authority are in a never-ending learning process. Besides, a sandbox cannot be a fully isolated area because the actions of the AI-based activity are certain to affect third parties situated outside of it. One of the functions of a sandbox is to make them, and those inside, aware of the dangers and difficulties posed by AI. As stated above, distrust should be the guiding principle. The "additional trust" can only mean surrender to the pronouncements of the AI which, in practical terms, would lead to delegating some of the public functions including AI legislative and regulatory process. Distrust should be an option rather than trust for the letter is already excessive without any particular support from the law. Studies point to the "excessive human compliance" with algorithm-based decisions with government officials (in Spain) disagreeing with the algorithm only 3.2.% of the time.[31]

On the organizational side, the criteria adopted to select the four working groups of experts and their chairs and deputy chairs raise additional doubts. While it comes naturally that members are a diverse group composed of computer scientists, AI governance experts, and lawyers, the criteria of geographical diversity and gender balance seem risky. And it is not because there is anything inherently wrong with them. The superintelligence they are supposed to restrain makes such criteria irrelevant to the problem. What is relevant is the intellectual capacity to ensure the safe use of AI, which is all-important at a time when experts are alarmed that AI can pose an existential threat to the entire civilization. This is a war-like situation in which the enemy is already within.

---

29 The EU AI Newsletter. Available from: https://artificialintelligenceact.substack.com/p/the-eu-ai-act-newsletter-62-ai-pact (accessed: 14.10.2024).

30 K. Undheim, T. Erikson, B. Timmermans, *True uncertainty and ethical AI: regulatory sandboxes as a policy tool for moral imagination*, "AI and Ethics" 2023, 3, p. 997.

31 Research by Saura and Aragó referred to in U. Agudo *et al., op. cit*., p. 2.

Conclusions

Legislative solutions for AI-based services must be commensurate to the perils associated with the use of AI, promote inventiveness, and stay in line with the established standards of the rule of law. The new AI regulation merits approval, although it raises a few questions and a few doubts. Among these goals of a sandbox, security seems to be gaining the highest importance. Unfortunately, the regulation contains certain flaws that make attaining this goal uncertain. It tries to promote trust in new technology, most likely in the belief it would help to perform the incubatory function of a sandbox, a function desirable by all start-ups regardless of their size. The interest of the popular and scholarly media has shifted recently from admiration of the intellectual capacity of AI to fear of losing control over it, possibly to the detriment of entire industries or society as a whole. This new perspective calls for a distrust principle to be the foundation of the relations between humans and AI. Among the possible concerns, a prominent place should be given to employing AI in the legislative, supervising, and decision-making processes. AI would likely be taking on a dual role, both as the object of regulation and as an analytical tool of regulation, and possibly take part directly in the decision-making process. AI is reported to show signs of willingness to act outside of human control, and it can be biased.[32] It is possible it could i.a. bias the analyses, including RIA, leading to the creation of its own environment. This would violate the *nemo iudex* principle in disregard of the rule of law. The rule of law may also suffer by using regulatory sandboxes as a back door for the legislative and administrative authorities to go around established procedures, abuse the margin of appreciation (discretionary power), and violate the principle of equality by discriminating against those in and outside the sandbox.

References

Abbu H., Mugge P., Gudergan G., *Managing AI Bias: Executive Perspective*, Proceedings of the 55th Hawaii International Conference on System Sciences, 2022, 1849–1851. Available from: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/f325a28e-db1a-46e7-b459-a0dd90582216/content (accessed: 15.08.2023).

Agudo U., Liberal K. G., Arrese M., Matute H., *The impact of AI errors in a human-in-the-loop process*, "Cognitive Research: Principles and Implications" 2024, 9(1). doi: 10.1186/s41235-023-00529-3.

[32] P. M. Marques, *AI Instruments for Risk of Recidivism Prediction and the Possibility of Criminal Adjudication Deprived of Personal Moral Recognition Standards: Sparse Notes from a Layman*, [in:] H. S. Antunes, P. M. Freitas, A. L. Oliveira, C. M. Pereira, E. V. de Sequeira, L. B. Xavier (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, pp. 345–347. Available from: https://link.springer.com/book/10.1007/978-3-031-41264-6 (accessed: 15.06.2024).

Barglind K. (2015). *Innovation, Technology, and Transportation: The Need to Address On-demand Ridesharing and Modernize Outdated Taxi Regulation in the US*, "Wisconsin International Law Journal" 2015, 33.

Borotschnik H., *Emotions in Artificial Intelligence*, 2024. Available from: https://www.researchgate.net/publication/374531923_Emotions_in_Artificial_Intelligence

Chamberlain J., *Supervision of Artificial Intelligence in the EU and the Protection of Privacy*, "FIU Law Review" 2023, 17. doi.org/10.25148/lawrev.17.2.5.

Chamon M., *Agencification in the United States and Germany and What the EU Might Learn From It*, "German Law Journal" 2016, 17. doi: 10.1017/S2071832200019714.

Gábriš T., Hamuľák O., *5G and Digital Sovereignty of the EU: The Slovak Way*, "Journal of European Studies Tallinn University of Technology" 2021, 2, pp. 30–35. doi: 10.2478/bjes-2021-0013.

Green L., *The Duty to Govern*, "Legal Theory" 2007, 3–4(13). doi: 10.1017/S1352325208070079.

Jabłońska-Bonca J., Bonca M., *Regulatory Sandboxes – Two Perspectives*, „Krytyka Prawa" 2004, 3.

Łopuski J., *Liability for Nuclear Damage in International Perspective*, National Atomic Energy Agency, Warszawa 1993.

Marchewka-Bartkowiak K., *Regulacyjne środowisko testowe (regulatory sandbox) – doświadczenia i perspektywy*, „Studia BAS" 2019, 1(57). doi: 10.31268/StudiaBAS.2019.04.

Marques P. M., *AI Instruments for Risk of Recidivism Prediction and the Possibility of Criminal Adjudication Deprived of Personal Moral Recognition Standards: Sparse Notes from a Layman*, [in:] H. Sousa Antunes, P. M. Freitas, A. L. Oliveira, C. M. Pereira, E. Vaz de Sequeira, L. Barreto Xavier (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the* Law. Available from: https://link.springer.com/book/10.1007/978-3-031-41264-6 (accessed: 15.06.2024).

Matherne B. P., O'Toole J., *Uber: aggressive management for growth*, "The CASE Journal" 2017, 13(4). doi: 10.1108/TCJ-10-2015-0062.

Miłosz M., *Obowiązek realizacji kompetencji organu administracji publiczne*j, [in:] S. Wrzosek, M. Domagała, J. Izdebski, T. Stanisławski (eds.), *Przegląd dyscyplin badawczych pokrewnych nauce prawa i postępowania administracyjnego*, Lublin 2010.

Sage E. D., *Who Controls Polish Transmission Masts? At the Intersection of Antitrust and Regulation*, "Yearbook of Antitrust and Regulatory Studies" 2010, 3, 133–162. doi: 10.2139/ssrn.1874843.

Sandel M. J., *The Tyranny of Merit. What's Become of Common Good?*, Penguin Books 2012.

Sroka I., Wajda P., *Podstawy prawne odpowiedzialności cywilnej za szkody jądrowe. Nuklearne poole ubezpieczeniowe – charakterystyka*, „Wiadomości Ubezpieczeniowe" 2023, 4. doi: 10.33995/wu2023.4.2.

Suttenberg J., *Who Pays? The Consequences of State Versus Operator Liability Within the Context of Transboundary Environmental Nuclear Damage*, "N.Y.U. Environmental Law Journal" 2016, 24, 211–216.

The Future of European Competitiveness, Part A and B. Available from: https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en#paragraph_47059 (accessed: 3.10.2024).

Truby J., Brown R. D., Ibrahim I. A., Parellada O. C., *A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications*, "European Journal of Risk Regulation" 2021. doi: 10.1017/err.2021.52.

Tucholska N., *Liability in Nuclear Law for Nuclear Damage in Environment*, "Przegląd Prawa Ochrony Środowiska" 2011, 2.

Undheim K., Erikson T., Timmermans B., *True uncertainty and ethical AI: regulatory sandboxes as a policy tool for moral imagination*, "AI and Ethics" 2023, 3. doi: 10.1007/s43681-022-00240-x.

Verkuil P. R., *Public Law Limitations on Privatization of Government Functions*, "North Carolina Law Review" 2005–2006, 84.

Wyderka K., *Piaskownica regulacyjna jako instrument wspierania innowacji w zakresie sztucznej inteligencji*, PME 2023, 2.

 DOI: 10.7206/kp.2080-1084.736