Acta Universitatis Sapientiae

Legal Studies

Volume 11, Number 2, 2022

Sapientia Hungarian University of Transylvania
Scientia Publishing House

Contents

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 5–18

DOI: 10.47745/AUSLEG.2022.11.2.01

Liability for Damage Caused by AI Entities

Zsolt György BALOGH
PhD, Associate Professor
Corvinus University of Budapest, Faculty of Business Administration (Hungary)
e-mail: zsolt.balogh@uni-corvinus.hu

Abstract. The age of big data and machine learning technologies brought the new flourishing of artificial intelligence research along with profound innovation in digital services. A global AI race is underway, and the EU seeks to play a determining role in it, by exploiting its scientific abilities and strengths. Beyond the commercial and technological interests, the EU is intent on preventing the damages and harms that can be caused by devices and systems using AI, which could undermine users' trust in this new and promising technology. The protection of users from AI-caused damage will consequently constitute a crucial factor of the global AI contest. European integration is about to elaborate a regulatory framework on civil liability related to AI applications.

Keywords: artificial intelligence, big data, regulation, European Union, risks

1. Introduction

Humanity's way of life has been profoundly transformed by the innovations of the information age and the technology giants that have made them massively available. While in the physical world climatic conditions are changing dramatically, devastating wars are breaking out, and a virus has swept across the globe, virtuality is infiltrating everyday life. People spend their time stuck to their screens, constantly refreshing endless streams of personalized content.

This is just one of the many applications of artificial intelligence (AI) and its impact on our lives. The sophisticated psychological trap of social networking is largely based on advanced profiling capabilities that harness AI. Over the past decades, AI has evolved into one of the most progressive, far-reaching, and challenging areas of computer science. A broad and enriching range of applications is emerging and, in parallel, we are facing more and more problems with the ethical, legal, and governance issues surrounding the use of AI.

Before the substantive discussion on legal liability issues of this technology, the attributes, capabilities, and functions of AI must be described. Consequently,

for the purpose of this legal survey, a definition and taxonomy of the AI systems will be inevitable.

2. Definition and Taxonomy

The recent proliferating technology-related literature offers a plethora of definitions and approaches on artificial intelligence per se and on functional AI systems. For the sake of the legal tract, only some – maybe arbitrary – approaches will be considered.

2.1. Scientific Definitions of AI

As no ultimate definition on AI can be found, one may consider certain elements of several interpretations. The concept of intelligence can be identified and measured by several attributes. The touchstone of intelligence can be human-like cognitive performance or an abstract, ideal rationality, that is, 'rightful thinking'. According to other definitions, intelligence would be the ability to conduct sophisticated thought processes and reasoning or engage in intelligent behaviour. These vectors, as described by Russell and Norvig,[1] delineate four main interpretations of AI as follows:

– thinking like a human being, that is, cognitive modelling;
– acting like a human being, that is, the ability for passing the Turing Test;
– thinking rationally, that is, the logic-based model;
– acting rationally, that is, the rational agent model.

Russel and Norvig, however, emphasize that the notion of intelligent agent[2] is the central concept of the aforementioned categories of artificial intelligence. The intelligent agent is designed to receive percepts from the environment and to perform actions. The famous 'Turing test', which Alan Turing himself called the 'Imitation Game',[3] constitutes a kind of touchstone of intelligent machine behaviour. Conducted according to some relevant criteria, the test basically implies a 'conversation' between a human party and a human or computer interlocutor in such conditions as to leave the human party unaware of whether s/he is interacting with a fellow human or a computer. At the end of the test, the human party is asked whether s/he thinks s/he has just communicated with another human or a computer. An AI system would be considered to have passed the Turing test if it would be indistinguishable from a human interlocutor by the human party. According to certain opinions, an AI chatbot called Eugene

1 Russell–Norvig 2021. 2.
2 Id. VII.
3 Bernhardt 2016. 157.

Goostman actually succeeded in passing the Turing test in 2014, though this information remains heavily controversial.[4] The intelligent agent that can have any chance of passing the Turing test shall require at least the following capabilities: natural language processing, knowledge representation, automated reasoning, and machine learning.

Considering the requirement for a rational – non-human – AI, which would not operate within the confines of a simple conversation, more abstract and exact approaches need to be implemented in the development process such as: mathematics, statistics, and several branches of formal logic as Boolean (propositional) logic, first-order logic, deontic logic, and fuzzy logic.

These are considered by now, among others, as the main disciplines of AI research and development. For the purposes of robotics[5] ('embodied AI'), however, even further disciplines and technologies need to be (and are being) developed such as: computer vision and face recognition, speech recognition, and affective computing for expressing (or more likely imitating or emulating) emotions.

It is evident that for the purpose of creating a practical AI system, several technologies must be developed and employed in conjunction, a problem that is sometimes overlooked.

2.1.1. Indeterministic Behaviour: A Crucial Challenge

A usual computer program is designed to operate – to behave – in a deterministic way. Every user expects a word processor, for example, to carry out its functions in a proper order without any 'creative' actions. In fact, such actions may even be considered as perturbing normal use. The unpredicted reaction of a program to input usually indicates a 'bug' in the code, an unforeseen error.

An AI application, however, is not a usual form of software. Some AI systems are based on deterministic algorithms, but most newly developed systems employ deep learning and related technologies that implement non-deterministic algorithms and require a good deal of arbitrary datasets to be educated (trained) on. Probabilistic functioning is an inherent attribute of these systems. This means that an AI system works in a non-deterministic way, and this crucial property imposes high security risks in the course of implementation and use of AI systems. These systems may produce answers in a less transparent and explicable way than a user would expect. In many cases, even the AI experts and developers cannot predict correctly and explain the conduct of the AI. An AI system can also be interpreted as a black box;[6] by giving the system an input,

4 Masnick 2014.
5 Häuselmann 2022. 47.
6 Tan 2022. 92.

the user will receive an output without any obvious reasoning or explanation. The mode of substantive operation of the system can be approached at a certain level only by inference engineering.

This is collateral, and rather undesirable side-effect of the highly sophisticated autonomous technologies used in creating AI systems. The lack of direct control over our machinery is an entirely new phenomenon in the history of technological civilization, and still we have no proper answers and policies for disentangling ourselves from this situation. The most threatening and alerting scenario in the evolution of AI and robotics would be the rise of self-aware AI and perchance superhuman intellect displayed by such systems. Academic discussions and papers[7] also warn of this opportunity beyond the realm of overabundant science fiction works. The security risk requires special care from developers and regulators when implementing this new technology.

There is a strong motivation and inevitable need to create proper controls and security provisions for the development and use of AI systems. As to the regulatory framework, this surely must soon be elaborated, in the form of new doctrines and norms beyond the habitual toolkit of today's law. The need to regulate the roles and liabilities of the providers and operators of the AI is becoming more evident, along with the technological evolution of these autonomous and intelligent systems.[8]

2.1.2. The Fields of Application of AI

The Dartmouth Conference[9] (Hanover, New Hampshire) of 1956 is claimed to be the founding event of AI research. Since that time, AI technologies and methods have grown very sophisticated, gave rise to many genuine fields of application, and percolated into several segments of social, economic, and personal activities. Some sectors that have implemented AI extensively include:[10] astronomy and other sciences, climatology, data- and cybersecurity, e-commerce, education, finance and banking (stock market management and forecast), gaming and entertainment, healthcare (diagnosing), household and personal assistance, manufacturing, robotics, social media platforms, and transport (navigation, traffic optimization, autonomous vehicles, etc.).

Most of the listed domains are closely related to personal activities and permit human involvement, so direct legal and liability issues may be concerned in respect of fundamental rights and freedoms of natural persons or the business interests of companies.

7 Totschnig 2020.
8 Custers–Fosch-Villaronga 2022. 10.
9 McCarthy et al. 1955.
10 JavaTpoint.AAI 2022a.

2.2. Definition of AI in the EU law

The Artificial Intelligence Act (hereinafter referred to as AIA) of the EU – technically a draft bill of a forthcoming EU Regulation – also attempts to define artificial intelligence for the purpose of constructing a regulatory framework. Article 3(1) of the AIA describes the notion of an artificial intelligence system – and not the abstract concept of AI – as follows (original emphasis): 'For the purpose[s] of the Regulation[,] *artificial intelligence system* (AI system) means software that is developed with one or more of the techniques and approaches listed in *Annex I* and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.'

The definition can be interpreted using the content of Annex I, which enumerates the relevant technologies as being:

– Machine learning approaches, including:
  - supervised,
  - unsupervised,
  - reinforcement learning, using a wide variety of methods, including
  - deep learning.

– Logic- and knowledge-based approaches, including:
  - knowledge representation,
  - inductive (logic) programming,
  - knowledge bases,
  - inference and deductive engines,
  - symbolic reasoning and expert systems.

– Statistical approaches:
  - Bayesian estimation,
  - search and optimization methods.

This interpretation tends to be as neutral as possible, that is, the legal concepts omit any appearance of the legislator taking sides in the scientific discussion on whether the technical criteria and quality of AI should be compared to the average human skills and capabilities or an abstract – mathematical and/or logical – rationality. The list of relevant technologies is also a substantive part of the definition. This agenda reflects the widely acknowledged scientific definition and taxonomy of AI.

3. Typology of AI

Speaking about AI, it is obvious to see that several levels of intelligence can be observed in this domain. Some systems work only within very limited abilities, while other ones may compete with skills of human experts – for instance, in

medical imaging, mostly implemented in diagnostics such as cancer diagnostics, pregnancy tests, electroencephalography, etc.

Considering the state of the art in AI development and having regard to some foresight, one may identify certain categories of AI applications. This taxonomy is necessary also for legal thinking, as the nature and abilities of the AI system shall determine the legal title, factors, and level of liability. The categories can be identified either on the basis of the functionalities of the AI system or according to their abilities.

3.1. AI Typology According to Functionalities

3.1.1. Reactive AI

Reactive AI is programmed to provide a deterministic output based on the input it receives. The applications based on reactive AI 'engines' (software and hardware used to implement an AI model) respond to identical situations in the same way every time, and they are not able to learn actions or conceive of past or future. These types of AI services cannot function beyond the tasks they were initially designed for. That makes them inherently limited and ripe for improvement. As to operable examples, we may consider some well-known applications as follows: Deep Blue – the chess-playing supercomputer by the IBM; spam-filtering utilities embedded into email servers; Google/YouTube/Spotify/Netflix recommendation engines; etc.

3.1.2. Limited-Memory AI

Artificial intelligence with limited memory learns from past experience and builds up empirical knowledge by observing the results of actions or newly generated data. This type of AI uses past observational information combined with pre-programmed information to make predictions and perform complex tasks. Obviously, machine learning capabilities are two steps up from the reactive AI mentioned in the previous point. Today, these systems are also extremely widespread. It may be noted that AI systems controlling autonomous vehicles are also a specific application of this type.

Limited-memory AI, as the name suggests, is still quite limited. The information that autonomous vehicles work with is ephemeral and not stored in the car's long-term memory, for example.

3.1.3. Theory of Mind AI

No industrial examples can be presented for the AI system included in this category – as yet. There are only a few scientific and technological experiments with some rudimentary elements of the decision-making capabilities equal – or uncannily similar – to humans. Machines with this cognitive AI will be able to understand and remember emotions and then adjust their behaviour based on these emotions when interacting with humans. An intelligent conversation with an emotionally intelligent robot that looks and sounds like a real human will be feasible with these machines. There are still many obstacles to the realization of consciousness-based AI, as the process of changing behaviour based on rapidly changing emotions in human communication is very elastic. This is difficult to imitate as we try to create more and more emotionally intelligent machines. Some humanoid robots, such as Sophia developed by Hanson Robotics in Hong Kong, can demonstrate some abilities of social interactions with human users. 'She' can recognize faces and respond to interactions with her own facial expressions.

3.1.4. Self-Aware AI

This category currently exists only in the world of science fiction, and there is no telling when this highly advanced form of artificial intelligence might emerge. At present, we do not have the necessary hardware, nor do we know the algorithms that could make such a machine work. This artificial intelligence is a machine that is self-aware and has its own emotions, not only having the ability to react – more or less – adequately to the actions and emotions of the people connected with it. This type of artificial intelligence, if and when it emerges, will not only be self-aware but will also have desires, needs, and emotions.

3.1.5. Superhuman AI

This category looks even beyond the realms of sci-fi, but in a particular way it is already a matter of scientific discussion. We can only have conjectures and surmises about such an entity and quite obscure premonitions regarding the implications of its emergence. It is predicted that the development of the superhuman AI is physically possible, and no reasons for its implausibility (at least in some distant future) are known. Joseph Carlsmith devoted a hefty paper to this scenario and predicts superhuman AI emerging with some likelihood by 2070.[11] In the same 'prophecy', he concludes that the permanent and unintentional disempowerment of all humans in such a scenario would be an existential catastrophe.[12]

11 Carlsmith 2021. 13.
12 Ibid.

3.2. Typology of AI Based on Capabilities

Another habitual and widely implemented taxonomy of AI is based on the capabilities of these systems. According to this approach, the following types of AI can be identified.

3.2.1. Weak AI, or Narrow AI

The weak (a.k.a. narrow) AI is a type of artificial intelligence that can perform a given task intelligently. Currently, these are the most common and available AI-supported or -operated systems. Weak AI cannot perform beyond its own domain or limitations, as it is only designed and trained for a specific task. Beyond a particular domain, the operation of the weak AI is unreliable and unpredictable.

Consider some operable examples as follows:

– Apple Siri is a narrow AI. Siri operates with a limited pre-defined range of functions.
– IBM Watson online soft-computing facilities also run under narrow AI. Watson's abilities include:
  • the expert system approach (logical structures called 'trees' designed to guide the user to a certain result),
  • machine learning,
  • natural language processing.
– Other narrow AI applications include:
  • chess – and other board game – player programs,
  • purchasing recommendation engines on e-commerce sites,
  • autonomous cars,
  • speech recognition and translation applications,
  • image recognition.

3.2.2. General AI a.k.a. AGI (Artificial General Intelligence)

General AI is a highly developed type of intelligent agent that could solve and carry out any intellectual task with human-like performance. An AGI system would think like a human on its own, likely even far exceeding human cognitive abilities. Currently, no such system exists, but this is a primary target of AI research and development. The timespan of this research effort is unpredictable. AI experts and knowledge engineers mostly agree that the AGI should have the capacities to represent knowledge, reason, develop strategy, decide under uncertainty (able to solve Bayesian problems), plan, learn (machine learning), communicate in natural language, and, finally, to integrate all these above-mentioned skills.

3.2.3. Strong, or Super AI

The strong, or super AI is a hypothetical concept referring to the level of machine intelligence that could surpass human skills and cognitive abilities. This would be an outcome of AGI. Pessimistic – or realistic – forecasts stipulate that this would impose a disastrous future for the mankind.

4. The Legal Risks of AI

After such a – partly futuristic – overview, we can now see what challenges and risks we can actually expect to face today in the context of the use of AI. What are the realistic risks, harms, and damage that AI systems can cause to individuals, groups of individuals, and society as a whole? What types of AI systems do we have any practical experience with?

Only the following categories of AI can be seen as extant systems and services: reactive AI, limited-memory AI, and weak/narrow AI. The other mentioned categories belong to the world of fantasy, and we must clarify that science fiction is not the genuine operational area of law. Legal thinking consequently shall concentrate on the challenges imposed by currently and or foreseeably operational AI systems. These are the lower class of intelligent agents but are also worth considering as sources of legal risks. Without the ambition to make a comprehensive list of legal interests jeopardized by AI systems, we may easily identify some fundamental categories. These are personality rights and property rights.

4.1. Personality Rights in Danger. Profiling and Web Scraping

Data protection law is a significant legal innovation of European legal culture. Within a few decades, this became a forefront of personality rights. National data protection authorities, NGOs, and civilian activists are combating the thirst for information of the modern state and Internet-related companies trading in personal and behavioural data such as social media platforms.

Digital technology – strengthened by AI capabilities – provides the data controllers with sophisticated tools and methods for monitoring and profiling society and private individuals alike. Never before in history has any state benefited from such an effective tool for controlling and manipulating society as the AI implementations we see in daily use even now. The profiling capabilities of social media providers, web stores, and government agencies are mostly based on AI algorithms. Alarming news on massive data breach incidents are regularly broadcast in the media. The Facebook–Cambridge Analytica data scandal illustrates the social and political

risks[13] of massive algorithmic profiling. GDPR and other related laws provide lawful treatment for these abuses and delicts. However, beyond these incidents, there is also the risk that the data processed and profiled by AI algorithms may lead to erroneous conclusions and thus to uninformed decisions. The law must provide an answer as to who and how is liable for the damage caused by such errors and misuse. Web scraping (mass gathering of online information for various purposes) by intelligent software agents is an increasingly widespread practice, also imposing special privacy risks.

4.2. Property Rights in Danger

Intelligent systems are used in several other fields of business – beyond data trading and social media. The modern financial system is also based on digital services. Banks and stock markets use intelligent agents to carry out financial operations. The banking business is regulated and protected by subtle, elaborate legal provisions.

Algorithmic trading – a.k.a. high-frequency trading – in stocks, however, is a relatively new phenomenon, and new challenges are imposed by AI-based algorithms employed in its course. Financial losses in this line of business can erode the livelihoods of families and undermine the prosperity of companies, causing huge damage. One of the most famous stock market incidents, probably caused by artificial intelligence algorithms, is the 2010 Flash Crash[14] on the Chicago Mercantile Exchange (CME). The investigation and interpretation of the causes is still ongoing.

Apart from AI-supported financial systems, one may meet with AI on the roads as well. The risk of autonomous vehicles has grown into a classical dispute topic of ethics and law. The harm caused by an erroneous, disoriented – or unethical – car can be significant and may cause personal harm, injury, or even death.

These are also challenging legal problems. Mainly, the question arises as to who will be held liable in cases like this. This is a new area where ethical considerations need to be taken into account before a legal framework can be established.

5. Ethical and Legal Doctrines on Liability for Damages Caused by AI

A robot is not a person and will not be one for a long time. When the age of self-aware AI, or strong AI, even superhuman AI comes – if ever –, we must reconsider this statement, but now is not the time. Therefore, the type of weak AI currently

13 Chan 2019.

14 Brush–Schoenberg–Ring 2015.

in existence obviously cannot be subject to any legal relationship since it has no legal capacity. Both the ethical and legal requirements for artificial intelligence are therefore imposed on the legally competent persons associated with the intelligent agent, namely: the developer, the service provider, and the user.

5.1. Ethical Framework of AI Liability

Transparency is the first and foremost among the ethical criteria concerning the development and operation of artificial intelligence. This means that it is inevitable that software developers are about to harmonize the algorithms they use, and – despite the fact that these are the most enshrined and confidential secrets of many businesses – they must stop using uncontrolled AI. In the scientific, philosophical, and legal disputes on the demanded framework regulation on AI, many further expectations are on the floor. Most of them are principles so abstract that extensive reasoning and interpretation will be needed to determine their exact meaning. The upper chamber of the British Parliament – the House of Lords – drafted an ethical standard[15] for the AI law of the UK. Five governing principles were laid down as the cornerstones of the forthcoming regulation as follows:

1. Artificial intelligence should be developed for the common good and benefit of humanity.
2. Artificial intelligence should operate on principles of intelligibility and fairness.
3. Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families, or communities.
4. All citizens should have the right to be educated to enable them to flourish mentally, emotionally, and economically alongside artificial intelligence.
5. The autonomous power to hurt, destroy, or deceive human beings should never be vested in artificial intelligence.

Translating moral standards into legal institutions is a non-trivial process with no clear outcome. EU law has replaced the notion of moral AI with the notion of 'trustworthy AI' and has assigned to it criteria that are now legally interpretable.

5.2. Trustworthy AI

Trustworthy AI[16] has three principles, which should be met throughout the system's entire life cycle: (1) lawfulness, (2) displaying of ethical behaviours, and (3) robustness. That is, trustworthy AI should be lawful, complying with all applicable laws and regulations, should be ethical, ensuring adherence to ethical principles and values,

15 House of Lords 2018.

16 European Commission High-Level Expert Group on AI (AI HLEG) 2019. 5.

and should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. These basic principles are transposed into seven further particular requirements to achieve trustworthy AI: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental well-being; (7) accountability.

The articulated concept of trustworthy AI is based on the European doctrine of fundamental rights and a corresponding set of ethical principles that are crucial in an AI context. The Ethics Guidelines developed at the behest of the European Commission emphasized the principles[17] of (1) respect for human autonomy, (2) prevention of harm, (3) fairness, and (4) explicability.

Explicability is probably the most problematic expectation for the AI developers though this is a crucial factor to set up and maintain users' trust in AI systems. The development and training processes of AI should be transparent, the capabilities and purpose of AI systems need to be openly communicated, and decisions need to be explained as far as possible to those directly and indirectly affected. The relevant fundamental rights in relationship with basic ethical principles should guarantee respect for human dignity, individual freedom, rule of law, democracy, equality, solidarity and freedom from discrimination and the fullest scale of citizens' right. AI developers and service providers granting products and services fuelled with AI capabilities must refrain from any practice and technological measure that could breach these fundamental values of law and ethics.

6. Conclusions

The development of AI still looks a long process, and we are just at the beginning of this long road. The legal and ethical issues concerning AI are still in the embryonic stage. The game is not over, and the stakes are very high. AI can be the gold standard of the future or can be a bane for mankind. The legal framework, the regulatory principles shall determine which scenario will be fulfilled. This is why the discussion and collaboration of software developers, knowledge engineers, and legal counsels will be inevitable in the development of artificial intelligence.

17 AI HLEG 2019. 12.

References

BERNHARDT, C. 2016. *Turing's Vision: The Birth of Computer Science*. Cambridge.

BRUSH, S.–SCHOENBERG, T.–RING, S. 2015. How a Mystery Trader with an Algorithm May Have Caused the Flash Crash. *Bloomberg*. https://www.bloomberg.com/news/articles/2015-04-22/mystery-trader-armed-with-algorithms-rewrites-flash-crash-story#xj4y7vzkg.

CARLSMITH, J. 2021. Is Power-Seeking AI an Existential Risk? *Open Philanthropy*. https://docs.google.com/document/d/1smaI1lagHHcrhoi6ohdq3TYIZv0eNWWZMPEy8C8byYg/edit#heading=h.pwdbumje5w8r.

CHAGAL-FEFERKORN, K. A. 2019. Am I an Algorithm or a Product? When Products Liability Should Apply to Algorithmic Decision-Makers. *Stanford Law & Policy Review* 30: 61–114.

CHAN, R. 2015. The Cambridge Analytica Whistleblower Explains How the Firm Used Facebook Data to Sway Elections. *Insider*. https://www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10.

CUSTERS, B.–FOSCH-VILLARONGA, E. 2022. Humanizing Machines: Introduction and Overview. In: *Law and Artificial Intelligence. Regulating AI and Applying AI in Legal Practice*. The Hauge–Berlin.

EUROPEAN COMMISSION. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (Artificial Intelligence Act – AIA) 2021. Brussels, 21.4.2021 COM(2021) 206 final 2021/0106 (COD). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206.

2022. Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) 2022. Brussels, 28.9.2022 COM(2022) 496 final 2022/0303 (COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496.

EUROPEAN COMMISSION HIGH-LEVEL EXPERT GROUP ON AI (AI HLEG). 2019. Ethics Guidelines for Trustworthy AI (EGTAI). Brussels. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

EUROPEAN PARLIAMENT. 2020. Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence (2020/2014(INL)). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0276.

EXECUTIVE OFFICE OF THE PRESIDENT – NATIONAL SCIENCE AND TECHNOLOGY COUNCIL. 2016. *Preparing for the Future of Artificial Intelligence*. Washington. https://obamawhitehouse.archives.gov/sites/default/

files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

HÄUSELMANN, A. 2022. Disciplines of AI: An Overview of Approaches and Techniques. In: *Law and Artificial Intelligence. Regulating AI and Applying AI in Legal Practice*. The Hauge–Berlin.

HINTZE, A. 2016. *Understanding the Four Types of AI, from Reactive Robots to Self-Aware Beings. The Conversation*. https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616.

HOUSE OF LORDS. 2018. *UK Can Lead the Way on Ethical AI, Says Lords Committee*. https://www.parliament.uk/external/committees/lords-select/ai-committee/news/2018/ai-report-published/.

JAVATPOINT.AAI. 2022a. *Application of AI*. https://www.javatpoint.com/application-of-ai.

2022b. *Types of Artificial Intelligence*. https://www.javatpoint.com/types-of-artificial-intelligence.

MARR, B. 2020. *Understanding the 4 Types of Artificial Intelligence (AI)*. https://www.linkedin.com/pulse/understanding-4-types-artificial-intelligence-ai-bernard-marr/.

2021. *What Are the Four Types of AI?* https://bernardmarr.com/what-are-the-four-types-of-ai/.

MASNICK, M. 2014. *No, a 'Supercomputer' Did NOT Pass the Turing Test for the First Time and Everyone Should Know Better*. https://www.techdirt.com/2014/06/09/no-supercomputer-did-not-pass-turing-test-first-time-everyone-should-know-better/.

MCCARTHY, J.–MINSKY, M. L.–ROCHESTER, N.–SHANNON, H. E. 1955. *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

RUSSELL, S.–NORVIG, P. 2022. *Artificial Intelligence: A Modern Approach*. Hoboken (New Jersey, USA).

TAN, J. M. E. 2022. Non-deterministic Artificial Intelligence Systems and the Future of the Law on Unilateral Mistakes in Singapore. *Singapore Academy of Law Journal* 34: 91–124.

TOTSCHNIG, W. 2020. Fully Autonomous AI. *Science and Engineering Ethics* 26: 2473–2485. https://doi.org/10.1007/s11948-020-00243-z.

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 19–34

DOI: 10.47745/AUSLEG.2022.11.2.02

The European Approach to Artificial Intelligence. Ethical and Regulatory Implications[1]

Kitti MEZEI
PhD, Research Fellow
Centre for Social Sciences (Budapest, Hungary),
Institute for Legal Studies
Assistant Professor
Budapest University of Technology and Economics (Hungary),
Faculty of Economics and Social Sciences, Business Law Department
Researcher
University of Public Service (Budapest, Hungary),
Eötvös József Research Centre, Cybersecurity Research Institute
e-mail: mezei.kitti@tk.hu

Anikó TRÄGER
PhD, Junior Research Fellow
Centre for Social Sciences (Budapest, Hungary),
Institute for Legal Studies
Assistant Lecturer
Budapest University of Technology and Economics (Hungary),
Faculty of Economics and Social Sciences, Business Law Department
e-mail: trager.aniko@gtk.bme.hu

Abstract. The development of artificial intelligence (AI) must ensure human-centred and ethical operations, transparency and respect for fundamental rights. In addition to its obvious benefits, AI also entails a number of risks such as opaque decision-making. The aim of this paper is to present and analyse in detail the legal environment for AI in the European Union, with a particular focus on the principles and directives, as well as the current and possible future legal framework, the draft EU AI Act. The article discusses the concept and framework of the EU AI Act on artificial intelligence. A separate chapter reviews the risk-based approach at the heart of the regulation. It provides a

1 The publication was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory and by the Ministry of Innovation and Technology NRDI Office within the framework of the FK_21 Young Researcher Excellence Programme (138965).

detailed analysis of the systems categorzed by risk, their requirements, and the regulatory solutions developed by the draft.

Keywords: artificial intelligence, EU AI Act, trustworthiness, risk-based approach, high-risk AI

1. Introduction

The application of Artificial Intelligence (AI) is expanding into ever more areas of life (e.g. it can improve healthcare, help law enforcement authorities fight crime more effectively, make transport safer, or even help detect fraud and cybersecurity threats, etc.). It is therefore undoubtedly one of the biggest challenges of our time, both from an economic and a regulatory perspective. This is illustrated by the publication by the EU Commission in 2020 of a White Paper on Artificial Intelligence, which forms the basis for specific regulation of AI development and applications at the EU level.[2] It sets out that AI can significantly impact our society, that it is necessary to build trust and confidence in it, and that the AI sector must be based on fundamental rights and values such as human dignity and privacy. Human-centred AI assumes technology that people trust because it aligns with the values underpinning human societies. Non-binding soft law solutions, such as ethics guidelines for AI, play a crucial role in establishing trustworthy AI, assessing risks, and managing the technology in a regulatory context.[3] These are embedded in regulation and can mitigate risks during the legislative process. Ethics by design is an approach to ensure that ethical requirements are appropriately considered in developing an AI system or technique. It aims to address ethical issues at the earliest stages of development rather than as an afterthought. In addition, this trend can have a positive cultural impact, particularly in the technology industry, where market leaders seek to get ahead of regulation rather than being left behind, designing their products and services to comply with legislation still in draft form.[4]

In the general design of AI regulation, four main ethical directions should be highlighted, as set out in the ethical guidelines of the High-Level Expert Group on AI: respect for human autonomy: do not control/manipulate humans, do not compromise democratic processes; prevention of harm: including resistance to unintended external influences that may result in harm; fairness: the development and use of AI systems should be fair; explainability: means transparency of

2 European Commission 2020a.

3 See the criticism of ethical principles in Héder (2020. 57) and Hagendorff (2020. 99).

4 An example of this is the Netherlands, which has already started to apply rules similar to the draft regulation, even though the regulation is not expected to enter into force until a later date. Bertuzzi 2022.

operation[5] (trusted AI systems can be traced and their decisions explained, in particular users should be informed that they have interacted with an AI system and also how the AI system works, what its capabilities are, in what ways and how reliably it uses the datasets provided to it). Other requirements include: human empowerment and human oversight; technical stability and security; data protection and management; diversity, non-discrimination, and equity; social and environmental well-being; accountability.[6]

Technology can be the target of regulation and a tool, even embedded in technology as a command. This encourages developers to address regulation by design at the early stages of development. For example, the White Paper states that AI systems are expected to have built-in safety and security mechanisms to ensure that any operation carried out by the system is demonstrably safe for the physical and mental well-being of the individuals involved. The European Union's regulation[7] points in this direction in several digital regulatory areas (e.g. data protection[8] and algorithmic trading[9]). Traditionally, technology vendors have tested their products ex-post after the risk has materialized. They should have taken measures to correct their processes and compensate for any damages if and when liability was found. This reactive model, which has always struggled to keep pace with technological developments, is becoming obsolete. Instead, legislators are encouraging companies to set up compliance teams around 'product advisors' and to take account of the harm and risks posed by a product at an early stage, to carry out an ethical and regulatory risk assessment. However, such regulation is flexible, requiring standards such as 'appropriate technical and organizational measures' that can be adapted to the company or product/service. Privacy by design and privacy by default are key concepts and are now the bases of

5 For example, in this context, a new draft transparency standard, IEEE P7001, is now available, one of the P70XX series of 'human standards' that are emerging from the IEEE Standards Association's global initiative on the ethics of autonomous and intelligent systems. P7001 aims to create a standard that has 'measurable, testable levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined'. P7001 is also generic in nature; it aims to be applicable to all autonomous systems, including robots (autonomous vehicles, assistive robots, drones, robotic toys, etc.) as well as software-only AI systems such as AI used in medical diagnostics, chatbots, facial recognition systems, etc. P7001 defines five different groups of stakeholders, and AI systems must be transparent to each group in different ways and for different reasons. Winfield et al. 2021.

6 European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG) 2019. 14.

7 See Hanani 2022. 137; Codagnone et al. 2022; Mökander et al. 2022. For a comparison with Chinese AI regulation, see also Roberts et al. 2021. 3659–3677.

8 With privacy by design, privacy safeguards must be built into products and services from the earliest stages of development. In other words, companies need to think about security measures at the design stage of data management processes before they start processing data. For example, pseudonymization, or encryption of personal data, is one way of ensuring compliance with *built-in* data protection.

9 Directive 2014/65/EU requires Member States to ensure that algorithmic trading systems do not create or contribute to disorderly trading conditions in the market and to address disorderly trading conditions that such algorithmic trading systems do create.

digital regulation. Engineers and developers need to address legal and regulatory requirements from the very beginning of the design of their digital products.[10]

Furthermore, in 2020, the European Parliament issued a report to the Commission with recommendations on the civil liability regime for AI.[11] In response, in September 2022, the Commission took the initiative to modernize the rules on the objective liability of manufacturers for defective products (from smart technology to pharmaceuticals).[12] The revised rules aim to create legal certainty for businesses, making investing in new and innovative products easier. They will ensure fair compensation in the event of damage caused by a defective product, including a digital or refurbished product. On the other hand, the Commission has proposed a targeted harmonization of national liability rules for AI for the first time.[13] A single set of rules would make it easier for victims of damage caused by AI to get compensation.[14]

The most important step forward in the comprehensive regulation of AI is the publication in April 2021 of the Commission's proposal for a draft Artificial Intelligence Act (hereinafter as AI Act), which contains important restrictions on AI systems used in or in connection with the EU.[15] The use of AI with specific characteristics, such as opacity due to the black box effect, complexity, dependence on data, and autonomous behaviour, may adversely affect several fundamental rights enshrined in the Charter of Fundamental Rights of the European Union. Because of these characteristics, both public authorities and the individuals concerned may lack adequate means to verify how a given algorithmic decision was made and whether the relevant rules have been respected. Therefore, the proposal aims to ensure a high level of protection of these fundamental rights and to address the different sources of risk through a clearly defined risk-based approach. This paper analyses the AI Act in detail.

2. The European Union's Draft AI Act

2.1. Scope of the AI Act and Definition of the AI System

The draft EU AI Act aims to implement a minimum set of horizontal rules applicable to all AI systems placed on the market or used in the EU. The new regulation would apply primarily to service providers established in the EU or third countries placing

10 82 Clarke 2022.
11 European Parliament 2020.
12 European Commission 2022a.
13 European Commission 2022b.
14 European Commission 2022c.
15 It should be noted that the AI Act should be read in conjunction with other major legislative packages, such as the Digital Services Regulation (DSA), the Digital Markets Regulation (DMA), and the Digital Governance Regulation (DGA), the first two of which primarily regulate large commercial online platform providers such as Google, Amazon, Facebook, and Apple (GAFA).

AI systems on the EU market or installing them in the EU and to users of AI systems located in the EU.[16] To prevent circumvention of the regulation, the new rules would also apply to providers and users of AI systems located in third countries if the output produced by these systems is used in the EU. However, the draft regulation would not apply to AI-based systems developed or used exclusively for military purposes, to authorities in third countries, international organizations, or to authorities using AI-based systems in the framework of international agreements on law enforcement and judicial cooperation. Another exemption has been added for people using AI for non-professional purposes, which would fall outside the scope of the AI regulation, except for the transparency obligations.[17]

The Commission proposes a technology-neutral definition of AI in Article 3(1) of the draft, which states that an AI system is an 'artificial intelligence system' meaning 'software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with'. Accordingly, the term AI system would refer to software-based technologies that include machine learning, logic and knowledge-based systems, and statistical approaches. This broad definition includes AI systems that can be used independently or as part of a product. An AI system can be designed to operate with varying degrees of autonomy. It can be used standalone or as part of a product, whether the system is physically integrated into the product (embedded) or serves the functions of the product without being integrated (non-embedded). The AI Act aims to be future-proof and cover current and future AI technology developments. To this end, the Commission would – using delegated acts (Article 4) – add new approaches and techniques for AI regulation to the list in Annex I as needed. Furthermore, Article 3 contains a long list of definitions, including the concepts of 'provider' and 'user' of AI systems, covering both public and private entities, as well as 'importer', 'distributor' and 'operator', 'sentiment recognition', and 'biometric categorization'.

2.2. The Risk-Based Approach

The use of AI, with its specific characteristics, can adversely affect several fundamental rights and the security of users. To address this, the AI Act adopts a risk-based approach, whereby AI applications are classified into risk classes, and

16 See Article 2. The AI Act would also apply to EU institutions, offices, bodies, and agencies acting as providers or users of AI systems.

17 Some members of the Council and the European Parliament would extend this by excluding from the scope of the regulation AI systems where national security issues are at stake. This would allow (autocratic) governments to use biometric mass surveillance or social scoring in the name and under the guise of 'national security' even if these are prohibited by the regulation. Bertuzzi 2022.

legal action is tailored to the specific risk level.[18] To this end, a distinction is made between unacceptable, high-risk, moderate-risk, and minimal-risk AI systems. Under this approach, AI applications would be regulated only to the extent strictly necessary to address specific risk levels.

2.2.1. AI Systems Falling into the Prohibited Category

With this in mind, the AI Act distinguishes a completely prohibited category (Title II), which includes the prohibition of facial recognition[19] (with exceptions)[20] in public places, subliminal manipulation, mass surveillance, or the unlawfulness of the social scoring system[21] (similar to the one used in China). All AI systems that clearly threaten people's safety, livelihoods, and rights are banned, from social scoring by governments to voice assistant games that encourage dangerous behaviour. Of these, subliminal manipulation has been the most criticized because the draft does not provide a precise definition of what should be understood by this or what cases might fall into this category. According to the literature, it generally refers to sensory stimuli that consumers cannot consciously perceive; for example, visual stimuli that last less than 50 milliseconds. However, most applications of AI will not be subliminal, as users will perceive it consciously. Thus, the AI Act in its current form still allows for many forms of AI-based manipulation.[22]

2.2.2. Moderate-Risk AI Systems

In addition, it identifies high-risk AI applications (Title III), for which it establishes binding rules, and other applications that are less risky (Title IV) but still deserve some attention, and it addresses the risks associated with these applications by supporting them with provisions to enhance transparency. These rules are contained in Article 52, which requires the AI to inform the person at all times that s/he is facing an AI. Systems capable of detecting emotions must inform the persons concerned, deepfake videos must be labelled, and it must be known that they are machine-forged moving images. These categories are neither prohibited nor high-risk in themselves. Interestingly, the AI Act classifies tools used by law

18 For more on this, see Mahler 2021.

19 For more on the regulation of facial recognition programmes in the EU, see Madiega–Mildebrath 2021.

20 The use of AI systems for the 'real-time' remote biometric identification of natural persons in publicly accessible locations for law enforcement purposes necessarily involves the processing of biometric data. The rules of the AI Act, based on Article 16 TFEU, which prohibit such use, subject to certain exceptions, should be applied as *lex specialis* to the rules on the processing of biometric data contained in Article 10 of Directive (EU) 2016/680, and therefore exhaustively regulate such use and the processing of the biometric data concerned.

21 See AlgorithmWatch 2022.

22 See Franklin et al. 2022. 35; Vergnolle 2021; Hacker 2021.

enforcement agencies to detect deepfake as high-risk, while deepfake content in general falls into the low-risk category. This is a curious discrepancy, which appears to be based on the assumption that deepfake technologies (which are mainly used in the private sector for the time being) constitute a lower risk than deepfake detection AI systems in the hands of state actors. However, under the AI Act, this labelling obligation does not apply to law enforcement. This means that when some law enforcement authorities use deepfake, they do not have to label it as such [Article 52(3)].[23] Biometric categorization systems – systems that biometrically group individuals according to categories such as 'gender, age, hair colour, eye colour, tattoos, ethnic origin, sexual or political orientation, based on their biometric data' – or emotion recognition systems, which are used in the context of Article 3(34), are not prohibited and are not included in the list of high-risk AI systems. Consequently, they fall into the category of AI systems of limited risk and are therefore covered by the provisions of Article 52(2) for both public and private actors, with the exception of law enforcement authorities. Finally, the draft leaves AI applications not falling into either category to regulation by codes of conduct, i.e. self-regulation. So, the AI Act does not contain specific rules for the use of AI, which is neither prohibited nor high-risk in itself, beyond the basic requirements, but would refer it to the so-called codes of conduct.[24] The use of codes of conduct is not new among the European Union's regulatory solutions. It is currently used, or more precisely required, in the field of media regulation, but the Digital Service Act[25] will also further strengthen its role in media regulation. In the case of the already cited DSA, the Regulation already makes it clear in the recitals[26] that the application of codes of conduct is to be made in the context of self- and co-regulation, similar to the current media regulation.

Although the self-regulation and co-regulation models have very similar elements, it is necessary to briefly distinguish between the two regulatory models. Self-regulation in principle does not have a mandatory nature. The actors in the field or some joint organization of such actors develop some professional-ethical standards. Those who consider it essential will voluntarily join this agreement and accept these standards as binding on themselves, possibly jointly handling complaints about them. Although there is no classical binding force of the state behind the regulation, if the participants mean what they say in their code of conduct, it can

23 Georgieva et al. 2022. 14.

24 Article 69(1) of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts (AI Act). See European Commission 2021.

25 European Commission 2020b.

26 Recital (68) of Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. European Commission 2020b.

significantly help achieve a high level of compliance with professional, ethical principles.[27]

Co-regulation generally builds on the self-regulation of organizations described above. However, in this case, the state is already involved in the regulation. The responsibility for compliance with and enforcement of the principles of the codes of conduct is shared between the state and the professional organizations involved in the regulation.[28] The classic legislative model is that the state (or, in this case, the European Union) creates a binding rule, and then the public authority enforces it. The stakeholders are only involved in the preparation and impact assessment of the legislation. By contrast, the central actor in the process of co-regulation, and thus in the drafting and applying of codes of conduct, remains the one to whom their content will otherwise be binding. Self- and co-regulation also provide a more flexible regulatory mechanism that considers regional specificities.

Going back to the text of the AI Act, its recitals[29] and the provisions of Article 69 only provide for the encouragement and support of the adoption of codes of conduct in general. This leads to the conclusion that the aim of EU legislation on regulating AI is essentially to promote self-regulation and that there is currently no thought of developing a co-regulatory model.

The AI Act provides in Article 69(1) that:

> The Commission and the Member States shall encourage and facilitate the development of codes of conduct aimed at promoting the voluntary application of the requirements set out in Chapter 2 of Title III to AI systems other than high-risk AI systems based on technical specifications and solutions which, in the light of the intended use of the systems, constitute an appropriate means of ensuring compliance with those requirements.

The quoted text of the regulation refers to the mandatory requirements for using high-risk AI. On this basis, the draft would aim to ensure that they are also applied to the highest possible degree in the case of lower-risk AI.

In this context, the Commission and the Council would therefore encourage AI actors and their organizations to adopt and implement codes of conduct setting out requirements in the areas of the risk management system, data, data governance, technical documentation, record keeping, transparency and provision of information to users, human oversight, accuracy, robustness and cybersecurity, in line with the AI Act.[30]

27 Examples of how self-regulation works can be found in the field of media regulation. On this, see Tófalvy 2013. 85–87.

28 Hegedűs 2015, Csink–Mayer 2012.

29 Recital (81) of the AI Act.

30 Title III, Chapter 2 of the AI Act.

In addition to the general support for self-regulation, the draft regulation also sets out, by way of example, a list of areas where it considers particularly important for non-high-risk schemes to develop a code of conduct in which they accept to be bound by more stringent provisions. Examples of such areas include the promotion of the voluntary application of requirements relating to environmental sustainability, accessibility for persons with disabilities, stakeholder participation in the design and development of AI schemes, and diversity of development teams to AI schemes based on clear objectives and key performance indicators to measure the achievement of those objectives.[31] As to who is entitled to adopt codes of conduct, the AI Act – similarly to the current models of media regulation – designates the regulated parties themselves, i.e. the individual providers of the AI systems or the organizations representing them, or both, included through the involvement of users and stakeholders and their representative organizations.[32] The draft regulation also states that 'codes of conduct may cover one or more AI systems, taking into account the similarity of the intended purpose of the relevant systems.'

Regarding codes of conduct, the draft also briefly states that the Commission and the Council will consider the specific interests and needs of small service providers and start-ups in encouraging and facilitating their development.[33]

The motivation behind the EU's move towards self-regulation is partly the time factor: AI, like the media, is a fast-moving field with many different areas. Classical legislative instruments are slow at the Member State level, but even more so in the EU. We should think here of the AI Act itself, which has been years in the making and is still only a draft. Furthermore, if market players can be involved in developing regulation based on their existing self-regulation and ethical principles, this will allow for significantly faster adaptation. Developers and professional organizations involved in self-regulation have the expertise and knowledge to develop and, where appropriate, monitor the principles. Greater acceptance and cooperation can be expected if they are involved in regulation. A further advantage could be that if AI developers move to codes of conduct under the Regulation, this could lead to more effective, detailed regulation than the current codes of ethics, which may or may not have any substance to them.[34]

In addition to the expected positive aspects, it is also necessary to briefly discuss the disadvantages of self-regulation. One of the main disadvantages of this model is that participation in it and adherence to codes of conduct is entirely voluntary. As a result, it ultimately lacks the classic binding force and enforceability. In line with this, it is also apparent that the AI Act – applying the risk-based approach here, too – has not opted for this regulatory model for higher-risk schemes but

31 Article 69(2) of the AI Act.
32 Article 69(3) of the AI Act.
33 Article 69(4) of the AI Act.
34 For more details on the criticisms of these, see: Ződi 2020; Larsson 2020. 437–451.

for 'classic' mandatory regulation, with the threat of heavy fines in the event of non-compliance.[35]

The question remains, then, how interested will AI developers – who are not required to apply the stricter rules of the Regulation – be to even adopt the much more stringent rules that are mandatory for high-risk AI. In the light of this development, it is also questionable whether the European Union will leave this area to self-regulation at all or whether it will move towards co-regulation, as in the case of the media, or whether it will return to the classic centralized regulatory solution but with less stringent rules for lower-risk areas.

2.2.3. High-Risk AI Systems

Proposed rules for high-risk AI are of interest because most of the provisions of the new regulation are built around this risk category. An AI system is considered high-risk either because it is a security component of an already tightly regulated product group (listed in Annex II, from toys through craft to medical instruments) or because it is used in an area where human rights are particularly at risk. The latter list includes two dozen specific applications in eight areas such as AIs for biometric identification of natural persons, AIs for the control of critical infrastructures (transport, gas, water, electricity), and some other AIs 'active' in various areas (such as recruitment, university admissions, credit assessment, and advice to judges). Indeed, the AI Act states that AI systems used in employment, management of workers and access to self-employment, in particular recruitment and selection of persons, decisions on promotion and dismissal, and the allocation of tasks to persons with a contractual relationship to work, as well as the monitoring or evaluation of such persons, should be considered as high-risk, as they may have a significant impact on the future career prospects and livelihood of these persons. A significant power imbalance characterizes law enforcement authorities' actions involving specific AI systems. They may lead to the surveillance, arrest, or privation of liberty of a natural person and other adverse effects on fundamental rights. In particular, if an AI system is not trained with good-quality data, does not meet adequate standards of accuracy or stability, or is not properly designed and tested before being placed on the market or otherwise put into service, it may select people in a discriminatory or otherwise unfair or unjust manner. It may also hinder the enforcement of important fundamental procedural rights such as the right to an effective remedy and a fair trial, as well as the rights to defence and the presumption of innocence, in particular, if such AI systems are not sufficiently transparent, explained, and documented. The AI systems used in migration management, asylum, and border management[36] affect people who are often in

35 Article 71 of the AI Act.

36 See Dumbrava 2021.

a particularly vulnerable situation and whose lives are affected by the outcome of the actions of the competent authorities. The accuracy, non-discriminatory nature, and transparency of the AI systems used in this context are therefore of particular importance in ensuring respect for the fundamental rights of the persons concerned, namely their rights to free movement, non-discrimination, privacy and protection of personal data, international protection, and due process. In previous compromises, the EU Council already moved towards curbing significant leeway for law enforcement. The new text extends the exemption to the four-eye principle, which requires at least two persons to verify a decision of a high-risk system. Moreover, public authorities using high-risk systems in law enforcement, migration, asylum and border control, and critical infrastructure have been exempted from registering on the EU database.[37]

Specific AI systems designed to administer justice and democratic processes should be considered high-risk, considering their significant impact on democracy, the rule of law, individual freedoms, and the right to an effective remedy and a fair trial. In particular, to address the risk of possible distortions, errors, and opacity, AI systems that aim to assist judicial authorities in researching and interpreting factual and legal elements and in applying the law to specific facts should be considered high-risk. However, this classification should not cover AI systems intended for purely ancillary administrative activities that do not affect the actual administration of justice in individual cases such as anonymization or pseudonymization of court decisions, documents or data, staff communications, etc., administrative tasks, or the allocation of resources. For the use of high-risk systems in this area, Member States might decide to appoint police forces or judicial authorities as market surveillance authorities. The text now specifies that such market surveillance activities should not affect the independence of the courts. Systems for pollution control have been removed from the list of high-risk use cases, while systems to calculate risks and pricing for insurance have been added, except if the provider is an SME.

The requirements for high-risk AIs in the regulation (Chapter 2) provide that risk assessment systems must always be established, implemented, documented, and maintained (Article 9). They must be operated in conjunction with appropriate data governance systems, and the data used for teaching the AI, validation, and testing must be 'clean' (Article 10). High-risk AIs must be accompanied by detailed documentation, and event logging systems must be associated (articles 11–12). Systems of this type must operate transparently and always retain human oversight and intervention (articles 13–14). They must also meet the requirements of accuracy, robustness, and cybersecurity (Article 15). Most of these requirements must be incorporated into the design of the high-risk AI system. In addition to the technical documentation to be prepared by the service provider, the other requirements

37 Bertuzzi 2022.

should be taken into account at the earliest stages of the design and development of the AIs. A new transparency obligation has been added, requesting the providers of systems susceptible of causing significant harm to include the expected output in the instructions for use when appropriate. For the quality management systems that high-risk AI providers will have to implement, new wording was introduced to align them with similar systems mandated under sectorial legislation.

Finally, it is worth addressing one of the fundamental rights most at risk when using AI systems, especially in the case of algorithmic decision-making: the right to equal opportunities and non-discrimination. The main cause of this is the incompleteness or error of the dataset used by the AI or used to train the AI or the inherent bias in the system. The bias in algorithmic decision-making that the problems mentioned above in the dataset may cause can lead to infringement without any intentionality or human awareness behind it. AI in decision-making can also produce discriminatory results if the system learns from biased training data and the AI Act imposes strict training data requirements.[38] Comprehensive and well-chosen teaching data (the examples used to train the AI) are key here. The role of the code producer also changes from being responsible for the programming (its being error-free) to be primarily responsible for the quality of the data and the correct choice of examples (see Article 10).

3. Concluding Thoughts

The AI Act is forward-looking, detailing the general requirements for high-risk AI systems (the so-called 'essential requirements'). In contrast, the detailed technical requirements will be defined mainly by European standards developed in the framework of European standardization. Although detailed technical standards have already played an important role in Chapter 5 of Title III, they are still largely missing. Their development will be crucial for the effective implementation and enforcement of the proposed AI Act. This observation can be made more generally concerning the implementation of the conformity assessment mechanism of the proposal. Conformity assessment of AI systems will be carried out according to technical rules defined entirely by notified bodies, i.e. private bodies that are supposed to be remunerated for their activities. Therefore, it is of the utmost importance to ensure that national authorities are given as much power as possible to democratically control how these organizations carry out their activities and how they implement the proposal's standards in concrete terms.

The mandatory requirements for high-risk AI systems are broadly based on the 'requirements for trustworthy AI' listed in the ethical guidelines of the High-Level Expert Group on Artificial Intelligence. They must be met before a system

38 Zuiderveen Borgesius 2018. 6.

can be placed on the market or put into service. These relate to data quality and management, documentation and record keeping, transparency and user information, human supervision, robustness, accuracy, and security. Introducing such mandatory requirements is a significant step forward in protecting against the harmful effects of AI systems. However, the proposal still needs to be significantly revised in terms of how high-risk systems are defined, and the requirements, which are currently based on a list, and the provisions are prescriptive.

By granting the notified body the right to have full access to teaching, validation, and testing data and to request access to source codes, the draft creates a tension between the need to regulate the activities of organizations responsible for the development of high-risk systems and the protection of the intellectual property of these organizations, in line with the freedom to conduct a business and the right to the protection of intellectual property, both of which are protected by the Charter of Fundamental Rights of the European Union. It is necessary to ensure that the know-how of undertakings is adequately protected, with appropriate confidentiality requirements, and that access requests are targeted and proportionate to the specific task.

A possible criticism is that it is difficult to predict the future use of AI systems and that it is too early to establish a definitive list of prohibited AI practices. The prohibition of subliminal manipulation under the AI Act provides a low level of protection. It only applies to a limited range of abuses and remains open to other non-subliminal but manipulative AI practices.[39]

References

ALGORITHM WATCH 2019. *Personal Scoring in the EU: Not Quite Black Mirror Yet, at Least if You're Rich*. https://bit.ly/3MAQBvM (accessed on: 14.11.2022).

BERTUZZI, L. 2022a. *EU Council Nears Common Position on AI Act in Semi-final Text*. https://bit.ly/3Hxyd6t (accessed on: 14.11.2022).

2022b. Once Bitten, Netherlands Wants to Move Early on Algorithm Supervision. *Euroactiv*. https://www.euractiv.com/section/digital/news/once-bitten-netherlands-wants-to-move-early-on-algorithm-supervision/ (accessed on: 19.11.2022).

CLARKE, O. 2022. *Legislators Worldwide Move to Adopt Regulation by Design*. https://bit.ly/3Tm4H6f (accessed on: 10.10.2022).

CODAGNONE, C. et al. 2022. *Identification and Assessment of Existing and Draft EU Legislation in the Digital Field. Study for the Special Committee on Artificial Intelligence in a Digital Age (AIDA)*. Luxembourg.

39 For a detailed analysis of the Artificial Intelligence Act and suggestions for amendments, see: Smuha et al. 2021; Ebers et al. 2021. 589.

CSINK, L.–MAYER, A. 2012. *Variációk a szabályozásra* [Variations on Regulation]. Budapest.

EBERS, M. et al. 2021. The European Commission's Proposal for an Artificial Intelligence Act – A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *Multidisciplinary Scientific Journal* 4. https://www.mdpi.com/2571-8800/4/4/43 (accessed on: 19.11.2022).

EUROPEAN COMMISSION. 2020a. *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (accessed on: 19.11.2022).

2020b. *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC COM/2020/825 Final*. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:825:FIN (accessed on: 19.11.2022).

2021. *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 Final*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 (accessed on: 19.11.2022).

2022a. *Proposal for a Revision of the Product Liability Directive. Brussels, 28.9.2022 COM(2022) 495 Final 2022/0302 (COD).* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0495 (accessed on: 19.11.2022).

2022b. *Proposal for a Directive Adapting the Rules on Non-contractual Civil Liability to Artificial Intelligence. Brussels, 28.9.2022 COM(2022) 496 Final 2022/0303 (COD)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496 (accessed on: 19.11.2022).

2022c. *New Liability Rules for Products and AI to Protect Consumers and Promote Innovation*. https://ec.europa.eu/commission/presscorner/detail/en/IP_22_5807 (accessed on: 19.11.2022).

EUROPEAN COMMISSION HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (AI HLEG) 2019. *Ethics Guidelines for Trustworthy AI*. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf (accessed on: 19.11.2022).

EUROPEAN PARLIAMENT. 2020. *Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence (2020/2014(INL))*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0276 (accessed on: 19.11.2022).

FRANKLIN, M. et al. 2022. Missing Mechanisms of Manipulation in the EU AI Act. *FLAIRS*: https://journals.flvc.org/FLAIRS/article/view/130723 (accessed on: 19.11.2022).

FUTURE OF LIFE INSTITUTE.

2022. *Manipulation and the AI Act.* https://futureoflife.org/wp-content/uploads/2022/01/FLI-Manipulation_AI_Act.pdf (accessed on: 19.11.2022).

GEORGIEVA, I. et al. 2022. *Regulatory Divergences in the Draft AI Act – Differences in Public and Private Sector Obligations*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2022)729507 (accessed on: 19.11.2022).

HACKER, P. 2021. Manipulation by Algorithms. Exploring the Triangle of Unfair Commercial Practice, Data Protection and Privacy Law. *European Law Journal* (forthcoming): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3835259 (accessed on: 19.11.2022).

HAGENDORFF, T. 2020. The Ethics of AI Ethics. An Evaluation of Guidelines. *Minds and Machines* 30. https://link.springer.com/article/10.1007/s11023-020-09517-8 (accessed on: 19.11.2022).

HANANI, R. J. 2022. The Politics of Artificial Intelligence Regulation and Governance Reform in the European Union. *Policy Sciences* 55. https://link.springer.com/article/10.1007/s11077-022-09452-8 (accessed on: 19.11.2022).

HÉDER, M. 2020. A Criticism of AI Ethics Guidelines. *Információs Társadalom* 4. https://inftars.infonia.hu/pub/inftars.XX.2020.4.5.pdf (accessed on: 19.11.2022).

HEGEDŰS, L. 2015. *Az ön- és társszabályozás vizsgálata egyes európai államok médiaigazgatásában* [Examining Co-regulation in the Light of Media Governance in Hungary and Abroad]. https://blszk.sze.hu/images/Dokumentumok/diskurzus/2014/2/heged%C5%B1s.pdf (accessed on: 19.11.2022).

LARSSON, S. 2020. On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society* 7. https://doi.org/10.1017/als.2020.19 (accessed on: 19.11.2022).

MADIEGA, T.–MILDEBRATH, H. 2021. *Regulating Facial Recognition in the EU*. https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/698021/EPRS_IDA(2021)698021_EN.pdf (accessed on: 19.11.2022).

MAHLER, T. 2021. Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal. *Nordic Yearbook of Law and Informatics*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4001444 (accessed on: 19.11.2022).

MÖKANDER, J. et al. 2022. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What Can They Learn from Each Other? *Minds and Machines* 32. https://link.springer.com/article/10.1007/s11023-022-09612-y (accessed on: 19.11.2022).

ROBERTS, H. et al. 2021. The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation. *AI & Society* 36. https://link.springer.com/article/10.1007/s00146-020-00992-2 (accessed on: 19.11.2022).

SMUHA, N. et al. 2021. *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence*

Act. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991 (accessed on: 19.11.2022).

TÓFALVY, T. 2013. Média a törvényen túl? [Media beyond the Law?]. *Médiaelmélet*. https://www.mediakutato.hu/cikk/2013_04_tel/06_media_onszabalyozas.pdf (accessed on: 19.11.2022).

VERGNOLLE, S. 2021. Identifying Harm in Manipulative Artificial Intelligence Practices. *Internet Policy Review*. https://policyreview.info/articles/news/identifying-harm-manipulative-artificial-intelligence-practices/1608 (accessed on: 19.11.2022).

WINFIELD, A. F. T. et al. 2021. A Proposed Standard on Transparency. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2021.665729 (accessed on: 19.11.2022).

ZŐDI, Zs. 2020. *On the Futility of Codes of Ethics in Regulating Artificial Intelligence*. https://bit.ly/3W6ynGc (accessed on: 01.12.2022).

ZUIDERVEEN BORGESIUS, F. Z. 2018 *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (accessed on: 01.12.2022).

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 35–64

DOI: 10.47745/AUSLEG.2022.11.2.03

Mitigating the Privacy Risks of AI through Privacy-Enhancing Technologies[1]

Barnabás SZÉKELY
LL.M, Certified Data Protection Officer
PrivacyPro Ltd. (Cluj-Napoca, Romania)
e-mail: szekelybarnabas@gmail.com

Abstract. The development and operation of an AI solution generally requires large amounts of data. This may involve processing of personal data, which implies privacy risks for the data subjects and the obligation to comply with data protection rules for data controllers. Privacy-enhancing technologies (PETs) can help enhance data collection and mitigate privacy risks posed by the development of AI solutions. In this context, this thesis proposes to present a set of emerging technologies that address privacy risks characteristic to machine learning models and enable privacy-preserving machine learning. The essay will highlight three state-of-the-art PET solutions: homomorphic encryption, secure multi-party computation, and differential privacy.

Keywords: artificial intelligence, data protection, privacy, European Union, differential privacy

1. Introduction

Artificial intelligence (AI) is becoming a key element for digital transformation, shaping the future of humanity in almost every industry and evolving at an accelerating pace. According to *The One Hundred Year Study on Artificial Intelligence* 2021 Study Panel Report led by Stanford University, the field of AI has made remarkable progress in almost all its standard sub-areas between 2016 and 2021. This includes vision, speech recognition, natural language processing, image and video generation, multi-agent systems, planning, decision-making, and integration of vision and motor control for robotics.[2] The speed of development can be linked to the recent advances in computing power and the increasing availability

1 The following study constitutes the first publication, in abridged form, of the author's LL.M dissertation submitted in the course of the Master's Programme in Artificial Intelligence for Public Services of the Madrid Polytechnic University (2022).

2 Stanford University – Littman–Ajunwa–Berger–Boutilier–Currie–Doshi-Velez–Hadfield–Horowitz–Isbell–Kitano–Levy–Lyons–Mitchell–Shah–Sloman–Vallor–Walsh 2021.

of vast swathes of data, also boosted by the evolution of AI investments. In 2019, investment in AI in the European Union (EU) grew by 64%, then by 37% in 2020, and the overall level of AI investments was estimated to reach €10.7 billion. If this trend is maintained, the EU will exceed its annual AI investment target of €22 billion by 2030. In the United States, the growth was 55% in 2019 and 50% in 2020, reaching €21.2 billion. On the contrary, in the United Kingdom, investment in AI grew at a higher rate in 2020 (46%) than in 2019 (40%).[3]

Through new products and services, AI is increasingly present in our daily lives. Besides the innovation, opportunities, and potential value to society, AI systems also pose a potential risk to the fundamental rights, health, and safety of citizens. Discrimination, privacy and data protection harms (for example, loss of confidentiality), lack of transparency, explainability and accountability became intensely discussed and debated issues of AI systems. As the High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission in June 2018 emphasizes in its Ethics Guideline on Trustworthy AI, privacy is a fundamental right particularly affected by AI systems.[4] The privacy implications of AI depend to a large extent on the specific use cases, the sensitivity of the training data, the social groups the system is deployed on, the overlapping legal requirements,[5] and social, cultural, and political aspects. In the second chapter of this thesis, titled *Artificial Intelligence vs. Privacy*, the main privacy and data protection risks raised by AI systems will be explored.

The development and use of AI systems often involve the processing[6] of personal data.[7] The General Data Protection Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR) is built in a technologically neutral manner and does not refer specifically to AI. In order to be able to face any technological evolution, GDPR regulates the processing of personal data regardless of the technology used,

3 European Commission Joint Research Centre – Evas–Sipinen–Ulbrich et al. 2022.

4 European Commission Directorate-General for Communications Networks, Content and Technology 2019.

5 The European Data Protection Board and the European Data Protection Supervisor (EDPS) have raised their concern that 'the (AI Act) Proposal is missing a clear relation to the data protection law as well as other EU and Member States law applicable to each "area" of high-risk AI system' listed in the Annex III of the Regulation. Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) 2021.

6 Article 4(2) GDPR: any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

7 Article 4(1) GDPR: any information relating to an identified or identifiable natural person (data subject). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person.

as the European Data Protection Board (EDPB) highlighted in a response to Sophie in't Veld's (Member of the European Parliament) letter on unfair algorithms. Also, the EDPB commented that any processing of personal data through an algorithm falls within the scope of the GDPR.[8] In conclusion, whenever the processing of personal data performed by an AI system falls within the territorial scope[9] of the GDPR, all provisions of the Regulation will apply to such processing. In the second chapter, the main challenges posed by GDPR requirements will be presented.

The GDPR, similarly to the Artificial Intelligence Act (AI Act),[10] applies a preventive risk-based approach. The basis of this approach is the 'data protection by design and by default principle' (DPbDD).[11] Data protection by design[12] requires the implementation of appropriate organizational and technological measures prior to and during the whole lifecycle of data processing activities. This ensures that privacy and data protection risks are identified and addressed in the early stages of the AI system's lifecycle, also that the data protection principles[13] and necessary safeguards are embedded in the AI system's entire lifecycle. Implementing these principles at a systemic level and ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed requires new technical approaches. The technologies that are designed to support the implementation of data protection principles are covered by the term Privacy-Enhancing Technologies (PETs). According to Borking and Raab, PETs 'are a coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the data system'.[14] The EDPB underlines that 'PETs that have reached the state-of-the-art maturity can be employed as a measure in accordance with the DPbDD requirements if appropriate in a risk-based approach' but in themselves do not necessarily satisfy the obligations under GDPR Art. 25 on data protection by design.[15] In the third chapter, an overview of emerging PETs that address the most common data security risks and challenges posed by big data and AI developments will be provided.

8 EDPB 2020b.

9 Article 3 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) 2016.

10 Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021.

11 Article 25 GDPR.

12 The concept was developed in the 1990s, brought forward by Ann Cavoukian, former Information and Privacy Commissioner of Ontario. At that time, the term Privacy by Design (PbD) was used.

13 Article 5 GDPR.

14 Borking–Raab 2001.

15 EDPB 2020a.

PETs are a promising set of techniques that can support privacy-preserving machine learning (PPML), facilitating the use of a powerful form of data analysis.[16] By 2025, 60% of large organizations will use one or more privacy-enhancing techniques in analytics, business intelligence, or cloud computing – as Gartner predicts.[17] In the third chapter of this thesis, three promising areas of PETs will be analysed in depth: homomorphic encryption, secure multi-party computation, and differential privacy. Also, real-world use cases will be discussed to demonstrate how these PETs contribute to privacy-preserving machine learning.

2. Artificial Intelligence vs. Privacy

As the penetration of AI is increasing, a growing number of sectors are transformed. Besides the benefits of AI, specific privacy and data protection risks arise in the case of AI systems that process large datasets of personal data or combine non-personal data that can lead to the re-identification of individuals.

2.1. Privacy and Data Protection Risks

Depending on the particular context, varying likelihood and severity of the risks, personal data processing could lead to physical, material, or non-material damage.[18] In recital 75, GDPR addresses among the potential consequences a broad range of harms and emphasizes that the processing of personal data may give rise to 'discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorized reversal of pseudonymization, or any other significant economic or social disadvantage'. The following subchapter will provide an overview of privacy risks in relation to compliance with data protection principles.

2.2. AI Meets the Data Protection Principles

The fundamental building blocks of the GDPR are the seven key data protection principles:[19]

a. Lawfulness, fairness, and transparency;
b. Purpose limitation;
c. Data minimization;
d. Accuracy;

16 The Royal Society 2019.
17 Gartner 2021.
18 Recital 75 GDPR.
19 Article 5 GDPR.

e. Storage limitation;
f. Integrity and confidentiality (security);
g. Accountability.

Regardless of the purposes and means of personal data processing, compliance with these principles is an essential requirement. Failure to comply with the principles at the heart of the GDPR can result in heavy fines. Infringements on the principles are subject to the highest tier of administrative fines, meaning financial penalties of up to €20 million or 4% of the total worldwide annual turnover, whichever is higher.[20] The principles that have particular relevance to AI systems are discussed in detail below.

2.2.1. Fairness

According to the EDPB, 'fairness is an overarching principle which requires that personal data shall not be processed in a way that is detrimental, discriminatory, unexpected or misleading to the data subject'.[21] Fair data processing presumes that data have not been collected or processed through unfair means, without the data subject's knowledge, or by misleading or deception of data subjects. Also, fairness implies that data is processed in ways that data subjects would reasonably expect and the continuous assessment of how the processing affects the interests of individuals.[22]

Fair processing requires that AI systems do not produce discriminatory effects. The AI HLEG's guidelines quoted earlier draw the attention that 'data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models.'[23] A freshly published study on auditing the quality of datasets used in algorithmic decision-making systems is pointing out that 'the possibility of obtaining biased AI outcomes is strongly related to the characteristics of the data and the quality of the data management process, including data gathering, cleaning, annotation and processing'.[24] If the datasets used for training are unbalanced or biased, the system may produce outputs that have discriminatory effects on individuals without objective justification. In order to mitigate these risks, high-quality training and testing data are necessary that are representative of the population the AI system is deployed on.

Frequent incorrect outputs of the AI systems would also breach the fairness principle. The performance of the trained model should be measured by statistical

20 Article 83(5) GDPR.
21 EDPB 2020a.
22 Information Commissioner's Office – The Alan Turing Institute 2020.
23 European Commission Directorate-General for Communications Networks (AI HLEG) 2019.
24 Panel for the Future of Science and Technology – European Parliamentary Research Service 2022.

accuracy measures such as precision, recall, accuracy, and F1 score. In some cases, where high-quality test data is unavailable or the output is subjective, measuring statistical accuracy would not be appropriate.

PETs such as secure multi-party computation (SMPC) and federated learning (FL) can facilitate compliance with the fairness principle. The technologies can be used to prevent and restrict data usage for purposes that negatively impact an individual. These technologies will be discussed in detail in the third chapter, titled *Privacy-Enhancing Technologies*.

2.2.2. Transparency

This requirement was explicitly included in the data protection legislation for the first time by the authors of the GDPR. Information about how personal data is collected, used, consulted, or otherwise processed should be transparent, easily accessible, and easy to understand. GDPR highlights that individuals should be made aware of implications, risks, safeguards, and rights in relation to the processing of personal data.[25] Recital 60 adds that data controllers[26] 'should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed'.

In the case of AI-assisted decision-making, it presumes to provide meaningful information about the logic involved, the significance, and the envisaged consequences of the AI decision. Where the decision is based solely on automated processing, information should be provided also about the right to object and the right to obtain human intervention.[27] This brings to life the obligation to explain the technical logic or reasoning behind a particular output of the AI system and the related human decision for AI-assisted decision making. The AI HLEG stated that 'technical explainability requires that the decisions made by an AI system can be understood and traced by human beings'. Also, they project that trade-offs might have to be made between enhancing a system's explainability and increasing its accuracy.

Transparency, human agency and oversight, and accountability play an important role as three key principles for trustworthy AI. AI HLEG underlines that if 'an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process'. Also, 'the explanation should be timely and adapted to the expertise of the stakeholder concerned'.[28]

25 Recital 39 GDPR.

26 Article 4(7) GDPR: the natural or legal person, public authority, agency or other body that, alone or jointly with others, determines the purposes and means of the processing of personal data.

27 Article 13(2) b) and f) GDPR.

28 AI HLEG 2019.

Explainability relies on the level of interpretability, which depends on the model or set of models used by the AI system. For example, the usage of support vector machine (SVM) models may result in low levels of interpretability. 'Black box' techniques such as artificial neural networks (ANNs) can produce very low levels of interpretability. This may also apply to random forest models in some cases.[29]

2.2.3. Purpose Limitation

The purpose limitation principle is considered the cornerstone of data protection and is strongly linked to other data protection requirements. Purpose limitation requires data to be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes. The principle implies that purposes for processing personal data should be determined from the outset of the data processing lifecycle, at the time of the collection of the personal data.[30] Processing data only for the purposes defined beforehand could be challenging for AI systems because the purpose may change as the model learns and develops.

Data supposed to be used as training data is frequently collected originally for other purposes. If the latter data processing purposes are incompatible with the original purpose, then the purpose limitation principle could be a barrier to the development of an AI system. When assessing if the purpose of further processing is compatible with the purpose for which the personal data was initially collected, the data controller should take into account the following:

> [A]ny link between those purposes and the purposes of the intended further processing; the context in which the personal data have been collected, in particular the reasonable expectations of data subjects based on their relationship with the controller as to their further use; the nature of the personal data; the consequences of the intended further processing for data subjects; and the existence of appropriate safeguards in both the original and intended further processing operations.[31]

Recital 50 states that irrespective of the compatibility of the purposes, further processing should be allowed if the data subject has given consent or the processing is based on Union or Member State law. The latter could apply when the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority. Also, further processing of data is considered to be compatible with the original purpose if it takes place in connection with scientific or historical research or for statistical and archival purposes in the public interest.

29 Information Commissioner's Office – The Alan Turing Institute 2020.

30 Working Party was set up under Article 29 of Directive 95/46/EC (WP29), Opinion 03/2013 on purpose limitation.

31 Recital 50 GDPR.

Scientific research purposes should be interpreted in a broad manner, including, for example, technological development and demonstration, fundamental research, applied research, and privately funded research.[32] In some cases, the development of artificial intelligence may be considered to constitute scientific research, but this presumes that the model learns something new – identifying trends or correlations – from the processed personal data.[33]

The data strategy of the European Commission encourages data exchange between the public sector and businesses, and reuse for data-driven innovation.[34] However, the complexity and uncertain application of the purpose limitation principle could hinder the reuse of personal data for AI-related technologies.

Similarly to the fairness principle, secure multi-party computation and federated learning are the two PETs which can play a role in satisfying the requirements of the purpose limitation principle.

2.2.4. Data Minimization

This principle requires identifying and processing the minimum amount of relevant and adequate personal data necessary to fulfil purposes. Minimization is referenced as an organizational measure for data protection by design and by default.[35] As AI systems generally involve the processing of large amounts of data, particularly during the training phase, at first sight, it may seem incompatible with the minimization principle. In some cases, it is also impossible to determine from the beginning which features of the training data may be relevant.

Data minimization in practice means preventing excessive data collection and using only the data necessary for the purposes of the processing. Instead of limiting data processing to a specific volume of data, it limits 'the amount of detail included in training or in the use of a model'.[36] The level of accuracy of the AI systems' output is the main factor in reducing the available data to the subset included in the final AI model.

Data minimization requires extensive testing. The predictive model built by the Norwegian Tax Administration that helps identify risk errors of tax returns is considered a good example of best practice. From the tested five hundred features, only the thirty that proved the most relevant were used.[37]

Data minimization is also about minimizing the risks of processing. Data controllers developing or using AI systems should assess the impacts on data

32 Recital 159 GDPR.
33 Norwegian Data Protection Authority (Datatilsynet) 2018.
34 European Commission 2020.
35 Article 25 GDPR.
36 Norwegian Data Protection Authority (Datatilsynet) 2018.
37 Ibid.

protection by performing a data protection impact assessment (DPIA)[38] and 'consider how to achieve the objective in a way that is least invasive for the data subjects'.[39] Risk reduction can be reached by reducing the degree of identification by perturbation, adding 'noise', or anonymization. Also, PETs such as synthetic data generation, federated learning, and differential privacy can be effective solutions for data minimization. The last one will be further explored in the next chapter.

2.2.5. Accuracy

The data accuracy principle requires that processed personal data is accurate and, where necessary, kept up to date. This was already present in Convention 108,[40] the first legally binding international instrument in the data protection field, and has been maintained after the GDPR replaced the Data Protection Directive 95/46/EC[41] in 2016. The principle means that the number of inaccurate data elements in training data should be limited. Also, hidden biases should be prevented and representativeness ensured in order to have accurate outputs. In a big data context, keeping personal data up to date could be a mission impossible to achieve. Accuracy and fairness principles together raise the standard for AI systems that make inferences about people. Such a system can be deployed only if it is sufficiently statistically accurate to fulfil its purposes.

Data accuracy has particular relevance for AI. As the French Data Protection Authority highlights in its report on ethical matters raised by AI, 'the matter of the quality of the data processed by algorithms and AI is the most straightforward. It is not difficult to understand that incorrect data or data that is quite simply out of date will lead to errors or malfunctions of varying gravity depending on the sector in question'.[42] Inaccurate data could have a significant impact on individuals in the deployment phase also, resulting in an erroneous output or unjustified decision.

2.2.6. Storage Limitation

The principle of storage limitation prohibits keeping personal data longer than needed for the purposes of the processing. In order to ensure that the personal data

38 According to the WP29 Guidelines on DPIA, endorsed by the EDPB, innovative use or applying new technological or organizational solutions or matching or combining datasets in a way that would exceed the reasonable expectations of the data subjects can trigger the need to carry out a DPIA. WP29, *Guidelines on Data Protection Impact Assessment* (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679 (WP 248 rev.01) 2017.

39 Norwegian Data Protection Authority (Datatilsynet) 2018.

40 Council of Europe, Convention for the protection of individuals with regard to the processing of personal data, opened for signature on 28 January 1981.

41 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1995.

42 French Data Protection Authority (CNIL) 2017.

are not kept longer than necessary, time limits should be established for erasure or for a periodic review.[43] Defining appropriate data retention periods assumes that we identified all the purposes of data processing in advance. This is undoubtedly challenging in the case of AI-based processing, as the purpose of processing may change during the development and due to the high level of data replication.

GDPR leaves room for exceptions: if the personal data is processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, the data may be stored for longer periods.[44]

2.2.7. Security

The principle focusing on confidentiality and integrity states that personal data must be processed in a manner that ensures their appropriate security, 'including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures'. Security requirements apply explicitly to data controllers and processors[45] also. Two sections of the GDPR, articles 32–36 are dedicated to security requirements. These include the newly introduced requirement to notify personal data breaches to the supervisory authority and in certain cases to the data subjects too.

In order to maintain security, organizations should evaluate the risks inherent in the processing and implement measures to mitigate those risks.[46] The measures put in place should ensure an appropriate level of security. This implies the implementation of risk-based technical and organizational measures both at the time of the determination of the means for processing and at the time of the processing itself. In order to implement measures which ensure a level of security appropriate to the risk, organizations should take into account the state of the art and the costs of implementation in relation to the risks and the nature of the personal data to be protected.[47] This requires that organizations developing, deploying, or using AI assess and mitigate the security risks personal data processing may raise.

In addition to the obligation to ensure confidentiality and integrity of data processing, Article 32 provides the ongoing availability and resilience of processing systems and services and the monitoring and testing of processing activities. AI HLEG stresses technical robustness, which is a critical component of achieving trustworthy AI. The expert group emphasizes that 'technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner

43 Recital 39 GDPR.
44 Article 5(1) e) GDPR.
45 Article 4(8) GDPR: a natural or legal person, public authority, agency or other body that processes personal data on behalf of the controller.
46 Recital 83 GDPR.
47 Article 32(1) GDPR.

such that they reliably behave as intended while minimizing unintentional and unexpected harm and preventing unacceptable harm'.[48]

Ensuring security of personal data implies more than preventing re-identification of data subjects, unauthorized disclosure of training data or model outputs, or inferences about individuals represented in the training data. As AI HLEG highlights, 'AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries' and should also 'have safeguards that enable a fallback plan in case of problems'.[49]

This presumes security capabilities against third-party malicious activities such as hacking, alteration of the training data, model poisoning or evasion, or model inversion attack. The next subchapter, titled *Attacks on AI Models*, provides an overview of the security issues raised by AI systems.

The EDPB and EDPS draw the attention in their joint report that the wording of Article 83 of AI Act[50] does not include a reference to changes in external risks. They recommend that 'a reference to changes of the threats-scenario, arising from external risks, e.g., cyber-attacks, adversarial attacks and substantiated complaints from consumers therefore should be included in Article 83 of the Proposal'.[51]

PETs can partially support the compliance of AI systems with the data security requirements. Employing differential privacy, homomorphic encryption, multi-party computation, federated learning or using synthetic data for training purposes can contribute to security goals and a privacy-preserving AI system. More on some of these technologies will be explained in the third chapter.

2.2.8. GDPR Fines

The importance of data protection principles is also reflected by the enforcement actions taken by the national supervisory authorities. The Hungarian Data Protection Authority (NAIH) imposed a fine of €665,000 in February 2022 for the unlawful use of artificial intelligence. A Hungarian bank applied an AI-powered emotion analysis on voice recordings of calls conducted between its customers and a call centre. The bank failed to comply with the transparency and purpose limitation principles. Further, the Authority considered the solution's inefficiency in predicting the customers' emotions accurately, as well as the risk of processing

48 AI HLEG 2019.

49 Ibid.

50 This Regulation shall apply to the high-risk AI systems, other than the ones referred to in Paragraph 1, that have been placed on the market or put into service before [date of application of this Regulation referred to in Article 85(2)], only if, from that date, those systems are subject to significant changes in their design or intended purpose.

51 EPDP 2022; Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) 2021.

data by AI. The decision of the Authority highlighted that the information, the data protection impact assessment (DPIA), and balancing test documentation provided by the bank were not in compliance with the GDPR.[52]

Also, Clearview, the controversial facial recognition service provider, has been heavily fined for infringing GDPR by scraping images of individuals from public web sources and generating biometric data. After the data protection authorities in the United Kingdom and Italy, the company has been hit with another sanction from the Hellenic Data Protection Authority.[53] The value of fines received totalled €49 million.

2.3. Attacks on AI Models

Given the complexity of big data processing, AI systems can pose specific threats in addition to the security issues associated with any IT system. Besides human errors or omissions, data breaches can also be caused by external attacks, which 'may target the data, the model or the underlying infrastructure, both software and hardware'.[54]

Big data systems are increasingly becoming targets of more elaborate and specialized attacks.[55] Attacks on AI systems are increasing constantly, but the AI industry is alarmingly unprepared for these.[56] The attacks resulting in data breaches can lead to financial losses in addition to privacy and data protection harms. The average data breach cost has climbed 12.7% in the last two years, reaching $4.35 million in 2022.[57]

This subchapter will concentrate on the threats that can target machine learning models and present potential security challenges for personal data.

2.3.1. Poisoning

Poisoning attacks target classification algorithms and are likely to occur in several stages of the lifecycle from data collection to monitoring. In such an attack, an adversary is able to manipulate data (insert malicious data into the training or validation data) or model (replacing model file with an altered one) in order to modify the algorithm's behaviour in a chosen direction at a later point in time.[58] This may cause intentional misclassification or discrimination affecting the model's accuracy, compromising the integrity and trustworthiness of the AI system.

Federated learning and homomorphic encryption are PETs discussed in relation to the prevention of poisoning.

52 Hungarian Data Protection Authority 2022.
53 Ibid.
54 AI HLEG 2019.
55 European Union Agency for Network and Information Security (ENISA) 2016.
56 Advisa – The Road to Secure and Trusted AI Report – The Decade of AI Security Challenges 2021.
57 IBM 2022.
58 ENISA 2016.

2.3.2. Model Inversion

We usually think in relation to machine learning models that the training dataset cannot be recovered from the trained model. As several studies have shown, even without access to the dataset used for training, the output of the machine learning algorithms can be extremely revealing.

The exception is provided by model inversion attacks, which aim at classification algorithms. This attack can take place when the attacker has access to certain personal data belonging to specific individuals included in the training data, and by observing the inputs and outputs of the machine learning model it can infer further personal information about those same individuals. An excess of training data and the possibility to repeatedly query the model can contribute to the success of these attacks, and in some cases the re-identification of the data subjects.

One of the earliest successful model inversion attacks were deployed on recommender systems, 'a demonstration that collaborative filtering systems, where item recommendations are generated for a user based on behavioural patterns of other users, can end up revealing the consumption patterns of individual users'.[59] Also, researchers demonstrated the possibility of reverse engineering on a medical model designed to predict the correct doses for an anticoagulant using patient data including genetic biomarkers. They proved that an attacker having access to some demographic information of the patients included in the training data could infer their genetic biomarkers from the model.[60] Model inversion proved effective for attacking Facial Recognition Technology (FRT) systems. Researchers could reconstruct the facial images associated with the individuals included in the training data and match these (by humans) with 95% accuracy.[61]

Model inversion attacks can produce unauthorized disclosure of some personal data processed for training, causing loss of confidentiality and affecting the trustworthiness of the system.

In order to deploy an AI system in a way which prevents model inversion attacks, we should use secure multi-party computation or synthetic data if possible.

2.3.3. Membership Inference

This attack targets regression and classification algorithms in the deployment phase. Membership inference attacks allow malicious third parties to determine whether a given individual was present in the training data or not.[62] The attack itself does not cause disclosure of personal data, but using the model in combination

59 Veale–Binns–Edwards 2018.
60 Fredrikson–Jha–Ristenpart 2015.
61 Ibid.
62 Shokri–Stronati–Song–Shmatikov 2017.

with other data about a particular individual could directly lead to data breaches.[63] For example, if patients' clinical records are used to train a model associated with a disease, attackers knowing that a certain patient's data was used to train the model could reveal that the patient has this disease.

Membership inference attacks exploit confidence scores specific to prediction models. The score is disproportionately high in a prediction about an individual that was in the training dat, because the model has seen the data before. This allows us to determine if the individual was in the training data.

Exposure to membership inference depends on the amount of information 'remembered' by the used machine learning algorithm from the training datasets and on the degree of overfitting the model.[64] The level of risks associated with the attack depends on the sensitivity that the information membership may reveal. For example, if the model is trained with data on a vulnerable population like people suffering from mental disorder, addictions, or HIV, a membership inference attack could have a high-risk impact.

Making inferences about individuals represented in the training data through black-box or white-box inference can lead to breach of the principle of confidentiality. Employing differential privacy during model training can provide defence against membership inference attacks. More about trade-offs and limitations in the next chapter.

EDPB underlines the importance of implementing appropriate safeguards to identify, measure, and mitigate the risks that are specific to some machine learning such as data poisoning, model inversion, and white-box inference. Also, it considers essential to put in place monitoring processes to monitor (logging and collecting information on accuracy and fairness) the AI systems once in use.[65] The conclusion of ENISA could be the right one for this chapter as well: 'AI permeates every aspect of our daily lives, and therefore it is of paramount importance to ensure the cybersecurity of AI to ensure that AI and the set of associated technologies will be trustworthy, reliable and robust.'[66]

3. Privacy-Enhancing Technologies

In the context of AI, the security dimension of data protection bears a leading role in managing threats in a multi-party ecosystem and implementing specific controls to ensure that the AI system is secure. This implies that the necessary technical and organizational safeguards are put in place in the design stage of new

63 Ibid.
64 Shokri–Stronati–Song–Shmatikov 2017.
65 EDPB 2022.
66 ENISA 2016.

AI applications.[67] This is the scope of the notion of data protection by design and default, introduced as a legal requirement by Article 25 of the GDPR.

3.1. AI with Privacy

In 2015, ENISA expressed the need for a conceptual shift from 'big data versus privacy' to 'big data with privacy', 'adopting the privacy and data protection principles as an essential value of big data'[68] – mainly due to the 'scale of big data processing which brings existing privacy risks into a whole new (and unpredictable) level'.[69] Therefore, following a data protection by design approach may ask for an innovative solution since AI systems can have multiple levels of data processing and different techniques. Several techniques and technologies were proposed, developed, and improved in the last four decades that aim to support the deployment and configuration of appropriate technical and organizational measures in order to satisfy specific data protection principles. The group of these emerging technologies and techniques is commonly known as Privacy-Enhancing Technologies (PETs). Although the implementation of PETs may be necessary to comply with the data protection legal requirements, these alone cannot ensure compliance. To achieve compliance, PETs should always be used in conjunction with data protection policy and governance systems and frameworks.

In relation to GDPR compliance, there are several techniques to enhance data protection for AI systems. The most common ones are homomorphic encryption, secure multi-party computation, differential privacy, and synthetic data. Applying these promising PETs alone or combined facilitates the privacy-preserving applications of AI. These will be highlighted in the following chapter, having a special focus on technologies supporting the compliance with the data security principle.

3.2. Homomorphic Encryption

In order to prevent unauthorized parties from accessing the processed personal data or to safely provide access, it is necessary to mask the data in three different states: at-rest, in-use, and in-transit. During the typical encryption techniques which secure data at-rest and in-transit, the original data once encrypted becomes obscured or unintelligible. The challenge was to develop an encryption technique which protects data in-use, while keeping it intelligible for processing.

In 1978, Ronald L. Rivest, Len Adleman, and Michael L. Dertouzos laid down the theoretical foundations of homomorphic encryption,[70] which allows computation

67 Ibid.
68 ENISA 2015.
69 Ibid.
70 Rivest–Adleman–Dertouzos 1978.

to be performed directly on encrypted data without requiring to decrypt it first. This method was further developed by Craig Gentry, who was the first to describe a construction for a fully homomorphic (typically asymmetric) encryption scheme.[71]

3.2.1. Benefits

Partially homomorphic encryption (PHE) allows only additions or only multiplications, somewhat homomorphic encryption can support a limited number of both additions and multiplications, and fully homomorphic encryption (FHE) enables multiple operations to be performed over encrypted data.[72] The end-result of the operation(s) remains also encrypted, and it can be decrypted by the owner of the key.[73] The result is equivalent to the results obtained working with the original unencrypted data directly.[74] In essence, homomorphic encryption ensures the secure outsourcing of specific operations on confidential data and safely provides access to them, being an important component of the defence against poisoning attacks.

3.2.2. Use Cases

Homomorphic encryption is a useful data protection by design measure in use cases when processing – for the time being, simple arithmetical operations such as addition and/or multiplication – is performed by a third party. For example, a cloud service provider as a processor can perform operations on behalf of the data controller without accessing the content of the personal data. A real-world use case is the collection of data from connected devices with the purpose of obtaining aggregated values. The aggregator service receives individual encrypted data and adds them up, resulting in the final data encrypted accumulated values.[75] Enabling privacy-preserving data aggregation, homomorphic encryption is especially suitable for smart meter systems.[76]

In the last few years, electronic homomorphic encryption has been extended to new use cases such as smart contracts, electronic voting,[77] genome privacy,[78] fraud detection,[79] or password breach monitoring,[80] and it has the potential to be used

71 Gentry 2009.
72 The Royal Society 2019.
73 Industry, Government and Academic Consortium to Advance Secure Computation 2017.
74 Spanish Data Protection Agency 2020.
75 Ibid.
76 Wang–Heb–Zhangc 2022.
77 AEPD 2020.
78 Industry, Government and Academic Consortium to Advance Secure Computation 2017.
79 Maass 2020.
80 Lauter–Kannepalli–Cruz Moreno 2021.

in a wide range of applications. It can also support privacy-preserving machine learning. For example, it can underpin privacy-preserving predictions.[81]

3.2.3. Limitations

Due to the diversity of operations performed on the encrypted data, FHE is currently inefficient besides the higher level of protection and utility. Application of FHE is held back currently since it is highly computationally intensive, suffers from bandwidth and latency issues, and running time can exponentially increase depending on security parameters.[82] PHE and SHE, on the other hand, provide good performance and protection, but with very limited utility.

As encrypted data is typically much larger, more storage and processing resources are needed to encrypt, store, and decrypt the data. Therefore, homomorphic encryption is extremely computationally expensive and impractically slow.[83] As cloud computing evolves, performance will increase, and this PET will become more accessible for commercial applications.[84]

AEPD highlights the risk arising from 'using the same key on the data that are going to be processed may entail a vulnerability in the encryption system'. Risk increases as the volume of the processed data and the period of access grows. AEPD stresses that 'the use of an additional encryption layer in communications is essential together with the need to minimise the information encrypted under the same key, which must be limited to the groups of data that operate with one another'.[85] Standardization of homomorphic encryption techniques is also a major ongoing challenge.

3.2.4. Potential

The level of maturity differs for variations of homomorphic encryption. There are products based on PHE offered on the market, SHE is piloted, but FHE is just at research level.[86] However, homomorphic encryption can already enable other PETs such as secure multi-party computation, private data aggregation, or federated machine learning.[87] The use of homomorphic encryption opens new doors for the secure processing of personal data 'such as servers based on the IoT, Cloud Computing and automated learning or Machine Learning'.[88]

81 The Royal Society 2019.
82 Ibid.
83 Ibid.
84 AEPD 2020.
85 Ibid.
86 The Royal Society 2019.
87 Ibid.
88 AEPD 2020.

Some companies started to use homomorphic encryption for end-to-end encrypted systems' data processing. Meta is experimenting with this technique to detect child sexual abuse material on end-to-end encrypted messaging platforms. However, it 'is not yet technically feasible to implement in messaging at scale' because it would take more than seven months to run on each message.[89]

3.3. Secure Multi-party Computation

Secure multi-party computation (SMPC) is a more mature PET enabling computation on encrypted data without losing data utility. Introduced in 1986 and the first prototypes developed in 2004,[90] SMPC is a subfield of cryptography concerned with enabling private distributed computations.

3.3.1. Benefits

SMPC allows 'computation or analysis on combined data without the different parties revealing their own private input'.[91] Lack of trust between two or more parties, also data protection restrictions (ex. appropriate legal basis, ensuring confidentiality) or technical constraints of data sharing between parties who intend to carry out analyses on their combined data, are addressed by this cryptographic technique providing data masking.[92] It is important that the original, distributed data existing across several parties does not need to be gathered to a central repository. Practically, SMPC enables operations to be performed on the input data of two or more parties, without revealing the input data of one party to the other parties, and ensures parties to jointly form the obtained results.[93] Unlike homomorphic encryption which protects data in storage and also during computing, SMPC supports only the latter.

SMPC can be used together with federated machine learning to leverage the benefits of stronger confidentiality with greater scale of computation.[94]

3.3.2. Use Cases

Commercial products for secure multi-party computation first appeared in 2010.[95] Some real-world applications preceded them. These permit auctions where participants could place bids without revealing them. A good example

89 Business for Social Responsibility 2022.
90 Malkhi–Nisan–Pinkas–Sell 2004.
91 The Royal Society 2019.
92 AEPD 2022.
93 Ibid.
94 Mugunthan–Polychroniadou–Byrd–Hybinette Balch 2019.
95 The Royal Society 2019.

is provided by Denmark, where SMPC was used to redistribute the country's EU-fixed production quota among sugar beet producers without the need for a central auctioneer or revealing commercially sensitive information.[96] This system allows bidders to identify the winner of the auction without revealing information related to the actual bid.

Pooling personal data from different governmental departments to gain insights for policy makers involves high privacy risks. SMPC can enable us to put information in a wider context without revealing citizens' personal data. For example, Estonia used SMPC to analyse encrypted income tax records and higher education records to determine if students who work during their studies are more likely to fail to graduate in time than fellow students who are focused exclusively on their studies.[97]

3.3.3. Limitations

Although SMPC is constantly evolving, like all PETs, it faces a number of challenges. The widespread use of the SMPC is limited by the relatively high costs of computation and bandwidth. However, where high-bandwidth settings are available (devices connected within a data centre), SMPC significantly outperforms FHE.[98] If all the participants outsource their computation to the same cloud provider, bandwidth costs can be reduced, 'but it requires a strong trust model'.[99] Also, data structures need to be standardized in order to perform data analysis with SMPC.

Since the output is revealed in the case of SMPC, 'the output must be controlled to limit what an adversary can infer about the private data from the output'.[100] This leakage can be addressed in the best way by combining SMPC with differential privacy.[101]

Beyond the SMPC execution itself, the protection of the cryptographic keys is a challenge that has to be tackled by involved parties.[102] Thus, parties implementing SMPC should have a high level of security capabilities.

3.3.4. Potential

SMPC is promising for operations that require large amounts of data as the training of machine learning models. SMPC can be used to allow private multi-party machine learning. Different parties send encrypted data to each other to train a

96 Bogetoft–Lund–Damgård–Geisler 2009.
97 Bogdanov–Kamm–Kubo–Rebane–Sokk–Talviste 2016.
98 Evans–Kolesnikov–Rosulek 2018.
99 Ibid.
100 Ibid.
101 Pettai–Laud 2015.
102 The Royal Society 2019.

machine learning model, eliminating the need for a trusted central authority that would perform the computation by gathering all the data and decrypting it.[103]

SMPC has a great potential for machine learning systems trained on health data. SMPC could facilitate the use and sharing of health data for R&D purposes, as it tackles the problem that data is distributed across several organizations, and gathering all the data to a central repository is rarely permitted by the data protection requirements.

By keeping the input data of each party private and providing the correct output to each of them, SMPC can support the compliance with the purpose limitation and prevent model inversion attacks. However, it is necessary to implement additional data protection measures to guarantee GDPR compliance.

3.4. Differential Privacy

While homomorphic encryption and SMPC deal with privacy during computation, differential privacy addresses privacy in disclosure.[104] This PET was introduced in 2006 by C. Dwork[105] and her team,[106] and it is based on the Law of Large Numbers.[107]

Differential privacy is a 'strong, mathematical definition of privacy in the context of statistical and machine learning analysis', and 'it is used to enable the collection, analysis, and sharing of a broad range of statistical estimates based on personal data'.[108] The differential privacy mathematically guarantees that the result of a differentially private analysis provides the same inference about any individual's personal data, regardless of whether that particular individual's personal data was included in the input to the analysis.[109] Differential privacy preserves the usefulness of data by allowing statistical analysis and identification of trends on larger datasets, but in a way that protects individuals' privacy by 'establishing data protection guarantees by design through the practical implementation of information abstraction strategies'.[110]

Depending on the stage when the data analysis is applied, differential privacy can be local (distributed) or global (centralized). In the case of local differential privacy, random noise is added at the data collection stage 'so that users get a "plausible deniability" type of guarantee with respect to data being collected about them'.[111] This may result in reducing accuracy by adding more noise than

103 Ibid.
104 Ibid.
105 Dwork 2006.
106 Dwork–McSherry–Nissim–Smith 2006.
107 Law of Large Numbers 2020.
108 Wood–Altman–Bembenek–Bun–Gaboardi–Honaker–Nissim–O'Brien–Steinke–Vadhan 2018.
109 Ibid.
110 AEPD 2021.
111 The Royal Society 2019.

the global approach, as adding noise at an early stage of the data lifecycle does not permit optimizing the amount of noise to a specific analysis.[112] Global differential privacy assumes that noise is added to the output, taking away the possibility to determine if a particular data record was included in the dataset used to produce the output.[113]

The adjustability of the amount of noise added to the original dataset is an important feature of differential privacy. By increasing the amount of noise, privacy risks decrease, but data utility may decline too. The challenge is to calculate the value of noise in a way 'that preserves the result within the utility range'.[114]

3.4.1. Benefits

Differentially private mechanisms can provide a way to query datasets containing private data while mitigating 'the risk of revealing whether a specific individual or organisation is present in a dataset or output'.[115]

One of the major benefits of differential privacy is the strong protection provided against membership inference attacks if the training process is differentially private.[116] Also, differential privacy is the best practice against re-identification attacks performed by combining different datasets.[117]

Another benefit of differential privacy is the possibility to quantify the privacy loss and compare it among different techniques. This enables the control and analysis of cumulative privacy losses when running multiple differentially private analyses on a particular dataset. Also, measuring privacy loss acquired by groups is possible.[118] Immunity to post-processing is also an important property of differential privacy. This allows to arbitrarily transform a differentially private output using some data-independent function, but without impacting its privacy guarantees.[119]

Dwork and her team have shown that differential privacy can improve generalization in machine learning algorithms.[120] In particular, 'if a differentially private learning algorithm has good training accuracy, it is guaranteed to have good test accuracy'.[121]

112 Ibid.
113 Ibid.
114 AEPD 2021.
115 Ibid.
116 Shokri–Stronati–Song–Shmatikov 2017.
117 Chin–Anne Klinefelter 2012.
118 Nguyen 2019.
119 Zhu–Van Hentenryck–Fioretto 2020.
120 Dwork–Feldman–Hardt–Pitassi–Reingold–Roth 2015.
121 Papernot–Guha Thakurta 2021.

3.4.2. Use Cases

Differential privacy is PET which has a wide range of applications from linear regressions and cumulative distribution functions to machine learning.[122]

After having implemented differential privacy for other services, the United States Census Bureau took the decision to replace data swapping, the previously applied disclosure avoidance mechanism, with differential privacy for the 2020 census. This was motivated by its goal 'to publish a specific, higher number of tables of statistics with more granular information than previously'[123] and at the same time to protect against emerging technology threats such as re-identification attacks.[124]

High-profile tech companies such as Apple, Google, Microsoft, and Uber also implemented differential privacy in practical applications. Apple used local differential privacy for collecting statistics from hundreds of millions of users in order to identify popular emojis, popular health data types, and media playback preferences in Safari.[125]

Google implemented differential privacy with a similar goal: such, to collect statistics from end-users in a privacy-preserving way.[126] Similarly, Uber in collaboration with the University of California implemented this method to perform analytics on user data and determine the average trip distance for users.[127]

3.4.3. Limitations

Adding noise to a dataset can cause accuracy and robustness issues. Especially for smaller datasets, this can harm utility. Practically, the trade-off of utility and privacy improves proportionally with the size of the dataset, where less noise is needed, as 'the more individuals included in a dataset, the harder it might be to identify that a specific individual was included'.[128]

In the case of local models, the utility is usually affected because the distributed data requires more noise to achieve differential privacy. In order to obtain highly accurate aggregate statistics, large datasets are essential. But working on large datasets does not automatically lead to great utility. The algorithms transforming a dataset to differentially private need to be designed to the specific use case to ensure that the output meets utility expectations.[129]

122 AEPD 2021.
123 The Royal Society 2019.
124 United States Census Bureau 2022.
125 Differential Privacy Team of Apple 2017.
126 Erlingsson–Pihur Korolova 2014.
127 Johnson–Near–Song 2018.
128 The Royal Society 2019.
129 Ács–Castelluccia 2014.

The privacy preservation effect of differential privacy is heavily dependent on the 'privacy budget', the 'quantitative measure of by how much the risk to an individual's privacy may increase by, due to that individual's data inclusion in the inputs to the algorithm'.[130] Setting the 'privacy budget' is key to ensuring privacy guarantees, and it requires expertise. It is crucial to take into consideration 'the statistical inferences that might happen after the release of results and how, for example, outsiders might be able to link data with side information'.[131]

For example, differential privacy could lose its privacy guarantees where differentially private data collection from the same individuals is continuous over time. It is not possible 'to collect differential privacy protected data from a community of respondents an indefinite number of times with a meaningful privacy guarantee'.[132] This is why the previously detailed user data collection performed by Apple and Google presents shortcomings.

3.4.4. Potential

Differential privacy provides great performance for datasets where the number of individuals is large but the weight of each individual to the output is limited. This privacy-enhancing technology contributes substantially to enable privacy-preserving machine learning.

Overfitting is a typical mistake in machine learning. This can be mitigated by achieving differential privacy, which 'goes hand in hand with preventing overfitting to particular examples'.[133]

Differential privacy supports a wide range of techniques used in statistics and machine learning such as classification, clustering, and also statistical disclosure limitation techniques such as synthetic data generation. The generated synthetic data retains statistical properties of the original data but at the same time protects against model inversion attacks.[134] We can state without doubt, differential privacy will have a central role in deploying privacy-preserving machine learning.

4. Conclusions

The adoption of AI and machine learning has skyrocketed since the pandemic. As AI systems are becoming increasingly widespread in the public and private sectors, the majority of organizations realize that privacy challenges can be hardly

130 The Royal Society 2019.
131 Ibid.
132 Domingo-Ferrer–Sánchez–Blanco-Justicia 2020.
133 The Royal Society 2019.
134 Ibid.

overcome. But despite the fact that privacy is considered the fourth most relevant AI risk after cybersecurity, regulatory compliance, and explainability,[135] 52% of business decision makers say that their company is not safeguarding data privacy through the entire lifecycle,[136] thus failing to meet an important condition for trustworthy AI. This creates exposure for the data involved in training, testing, or deploying the system. Attacks targeting machine learning models can exploit these vulnerabilities in novel ways, as has been shown in the second chapter, increasing privacy risks for individuals and the odds of a hefty GDPR fine.

Due to the volume of the data, the complexity of the processing, and the unforeseen consequences for data subjects, applying the data protection principles in a machine learning context is far from straightforward. Practically, data protection compliance would require translating these principles into concrete requirements and system design specifications and then finding and implementing appropriate technical and organizational measures throughout all of the stages of the data processing lifecycle. As GDPR asks for state-of-the-art and risk-based safeguards implemented from the design phase, innovative solutions and approaches are needed, which are better suited to the personal data processing performed by machine learning models in order to unlock the full potential of these data-driven AI technologies.

Privacy-enhancing technologies specifically tailored to mitigate privacy risks characteristic to machine learning models were presented in the third chapter. Homomorphic encryption, secure multi-party computation, and differential privacy not only provide shield against attacks targeting machine learning models but can bring a significant contribution to comply with the fairness, purpose limitation, and data minimization principles besides the data security principle. PETs enable multiple applications and bring new possibilities for data analysis.[137] These technologies evolving at an accelerating pace since 2000 could open the horizon for privacy-preserving machine learning. However, PETs do not transform personal data processing compliant with data protection regulations in one fell swoop, and they should be used together with process controls, high-standard policy, and data governance systems.

At this stage, significant barriers are still present, PETs being limited by high computational demand, data-interoperability, data utility, and accuracy issues. Security risks resulting from reverse engineering are also highly debated. Maturity level and early-stage use of PETs also make the road to market-wide penetration meandering. This could be enhanced by promoting good practices, initiating standardization, and developing certification mechanisms for mature PETs by the responsible bodies. These all point to the need of further research and development

135 McKinsey 2021.
136 IBM 2022.
137 The Royal Society 2019.

to explore the potential benefits and impact of PETs on processing personal data by AI systems.

Hopefully, the pace of innovation in this field will be maintained, and in the near future we will be able to enjoy the benefits of AI systems deployed by private and public organizations together with top-notch privacy safeguards.

References

ÁCS, G.–CASTELLUCCIA, C. 2014. *A Case Study: Privacy Preserving Release of Spatio-Temporal Density in Paris*. http://www.crysys.hu/~acs/publications/AcsC14kdd.pdf.

AEPD (SPANISH DATA PROTECTION AGENCY (AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS). 2020. *Encryption and Privacy III: Homomorphic Encryption*. https://www.aepd.es/en/prensa-y-comunicacion/blog/encryption-privacy-iii-homomorphic-encryption.

2021. *Anonymisation and Pseudonymisation (II): Differential privacy*. https://www.aepd.es/en/prensa-y-comunicacion/blog/anonymisation-and-pseudonymisation-ii-differential-privacy.

2022. *Privacy by Design: Secure Multi-Part Computation: Additive Sharing of Secrets*. https://www.aepd.es/en/prensa-y-comunicacion/blog/privacy-by-design-secure-multi-part-computation-additive-sharing-secrets.

BOGDANOV, D.–KAMM, L.–KUBO, B.–REBANE, R.–SOKK, V.–TALVISTE, R. 2016. *Students and Taxes: A Privacy-Preserving Study Using Secure Computation*. https://www.researchgate.net/publication/302065845_Students_and_Taxes_a_Privacy-Preserving_Study_Using_Secure_Computation.

BOGETOFT, P.–LUND, D. H.–DAMGÅRD, I.–GEISLER, M. 2009. *Secure Multiparty Computation Goes Live*. https://www.researchgate.net/publication/220796917_Secure_Multiparty_Computation_Goes_Live.

BORKING, J.–RAAB, C. 2001. *Laws, PETs and Other Technologies for Privacy Protection, Journal of Information, Law and Technology*. https://www.researchgate.net/publication/220667925_Laws_PETs_and_Other_Technologies_for_Privacy_Protection.

BUSINESS FOR SOCIAL RESPONSIBILITY 2022. *Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption – Executive Summary*. https://about.fb.com/wp-content/uploads/2022/04/BSR-E2EE-HRIA-Executive-Summary.pdf.

CHIN A.–KLINEFELTER, A. 2012. *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2062447.

COUNCIL OF EUROPE 1981. *Convention for the Protection of Individuals with Regard to the Processing of Personal Data, Opened for Signature on 28 January 1981*. https://rm.coe.int/convention-108-convention-for-the-protection-of-individuals-with-regar/16808b36f1.

DIFFERENTIAL PRIVACY TEAM OF APPLE 2017. *Learning with Privacy at Scale*. https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf.

DOMINGO-FERRER, J.–SÁNCHEZ, D.–BLANCO-JUSTICIA, A. 2020. *The Limits of Differential Privacy (and Its Misuse in Data Release and Machine Learning)*. https://arxiv.org/pdf/2011.02352.pdf.

DWORK, C. 2006. *Differential Privacy*. https://audentia-gestion.fr/MICROSOFT/dwork.pdf.

DWORK, C.–FELDMAN, V.–HARDT, M.–PITASSI, T.–REINGOLD, O.–ROTH, A. 2015. *Generalization in Adaptive Data Analysis and Holdout Reuse*. https://arxiv.org/pdf/1506.02629.pdf.

DWORK, C.–MCSHERRY, F.–NISSIM, K.–SMITH, A. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis*. https://link.springer.com/content/pdf/10.1007/11681878_14.pdf.

EDPB 2020a. *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, Version 2.0, 2020*. https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf.

2020b. *Response Letter to Sophie in't Veld*. https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_letter_out2020_0004_intveldalgorithms_en.pdf.

2022. *Guidelines 05/2022 on the Use of Facial Recognition Technology in the Area of Law Enforcement*. https://edpb.europa.eu/system/files/2022-05/edpb-guidelines_202205_frtlawenforcement_en_1.pdf.

ERLINGSSON, Ú.–PIHUR, V.–KOROLOVA, A. 2014. *RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response*. https://arxiv.org/pdf/1407.6981.pdf.

EUROPEAN COMMISSION. 2020. *A European Strategy for Data*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066

EUROPEAN COMMISSION DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS (AI HLEG). 2019. *Content and Technology, Ethics Guidelines for Trustworthy AI*. https://data.europa.eu/doi/10.2759/346720.

EUROPEAN COMMISSION JOINT RESEARCH CENTRE–EVAS, T.–SIPINEN, M.–ULBRICH, M. et al. 2022. *AI Watch: Estimating AI Investments in the European Union*. https://data.europa.eu/doi/10.2760/702029.

EUROPEAN UNION AGENCY FOR NETWORK AND INFORMATION SECURITY (ENISA).

2015. *Privacy by Design in Big Data – An Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics*. https://www.enisa.europa.eu/publications/big-data-protection/@@download/fullReport.

2016. *Big Data Threat Landscape and Good Practice Guide*. https://www.enisa.europa.eu/publications/big-data-security/at_download/fullReport.

2020. *AI Cybersecurity Challenges – Threat Landscape for Artificial Intelligence*. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/at_download/fullReport.

EVANS, D.–KOLESNIKOV, V.–ROSULEK, M. 2018. *A Pragmatic Introduction to Secure Multi-Party Computation*. https://www.cs.virginia.edu/~evans/pragmaticmpc/pragmaticmpc.pdf.

FREDRIKSON, M.–JHA, S.–RISTENPART, T. 2015. *Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures*. https://rist.tech.cornell.edu/papers/mi-ccs.pdf.

FRENCH DATA PROTECTION AUTHORITY (CNIL) 2017. *How Can Humans Keep the Upper Hand – The Ethical Matters Raised by Algorithms and Artificial Intelligence*. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.

GARTNER 2021. *Top Strategic Technology Trends for 2022: Privacy Enhancing Computation*. https://www.gartner.com/doc/reprints?id=1-27VY7GL1&ct=211103&st=sb.

GENTRY, C. 2009. *Fully Homomorphic Encryption Using Ideal Lattices*. https://www.cs.cmu.edu/~odonnell/hits09/gentry-homomorphic-encryption.pdf.

HELLENIC DATA PROTECTION AUTHORITY. 2022. *Press Release about Fining Clearview AI Inc., 20 July 2022*. https://edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros_en.

HUNGARIAN DATA PROTECTION AUTHORITY.

2022. *Press Release about a Fine Imposed in Connection with the Use of Artificial Intelligence, 20 May 2022*. https://edpb.europa.eu/news/national-news/2022/data-protection-issues-arising-connection-use-artificial-intelligence_en.

IBM. 2022. *Cost of a Data Breach – Report*. https://www.ibm.com/downloads/cas/XZNDGZKA.

IBM–MORNING CONSULT. 2022. *IBM Global AI Adoption Index 2022*. https://www.ibm.com/downloads/cas/GVAGA3JP.

INDUSTRY, GOVERNMENT AND ACADEMIC CONSORTIUM TO ADVANCE SECURE COMPUTATION. 2017. *Homomorphic Encryption Standardization, Basics of Homomorphic Encryption*. 2017. https://homomorphicencryption.org/introduction.

INFORMATION COMMISSIONER'S OFFICE – THE ALAN TURING INSTITUTE. 2020. *Explaining Decisions Made with AI*. https://www.pdpjournals.com/docs/888063.pdf.

JOHNSON, N.–NEAR, J. P.–SONG, D. 2018. *Towards Practical Differential Privacy for SQL Queries*. https://arxiv.org/pdf/1706.09479.pdf.

LAUTER, K.–KANNEPALLI, S.–MORENO, R. C. 2021. *Password Monitor: Safeguarding Passwords in Microsoft Edge*. https://www.microsoft.com/en-us/research/blog/password-monitor-safeguarding-passwords-in-microsoft-edge.

MAASS, E. 2020. *Fully Homomorphic Encryption: Unlocking the Value of Sensitive Data While Preserving Privacy*. https://securityintelligence.com/posts/fully-homomorphic-encryption-next-step-data-privacy.

MALKHI, D.–NISAN, N.–PINKAS, B.–SELL, Y. 2004. *Fairplay – A Secure Two-Party Computation System*. https://www.usenix.org/legacy/event/sec04/tech/malkhi/malkhi.pdf.

MCKINSEY 2021. *Global Survey: The State of AI in 2021*. https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021.

MUGUNTHAN, V.–POLYCHRONIADOU, A.–BYRD, D–HYBINETTE BALCH, T. 2019. *SMPAI: Secure Multi-party Computation for Federated Learning*. https://www.jpmorgan.com/content/dam/jpm/cib/complex/content/technology/ai-research-publications/pdf-9.pdf.

NGUYEN, A. 2019. *Understanding Differential Privacy – From Intuitions behind a Theory to a Private AI Application*. https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a.

NORWEGIAN DATA PROTECTION AUTHORITY (DATATILSYNET). 2018. *Artificial Intelligence and Privacy*. https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

PANEL FOR THE FUTURE OF SCIENCE AND TECHNOLOGY AND EUROPEAN PARLIAMENTARY RESEARCH SERVICE. 2022. *Auditing the Quality of Datasets Used in Algorithmic Decision-Making Systems*. https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf.

PAPERNOT. N.–GUHA THAKURTA, A. 2021. *How to Deploy Machine Learning with Differential Privacy*. https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy.

PETTAI, M.–LAUD, P. 2015. *Combining Differential Privacy and Secure Multiparty Computation*. https://eprint.iacr.org/2015/598.pdf.

RIVEST, R. L.–ADLEMAN, L.–DERTOUZOS, M. L. 1978. *On Data Banks and Privacy Homomorphisms*. https://people.csail.mit.edu/vinodv/6892-Fall2013/RAD78.pdf.

SHOKRI, R.–STRONATI, M.–SONG, C.–SHMATIKOV, V. 2017. *Membership Inference Attacks against Machine Learning Models*. https://arxiv.org/pdf/1610.05820.pdf.

STANFORD UNIVERSITY–LITTMAN, M. L.–AJUNWA, I.–BERGER, G.–BOUTILIER, C.–CURRIE, M.–DOSHI-VELEZ, F.–HADFIELD, G.–HOROWITZ, M. C.–ISBELL,

C.–KITANO, H.–LEVY, K.–LYONS, T.–MITCHELL, M.–SHAH, J.–SLOMAN, S.–VALLOR, S.–WALSH, T. 2021. *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100). Study Panel Report.* http://ai100.stanford.edu/2021-report.

THE ROYAL SOCIETY. 2019. *Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis.* https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf.

UNITED STATES CENSUS BUREAU. 2020. *Census Data Products: Next Steps for Data Releases*. https://www.census.gov/newsroom/blogs/random-samplings/2022/04/2020-census-data-products-next-steps.html

VEALE, M.–BINNS, R.–EDWARDS, L. 2018. *Algorithms That Remember: Model Inversion Attacks and Data Protection Law*. https://arxiv.org/pdf/1807.04644.pdf.

WANG, B.–HEB, S–ZHANGC, S. 2022. *Privacy Protection Data Aggregation Scheme with Batch Verification and Fault Tolerance in Smart Grid Communication*. http://166.62.7.99/assets/default/article/2022/04/19/article_1650361841.pdf.

WOOD, A.–ALTMAN, M.–BEMBENEK, A.–BUN, M.–GABOARDI, M.–HONAKER, J.–NISSIM, K.–O'BRIEN, D. R.–STEINKE, T.–VADHAN, S. 2018. *Differential Privacy: A Primer for a Non-technical Audience*. https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1058&context=jetlaw.

ZHU, K.–VAN HENTENRYCK, P.–FIORETTO, F. 2020. *Bias and Variance of Post-processing in Differential Privacy*. https://arxiv.org/pdf/2010.04327.pdf.

*** *Advisa – The Road to Secure and Trusted AI Report – The Decade of AI Security Challenges.* 2021. https://adversa.ai/download/1220.

*** *Directive 95/46/EC (WP29), Opinion 03/2013 on Purpose Limitation.* 2013. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

*** *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data.* 1995. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046.

*** *Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).* 2021. https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

*** *Law of Large Numbers.* 2020. https://encyclopediaofmath.org/index.php?title=Law_of_large_numbers.

*** *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).*

and Amending Certain Union Legislative Acts. 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

*** *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation).* 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679.

*** *WP29, Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679 (WP 248 rev.01).* 2017. https://ec.europa.eu/newsroom/article29/items/611236/en.

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 65–80

DOI: 10.47745/AUSLEG.2022.11.2.04

Some Remarks on the 'AI Judge' in the Context of Recent European Union Regulatory Action

János SZÉKELY
PhD, Senior Lecturer
Sapientia Hungarian University of Transylvania, Department of Legal Science
ORCID: 0000-0003-4254-2054
e-mail: szekely.janos@kv.sapientia.ro

Abstract. The utilization of artificial intelligence (AI) as an aid during adjudication is no longer a future prospect but a reality. While current implementations of the technology are as of yet far removed from the future science fiction would have us fear, the prospect of the 'AI judge' must now be seriously considered. In our analysis, we investigate whether such a prospect would be compatible with fundamental rights and proposed EU norms set to govern the use of AI technology. We also examine the ethical requirements for utilizing AI, including in the judicial domain. We find that in lack of a possibility of granting a reasoned decision, in the course of a transparent procedure AI fails to meet the basic requirements that would allow its use under current and predictable future regulatory conditions during adjudication in the European Union. We further find that the shortcomings of the technology and the regulatory environment would hinder the accountability required for implementing the 'AI judge'. We conclude that the specific needs of adjudication have not been duly considered during the preparation of EU instruments in the field of AI, and further regulation as well as research will be necessary.

Keywords: artificial intelligence, ethics, courts, fundamental rights, fair trial, European Union

1. Introductory Remarks

Artificial intelligence (AI) has become something of a catchphrase in almost all disciplines of science and even the arts in recent years. That it holds great promise and equally great risks stands beyond any doubt. Yet, until lately, national legislatures as well as international organizations have been reluctant to propose regulation that would either direct or hinder the development or deployment of

AI to certain tasks; compulsory rules have only been formulated and implemented regarding personal data protection.[1]

This 'silence of the legislator' is rapidly becoming untenable. The development of AI applications now appears to have definitively left behind the era of false starts and sudden stops that have plagued the technology in the past.[2] Practical and economically viable AI solutions are already in use, or they are, at the very least, rapidly emerging. Thus, the period of 'salutary neglect' by regulators is at its end: on both shores of the Atlantic, they now aim to set forth new AI laws as soon as possible.[3] In this drive to legislate, the European Commission has tabled two regulatory drafts to constitute the cornerstones of European AI law: the proposed Artificial Intelligence Act[4] (AIA) and the proposed AI Liability Directive.[5] These proposals, still in the course of development and subject to public debate, aim to channel AI development in EU Member States according to the precautionary principle and implement fault-based liability aided by some presumptions of misconduct and causation in case AI systems should cause damage during their intended or unintended functioning.

One of the possible implementations in which AI presents great potential is that of dispute resolution, either as an instrument for aiding judicial decision-making in one way or another or as an autonomous AI adjudicator.[6] In our study, we aim to analyse the compatibility of these two forms, and specifically the 'AI judge' with the way the European Commission, a body of the European Union, currently envisages AI regulation in order to predict the possible future(s) of AI-based or AI-aided dispute resolution in the European Union.

2. Principles of AI and the 'AI Judge'

During the development of the AIA, the European Commission convened the High-Level Expert Group on Artificial Intelligence (AI HLEG), which would develop the main guidelines for the proposed regulation. The AI HLEG formulated a 'European' approach to AI based on three guiding principles: 1. compliance with the law, 2. fulfilment of ethical principles, and 3. robustness. These principles were then expanded into the following list of assessment criteria: 1. human agency and oversight; 2. technical robustness and safety; 3. privacy and data governance;

1 di Carlo–De Bondt–Evgeniou 2021.
2 Francesconi 2022.
3 Casovan–Shankar 2022.
4 European Commission, Directorate-General for Communications Networks, Content and Technology 2021.
5 European Commission 2022.
6 See Szekely 2019.

4. transparency; 5. diversity, non-discrimination, and fairness; 6. societal and environmental well-being; 7. accountability.[7]

In grounding EU AI law in principles such as these, the AI HLEG did not propose any particular rules for AI-based adjudication. Still, some basic conclusions can already be drawn from the list of principles and requirements regarding the future regulatory landscape of AI-based adjudication. Firstly, the principle of compliance with the law as outlined by the AI HLEG seems to mirror some of the guarantees associated with a fair trial (compliance with a procedure conducted according to the law by a court that is itself established under the law).[8] Secondly, through requirements such as human agency and oversight, as well as transparency, non-discrimination, fairness, and accountability, the AI HLEG set forth the framework with which AI-based adjudication would need to take place. This framework is not all that distant from that found in other international instruments such as the European Convention on Human Rights[9] (specifically articles 6 and 14) or the Charter of Fundamental Rights of the European Union[10] (articles 21 and 47).

We should note here that the AI HLEG, as opposed to the drafters of these other instruments, did not limit itself to only enumerating desiderata without presenting how they should actually be achieved and when they are considered to have been achieved. In fact, it also elaborated a document with the title *Ethics Guidelines for Trustworthy AI*[11] and another one with the title *Assessment List for Trustworthy Artificial Intelligence (ALTAI)*.[12] Both are useful for exploring the notions that underpin the AIA and in part also the AI Liability Act.

2.1. Desiderata of an Ethical and Trustworthy AI in the AI HLEG Preparatory Documents

Observance of ethical principles during the development of AI systems and the coordinates of their trustworthiness according to the AI HLEG must be subject to constant monitoring and later to review during the operation of the AI system. Ethical principles according to the AI HLEG are deemed to have been adequately considered[13]

7 The Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 2019.

8 Guide on Article 6 of the European Convention on Human Rights. Right to a Fair Trial (Civil Limb) 2022. 58–59. See the case-law of the European Court of Human Rights as cited in the Guide, for example, in cases *Guðmundur Andri Ástráðsson v Iceland* (GC), 2020, §§ 207 and 211 as well as *Pasquini v San Marino*, 2019, §§ 103 and 107 (court established by law, judges appointed according to the law); *Xero Flor w Polsce sp. z o.o. v Poland*, 2021, §§ 245–251 (court operating with impartiality and independence according to the law).

9 European Convention on Human Rights 1950.

10 Charter of Fundamental Rights of the European Union 2012.

11 Independent High-Level Expert Group on Artificial Intelligence 2019.

12 Independent High-Level Expert Group on Artificial Intelligence 2020a.

13 Independent High-Level Expert Group on Artificial Intelligence 2019. 9.

if the 'moral and legal entitlements' (specifically those constituted by fundamental rights enshrined in EU Treaties and the EU Charter) have been considered during its development. Thereby, any AI development must by definition be human-centric in keeping with the higher-order principles of human dignity, freedom of the individual, respect for democracy, justice and the rule of law, equality, non-discrimination, solidarity, as well as respect for citizen's rights.[14] These principles are translated by the AI HLEG into the world of AI by the desiderata of respect for human autonomy, prevention of harm, fairness, and explicability.[15] In this context, human autonomy is understood as being respected when AI is not utilized to manipulate, coerce, or otherwise direct the behaviour of humans to an 'unjustifiable' degree, and it must ensure human oversight of the operation of AI systems. Such systems must be designed not to cause harm to human beings, or to 'exacerbate' harms already caused, and they should be designed so as to prevent possibilities of abuse. Furthermore, AI must be developed with fairness. The AI HLEG distinguishes here between substantive (material) and procedural (formal) meanings of fairness. In the material sense, fairness would demand an equitable distribution of gains and risks, as well as the desiderata of proportionality between means and ends, while ensuring non-discrimination (both as bias and as stigmatization). Importantly for AI adjudication, procedural fairness would entail a right to an effective remedy against adverse AI outputs, even if they have been authorized by a human operator. This latter desideratum strongly associates procedural fairness with explicability. It is here that we find the Achilles heel of the ethics-based approach by the AI HLEG, as on the question of explicability a major compromise takes place. The desideratum of explicability is defined thus:

> (…) processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.[16]

14 Independent High-Level Expert Group on Artificial Intelligence 2019. 10–11.
15 Id. 12–13.
16 Id. 13.

Simply put, explicability may be considered as observed when 'the system as a whole respects fundamental rights'. The question may justly be asked: what does the AI HLEG mean by a system which, while not subject to explicability due to the very mechanism of its operation, still respects fundamental rights? To illustrate the problem, we would like to point out that the way things stand, as in the case-law of the European Court of Human Rights, the right to a reasoned (i.e. 'explicable') decision is in and of itself a fundamental right.[17] It is also known that AI may be non-transparent by its very nature, i.e. unable to provide reasons for even correct outputs (a problem referred to by the AI HLEG). This problem is known in the literature as 'opacity'.[18]

The tension between the principle of human autonomy and prevention of harm is emphasized by the AI HLEG;[19] however, the expert group seems to have ignored the fundamental synergy between explicability and fairness (constituting two facets of the right to a fair trial in the judicial context) when considering the possibly opaque way in which AI operates.[20] Even more disturbingly, the fall-back solution proposed by the AI HLEG for cases of opaque AI operation, called 'other explicability measures', also fails to address this issue, as no amount of traceability and auditability of an AI system will result in obtaining a true 'reasoning' from a 'black box' AI. Such measures may help with attaining output-based legitimacy[21] of the AI adjudicator (by e.g. verifying during an audit that a human judge would have reached a similar solution or that the solution is in keeping with the prevailing case-law); however, this will do nothing to ensure a reasoned decision. It seems that while for some other applications development and use of non-transparent AI may be considered ethical, based on the de minimis set of fall-back measures, these are not apt for solving the tension between explicability and fairness in the case of adjudication: fairness in the procedural sense is unattainable without explicability. This results prima faciae in an incompatibility of opaque AI with applications in the field of adjudication.

2.2. The ALTAI Assessment Criteria and Their Potential Impact on AI Used for Adjudication

The ALTAI assessment criteria also emphasize respect for fundamental rights (even if the right to a fair trial is not mentioned).[22] The first assessment criterion (Requirement #1) of the list, which verifies conditions of human agency and

17 Guide on Article 6 of the European Convention on Human Rights. Right to a Fair Trial (Civil Limb) 2022. 96–97; Fink 2021.

18 See Wischmeyer 2020.

19 Independent High-Level Expert Group on Artificial Intelligence 2019. 13.

20 For a detailed presentation of the most significant problems posed by lack of explicability in AI adjudication, and the requirement of explainable AI (xAI), see Deeks 2019. For a critique of explicability as understood by the European legislator (i.e. in senses other than a fully human-readable, clearly reasoned decision), see Edwards–Veale 2017. 65 et seq.

21 Chesterman 2021. 275.

22 Independent High-Level Expert Group on Artificial Intelligence 2020a. 5–6.

oversight, asks, inter alia, whether the AI system may generate the overreliance of human operators (a risk known as automation bias).[23] This criterion is based on the foregone conclusion that automation bias exists as a phenomenon and may constitute a problem in the realm of AI use.[24] Already in this stage of assessment, the possible 'black box' operational model, the inability of AI to produce the reasoning for its output poses problems. Evaluation of the risk of automation bias is strongly linked to the output-based legitimacy of the AI system, whereas, as recent research has pointed out, human operators tend to be biased by automated systems (just as they tend to do in case of human advice) selectively,[25] i.e. when the advice given is in line with their pre-existing biases. This manifestation of the automation bias is more difficult to guard against when no reasoning for AI output is present.

As part of Requirement #1 of the ALTAI assessment criteria, human oversight of the AI system also must be evaluated.[26] Four situations of such oversight are considered as possible by the AI HLEG: fully autonomous systems, human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command systems (HIC).[27] By projecting these operational models onto possible AI implementations in adjudication, we may find that full automation is not an option (as the human factor must be relied upon at least during the enforcement phase of some judicial decisions), and HOTL solutions would make it impossible for the human factor to be effectively involved in the adjudication activity undertaken by the AI outside the design and operational monitoring phases of implementation. HITL and HIC solutions are the most likely compliance options with Requirement #1.

The main ALTAI criterion that must be considered when contemplating AI adjudication is Requirement #4, which refers to transparency.[28] This is actually constituted of a subset of several coordinates for evaluation, namely: traceability, 'explainability' (the exact wording used in ALTAI Requirement #4, for which we shall use the notion of 'explicability' in line with the terminology in the AI HLEG Ethics Guidelines for Trustworthy AI), and communication. Traceability refers to the ability to document the source of the data, the content of the procedures (models)

23 For a discussion on automation bias, see Skitka–Mosier–Burdick 1999.

24 For a recent discussion on automation bias during use of AI applications, see Zuiderveen Borgesius 2020.

25 Alon-Barkat–Busuioc 2022.

26 Independent High-Level Expert Group on Artificial Intelligence 2020a. 8.

27 'Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system. Human-on-the-loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal, and ethical impact) and the ability to decide when and how to use the AI system in any particular situation. The latter can include the decision not to use an AI system in a particular situation to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by an AI system.' Independent High-Level Expert Group on Artificial Intelligence 2020a. 8.

28 See Independent High-Level Expert Group on Artificial Intelligence 2020a. 14–15.

used in generating a given output by the AI system, and the quality assessment of the output, as well as the logging of outputs in the form of AI decisions or recommendations. The AI HLEG ALTAI criteria define explicability in a manner identical to that found in the *AI HLEG Ethics Guidelines for Trustworthy AI*, which have been presented above, but for its assessment they only provide for the AI operator to state whether it explained (whenever possible) the reasons for the AI output to the party these have later affected. Finally, the AI operator must communicate to the party subject to the activity of an AI that the party's interlocutor is an AI entity. This last criterion in the case of adjudication should involve a prior warning to parties that their case will be subject to automated adjudication or an AI entity will provide feedback to the adjudicator regarding the solution that is to be given.

Requirement #5 of the ALTAI criteria emphasizes diversity, non-discrimination, and fairness. Under this criterion, avoidance of unfair bias must be achieved through taking the necessary measures both in algorithm design and when compiling the dataset used for 'training' the AI, which must be representative of the persons subjected to the operation of the AI. Avoiding bias also involves monitoring outputs for detecting potential bias. When developing and monitoring AI, fairness must be assessed after considering several definitions of fairness and consulting the 'impacted communities' regarding the operation of the AI (an action which must be assessed not just as an element of fairness but also when verifying that stakeholder participation in AI development took place). The requirement for considering several definitions of fairness is problematic when AI is to be employed as an autonomous adjudicator or (more likely) a guide for a human judge. As aptly observed by John Rawls when developing his theory of justice,[29] fairness is by no means a unitary concept and may even depart from the 'contractarian' view of equality between the parties. Utilitarian and intuitionist views of justice (mainly manifested in the prioritization – based on various criteria – of interests to be conserved, before those that must be disregarded when faced with the need to resolve a problem such as a judicial dispute) may even collide.[30] A recent meta-analysis[31] of studies regarding measures taken to ensure the fairness of AI systems has also shown that this is more difficult to attain than simply filtering protected attributes (e.g. ethnicity, individual, social, and economic attributes, etc.) from the data and would require contrasting individual aspects of fairness with the currently engrained notion of collective fairness, something which cannot be done without a yet non-existent definition drawn from international human rights (case-)law. The ALTAI criteria impose considering several definitions of fairness, whereas they ignore the possibility that for some AI applications, such as adjudication,

29 See Rawls 1999. 15 et seq.
30 See Rawls 1999. 25–45.
31 Varona–Lizama-Mue–Suárez 2021.

a unitary definition (along the lines of the ideas present in the AI HLEG Ethics Guidelines for Trustworthy AI, perhaps defined in human rights case-law) should be utilized and, in fact, also standardized.

Requirement #7 of the ALTAI criteria imposes accountability on AI systems and their operators.[32] Auditability as the first component of accountability is considered as conceptually equivalent to traceability (ensuring sufficient documentation of the AI in its design and implementation to prevent and detect unintended functions), which is complemented by the possibility of independent third-party review in the form of an audit. It is here that we must consider the way in which 'auditability' during the judicial process is implemented: by means of the appeals process available to the parties. This type of specific 'auditability' is inexorably linked to the explicability of the output, not simply the functional oversight of the AI design and operation, as without a human-readable reasoning, no 'audit' of the legality of judicial decisions can be undertaken. Risk management is also paramount in the ALTAI criteria. This requires monitoring and reporting on potentially hazardous consequences of AI operation, with the involvement of third parties (including members of civil society). This again seems to contrast with the basic characteristic of the administration of justice as a highly specialized activity undertaken by highly specialized personnel engaged in providing a public service while also exercising public authority (imperium).

Keeping the problems identified above in view, it is in no way surprising that the AI HLEG when analysing regulatory requirements for AI in the law enforcement and justice sectors in another document, titled *Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence*,[33] concluded that even though there is a marked need for such applications, including during the organization of case material and the supervision of judicial outcomes to detect bias…

> deployment at greater scale generates risks and opportunities that are not yet fully understood. More research, scrutiny and deliberation are needed prior to formulating legal, ethical or policy guidance. It would be, therefore, important to launch a wide-spread policy debate in Europe (and beyond) on the development, use and impact of AI-assisted and AI-enabled decision-making systems in justice and law enforcement.[34]

The AI HLEG practically abdicated from considering specific policies for such systems at the current moment. This procrastination did not prevent the European Commission from proposing regulation for such situations in the AIA (even if the regulation – as we shall see – is rather scant).

32 Independent High-Level Expert Group on Artificial Intelligence 2020a. 21–22.
33 Independent High-Level Expert Group on Artificial Intelligence 2020b. 10–11.
34 Id. 11.

3. Principles Governing the 'AI Judge' as Reflected in the Language of the AIA

The European Commission recognized in the very text of the AIA, in Recital (3), that AI may contribute greatly to the administration of justice inter alia. In Recital (40) of the AIA, however, the Commission proposed that such systems should be considered as high-risk AI…

> considering their potentially significant impact on democracy, rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial. In particular, to address the risks of potential biases, errors and opacity, it is appropriate to qualify as high-risk AI systems intended to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts.

In the very same recital, the Commission attached a caveat to this proposal, stating: 'Such qualification should not extend, however, to AI systems intended for purely ancillary administrative activities that do not affect the actual administration of justice in individual cases, such as anonymisation or pseudonymisation of judicial decisions, documents or data, communication between personnel, administrative tasks or allocation of resources.'

Thus, the Commission instituted a two-tiered approach to AI regulation when administration of justice is concerned. AI used during adjudication (finding facts, interpreting and applying the law to them) or which assists such activities must be deemed high-risk, while AI used only in organizing the administration of justice is to be considered low-risk. We tend to find this approach imperfect. The activity of (civil) courts during adjudication far exceeds finding and interpreting facts, and then applying the law, as has been shown[35] in the literature, also involves considering the parties' submissions, organizing admissible evidence, or documenting the case based on legal literature, prevailing doctrine, and case-law in a fundamentally adversarial procedure. These activities all influence the outcome of adjudication, and any AI involved in them, even in a purely 'ancillary administrative' capacity, should be considered as high-risk.

Annex III of the AIA (which lists high-risk AI implementations according to Article 6(3) of the regulation) at point 8(a) specifically refers to 'AI systems intended to be used by a judicial authority or on their behalf to interpret facts or the law and to apply the law to a concrete set of facts'. The same critique that may be levelled against Recital (40) is also applicable in this case. Article 6(3) of the Regulation imposes considering AI implementations listed in Annex III as high-

35 See Gerber 2002.

risk 'unless the output of the system is purely accessory in respect of the relevant action or decision to be taken and is not therefore likely to lead to a significant risk to the health, safety or fundamental rights'. Whether any given output is accessory is set to be decided by the European Commission through the adoption of implementing instruments to the Regulation. Therefore, some clarity of whether any given AI applied to adjudication is or is not high-risk shall only be achieved once these norms have also been adopted.

Both Recital (40) and the context of Annex III of the AIA seem to hint that AI participation during adjudication was considered by the Commission as a possibility subject to regulation, even though the AI HLEG in its Sectoral Considerations on the *Policy and Investment Recommendations for Trustworthy Artificial Intelligence* specifically stated that further research is necessary before guidance on the topic (including its regulation) may be offered.

Chapter 2 (articles 8–15) of the AIA sets forth the specific rules for utilizing high-risk AI implementations. Analysing those rules contained in this chapter that are of interest to our inquiry, we may quickly ascertain that their content is only partly reminiscent of those contained in the AI HLEG preparatory documents.

In order to ensure fairness, for example, Article 10(3) of the AIA imposes that datasets used must be statistically representative. Article 10(4) furthermore requires that these consider the purpose of the AI as well as the location and the 'behavioural or functional setting' in which it will be deployed. Article 10(5) makes bias monitoring compulsory for providers of the AI implementation in the measure required to assure intended functioning, a purpose for which processing of data that would be otherwise prohibited by the GDPR is allowed.

Article 13 (with the marginal title *Transparency and Provision of Information to Users*) of the AIA markedly departs from the exigence of transparency as proposed by the AI HLEG. The text only requires that during the design and operation of the AI system interested parties be informed of the major characteristics of this system, of its abilities and limitations as well as the identity of the operator. No requirement of explicability is set forth at all. This solution, called by some authors in its version present in the GDPR as a 'transparency fallacy',[36] is in no way compatible with the notion of a reasoned decision, as it only provides possible technical information on 'how' but not on 'why' the AI has reached a given decision.

Human oversight (Article 14) was regulated in the AIA; however, the human 'overseer's' abilities to monitor and influence the AI system are permitted to be limited by the nature of the system (by the 'as appropriate to the circumstances' clause in Article 14(4) first sentence of the AIA). Human oversight is thus mainly relegated to monitoring, constant awareness of possible automation bias, and interpreting outputs (Article 14(4)(a) to Article 14(4)(c)). The human 'overseer' must also have the ability not to stop utilizing the AI system at any time 'or otherwise

36 Edwards–Veale 2017. 65–67.

disregard, override or reverse the output of the high-risk AI system' (Article 14(4)(d)) and 'to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure' (Article 14(4)(e)). These two requirements impose the HIC model of human oversight to AI. As such, they seem to exclude any AI with a function in adjudication from operating independently of a highly human-qualified factor (such as a judge) who oversees its activity, as the AI's output must remain under human control. We should keep in mind here that this requirement, just as the rules cited above, is enacted subject to the 'as appropriate to the circumstances' clause and may as such be dispensable.

Record keeping (Article 12), as well as the existence of an appropriate risk management system (Article 9), based on ex ante analysis of risks and on monitoring operation, are also provided for; these rules are, however, not central to our inquiry.

4. Accountability for Malfunctions of the 'AI Judge'

Accountability for the malfunctions of an AI (manifested as erroneous output or lack of output) is underscored several times by the AI HLEG documents and the vast majority[37] of similar works. Yet these documents fail to elaborate on the practical ways in which accountability should be achieved, apart from emphasizing pre-emptive measures and monitoring, with little to no mention of ex post facto liability issues. In fact, a recent meta-analysis of concepts used in most, if not all, international policy documents on AI to date (2022) showed that while the notion of 'accountability' is quasi-ubiquitous,[38] this is not correlated with 'liability'. The latter is in fact conspicuously absent from among the regulatory priorities in the field of AI.

This holds all the truer for AI employed during or for the purposes of adjudication. A good example for this phenomenon is the *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment* developed by the European Commission for the Efficiency of Justice (CEPEJ). This evidently partisan policy document, which aims to represent exclusively the interest of the judiciary in AI development, mentions the notion of 'liability' only twice: once in the context of European Court of Justice case-law, which requires that Member States compensate damage resulting from the egregious breach of Community law by their courts, and another time, when discussing the risks of liability presented by AI in the case of judges who chose to decide against predictive algorithms.[39] The main form in which accountability is manifested under civil law would, of

37 Lupo 2022. 628.
38 Lupo 2022. 627 et seq.
39 European Commission for the Efficiency of Justice 2019. 24, 56.

course, be civil (i.e. pecuniary) liability, yet this concept seems somehow alien to most AI policy documents.

The proposal for an AI Liability Directive that was tabled by the European Commission aims to treat non-contractual civil liability issues arising out of the operation of AI systems. The scope of the proposed directive (Article 1) is not limited to liability between subjects of private law, and therefore its use, if adopted, will be conceivable in cases when liability for the erroneous results of AI-aided or AI-generated outcomes during adjudication are concerned. The AI Liability Directive primarily imposes obligations on the 'provider' of an AI system (Article 2(3) of the directive). The provider is defined in Article 3(2) of the AIA as being 'a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge'.

In some cases, the AI Liability Directive may also be employed against the 'user' of an AI system (Article 2(3) of the directive), as defined in Article 3(4) of the AIA: 'any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity'. It is this latter situation in which most AI systems in the administration of justice will be utilized.

The European Commission, when drafting the AI Liability Directive, could have opted for a model of strict, or 'no-fault', liability or for imposing compulsory insurance on AI system providers and users for cases when damage is caused in a non-contractual setting. These policy options, though considered during the impact assessment of the Directive by the Commission (and explicitly referred to in the Explanatory Memorandum published as the introductory part of the proposed Directive), were discarded in favour of maintaining the 'fault-based' liability model, which constitutes the general rule of non-contractual liability in most Member States.

The difficulty posed by this model is that it requires that both proof of fault and proof of causation between the fault and any damage be provided by the aggrieved party. In the case of AI systems, such proof is near-impossible due to the substantial number of participants during the development of the systems, the proprietary nature of some, if not most, components, and its characteristic autonomy that makes some AI outputs (lack of output) less than perfectly predictable. To ease the evidentiary requirements and facilitate compensation for damage caused by AI systems, the AI Liability Directive institutes two presumptions and a requirement to provide evidence for parties providing or operating high-risk AI systems.

As a first measure, Member States are required to ensure that their courts may order disclosure of evidence by the defendant if a high-risk AI is suspected of having caused damage, such evidence was already requested by the aggrieved party, that party has undertaken all reasonable measures to gather the evidence

of its own effort and has demonstrated to a sufficient degree that the defendant is likely to be in possession of the relevant evidence (Article 3(1) and Article 3(2)). Any evidence disclosed must be proportionate to the claim, and the court must preserve the confidentiality of proprietary information (Article 3(4)).

In case the defendant does not comply with the court order for disclosure of evidence, it may be presumed not to have complied with its duty of care imposed by the AIA (as per Article 3(5), Article 4(2) and Article 4(3) of the AI Liability Directive, as well as Article 10(2), Article 10(4), Article 13 to Article 16(a), Article 16(g), Article 21 and Article 29 of the AIA).

Demonstrated or presumed non-compliance with the duty of care in the form of the breach of AIA provisions enumerated above also results in a presumption of causation between the fault of the defendant and the damaging output (lack of output) by the AI system if this causality link is otherwise 'reasonably likely', and it has been demonstrated that the AI's output (lack of output) was the cause of the damage suffered (Article 4(1) of the AI Liability Directive). Thus, in a way relevant to our inquiry, in cases when the AI system was not developed and trained in a way that insures non-biased output, or that do not comply with the transparency or human oversight, requirements laid down in the AIA presumption of causation between these factors and the damaging output (lack of output) may be presumed.

We may observe here that due to the view enshrined in the AIA regarding transparency, the lack of sufficient reasons for an AI output may not be invoked by the claimant to benefit from the presumption of causation between the fault of the provider (or the user) and the AI output regulated by the AI Liability Directive, even if the right of the aggrieved party to receive such a reasoned decision constitutes a fundamental right. Also, in lack of a reasoned decision, as we have shown, no proper human oversight of the 'AI judge' may be achieved. Only this latter situation may be invoked as a basis for employing the presumption of causation laid down in the AI Liability Directive. Therefore, whereas other fundamental rights, such as non-discrimination, are better positioned to be protected by the AI Liability Directive, such protection is less clear in the case of the right to a reasoned decision.

5. Concluding Remarks

In our study, we have attempted to analyse the currently proposed regulatory framework in the European Union for the field of AI implementations in the light of the prospect of AI-aided or AI-generated adjudication, the so-called 'Ai judge'. We have found that preparatory documents, such as those drafted by the AI HLEG, have identified the risks posed by AI systems in a general manner and have not duly concentrated on ensuring a fair trial in case such systems would be employed for adjudication. Specifically, the basic right to a reasoned decision, something that as of

yet seems to exceed the abilities of AI systems, has not been adequately considered among the various exigences of AI transparency even if the lack of such a decision (even of an administrative, not just judicial nature) may make exercise of judicial remedies only an illusory possibility, also affecting human oversight of AI outputs. We have also observed that the regulatory proposals by the European Commission in the form of the AIA and the AI Liability Directive do not tend to provide solutions for the problem of a lack of sufficient reasoning. We, therefore, consider that along with future research in this field, regulation is also necessary that specifically deals with the implications of AI use during the administration of justice.

References

ALON-BARKAT, S.–BUSUIOC, M. 2022. Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory* 2022/February: https://doi.org/10.1093/jopart/muac007.

CARLO, C. Di–DE BONDT, M.–EVGENIOU, T. 2021. AI Regulation Is Coming. *Harvard Business Review* 2021/September–October. https://hbr.org/2021/09/ai-regulation-is-coming.

CASOVAN, A.–SHANKAR, V. 2022. A Framework to Navigate the Emerging Regulatory Landscape for AI. *OECD. AI Policy Observatory*. https://oecd.ai/en/wonk/emerging-regulatory-landscape-ai.

CHESTERMAN, S. 2021. Through a Glass, Darkly: Artificial Intelligence and the Problem of Opacity. *The American Journal of Comparative Law* 69(2): 271–294. https://doi.org/10.1093/ajcl/avab012.

DEEKS, A. 2019. The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review* 119(7): 1829–1850.

EDWARDS, L.–VEALE, M. 2017. Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review* 16(18): 18–84.

EUROPEAN COMMISSION. 2022. *Proposal for a Directive of the European Parliament and of the Council on Adapting Non-contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) COM(2022) 496 Final*. https://ec.europa.eu/info/sites/default/files/1_1_197605_prop_dir_ai_en.pdf.

EUROPEAN COMMISSION, DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY. 2021. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM/2021/206 Final)*. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.

EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE. 2019. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment*. Strasbourg: Council of Europe. https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c.

FINK, M. 2021. The EU Artificial Intelligence Act and Access to Justice. *EU Law Live* 2021/May. https://eulawlive.com/op-ed-the-eu-artificial-intelligence-act-and-access-to-justice-by-melanie-fink/.

FRANCESCONI, E. 2022. The Winter, the Summer and the Summer Dream of Artificial Intelligence in Law. *Artificial Intelligence and Law* 30(2): 147–161. https://doi.org/10.1007/s10506-022-09309-8.

GERBER, D. J. 2002. Comparing Procedural Systems: Toward an Analytical Framework. In: *Law and Justice in a Multistate World: Essays in Honor of Arthur T. Von Mehren*. Ardsley.

INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE. 2019. *Ethics Guidelines for Trustworthy AI*. Brussels. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

2020a. *Assessment List for Trustworthy Artificial Intelligence*. Brussels. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

2020b. *Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence*. Brussels. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

LUPO, G. 2022. The Ethics of Artificial Intelligence: An Analysis of Ethical Frameworks Disciplining AI in Justice and Other Contexts of Application. *Oñati Socio-Legal Series* 12(3): 614–653.

RAWLS, J. 1999. *A Theory of Justice. Revised Edition*. Cambridge (Massachusetts, USA).

SKITKA, L. J.–MOSIER, K. L.–BURDICK, M. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies* 51(5): 991–1006. https://doi.org/10.1006/ijhc.1999.0252.

SZEKELY, J. 2019. Lawyers and the Machine. Contemplating the Future of Litigation in the Age of AI. *Acta Universitatis Sapientiae, Legal Studies* 8(2): 231–244.

THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS. 2019. *Building Trust in Human Centric Artificial Intelligence*, (COM(2019)168). https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence.

VARONA, D.–LIZAMA-MUE, Y.–SUÁREZ, J. L. 2021. Machine Learning's Limitations in Avoiding Automation of Bias. *AI & Society* 36(1): 197–203. https://doi.org/10.1007/s00146-020-00996-y.

WISCHMEYER, T. 2020. Artificial Intelligence and Transparency: Opening the Black Box. In: *Regulating Artificial Intelligence*. Cham. 75–101. https://doi.org/10.1007/978-3-030-32361-5_4.

ZUIDERVEEN BORGESIUS, F. J. 2020. Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence. *The International Journal of Human Rights* 24(10): 1572–1593. https://doi.org/10.1080/13642987.2020.1743976.

*** Charter of Fundamental Rights of the European Union. 2012. Official Journal C 326 26.10.2012: 391–407.

*** European Convention on Human Rights. 1950. https://www.echr.coe.int/documents/convention_eng.pdf.

*** *Guide on Article 6 of the European Convention on Human Rights. Right to a Fair Trial (Civil Limb)*. 2022. Council of Europe – European Court of Human Rights. https://www.echr.coe.int/documents/guide_art_6_eng.pdf.

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 81–97

DOI: 10.47745/AUSLEG.2022.11.2.05

The Limits of Protection of Human Rights in Warfare Led by AI

Csongor Balázs VERESS
PhD student at National University of Public Service (Budapest, Hungary)
Research Fellow at Mathias Corvinus Collegium (Budapest, Hungary)
e-mail: csongor.veress@mcc.hu

Abstract. We live in a world of increasing geopolitical tensions, and we are witnessing a military artificial intelligence arms race. In this competition, the goal is to develop lethal autonomous weapons systems, sometimes called 'slaughter-bots', which use AI to recognize, select, and eliminate human targets without any person's assistance. The protection of human rights is difficult in all wars, especially in today's warfare. Innovative technologies offer a way to accomplish political aims, for instance, in the grey zone between war and peace. New technological advances, however, may provide choices to better recognize, comprehend, protect against, and counter hybrid threats. Therefore, it is crucial for academia and industry to have a thorough grasp of the ramifications of cutting-edge technology in a hybrid warfare/conflict setting as well as for political, civilian, and military leaders and decision-makers. In my study, I attempt to examine the role of AI in modern warfare, the weapons used in it, and how human rights can be protected in such circumstances. The morality of giving robots the authority to choose who lives and who dies on the battlefield is discussed in relation to lethal autonomous weapons systems.

Keywords: artificial intelligence, lethal autonomous weapons, hybrid warfare, cyber warfare

1. Introduction. Some Thoughts on Modern Warfare

We live in the age of fourth-generation warfare, also known as hybrid warfare. There is a distinct hybrid manner of fighting that now pervades all armed conflict. Its focal point is not largely in the military sphere in contrast to military-centric dynamic combat. The practical form of hybrid warfare can be unexpectedly fresh and vary from case to case, yet it is far from original in its essence. Three important traits and their hybrid orchestration help to distinguish this warfare in the limited meaning, which is of a strategic nature:

1. Focusing the decision of war/conflict primarily on a broad spectrum of non-military centres of gravity.
2. Operating in the shadow of various interfaces, such as between war and peace, friend and foe, internal and external security, civil and military domains, state and non-state actors.
3. Utilizing a creative combination, hybrid orchestration and the parallel use of different civil and military, regular and irregular, open as well as covert means, methods, tactics, strategies and concepts of warfare, thereby creating 'ever-new' mixed hybrid forms.[1]

It must be stressed that hybrid warfare has the ability to involve all escalation stages even if hybrid warfare actors typically deploy innovative and indirect techniques of limited conflict and a restricted use of armed force. The game will always involve friction and uncertainty, and the use of force that is thought to be reasonable may escalate. In order for hybrid warfare participants to achieve their political objectives, a military decision as such is not always necessary due to its concentration on a wide range of non-military centres of gravity. While pursuing a decision on a non-military centre of gravity, the hybrid warfare actor may be able to stop its opponent from determining the conflict on the military battlefield. In this situation, morale and legitimacy can be effective weapons.

Due to its enormous potential for surprise and offensive action, especially against militarily stronger adversaries, hybrid warfare often favours the offensive. By acting covertly in the murky spaces between interfaces, concealing or credibly denying an actor's intention and position as a combatant, and using force sparingly and only as a last resort, this builds on the ability to ambiguously resolve conflicts. Hybrid actors have the ability to create new situations that are nearly impossible to reverse later without expending excessive effort by employing long-term, indirect, or veiled 'salami tactics' or, alternatively, by carrying out swift, unexpected offensive operations and obtaining a fait accompli. As a result, the offensive capability of hybrid warfare poses a unique challenge to the defender: being caught off guard and not even realizing that one is under hybrid attack until it is too late.

Future conflicts will be waged differently as a result of a technologically advanced and interconnected globe. Cross-domain connectivity and the virtualization of functions in military forces and societies are being driven by the tremendously dynamic, continuing technological race. This merges real life and virtual reality, as well as personal and professional lives. The success of the Russian and Chinese rise in military strength across all operational domains – space, cyber, air, sea, and land –, which lies at the absolute centre of their excellence in hybrid warfare, has been largely attributed to the combination of the potential of new technologies and the subsequent development of operational concepts. In the areas of anti-access and area denial, they have developed capabilities such as ballistic and cruise

1 Schmid–Thiele 2019. 213.

missiles, offensive cyber weapons, and electronic warfare. They are now the West's technological and military opponents, and they are starting to gain an advantage.

Hybrid techniques and tools are accelerated by new technologies. They aid to increase the range of hybrid players' activities and their chances of success by enhancing the initial conditions for hybrid action and growing their toolkits. New technical advancements may provide choices to better recognize, comprehend, defend against, and defeat hybrid threats at the same time. It is crucial for political, civilian, and military officials and decision-makers, as well as for academia and industry to develop a shared and thorough knowledge of the consequences of new technologies in a fourth-generation conflict context in order to prevent, deter, and, if necessary, outmanoeuvre hybrid opponents.

There can be as much as 19 technologies that are relevant for the evolution of hybrid challenges, namely:

> 5G; additive manufacturing; artificial intelligence; autonomous systems; biotechnology; cloud computing; communication networks; cyber and electronic warfare; distributed ledger; directed energy; extended reality; hypersonics; the Internet of things; microelectronics; nano-materials; nuclear modernization; quantum sciences; space assets; and ubiquitous sensors. These emerging technologies are likely to drive developments in hybrid conflict/warfare in the coming years. Seven of these technologies would appear to have a prominent role and have been examined in more depth: 5G; artificial intelligence; autonomous systems; cyber and electronic warfare; extended reality; quantum sciences; and space.[2]

Artificial intelligence enables data to reach its full potential. It is the top choice for strategic anticipation, enhancing judgment and situational awareness, targeting, and serving as a crucial enabler of human–machine teaming. Autonomous systems have emerged as a crucial tool for both virtual and real-world applications. They will inevitably bring a large number of people to the area of combat. Some of them will serve in crucial missions as disposables. Electronic and cyber warfare are important facilitators of hybrid threats. Spectrum warfare is a new, essential capacity for mission success that combines cyber and electronic warfare.

In essence, hybrid warfare and conflict are nothing new. However, technological developments point to a significant expansion of the spectrum of hybrid dangers. They provide new opportunities for aggression and the use of force in a hybrid warfare/conflict setting because of their disruptive potential. Hybrid techniques and tools are accelerated by new technologies. They aid to increase the range of hybrid players' activities and their chances of success by enhancing the beginning conditions for hybrid action and growing their toolkits. Modern technology offers

2 Schmid–Thiele 2021.

a means of achieving political objectives at numerous interfaces, such as the line between war and peace. New technology advancements, however, may provide choices to better recognize, comprehend, protect against, and counter hybrid threats. Therefore, it is crucial for industry and academia to have a thorough knowledge of the ramifications of new technologies in a hybrid warfare/conflict context as well as for political, civilian, and military leaders and decision-makers.

2. The Use of Artificial Intelligence in Hybrid Warfare

What category of technology does artificial intelligence fall under? Machine learning and neural networks are two AI techniques that are utilized by researchers, businesses, and governments. Artificial intelligence is the use of machines, or computers, to replicate actions that are supposed to require human intelligence. Existing research on the trajectory of AI technology development indicates that there is a large deal of ambiguity, even among AI researchers, concerning the potential rate of advancements in AI.[3]

How will the development of artificial intelligence alter the way battles are fought?

> The answer, of course, depends. And it mainly depends on what type of wars are being fought. AI could very well change the fundamental nature of conventional conflicts between states. Technologies enabled by AI could become so powerful and ruthless that war as we know it becomes too deadly and costly to contemplate. But what about the shadow wars? What about irregular wars between states, non-state groups, and proxies? In other words, how will AI affect the type of wars that the United States is most likely to fight?[4]

In irregular warfare, where information and understanding supremacy might prove decisive by enhancing the speed, precision, and effectiveness with which information is used in these battles, AI will drive an evolution. However, developments in AI over the next ten years are unlikely to be revolutionary, especially in a type of warfare where people have historically outperformed hardware.

Armed forces all across the world are accelerating research and development due to the promise of AI, which includes its capacity to quickly and accurately improve everything from logistics and battlefield planning to human decision-making. Why militaries are interested in AI is illustrated by three possible application areas. The first issue is that many modern militaries confront the same data difficulty as businesses or the government at large – there is frequently too much data, and it

3 Horowitz 2018.

4 Egel–Robinson–Cleveland–Oates 2019.

is difficult to process it quickly enough. Narrow AI applications for information processing have the ability to accelerate the process of data interpretation, freeing up human labour for more complex tasks. Second, senior military and civilian authorities think that the pace of warfare is accelerating, from hypersonics to cyberattacks. Speed is about decision-making. For example, an aircraft piloted by AI and liberated from the restrictions of protecting a human pilot might exchange many of the benefits of having human pilots in the cockpit for speed and manoeuvrability, exactly like with remotely piloted systems. Third, AI might make a range of new military strategies for use in combat possible. But AI faces challenges when used as military adoption. AI systems are educated for highly specific tasks, such as playing chess or analysing photos, according to their specific nature. However, because of friction and what is called the 'fog of war', the environment quickly changes throughout combat, and AI systems may be unable to adapt.[5]

My research conducted in the military and defence fields aimed to learn how and where AI is now used by militaries and intelligence agencies around the world, as well as the potential benefits AI may soon bring to the industry. AI is useful in military and defence organizations for: 'Autonomous Weapons and Weapons Targeting; Surveillance; Cybersecurity; Homeland Security Logistics and Autonomous Vehicles'.[6] Computer vision is now used by autonomous weapon platforms to recognize and track targets. In order for a weapon to be considered autonomous, it must be able to recognize and track targets within the area it has been sent to protect. So, how can we be sure that this will work perfectly? How can anybody assure us that these weapons will not shoot civilian targets? Theoretically, nowadays there are no autonomous weapon platforms that are being designed to fire its ordnance without the express approval of a monitoring operator. There is, however, a significant demand for AI cybersecurity solutions. Due to the significant amount of danger involved with data breaches in military and defence networks, this seems reasonable in terms of cybersecurity. Machine learning appears to be used by a number of AI businesses and defence contractors to provide security products that can recognize and foresee threats before they have an impact on networks. Threats to cybersecurity arise in a variety of forms and dimensions. Artificial intelligence has the potential to significantly contribute to preventative actions.

Some theories about AI can be exaggerated, while others warn that a dystopian future with killer machines is more likely than we believe. An 'enabler' rather than a weapon, AI is viewed from a more measured perspective. When it comes to national security and defence, AI is best understood as a collection of tools and software that may assist militaries in resolving specific problems during a variety of tasks. Urban warfare is one of the biggest problems the American military is now

5 Horowitz 2018.
6 Roth 2019.

confronting. The U.S. military still has to work on how it prepares, equips, and organizes for operations in crowded urban settings – urban warfare specialists have noted. Additionally, as cities become progressively bigger and more complicated, the U.S. military will struggle to maintain its technological and operational edge in these crowded, disputed areas. Therefore, urban environments offer a useful test case for assessing the advantages, dangers, and ramifications of AI on the battlefield.

The speed and accuracy of decision-making on the urban battlefield will enhance with AI-enabled intelligence, surveillance, and reconnaissance. Because of the massive volumes of data that cities generate, intelligence, surveillance, and reconnaissance are one of the fields with most potential for AI applications in urban combat. Military and intelligence analysts can now access thousands of publicly available datasets for insights into the demographic, social, economic, and logistical characteristics of cities and their inhabitants thanks to advancements in high-fidelity sensing, image recognition, and natural language processing.

Automated intelligence processing has the potential to change the game. Currently, experts spend hours combing through pictures and videos taken by unmanned aerial vehicles. Urban warfare success depends on having accurate and timely intelligence about the enemy's capabilities, whereabouts, and actions as well as the topography, infrastructure, and people of the city. However, the sheer volume of information is disorienting. In addition to enhancing battlefield decision-making, AI has the ability to lessen the danger of casualties, 'friendly fire', and collateral damage in urban conflict. Some AI and machine learning developments might actually prolong urban violence. Technologies with AI capabilities will enhance force protection and sustainment, boosting survivability and lowering military casualties. Soldiers who are well equipped and protected can fight longer. For example, the use of drones has allowed the United States to carry out deadly counter-terrorism operations globally for almost 20 years now. Because drone strikes do not threaten the lives of U.S. military personnel and are thought to be relatively inexpensive, American public opinion continues to be mostly favourable despite conflicting information regarding their effectiveness and multiple instances of civilian casualties. This cycle might continue with AI-enabled devices and tools that improve urban combat survivability.[7]

When it comes to supporting human decision-making, artificial intelligence has emerged as one of the most crucial technologies for any country. Data-driven and algorithm-driven, AI will change practically every element of life, from how people are educated through how they earn a living to how they protect against attacks in almost every field. AI outcomes in the machine learning branch of AI are mostly influenced by training data. Trained algorithms currently function as 'black boxes'. As a result, these algorithms may exhibit opacity, flaws, or intentional manipulation. Making sure that AI development and integration processes are transparent, understandable, and verifiable will be essential.

7 Konaev 2019.

One of the fundamental technologies of the digital age is AI. Nowadays, established economic sectors are under pressure to change as a result of digitalization. As new added value arises by combining data with AI systems, this pressure will increase. The development of big data and AI technology is currently accelerating dramatically, with significant potential ramifications for business and industry, politics and society, as well as a variety of military applications. Warfare will become more sophisticated as a result of AI. The future of warfare and the effects of AI and machine learning in the military can best be seen as a set of enabling technologies that will be used across the majority of the military realm. It will considerably help push the boundaries in grey zones in hybrid warfare. AI presents a wide range of opportunities for enhancing the skills of those with great abilities in this area, resulting in a set of technologies and applications that can assist military in overcoming real-world problems during a variety of operations. Better cost-efficiency, lessening the burden on humans, and enhanced cyber capabilities are just a few of them. Instead of being used by humans as tools, AI-driven autonomous tools will become 'useful teammates' for them.

By accelerating the speed, accuracy, and efficacy with which information is used and rendered usable, AI technology will lead to an evolution in hybrid warfare where information superiority and awareness can prove crucial. AI will make it possible for group behaviours to be imitated, influenced, and changed in hybrid conflicts, hence influencing their social and economic repercussions. Artificial intelligence is a top priority for armed forces and intelligence services in coping with hybrid warfare eventualities because of its potential to streamline complex operations and make them more effective. For instance, it will become far more difficult to conceal soldiers, proxies, or their equipment, as facial recognition, biometrics, and behavioural signature identification technology becomes more commonplace. A nation-state can do a lot to combat a hybrid insurgency if it has an extensive intelligence-gathering, processing, and exploitation apparatus powered by AI. 'Aggressors will increasingly have the opportunity not merely to spread disinformation or favorable narratives or to damage physical infrastructure, but to skew and damage the functioning of the massive databases, algorithms and networks of computerized or computer-dependent things on which modern societies will utterly depend.'[8]

Over the next 10 to 15 years, artificial intelligence will be a key enabler of innovation, impacting every business as well as the nature of daily activities for private persons in their real world and online. Furthermore, as practical, real-world applications of AI have just recently begun to emerge, it is possible to overestimate the degree of AI's penetration into powerful economies or what the technology could do in the hands of a social manipulator. It is crucial to keep in mind that the pace of AI development or its potential applications over the coming ten years should not be overstated when assessing the implications for social manipulation and virtual

8 Thiele 2020. 11.

societal violence. Nevertheless, they are probably important.[9] 'Military decision making plays a key role across different domains – land, maritime, air, space, and cyber – and across organizational levels – strategic, operational, tactical, technical.'[10]

3. Cyber Warfare

It is widely known that computers are getting quicker and more commonplace. Every day, machines are becoming more powerful and smaller. Robotics, nano- and biotechnology, artificial intelligence, distributed ledgers, sensor technologies, and 5G are all examples of fundamental advancements. Prototypes, vehicle and weapon pieces, and other items are produced using additive manufacturing techniques (the better today's computers are, the more they contribute to enhancing future computers). As the vast potential of artificial intelligence is becoming more and more significant, its adoption by various actors is brisk, as are governmental authorities' and criminal actors' use of it. We may anticipate that a wide range of technology will contribute to hybrid warfare and its goals. To find hybrid operations, real-time analytics and anomaly detection will be key components. The ability to process online data streams will be crucial to a situational awareness, which alerts to certain actions, flags complex events, or highlights new developments. Sensors (the Internet of Things), people (social media), systems (logs), mobiles (locations), etc. are generating continuous and/or event-driven data.

Data now underpins any nation's might, both economically and militarily. Remotely connected robotics combines computers and automation in novel ways through data and communication networks. Data is the new oil in a world where connectivity is always present. And the new oil rigs are networks. Similar to how crude oil must be refined to produce useful goods like gasoline, data must be refined to provide knowledge that can be put to use.

The Internet, a massive global network for information transfer and a complicated, difficult-to-understand 'system of systems', is the backbone of the information or digital age. Digital transformation has had a significant impact on all facets of society, including business, the economy, and governmental sectors like security and defence.

4. LAWs against Human Rights

Lethal autonomous weapons (LAW), or 'killer robots', are drones (as opposed to the sci-fi concepts of humanoid robots, which are still very difficult to create and power), which would make up the majority of the proposed AI-driven weapons. They might

9 Mazarr–Bauer–Casey–Heintz–Matthews 2019. 68.
10 Kerbusch–Keijser–Smit 2018.

be produced at a lower cost and in a considerably smaller size than the current military drones. Although there are ways to use AI to lessen the collateral damage and negative effects of war, fully autonomous weapons would also introduce a number of new legal, ethical, practical, and strategic issues. For this reason, scientists and activists have urged the United Nations and other international governments to consider a pre-emptive ban. The simplest defence of autonomous weapons from a military standpoint is that they provide a plethora of new capabilities. There can only be a certain number of drones in the sky at once if each one must be piloted by a human who decides when the drone should shoot. Fully autonomous weapons will make it easier and cheaper to kill people. Some experts are in favour of LAWs:

> The most interesting argument for autonomous weapons (…) 'is that robots can be more ethical'. Humans, after all, sometimes commit war crimes, deliberately targeting innocents or killing people who've surrendered. And humans get fatigued, stressed, and confused, and end up making mistakes. Robots, by contrast, 'follow exactly their code' (…). Unlike human soldiers (…) machines never get angry or seek revenge.[11]

Fortunately, the current consensus on 'killer robots' among legal and military experts is that they would do greater harm. Some say that robots designed to follow the laws of war would not take morality into account. Soldiers occasionally go well beyond what the law allows them to do. However, other times they do better since they are human and are subject to moral as well as legal imperatives robots could not be subjected to. You can prevent both types of errors by collaborating between humans and machines, as they make distinct sorts of errors. Weapons may be created that are programmed to understand the laws of war and, as a result, will disobey any orders from humans that break those laws. These weapons would also not have the power to murder without human intervention.

5. The Possible Protection Afforded by Human Rights against Autonomous Weapons Systems

The rise of autonomous weapons systems, or weapons that allow computers to operate them rather than humans, has received a lot of attention in recent years. Artificial intelligence and other technologies have advanced significantly during the last ten years. These will enable the creation and use of completely autonomous weapons systems that, when activated, select, attack, destroy, or injure human targets while functioning effectively without direct human supervision. 'These

11 Piper 2019.

weapons systems are often referred to as lethal autonomous robotics (LARs), lethal autonomous weapons systems (LAWS) or, more comprehensively, autonomous weapons systems (AWS).'[12] The rapid development of these weapons systems raises extremely serious concerns regarding human rights, undermining the right to life, the prohibition of torture, and other forms of ill-treatment, the right to personal security, and other human rights. It could also fundamentally alter how military operations are conducted. Autonomous weapons systems can be made to have deadly or less deadly consequences and can be utilized in armed conflict and/or military scenarios. As they spread, it is likely that private companies, people, and armed non-state groups will start using them. Autonomous means weapons that can choose targets and launch an assault without real or effective human control that can guarantee the proper use of force. Such systems could have a negative impact on a person's human rights since they would employ violence (including less-than-lethal force) against persons. It is urgently necessary to pay attention to and consider the questions surrounding the development and potential use of autonomous weapons systems outside of armed conflict (and the ability of such systems to abide by human rights laws), as these issues are at least as challenging as those pertaining to their use on the battlefield. Only then will concrete steps that address this significant area of international law be taken.

There are five important human rights problems to take into account in the present autonomous weapons systems discussion: 1) the scope of the Convention on Certain Conventional Weapons does not cover non-conflict situations; 2) autonomous weapons systems will not be able to comply with relevant international human rights law and policing standards; 3) developments in existing semi-autonomous weapons technology pose fundamental challenges for the international human rights law framework; 4) in the absence of a prohibition, autonomous weapons systems must be subject to independent weapons reviews; 5) autonomous weapons systems will erode accountability mechanisms.[13] The issues identified are by no means exhaustive but rather seek to elicit the principal concerns around the potential use of autonomous weapons systems in military operations.

It would be fundamentally incompatible with international human rights legislation to utilize autonomous weapons systems, including less-than-lethal robotic weapons, in military operations since it would result in unjustified murders, injuries, and other human rights crimes. Additionally, the use of autonomous weapons systems would make it extremely difficult to hold people accountable for grave transgressions, and it could further cement the impunity for crimes against international law. Therefore, there is need for a preventative prohibition on the development, transfer, deployment, and use of autonomous weapons systems,

12 Amnesty International 2015.
13 Ibid.

including fully autonomous systems that use less-than-lethal weaponry but have the potential to kill or seriously injure people.

A key tenet of international human rights law is that no one's life may be taken arbitrarily. The right of everyone 'to life, liberty, and security of person' is upheld by the Universal Declaration of Human Rights (UDHR, Article 3). Every human being has the intrinsic right to life, according to Article 6(1) of the International Covenant on Civil and Political Rights (ICCPR). The law must defend this right. No one's life may be taken unlawfully. International human rights law states that this clause cannot ever be altered, waived, or suspended, not even 'in time of public emergency which threatens the life of the nation'. The right not to be arbitrarily deprived of one's life is therefore in theory applicable even in situations of outright armed conflict; however, in such areas, the definition of 'arbitrary' is typically decided by the provisions of international humanitarian law. Article 9 of the ICCPR safeguards the right to liberty and security of the person. This means that a person's freedom cannot be arbitrarily or unjustly taken away, and arbitrary detention or arrest is forbidden. The Human Rights Committee recently stated that the right to security of person 'protects individuals against intentional infliction of bodily or mental injury, regardless of whether the victim is detained or not. For instance, when they do bodily harm without justification, representatives of the States parties infringe the right to personal security.'[14] They further state that actors during conflicts 'should also prevent and redress unjustifiable use of force in [the] military and protect their populations against abuses by private security forces, and against the risks posed by excessive availability of firearms'.[15] Autonomous weapons systems could potentially be used to transgress laws against torture and other cruel, inhumane, or humiliating treatment or punishment. Similar to the ban on unlawful killings, deprivation and torture are also forbidden in all situations, including armed combat, and cannot ever be excused. No matter which international treaties a state has ratified, this prohibition is an absolute requirement of international law that must be abided by all parties.[16]

In particular, with appropriate attention to the protection of the rights to life and security of person, as well as the avoidance of torture and other ill-treatment, the international community has developed guidelines to help guide nations in guaranteeing human-rights compliant use of force in military. To be able to conduct lawful military operations, autonomous weapons systems would need to be able to effectively assess the degree to which there was an imminent threat of death or serious injury, correctly identify who posed the threat, consider whether force is necessary to neutralize the threat, be able to identify and use means other than force, be able to deploy different modes of communication and policing weapons

14 Amnesty International 2015.
15 Ibid.
16 Ibid.

and equipment, etc. To compound matters, each case would necessitate a distinct and one-of-a-kind answer, which would be incredibly difficult to reduce to a sequence of sophisticated algorithms. Without meaningful and effective human control and judgment, it is impossible for autonomous weapons systems to comply with these rules, especially in uncertain and ever-changing circumstances.[17]

Under Article 36 of the 1977 Additional Protocol I to the four Geneva Conventions of 1949 (henceforth, Article 36) and in accordance with international humanitarian law and other relevant international law, States Parties to the Convention are required to assess whether a new weapon, means, or technique of conflict is legitimate: 'In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.'[18]

For weapons and their usage to be compliant with international law, national legislation, and pertinent international and national standards, transparent weapons evaluations are therefore essential. In debates over autonomous weapons systems, an increasing number of states have claimed that Article 36 may offer a means to ensure that such systems will abide by international humanitarian law without the necessity for banning them. Although the discussion of and involvement with Article 36 is useful, it is insufficient for a number of reasons.

Firstly, it is unclear from Article 36 how the examination of weapons should be conducted. It is known that only a small number of nations have formal evaluation processes in place for new weapons. States that have created formal review processes have done so using various levels of specificity and according to various criteria. Additionally, there is sometimes a lack of openness and transparency in how, when, and how often states perform evaluations of their weapons.

Secondly, a review of firearms under Article 36 does not apply to all equipment and weapons, and it does not look at how they may be employed in policing and military activities. It is legal to employ some weapons in armed combat but not in policing, and the opposite is also true. As a result, such weapon assessments would not include certain dangerous and less-than-lethal autonomous weapons systems. Despite the fact that Article 36 also calls for States Parties to consider whether new weapons, means, and methods of warfare are permissible under any other rule of international law applicable to the High Contracting Party', which inevitably entails a review of compliance with international human rights law, this would only apply to military operations during armed conflicts. Therefore, as arguments about the application of Article 36 go on, states, civil society groups, technical, legal, and other experts who are now looking into the topic of autonomous weapons

17 Ibid.

18 United Nations 1949.

systems must fill this vacuum 'on the fly', i.e. in the absence of a restriction on such systems.

Important questions regarding individual criminal liability and responsibility for human rights breaches are raised by the creation, implementation, and usage of autonomous weapons systems. In order to hold those responsible accountable, every fatality and serious injury that occurs during a military operation must be the subject of a mandatory documentation and investigation under judicial oversight. In order for this to happen, there must be a comprehensive and open mechanism in place to hold military officers responsible for their choice to use force. This necessitates the presence of a third-party accountability mechanism with the authority to conduct thorough, fair, and impartial inquiries. States are required to respect the ban on the arbitrary taking of life and to take all necessary steps to stop, look into, punish, and make amends for the harm caused by private individuals or entities violating human rights. States are also required by international human rights legislation to investigate claims of human rights abuses and prosecute the offenders as part of the right to an actual solution, which is a right that is applicable at all times. It is impossible to bring a robot to court in the case of fatal and less-than-lethal autonomous weapons systems use. Instead, those engaged in the development, production, and operation of Autonomous Weapons Systems, as well as higher-ranking officers and political figures, could be held liable. However, given the numerous variables autonomous weapons systems may encounter, none of these players could possibly predict how it will respond in any given situation. Furthermore, without efficient human control, higher officers would not be able to stop an autonomous weapons system from engaging in illegal activities or discipline it for misbehaviour.

In addition, it is doubtful that autonomous weapons systems could adhere to global norms governing the use of force given the state of technology at the time and the impossibility that it could ever achieve the human levels of discretion necessary in the legal conduct of military action. Particularly questionable is the ability of autonomous weapons systems to uphold the fundamental human rights standards of legality, necessity, and proportionality. Without meaningful and efficient human management, deadly and less-than-lethal autonomous weapons systems would not be able to accurately assess complicated military scenarios and adhere to international standards that forbid the use of lethal force unless in defence against an immediate threat of serious harm or death. The issues that autonomous weapons systems bring up are not adequately covered by current international humanitarian law. The issues presented could be addressed by a new treaty to regulate systems of weapons in accordance with international humanitarian law, morality, international human rights legislation, responsibility, and security. Governments have been encouraged to begin negotiating a new

international treaty on killer robots.[19] By incorporating the following components, such an instrument would address the ethical, security, accountability, and legal issues that such systems raise:

– A broad scope that covers all weapons systems that select and engage targets on the basis of sensor inputs—that is, systems in which the object to be attacked is determined by sensor processing, not by humans;
– A general obligation to retain meaningful human control over the use of force;
– A prohibition on the development, production, and use of weapons systems that by their nature select and engage targets without meaningful human control;
– A prohibition on the development, production, and use of autonomous weapons systems that target people; and
– Positive obligations to ensure other autonomous weapons systems cannot be used without meaningful human control.[20]

During armed conflict, the rules of international humanitarian law must still dominate. For example, the international humanitarian law that governs armed conflicts contains an implicit need for human judgment. The principles of distinction, proportionality, and military necessity enshrined in international treaties like the 1949 Geneva Conventions and deeply rooted in international customary law specifically contain this need as an implicit part. International human rights law, which guarantees certain human rights for all persons regardless of their national origins or local regulations, also implicitly upholds similar concepts. Autonomous weapons systems raise a host of ethical and social concerns, including issues of asymmetric warfare and risk redistribution from combatants to civilians and the potential to lower the thresholds for nations to start wars.[21] There is a separate concern that such systems may not have an identifiable operator in the sense that no human individual could be held responsible for the actions of the autonomous weapons system in a given situation.[22]

The main technical question is still whether it is possible to create a robot that can recognize legitimate targets, such as military objectives, combatants, and civilians directly engaged in hostilities on the one hand and those protected from attacks by international humanitarian law, such as civilians and civilian objects, specially protected objects like cultural properties on the other. While it might be hard to program autonomous weapons for every scenario that can arise in a battle, could it be possible for them to 'learn'? It will be difficult to implement some components of international humanitarian law in a computer program, such as the concept of direct participation in hostilities. What constitutes an international

19 Human Rights Watch 2021a.
20 Human Rights Watch 2021b.
21 Asaro 2008.
22 Sparrow 2007.

armed war and a non-international armed conflict are the real issues at hand. What level of violence must occur before there is an armed conflict between a state and a non-state actor? These inquiries are not unique to robots, and human beings must provide the answers even when autonomous weapons are utilized.

6. Conclusions

In today's world, it is very difficult to protect human rights and privacy, as AI may be used as a weapon of hybrid warfare. Robotization of battlefields and the subsequent creation and integration of artificial intelligence in traditional weapons platforms will advance quickly. For example, when a conquering army wants to take a major city instead of troops fighting in the urban area, it is easier to send dozens of small drones with simple instructions: 'Shoot everyone holding a weapon.'[23]

The political and economic justification for waging wars rather than preventing them may change as a result of developments in autonomy and artificial intelligence that help lower the danger to military troops and enhance sustainment. Therefore, even while AI has the potential to exacerbate politics and war, this is not because it will aid in the automation of human labour. Instead, it is because AI will alter how various human groups – such as the military, civilian leadership in charge of making decisions about the use of force, and the general public that either approves of or criticizes these actions – relate to one another. In the AI era, war will still be politics carried out in a different way. But as war becomes less expensive as a result of technology and political leaders no longer have to pay as much for their actions, why cease fighting? Technological advancements that potentially alter civilian–military relations, political authority, and the methods used to conduct war have significant ethical ramifications. And it is these questions that should keep us up at night, not sensationalized images of self-aware terminators and self-driving drone swarms.[24]

Half a dozen countries are at various phases of developing lethal autonomous weapons systems or robotic weapons such as self-driving robotic vessels designed to travel thousands of miles to find and destroy submarines and sea mines with not even one crew member present. However, despite the fact that the militaries of developed nations are competing to develop lethal autonomous weapons systems to carry out a variety of tasks on the battlefield, a significant portion of robotic engineers, ethical analysts, and legal experts are adamant that robotic weapons will never meet the standards of distinction and proportionality required by the laws of war and will, therefore, be illegal.

23 Piper 2019.
24 Konaev 2019.

References

AMNESTY INTERNATIONAL. 2015. *Autonomous Weapons Systems: Five Key Human Rights Issues for Consideration*. https://www.amnesty.org/en/wp-content/uploads/2021/05/ACT3014012015ENGLISH.pdf (accessed on: 10.09.2022).

ASARO, P. 2008. How Just Could a Robot War Be? In: *Current Issues in Computing and Philosophy*. Amsterdam: 50–64.

2012. On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making. *International Review of the Red Cross*: https://international-review.icrc.org/sites/default/files/irrc-886-asaro.pdf (accessed on: 10.09.2022).

BÄCHLE, T.–BAREIS, J. 2022. "Autonomous Weapons" as a Geopolitical Signifier in a National Power Play: Analysing AI Imaginaries in Chinese and US military policies". *European Journal of Futures Research*: https://eujournalfuturesresearch.springeropen.com/counter/pdf/10.1186/s40309-022-00202-w.pdf (accessed on: 10.09.2022).

EGEL, D.–ROBINSON, E.–CLEVELAND, C.–OATES, C. 2019. AI and Irregular Warfare: An Evolution, Not a Revolution. *War on the Rocks*: https://warontherocks.com/2019/10/ai-and-irregular-warfare-an-evolution-not-a-revolution/ (accessed on: 10.09.2022).

HEYNS, C. 2016. Human Rights and the Use of Autonomous Weapons Systems (AWS) during Domestic Law Enforcement. *Human Rights Quarterly* 38: 350–378.

HOROWITZ, M. 2018. The Promise and Peril of Military Applications of Artificial Intelligence. *Bulletin of the Atomic Scientists*: https://thebulletin.org/2018/04/the-promise-and-peril-of-military-applications-of-artificial-intelligence/ (accessed on: 10.09.2022).

HUMAN RIGHTS WATCH. 2020. Stopping Killer Robots: https://www.hrw.org/report/2020/08/10/stopping-killer-robots/country-positions-banning-fully-autonomous-weapons-and (accessed on: 10.09.2022).

2021a. Killer Robots: Negotiate New Law to Protect Humanity: https://www.hrw.org/news/2021/12/01/killer-robots-negotiate-new-law-protect-humanity (accessed on: 10.09.2022).

2021b. Crunch Time on Killer Robots: https://www.hrw.org/news/2021/12/01/crunch-time-killer-robots (accessed on: 10.09.2022).

HUSAIN, A. 2021. AI Is Shaping the Future of War. In: *National Defense University Press*. https://ndupress.ndu.edu/Media/News/News-Article-View/Article/2846375/ai-is-shaping-the-future-of-war/ (accessed on: 10.09.2022).

KERBUSCH, P.–KEIJSER, B.–SMIT, S. 2018. *Roles of AI and Simulation for Military Decision Making*. https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-160/MP-IST-160-PT-4.pdf (accessed on: 10.09.2022).

KONAEV, M. 2019. With AI, We'll See Faster Fights, but Longer Wars. In: *War on the Rocks*: https://warontherocks.com/2019/10/with-ai-well-see-faster-fights-but-longer-wars/ (accessed on: 10.09.2022).

MAZARR, M.–BAUER, R.–CASEY, A.–HEINTZ, S.–MATTHEWS, L. 2019. *The Emerging Risk of Virtual Societal Warfare*. Santa Monica.

PIPER, K. 2019. Death by Algorithm: The Age of Killer Robots Is Closer than You Think. *VOX*: https://www.vox.com/2019/6/21/18691459/killer-robots-lethal-autonomous-weapons-ai-war (accessed on: 10.09.2022).

ROTH, M. 2019. Artificial Intelligence in the Military – An Overview of Capabilities. *Emerj Artificial Intelligence Research*: https://emerj.com/ai-sector-overviews/artificial-intelligence-in-the-military-an-overview-of-capabilities/ (accessed on: 10.09.2022).

SASSÓLI, M. 2014. Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified. *International Law Studies – US Naval War College*: https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1017&context=ils (accessed on: 10.09.2022).

SCHMID, J.–THIELE, R. 2019. Hybrid Warfare – Orchestrating the Technology Revolution. In: *NATO at 70: Outline of the Alliance Today and Tomorrow*. Bratislava: 211–226.

2021. *Hybrid Warfare: Future & Technologies*. Wiesbaden.

SPARROW, R. 2007. Killer Robots. *Journal of Applied Philosophy* 24: 62–77.

THIELE, R. 2020. Artificial Intelligence – A Key Enabler of Hybrid Warfare. *Hybrid CoE Working Paper* 6: https://www.hybridcoe.fi/wp-content/uploads/2020/07/WP-6_2020_rgb-1.pdf (accessed on: 10.09.2022).

UNITED NATIONS. 1949. *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol 1)*. https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-additional-geneva-conventions-12-august-1949-and (accessed on: 10.09.2022).

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 98–112

DOI: 10.47745/AUSLEG.2022.11.2.06

A General Overview of Artificial Intelligence and Its Current Implications in Civil Law

Emőd VERESS
DSc, Professor
Sapientia Hungarian University of Transylvania (Cluj-Napoca, Romania),
Department of Law
University of Miskolc (Hungary), Faculty of Law, Department of Commercial Law
ORCID: 0000-0003-2769-5343
e-mail: emod.veress@sapientia.ro

Abstract. In the course of this study, the author briefly presents some of the major issues raised by the prospect of artificial intelligence (AI) development in the field of civil law. Firstly, problems posed by possible AI agents acting for a natural or legal person (principal) are analysed, with the conclusion that as of yet liability for damage caused by the AI both to the parties of the juridical act concluded by the artificial agent and to any third parties remains with the owner or operator of the AI, with all the injustices this situation entails. Secondly. situations of liability for damage caused by use of an AI system for aiding decision-making are presented. It is shown that liability gaps exist in such situations due to lack of regulation. Thirdly, the possibility of AI-held (mostly non-pecuniary) intellectual rights is analysed, which in the light of current regulation and recent foreign case-law seems excluded. Finally, the possibility of granting legal personality to AI systems is raised as a possible solution to the aforementioned dilemmas. It is shown that this would be only an apparent solution, while legal personality for AI would entail greater risks, and is therefore to be avoided. It is concluded that further research and regulation may be necessary to resolve the problems that were identified.

Keywords: artificial intelligence, civil law, civil liability, artificial agent, intellectual property, legal person

1. Introduction

The topic of artificial intelligence (AI) has come to the fore in recent legal literature. Numerous current discussions are aimed at exploring the implications of new technology in law enforcement, public administration, and justice. The problems

presented by AI are indeed diverse: its complexity and opacity[1] (the inability of some AI applications to give reasons for their actions or inactions the way human beings would and the secrecy resulting from data protection requirements, or the proprietary nature of some of its elements), the existential threat that it may pose to humanity,[2] and the risks it already poses to human rights (be it in the form of mass surveillance or predictive functions[3] used in law enforcement) all must be considered. The legal and scientific literature in fact seems overwhelmed by these topics, whereas AI may also lead to major new developments in several fields of civil law, specifically in the domains of obligations (agency and contractual as well as non-contractual liability) and rights in rem such as intellectual property law, which also raise the prospect of legal personhood.

AI differs from most, if not all, previous technological leaps by its very nature: it is capable of autonomous, even independent action which, unlike in the case of machines or animals, is purpose-oriented. The basic modality in which AI operates presupposes the existence of pre-set goals usually determined by a human operator, data which is input into the AI algorithm from a pool that is provided by humans, and a processing mechanism which based on the input data and the pre-set goals is able to result in desirable outputs. There are several ways to achieve the outputs, which presuppose a greater or lesser extent of human oversight and interaction and which are more or less opaque depending on the particular technological solution used.[4] The exact way in which AI generates the output may be impossible to know or influence even for the most skilled operator due to the nature of the technical solution employed, making it the product of 'machine thought' combined with human-designed elements.

It must also be kept in mind here that the outputs an AI can generate are varied and diverse, unlike the products of any other technology. An AI may produce real-world consequences in ways only human beings could in the past, e.g. by directing vehicles, deciding and conducting business transactions, intelligently conversing, even playing games with humans, providing vital advice to human decision-makers such as doctors and judges, giving medical care or even creating art. There is a myriad of human–AI interactions that set AI apart from any other past technology and demand legal solutions.[5] All this stems from AI being a universal technology, even if artificial general intelligence (AGI), the human-like artificial intellect of science-fiction fame is most definitely a future, or even impossible development.

The functional similarity and possible intertwining between AI outputs and human actions or omissions creates an entirely new landscape to which a regulatory

1 For the meaning and implications of 'opacity' in the context of AI, see Burrell (2016), Chesterman (2021), and Wischmeyer (2020).

2 For an exploration of the long-term risks presented by AI development, see Harari (2017).

3 See Zuiderveen Borgesius 2020, Citron–Pasquale 2014, Knobloch 2018, Bertolini–Episcopo 2021.

4 For a non-technical description of the methods and processes (mainly machine learning and neural networks) on which AI implementations are based, see Boden (2018).

5 See Surden 2019. 1335–1337.

response is necessary, as existing rules cannot be shoehorned to fit new realities. Major problems are posed by AI acting as an artificial agent in legal transactions, as adviser (or replacement) of human decision-makers, and as creator of products protected by intellectual property rights or, indeed, other rights in rem. In the following, I aim to analyse some of these problems and draw some conclusions on regulatory priorities in future civil law norms.

2. AI as an Agent for a Principal

AI agents are an ever more widespread reality.[6] They are employed in economic transactions of various types and are able to interact with human counterparties;[7] they are also able to conclude their transactions based on purposeful autonomous reasoning.[8] These 'artificial agents' are not simple pieces of software executing human-generated orders for buying and selling whatever asset is being traded on a given market but are able to decide whether to conclude a transaction, and under what circumstances to do so. They exercise contractual freedom in a way that is not directly (or even indirectly) determined by their human 'principal' and are prone to the kind of risks any 'human' agent would be: concluding transactions that are egregiously disadvantageous to one of the parties, and therefore infringing on an obligation of contractual loyalty and equilibrium of performances, or exceeding their powers, conducting transactions outside the scope of their activity or failing to conduct transactions that are considered rational by the parties, leading to issues of liability.[9] The principal–agent relationship and also the agent–counterparty relationship must be considered in these situations. If the regulator is to address these issues, either some analogy must be found between the situations autonomous agents create and regulatory models that are already known, or, in the absence of such an analogy, new regulatory models must be invented.

The first and one of the worst problems legal science must contend with when analysing autonomous agents is one inherent to the theory of juridical acts: do these 'machines' possess capacity and will to conclude a juridical act (e.g. a contract) on behalf of another person under the law?[10] In fact, should they even be considered as an agent at all (as under most systems of law only a person may be an agent)?[11] After all, human agents are entitled to act on behalf of another based on a contract between them and their principals.

6 See Milana–Ashta 2021.
7 See March 2021.
8 See Kuo et al. 2021.
9 See Pagallo 2013. 89 et seq.
10 Chopra–White 2011. 29.
11 Tanna–Dunning 2022. 138.

Capacity is understood as a personal competence by the party to act reasonably in the conclusion of juridical acts, weighing advantages against disadvantages and deciding in a rational way for, or against, concluding the act; this definition of capacity is inexorably linked to that of will, as the formation of will requires an ability to conduct a reasoning, the way in which contractual will forms according to the will theory of contract.[12] In the case of an agent acting for a principal, the latter is the one who shall usually pre-determine the elements on which reasoning shall be based (e.g. the type and quantity of assets to be bought or sold, the price range, the date of delivery, etc.) and, as the case may be, even the party with whom the contract is to be concluded. Still, even the basic elements of the contract, such as quantity or price, may be left to the discretion of the agent. Oftentimes they are, specifically when the principal is counting on the acumen of the agent to obtain a better deal. It is in such circumstances that the artificial intelligence agent excels. It may be able to identify and adjust for future predictable circumstances, which are likely to result in an advantage for the principal.

The question is whom we deem liable (if liability can even be apportioned) when the artificial agent concludes a juridical act that is 1. prejudicious to the principal and/or prejudicious to the counterparty, 2. not prejudicious to any party but for some reason also undesired, 3. prejudicious to a third party, or 4. contrary to the law and therefore null and void or subject to similar punitive measures that would at least partly rob it of its efficacy.

The first hypothesis of the first situation is apparently the simplest: since AI currently does not benefit from personhood under the law, it cannot be held liable by the principal for acting in a way that was prejudicious to the former and possibly advantageous to the counterparty. Indeed, if the artificial agent is the product of the principal, this solution should stand. After all, as a rule (which bears some exceptions), no one may claim liability for damage caused to himself or herself by their own tools. Still, AI applications may be developed by third parties then licensed, loaned, or otherwise ceded to the principal. This may take place free of charge, e.g. for testing purposes, when the principal may even assume the risk of malfunctions, as is the case during so-called 'beta-testing', a standard practice when developing information technology applications. It may also take place for some fee, as a service supplied to a client. In the first case, the principal is unlikely to benefit from any liability for damage caused or may benefit only from forms of liability, such as for tort, if the supplier of the AI agent is unable to disclaim liability. In the second case, or in situations when liability cannot be disclaimed (cases of egregious negligence or bad faith by the supplier), questions of contractual liability may arise between the contracting party and the AI supplier.

The second hypothesis of the first situation above is also relatively clear: the AI is not a person, and contractual liability is of a strict character in comparative

12 van der Kaaij 2019. 39.

law,[13] just as some forms of tort[14] law. Therefore, any non-performance by the principle will render it liable to the counterparty.

A problem also known in the case of human agent is constituted by the situations when the agent out of error, or even with bad faith, concluded a contract on behalf of the principal which is per se not disadvantageous to either of the parties but which was not desired by the principal, and therefore the agent acted outside the bounds of the mandate received. The problem of unforeseeable contracting by an AI agent has been discussed in the literature,[15] with the conclusion that in such situations, as the AI system is not a person, it will be legally indistinguishable from the entity which 'employed' it. Therefore, the contract concluded by AI, unlike in the case of a human agent acting outside the scope of his or her mandate, will remain valid and will bind the principal, resulting in liability if non-performance occurs.

If third parties have also suffered some form of damage due to the actions of the AI agent, non-contractual liability must also be considered. In cases of damage caused by an agent under current norms, the principal may be held liable, perhaps with the possibility of a later action against the agent, or the agent may be held liable alone; in situations when damage was caused by the AI agent, however, the principal will have to assume liability alone, perhaps complemented by a later claim against the supplier of the AI system.[16]

A last scenario that may occur is when the contract is concluded between the parties through the AI agent as an intermediary but its efficacy is compromised by the actions or omissions of the AI agent. In such circumstances, the parties may even desire the continuation of the contract, while its being null and void or, as the case may be, avoidable or otherwise unenforceable may result in damage to all of them. As the law stands, fault for the inefficacy of the contract is attributed to the party who caused such inefficacy[17] if the counterparty was unaware of the reasons for it at the moment the contract was concluded. Thus, the reasoning according to which the AI agent is a tool of the principal must again be considered and the principal alone held liable for the inefficacy (with an eventual possibility for submitting a further claim against the provider of the AI system should the damage caused not be subject to a valid disclaimer).

These questions, although known and discussed in the literature, have not yet prompted regulatory action. In fact, contractual liability in case of AI agents is not discussed as a specific topic, not even in the context in which the European Commission has already proposed[18] regulating non-contractual liability in the case of AI. The question arises as to whether extant norms of contract law may be

13 See Menyhárd et al. 2022.
14 See Dam 2013.
15 Tanna–Dunning 2022. 139.
16 Ibid.
17 See Menyhárd et al. 2022.
18 See European Commission 2022b.

sufficient to resolve such situations, whether in fact new regulation is not even necessary.

As we have seen, what sets apart the problems posed by AI agents from those of human agents is their lack of personhood under the law: as things stand, they are simple tools employed for a given purpose, which renders the principal as the sole party responsible for any damage caused by the AI agent, even when its actions are unpredictable or incomprehensible to the principal. The principal may possibly claim damages from the supplier of the AI, thereby pushing the liability issue along the supply chain. While in the realm of contractual liability such a solution seems acceptable at first glance, the fairness of imposing strict liability (between parties) for the actions of an AI may be disputed. After all, the principal in such situations may be held liable for circumstances it is unable to foresee, whereas the occurrence of such circumstances in other elements of the contractual relationship, such as performance (e.g. frustration of performance), would otherwise exempt the party from liability, as seen in comparative law.[19] Even worse, liability towards third parties would be non-contractual; so, whenever the fault of the principal cannot be proven, the supplier of the AI system cannot be held liable. In such cases, a liability gap will result.[20] Such gaps may discourage the use of AI and contacting when the counterparty is aware of AI contribution to the conclusion of the contract. For this reason, strict liability regimes should be adapted and compulsory insurance considered in the case of AI agents.[21]

The problems posed by the AI agent could also, in theory, be treated by granting personhood to the AI entity involved. This in fact would result in the AI itself 'supporting any liability from its own assets in cases when its actions caused damage to another, very much like a human agent would'. Such a prospect has been proposed;[22] however, no consensus has been reached on the matter.

3. Liability for AI Acting as an Aid to Human Decision-Makers

AI applications are already being used as an aid to human decision-making. In this capacity, AI is usually utilized in conjunction with a human controller, or supervisor, who may, depending on the solutions used, either influence or even overrule the AI decision (a solution known as 'human in the loop'[23]).

19 See Veress et al. 2022.
20 Allen 2022, De Conca 2022.
21 See Allen 2022. 155–157.
22 See, for example, Rab 2022. 370–371.
23 Church–Cumbley 2022. 189.

AI-aided human decision-making under this concept raises problems of civil liability. To illustrate this, let us consider the following hypothetical situation: an AI used for medical diagnostics (an activity for which several diagnostic tools[24] are already in existence) detects the presence of a tumour, which would require medical action. In this situation of human–AI interaction, the doctor as the human factor may choose to overrule the AI and set up another diagnosis or may confirm the AI diagnosis resulting in the necessity for long-term treatment of the patient, with numerous side-effects. Two questions are inevitably raised here: 1. what happens if the AI was wrong and damage was caused by confirming it, and 2. what happens if the AI was right, and damage was caused by overruling it.[25]

Based on the fault-based liability model applied to non-contractual liability (liability for tort) by the law of obligations in most civil law and also common law jurisdictions, for the aggrieved party to be able to claim damages, he or she must demonstrate the existence of an illicit conduct (fault) on behalf of the tortfeasor, the existence of damage, and the causality between that fault of the tortfeasor and the damage caused. It is the demonstration of fault and of causality between the fault and the damage caused that is most relevant to our inquiry.

In the first situation (the AI was wrong, and the human confirmed the decision), it is for the aggrieved party to prove that the human factor was at fault, based on all information available to him or her at the moment he or she decided to accept the AI diagnosis. Setting aside numerous difficulties involved in proving malpractice, we would like to focus here on one aspect, called 'automation bias'.[26] In situations of machine-influenced medical decision-making, the human factor interpreting diagnostic results tends to accept these results more readily than to overrule them. This is due to the 'comply or explain' logic, in which the human operator feels he or she must provide a reasoned decision when overruling the machine, while no such reasoning, beyond the existence of the automated advice is necessary when accepting an AI-generated diagnosis. In the latter case, the human factor (doctor) can already defend against future claims for liability by simply invoking the machine decision and the high degree of confidence awarded to it in the medical profession. This reasoning is legally correct, as, unless the aggrieved party manages to demonstrate that the human factor had adequate reason to overrule the AI, a probatio diabolica in its own right in malpractice cases, any non-contractual fault-based liability will be very difficult to invoke, as the human factor would be considered as having acted diligently. Here, civil law tends to reinforce the automation bias.

Similarly, in the second situation (the AI was right but was mistakenly overruled by the human factor), proof of fault may be provided more easily by the aggrieved

24 See Gupta–Prasanna–Raghunath 2021.
25 For a more in-depth analysis of similar hypotheses, see Neri et al. 2020.
26 See Bond et al. 2018.

party, who can invoke the fact that the AI proposed a certain diagnosis, and the AI generally tends to be right. In this case, and contrary to the desired effects of the burden of proof imposed on the aggrieved party, it will be the alleged tortfeasor (the human factor interpreting the AI result) who will face an 'uphill battle' as the burden of proof may be inverted after the aggrieved party invoked the AI diagnosis, so it will be up to the alleged tortfeasor to demonstrate that he or she had adequate reason to believe that the AI was wrong. In this case, it will be the human supervisor of the AI who will find himself or herself in a disadvantaged position.

In both cases, it seems that relying on the AI has massive evidentiary benefits to the party invoking the results of AI advice (be it the alleged tortfeasor or the aggrieved party). Simply put, the AI output will be the most easily obtained evidence in the case. This in itself tends to discourage overruling the AI, as any doctor to do so would have to explain why he or she opted not to comply with AI advice, a strenuous and risky task in case the doctor would later need to demonstrate his or her lack of fault for the damaging outcome. This favours rational optimization, a phenomenon in behavioural law and economics,[27] when relying on the AI has net advantages over overruling it in view of any trial aimed at holding the human supervisor of the AI liable. This phenomenon should be considered as one factor in strengthening the automation bias, not as a subconscious reliance but as a rational behaviour of the human factor called upon to supervise the AI.

The proposed AI Liability Directive[28] for regulating non-contractual liability on a European level in the case of AI causing damage does very little to combat this problem. The complex system of presumptions it employs does not alleviate the evidentiary benefits of relying on an AI output as opposed to overriding it, as the directive only addresses situations of fault during AI development, and not those which occur during its use in hypotheses such as the above in conjunction with a human supervisor, which would remain subjected to domestic rules on non-contractual malpractice liability.

The only true modality of avoiding utility-maximizing behaviour in relying on AI output and avoiding challenging it would be for the AI itself (or another person than the one called upon to confirm or overrule the AI output) to be somehow held liable for the results of its output. One way of doing this is implementing a so-called 'human in command' model of AI supervision, whereby the human factor is not 'in the loop' as a co-decider along with the AI but simply filters out egregiously mistaken outputs and otherwise refrains from examining AI decisions on their merits. This solution, favoured by the proposed EU Artificial Intelligence Act[29] (AIA), posits that any action taken by the AI must be under human control, without requiring the human to examine the AI decision on its merits. This shifts liability

27 Zamir–Teichman 2018. 589 et seq.
28 European Commission 2022b.
29 European Commission 2022a.

from the human in command of the AI to the system's provider (manufacturer) or user (e.g. the medical establishment where a medical AI is operated). This solution helps discouraging over-reliance on AI output to any human factor (even when 'human-in-the-loop' and 'human-in-command' models are applied concomitantly) and helps incentivize AI manufacturer and institutional users to ensure that AI output is reliable, which leads to the development of what some authors have termed 'trustworthy AI'.[30]

Another possible solution would be constituted by granting AI itself some form of (even limited) legal personality, thereby ensuring that the AI itself remains liable for any damage caused. In this model as well, the human supervisor would not have any advantage in not overriding the AI's output, as he or she would not be held personally liable. This solution could also be attained by instituting compulsory insurance for some damage caused by AI outputs. The latter two options, in a different context, were examined by the framers of the AI Liability Directive and the AIA (as results from the early drafts of these instruments); however, neither option was implemented. Especially legal personality for AI proved to be an untenable proposal in the face of opposition towards this prospect, as it would entail more disadvantages than possible advantages.[31]

4. AI and Intellectual Property

Clarity in the rules governing intellectual property, and especially regarding the regime of intellectual property rights in rem,[32] is crucial in order to ensure the development of technology (specifically software) and the furtherance of both sciences and arts. AI systems today are capable of developing software,[33] writing poetry,[34] and creatively generating images reminiscent of the work of human artists.[35] In this context, a myriad of problems arise as to the authorship and, consequently, oftentimes also the 'ownership' of the intellectual property produced by AI.[36]

As we have seen above, AI is not a person, therefore it can claim neither authorship nor ownership of the products resulting from its actions. This leaves open the possibility that AI may in fact act in the benefit of some 'classical' legal person such as a corporation, as a tool, rather than as an author, and therefore any intellectual property rights in rem should rest with the operator of the

30 Thiebes–Lins–Sunyaev 2021.
31 See Floridi–Taddeo 2018.
32 See Rahmatian 2011.
33 Provan 2021.
34 See, for example, the *Poem Generator*.
35 See, for example, *DALL-E 2*.
36 Ihalainen 2018.

AI system. The operator, however, may not be identical with the developer of the AI systems or its owner, so the question immediately arises as to whether these latter persons may claim any rights over the resulting intellectual property asset. Furthermore, a more complex and equally important question can be raised regarding the 'originality' of the work produced, a key aspect of intellectual property law, as originality has in the past been thought of as a specifically human contribution to activities of artistic and technical creation. This problem (also called 'agency' in the literature, although not in the same sense as the agency contract to which we have referred to above) is centred around the AI acting as an autonomous agent[37] which could, in theory, make it a 'creator' in the meaning of artificial intelligence but may also make it a tool for 'intellectual property trolling'.[38] In such cases, the AI-generated content is abused either to formulate claims of intellectual property infringement or to force concessions from owners of intellectual property, especially in cases when copyright is concerned due to claimed similarity of works. After all, AI is only a 'derived' creator, as it acts based on the results of machine learning, and whether artificial creativity may be equated with human creativity is still uncertain.

To date, the specifically original human factor in creativity has been considered a major obstacle in granting intellectual property rights to AI for works produced by it (as stated by the U.S. Copyright Office in its review in the Thaler case).[39]

Therefore, according to the as-of-yet meagre case-law,[40] intellectual property rights cannot be granted to non-human intellects. This raises a specific problem, beyond attribution of a work (which was at stake in the Thaler case), as in rem intellectual property rights have a specific pecuniary content resulting from the exclusivity of use and reproduction granted to the copyright or other intellectual property right owner. As the case law stands, even if AI would be granted legal personality, meaning that it would be able to hold assets and have liabilities in its own name, its specifically non-human nature would make it impossible for it to be considered the author of the intellectual property it 'owns', as it resulted from the AI's actions. This would inevitably lead to confusion, as the author or creator of a work protected as a rule by regulations on intellectual property is considered its primordial owner. Here, the hypothetical legal person AI would come to exercise property rights.[41]

37 Gervais 2020.

38 Ihalainen 2018.

39 See Recker (2022) for some information on a recent U.S. Copyright Office ruling in the *Thaler* case, in which the request for registering an AI as the author of a work was denied on grounds of lack of 'human authorship'.

40 For this case-law, see Free 2022. 233–234.

41 For such proposals, see Davies 2011 and Brown 2021.

5. AI as a Legal Person

The above-described problems posed by various AI implementations, such as the issues concerning AI agents, AI as an aid to human decision-makers, as well as AI taken as a creator of intellectual property, all converge towards the problem of legal personality for (at least some) AI implementations. This possibility has been examined in the literature.[42] The consensus of most authors on the topic is that some form of legal personhood may be awarded to AI in the future. The authors state, among other considerations, that the legal difficulties caused by the problems involving agency, effects of AU decisions, and intellectual property rights for AI-generated content would be solved by granting legal personhood to AI systems. These would have rights and obligations, hence would own assets and would be subjected, if need be, to regimes of civil liability. In this context, a regime similar to that of corporate legal persons would become applicable to AI entities.

This solution, which was initially even considered by the European Commission when drafting the AIA, was strongly contested[43] by other authors. The main arguments for this position referred to the fact that, as is the case with legal persons currently in existence, ultimately a human being or group of human beings and not the AI would have to bear the consequences of the AI's actions or inactions. Furthermore, by deliberately underfunding legal persons constituted by AI systems, liability for damage caused would be avoided, and a moral hazard would result, which would run contrary to the desideratum of creating 'trustworthy AI',[44] which is contingent upon a high degree of accountability for AI developers and operators. As things stand, such operators and developers of AI systems are, of course, humans. Finally, a good deal of criticism resulted from the lack of any obvious advantage that would result from granting legal personality to AI, as the issues of agency and liability may be resolved based on compulsory insurance and respect for the precautionary principle. Considering these reasons, the AIA proposal was finalized and published without legal personality for AI mentioned in its text.

6. Conclusions

In my study, I have outlined some of the 'neuralgic points' of the interaction between AI technology and civil law. As is apparent, these points and the potential problems of legal science and doctrine they entail are far removed from the specifically public law issues, or at the very least issues concerning both private and public law, which are much more abundantly referred to in the literature, especially the problems of

42 See Andrade et al. 2007, Calverley 2008, Kurki 2017, Solaiman 2017, Schirmer 2020, Mik 2021.
43 Floridi–Taddeo 2018, Jowitt 2020.
44 Thiebes–Lins–Sunyaev 2021.

bias, e.g. in criminal and in administrative adjudication or during assessment of job applications. I believe that the consequences of AI technology in the field of private law should not be overlooked, as most daily interactions with AI systems will occur in the context of private law relationships: while concluding contracts, while working, travelling, or even staying at home. Even if the need for regulating AI–human interactions in the domain of private law seems less stringent, it is likely to increase exponentially in the future; therefore, we propose conducting further research to determine the optimal legal regime for these interactions. A good starting point for this research would be to assess the efficiency of the AI Liability Directive as the newest proposed addition to European Union private law, once the directive enters into force.

References

ALLEN, J. G. 2022. Agency and Liability. In: *Artificial Intelligence. Law and Regulation*. Cheltenham (UK)–Northampton (USA): http://dx.doi.org/10.4337/9781800371729.

ANDRADE, F.–NOVAIS, P.–MACHADO, J.–NEVES, J. 2007. Contracting Agents: Legal Personality and Representation. *Artificial Intelligence and Law* 15(4): 357–373. https://doi.org/10.1007/s10506-007-9046-0.

BERTOLINI, A.–EPISCOPO, F. 2021. The Expert Group's Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies: A Critical Assessment. *European Journal of Risk Regulation* 12(3): 644–659. https://doi.org/10.1017/err.2021.30.

BODEN, M. A. 2018. *Artificial Intelligence. A Very Short Introduction*. Oxford.

BOND, R. R.–NOVOTNY, T.–ANDRSOVA, I.–KOC, L.–SISAKOVA, M.–FINLAY, D.–GULDENRING, D. et al. 2018. Automation Bias in Medicine: The Influence of Automated Diagnoses on Interpreter Accuracy and Uncertainty When Reading Electrocardiograms. *Journal of Electrocardiology* 51(6, Supplement): 6–11. https://doi.org/10.1016/j.jelectrocard.2018.08.007.

BROWN, R. D. 2021. Property Ownership and the Legal Personhood of Artificial Intelligence. *Information & Communications Technology Law* 30(2): 208–234. https://doi.org/10.1080/13600834.2020.1861714.

BURRELL, J. 2016. How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms. *Big Data Society* 3: https://doi.org/10.1177/2053951715622512.

CALVERLEY, D. J. 2008. Imagining a Non-Biological Machine as a Legal Person. *AI & Society* 22(4): 523–537. https://doi.org/10.1007/s00146-007-0092-7.

CHESTERMAN, S. 2021. Through a Glass, Darkly: Artificial Intelligence and the Problem of Opacity. *The American Journal of Comparative Law* 69(2): 271–294. https://doi.org/10.1093/ajcl/avab012.

CHOPRA, S.–WHITE, L. F. 2011. *A Legal Theory for Autonomous Artificial Agents*. Ann Arbor (USA).

CHURCH, P.–CUMBLEY, R. 2022. Data and Data Protection. In: *Artificial Intelligence. Law and Regulation*. Cheltenham (UK)–Northampton (USA): 163–238. http://dx.doi.org/10.4337/9781800371729.

CITRON, D. K.–PASQUALE, F. A. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89(8): 1–34.

DAM, C. V. 2013. 297. Strict Liability. In: *European Tort Law*. Oxford: https://doi.org/10.1093/acprof:oso/9780199672264.003.0010.

DAVIES, C. R. 2011. An Evolutionary Step in Intellectual Property Rights–Artificial Intelligence and Intellectual Property. *Computer Law & Security Review* 27(6): 601–619. https://doi.org/10.1016/j.clsr.2011.09.006.

DE CONCA, S. 2022. Bridging the Liability Gaps: Why AI Challenges the Existing Rules on Liability and How to Design Human-Empowering Solutions. In: *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*. The Hague: 239–258. https://doi.org/10.1007/978-94-6265-523-2_13.

EUROPEAN COMMISSION. 2022. *Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) COM(2022) 496 Final*. https://ec.europa.eu/info/sites/default/files/1_1_197605_prop_dir_ai_en.pdf.

FLORIDI, L.–TADDEO, M. 2018. Romans Would Have Denied Robots Legal Personhood. *Nature* 557: 309. https://doi.org/10.1038/d41586-018-05154-5.

FREE, R. 2022. Intellectual Property. In: *Artificial Intelligence. Law and Regulation*. Cheltenham (UK)–Northampton (USA): 213–238. http://dx.doi.org/10.4337/9781800371729.

GERVAIS, D. 2020. Is Intellectual Property Law Ready for Artificial Intelligence? *GRUR International* 69 (2): 117–118. https://doi.org/10.1093/grurint/ikz025.

GUPTA, P. K.–PRASANNA, D. V.–RAGHUNATH, S. S. 2021. How Artificial Intelligence Can Undermine Security: An Overview of the Intellectual Property Rights and Legal Problems Involved. In: *Applications in Ubiquitous Computing*. Cham. 37–58. https://doi.org/10.1007/978-3-030-35280-6_3.

HARARI, Y. N. 2017. *Homo Deus. A Brief History of Tomorrow*. London: Vintage.

IHALAINEN, J. 2018. Computer Creativity: Artificial Intelligence and Copyright. *Journal of Intellectual Property Law & Practice* 13(9): 724–28. https://doi.org/10.1093/jiplp/jpy031.

JOWITT, J. 2020. Assessing Contemporary Legislative Proposals for Their Compatibility with a Natural Law Case for AI Legal Personhood. *AI and Society* January 2020: https://doi.org/10.1007/s00146-020-00979-z.

KNOBLOCH, T. 2018. *Vor Die Lage Kommen: Predictive Policing in Deutschland*. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/Graue Publikationen/predictive.policing.pdf.

Kuo, C.-H.–Chen, C.-T.–Lin, S.-J.–Huang, S.-H. 2021. Improving Generalization in Reinforcement Learning-Based Trading by Using a Generative Adversarial Market Model. *IEEE Access* 9: 50738–50754. https://doi.org/10.1109/ACCESS.2021.3068269.

KURKI, V. A. J. 2017. Why Things Can Hold Rights: Reconceptualizing the Legal Person. In: *Legal Personhood: Animals, Artificial Intelligence and the Unborn*. Cham. 69–89. https://doi.org/10.1007/978-3-319-53462-6_5.

MARCH, C. 2021. Strategic Interactions between Humans and Artificial Intelligence: Lessons from Experiments with Computer Players. *Journal of Economic Psychology* 87(December): 102426. https://doi.org/10.1016/j.joep.2021.102426.

MENYHÁRD, A.–HULMÁK, M.–BALLIU, A.–STEC, P.–VERESS, E.–DUDÁS, A.–HLUŠÁK, M. 2022. Damages. In: *Contract Law in East Central Europe*. Miskolc–Budapest. 419–460. https://doi.org/10.54171/2022.ev.cliece.

MIK, E. 2021. 419 AI as a Legal Person? In: *Artificial Intelligence and Intellectual Property*. Oxford: https://doi.org/10.1093/oso/9780198870944.003.0020.

MILANA, C.–ASHTA, A. 2021. Artificial Intelligence Techniques in Finance and Financial Markets: A Survey of the Literature. *Strategic Change* 30(3): 189–209. https://doi.org/10.1002/jsc.2403.

NERI, E.–COPPOLA, F.–MIELE, V.–BIBBOLINO, C.–GRASSI, R. 2020. Artificial Intelligence: Who Is Responsible for the Diagnosis? *La Radiologia Medica* 125(6): 517–521. https://doi.org/10.1007/s11547-020-01135-9.

PAGALLO, U. 2013. *The Laws of Robots. Crimes, Contracts and Torts*. Dordrecht–Heidelberg–New York–London.

PROVAN, G. 2021. Using Artificial Intelligence for Auto-Generating Software for Cyber-Physical Applications. In: *Artificial Intelligence Methods for Software Engineering*. New Jersey–London–Singapore–Beijing–Shanghai–Hong Kong–Taipei–Chennai–Tokyo: 211–240.

RAB, S. 2022. Telecoms and Connectivity. In: *Artificial Intelligence. Law and Regulation*. Cheltenham (UK)–Northampton (USA): 355–377. http://dx.doi.org/10.4337/9781800371729.

RAHMATIAN, A. 2011. Modern Studies in Property Law. In: *Intellectual Property and the Concept of Dematerialised Property*. Oxford: 361–383. https://ssrn.com/abstract=1917950.

RECKER, J. 2022. U.S. Copyright Office Rules A.I. Art Can't Be Copyrighted. *Smithsonian Magazine*. https://www.smithsonianmag.com/smart-news/us-copyright-office-rules-ai-art-cant-be-copyrighted-180979808/.

SCHIRMER, J.-E. 2020. Artificial Intelligence and Legal Personality: Introducing "Teilrechtsfähigkeit": A Partial Legal Status Made in Germany. In: *Regulating Artificial Intelligence*. Cham. 124–141.

SOLAIMAN, S. M. 2017. Legal Personality of Robots, Corporations, Idols and Chimpanzees: A Quest for Legitimacy. *Artificial Intelligence and Law* 25(2): 155–179. https://doi.org/10.1007/s10506-016-9192-3.

SURDEN, H. 2019. Artificial Intelligence and Law: An Overview. *Georgia State University Law Review* 35(4): 1305–1337.

TANNA, M.–DUNNING, W. 2022. Commercial Trade. In: *Artificial Intelligence. Law and Regulation*. Cheltenham (UK)–Northampton (USA): 133–145. http://dx.doi.org/10.4337/9781800371729.

THIEBES, S.–LINS, S.–SUNYAEV, A. 2021. Trustworthy Artificial Intelligence. *Electronic Markets* 31(2): 447–464. https://doi.org/10.1007/s12525-020-00441-4.

VAN DER KAAIJ, H. D. S. 2019. *The Juridical Act. A Study of the Theoretical Concept of an Act That Aims to Create New Legal Facts*. Cham. https://doi.org/10.1007/978-3-030-15592-6.

VERESS, E.–HULMÁK, M.–BALLIU, A.–TOMCZAK, T.–DUDÁS, A.–HLUŠÁK, M. 2022. Changes in Circumstances: Frustrated Contracts and Legislative or Judicial Modification of the Contract. In: *Contract Law in East Central Europe*. Miskolc–Budapest. 419–460. https://doi.org/10.54171/2022.ev.cliece.

WISCHMEYER, T. 2020. Artificial Intelligence and Transparency: Opening the Black Box. In: *Regulating Artificial Intelligence*. Cham. 75–101. https://doi.org/10.1007/978-3-030-32361-5_4.

ZAMIR, E.–TEICHMAN, D. 2018. *Behavioral Law and Economics*. Oxford.

ZUIDERVEEN BORGESIUS, F. J. 2020. Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence. *The International Journal of Human Rights* 24(10): 1572–1593. https://doi.org/10.1080/13642987.2020.1743976.

*** *DALL-E 2* 2022. https://openai.com/dall-e-2/.

*** *Poem Generator*. https://www.poem-generator.org.uk/.

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 113–124

DOI: 10.47745/AUSLEG.2022.11.2.07

The Status and Role of Law and Regulation in the 21st-Century Hybrid Security Environment[1]

Ádám FARKAS
PhD, Senior Research Fellow
Faculty of Military Science and Military Officer Training of the University of Public Service (Budapest, Hungary)
e-mail: farkas.adam@uni-nke.hu

Abstract. The author examines in detail the notion of 'lawfare', and its various interpretations, which lead to the use of the regulatory environment by some actors in order to achieve strategic objectives, including during geopolitical competition and armed conflict. The author finds that through lawfare, hybrid warfare may be conducted by using (and as the case may be, even abusing) the rule of law to the advantage of one of the actors. The author concludes that it is necessary to consider 'lawfare' in its various forms as an element of the legal environment, for the purpose of ensuring national security.

Keywords: lawfare, hybrid threat, legitimacy, legal resilience, defence and security law

1. Introductory Remarks

In respect of maintaining security, setting security objectives, and defence as a set of security activities, law plays an outstanding role in the Transatlantic zone though there were major shortfalls and objectives failed to be met in this area in the past decades. A constitutional state, in which the exercise of governmental power is constrained by the law and which guarantees the exercise, evolution, and development of the rights of individuals and society as a solid foundation, is inconceivable without proper and modern legal bases.

In other words, it follows from the very rule of law that the law also has a prominent role to play in terms of security and defence. Nevertheless, it is worth establishing a general connection between this and the extremely broad concept of complex security and the hybrid threats that have become dominant in the early 21st century. By establishing this connection, we can also highlight in connection with the foregoing that the shared horizon of security and regulation is much broader

1 This work was supported by the TKP2020-NKA-09 project financed under the Thematic Excellence Programme 2020 by the National Research, Development and Innovation Fund of Hungary.

than we have thought it is while identifying defence as a totality of activities guaranteeing active security and typically linked to the monopoly of the state to exercise legitimate coercion. A regulatory framework for safety in the broadest sense includes all the rules that must be complied with in the course of the various activities such as transport, industrial production, farming, healthcare, research, etc. in a manner that the performance of the activity concerned does not have an effect that threatens or erodes security. However, the horizon of security and justice is still far broader than what we would normally – prima facie – associate it with in the context of policing, national defence, and national security. In line with the foregoing, there may be a number of potential security hazards or threats in the various sectors of security that are, upon reaching a certain level, connected to security-related activities and may require specific sectoral actions, regulations, and cross-sectoral coordination both below and above this level in order for significant losses to be avoided. On the one hand, this is precisely the core of the hybrid threat, as it relies on the growing weight of non-military factors in the competition for power, which has historically been traditionally military. This equally builds on the development of technology, the exposure of affluent and consumer societies to technology, and the multiple exploitability of a highly differentiated technological environment.

The use of non-military factors in power struggles cannot be considered a completely new phenomenon in history, as the illegitimate use of various acts of sabotage or of natural resources and the related structures has long shown that critical infrastructures of society can be used for warlike purposes without any traditional armed resistance with an openly offensive intent. However, the multi-stage revolutionary development of industrial societies and subsequently technology has led to a significant rise in demand for comfort and heightened expectations in mainstream Transatlantic societies through welfare and consumer lifestyles. This kind of development has made many dependent on innovations granting a higher level of comfort and has also significantly increased society's exposure to increasingly technology-based services. In addition to the proliferation of everyday necessities, it is important to highlight the explosion and global spread of information technologies. They have also made it possible to access information and, where appropriate, to influence individuals and societies with false or distorted information, which can clearly be used to prepare the way for or even to increase the effectiveness of hard-power measures. This can be identified as an extremely significant change in the hybrid environment in which military and non-military elements have operated in recent decades, just as the legal aspects of this development, on the effectiveness and modernity of which social legitimacy in modern states is also based to a considerable extent. Accordingly, as the technology, research, production and services provision become increasingly

differentiated, the scope and differentiation of state and social regulation increase, which has a similar effect on the content, regulation, and guarantee of security.

Thus, security and defence must be interpreted in the same context as the legal system and the functioning of the state as a whole: they have to converge with the dynamic changes of the environment in the broadest sense. Following clearly from the essence of the rule of law, it is not only a development competition but also a new kind of security vulnerability if potential regulatory gaps can be exploited by adversaries for their own ends. From a different perspective, however, regulation also has a key role to play in maintaining and strengthening security and, in this context, in defence as an activity for historical and functional reasons. Security and defence systems must be established as a coherent whole of subsystems capable of rapid, efficient, and drillable responses. Regulation has played a historically prominent role in this. It is no coincidence that in the history of the armed forces, various regulations and rules trace their history back to the dawn of organized societies. They are also a yardstick of development regarding their importance for defence, social functioning, and the state. According to Niccoló Machiavelli:

> But if they should consider the ancient institutions, they would not find matter more united, more in conformity, and which, of necessity, should be like to each other as much as these (civilian and military); for in all the arts that are established in a society for the sake of the common good of men, all those institutions created to (make people) live in fear of the laws and of God would be in vain, if their defence had not been provided for and which, if well arranged, will maintain not only these, but also those that are not well established. And so (on the contrary), good institutions without the help of the military are not much differently disordered than the habitation of a superb and regal palace, which, even though adorned with jewels and gold, if it is not roofed over will not have anything to protect it from the rain.[2]

Naturally, the concept of reliance solely and predominantly on military force has become obsolete, but not so the essence of the message: the state and its rules as well as individual and social security cannot stand the test of time without defence.

It is therefore worthwhile to make its legal role in this 21st-century hybrid security environment a priority issue for investigation, beyond the question of exploitability of legal gaps and conflicts, in a more complex dimension. In this respect, a number of valuable works and findings have been produced in the last decade on hybrid threats, but I believe that it is important to draw even more attention to a theoretical, systemic approach to the issue, thus strengthening the

2 Machiavelli 2001. 6.

reception of a novel – security-driven – approach to law in general legislative, legal, and jurisprudential thinking. To this end, in this paper I would like to emphasize three aspects: (1) the importance of defence regulation in the rule of law, adapting to a changing environment; (2) the role of regulation in the functioning of defence and security organizations; (3) the question of the strategic applicability of law as an instrument of influence and warfare.

2. The Place and Role of Modern and Adaptive Defence Regulation in the Dimension of the Rule of Law

First of all, it should be noted that defence regulation tracking changes and development in the world is also of key importance from the perspective of the rule of law because it is not the principle of 'everything is allowed that is not forbidden'[3] that is applied by state organizations and, in particular, law enforcement organizations[4] but rather the need for operation that the powers granted by law allow, i.e. the requirement of constitutional defence. This is one aspect of the state's self-restraint. Based on the predictable, efficient, and foreseeable operation of organizations of defence and security, which is also expected by the civil society, this is also the basis of order, stability, and, hence, the ability to exercise individual rights, social development, and economic growth. Regulation, especially efficient regulation, is, therefore, a fundamental guarantee from the perspective of the functioning, controllability, and, ultimately, reliability of the state. It is, therefore, no mere coincidence that this approach looks back on impressive history in respect of civilian and military relationships as well as the relationship between law enforcement and public administration.

Given that, despite the difficulty of its definition in detail, the rule of law is a set of minimum requirements whose main components, the states in the Transatlantic region, are familiar with it and accept it, it is easy to realize that an appropriate regulation is also important in an international context. However, it is worth supplementing this topic with a brief proposition, namely one that focuses on real globalization and its real-time interactions with the states outside the Transatlantic area. The systemic foundations of globalization based on physical and real-time interactions have been laid by global capitalism, and its structure has been made complete by technological development. However, due to its nature, capitalism assumes the existence of a multitude of contractual relationships, whether or not the parties concerned are advocates of the Western concept of law, i.e. those tenets that insist on the guarantees the rule of law provides. As a result, it is safe

3 Patyi 2015.

4 Patyi 2016, Farkas–Till 2016, Farkas 2018b.

to say that even in states that do not (or do not fully) agree to the rule of law in its Transatlantic sense, regulation complying with Western requirements has strengthened as it serves as a basis for trust in business. Furthermore, formally, mechanisms of legal protection are undergoing development in various parts of the world because such development is a guarantee needed for business relations and ensuring labour mobility. This necessarily affects the functioning of the defence and security organizations of the individual states, i.e. certain predictability and guarantee minimums, albeit at a varying level of authorization, are emerging worldwide, with the exception of autocracies, pseudo-states, and failed states.

Conversely, if the regulation of defence and security functions are not sufficiently up to date, consistent, stable, and predictable, trust in the state can erode. Such loss of confidence can also be interpreted in relation to individuals and groups constituting a nation, other states in federal association or partnership with the state, and actors with business interests or plans in connection with the state.

Outdated, inconsistent, and inadequately enforced regulations may:

a) weaken the state's ability to adequately respond to newer and more complex threats and crises or even lead to the lack of such ability,
b) make the state's responses to various contingencies unpredictable or at least uncertain,
c) and ultimately provide for a reasonable possibility of abuse by the state or its institutions.

The existence of any one of these uncertainties can weaken the sovereignty of the state concerned and undermine its stability as well as economic and social attractiveness, and ultimately lead to a crisis in that state if they materialize and get out of control.

However, unusual as it may be, at a European level, currently, the impact on economic confidence should be highlighted in connection with the importance of the modernity, consistency, and predictability of defence and security regulation. The underlying reason for such focus is that economic prosperity including the growth in investments and innovation and their establishment and operation is hard to envisage in a state where there are embarrassing questions about fundamental security issues or uncertain solutions to specific crises.

Appropriate regulation reflects, in addition to trust, the state's readiness and professionalism related to security, which is key to development and the trust needed for it in all respects. Creating such trust is not a prerogative of large and medium-sized powers, as even small states have solid defence and security systems capable of inspiring trust. A European and an Asian example is Switzerland and Singapore respectively, where a broad interpretation of security is combined with a corresponding complex defence system.

3. The Importance of Regulation for the Effective Functioning of Security and Defence Organizations

Furthermore, it should be stressed that regulation is essential for all well-structured and well-prepared defence organizations, i.e. optimal regulation is a precondition for efficiency in the performance of tasks, a guarantee for the observance of the rule of law. Without predefined protocols, there is no hierarchy or authority, i.e. no system of command and control can be built. Without proper regulation, the performance of specific tasks cannot be planned, as the proper structuring and the subsequent definition of and the accountability for tasks is also based on regulation. Ultimately, without effective regulation, military forces cannot be prepared either, as its precondition is a well-defined order of operation to be followed in certain cases that can only take its final, fathomable, and required form in regulation due to the complexity of the systems in question. In other words, only a well-regulated and modern defence system can be a good and effective defence system. Consequently, an incomplete, outdated, and inconsistent regulation can have a direct negative impact on the effectiveness of defence forces, and thus on individual and societal security.

However, this principle is not of legal origin, as the regulation of armed forces, which is historically the institutional basis of defence, preceded legal regulation in the modern sense and was typically below the level of legal regulation until the development of civil (rule of) law. However, this did not mean the under-regulation of functions. The importance and fundamental significance of regulation stems from the nature of the organization of defence. This is well reflected in the fact that one of the cornerstones of military science was the analysis of military history and, as part of it, the organization, regulation, and management of armies and defence systems, which served as a basis for outstanding theoretical summaries, i.e. military theory in analysing the works of thinkers laying down principles and a series of related analyses by representatives of related sciences.[5] This is a tradition in the historically dominant military dimension of defence, which was adopted by law enforcement science and then by research dealing with national security functions.

Therefore, the fundamental role of regulation in relation to defence stems from the need for organization during defence itself, which forms its inherent nature. Thus, in this respect, the fact that in the Europe of the 17th–19th centuries one of the main impacts of the development of defence infrastructure and armies was exerted by regulation through military orders did not follow primarily from the development of the state and regulation but rather from the traditions of military organization and the sciences that assisted it. From this point of view, the fact that as a result of the evolution of the civil (rule of) law increasingly important legal

5 Szendy 2017; Forgács 2017, 2020; Bellamy 2016.

frameworks and bases have been created to regulate the armed forces and national defence is not a new phenomenon; this, however, does not mean that the law relying on optimal legislation and assisting professionals restricts defence measures by regulation but rather that the need of military organization for being regulated and the rule of law have created rules representing various levels of hierarchy and synergy. An excellent example of this in Hungary was the multi-stage process of national defence regulation and development at the turn of the 20th century,[6] which encouraged the development of law enforcement and the independence and, later, regulation of national security functions even if statutory regulation was rather delayed due to the vicissitudes of Hungary's history.

4. The Strategic Possibility of Using Law as an Instrument of Influence and Warfare

Thirdly, in the security environment of the 21st century, the use by state and non-state actors of what is called lawfare,[7] i.e. law as a tool of warfare, as a tool of strategic influence, is also a serious challenge to the rule of law and security. Although this phenomenon is not new, it has become a tool of strategic importance and easier to prepare due to the availability of rules through new threats and the digitalization of state functions and – within that – regulation. Growing importance is best reflected by the legal implications of drone warfare[8] and hybrid threats[9] that have emerged in recent years. In his study on the topic, Orde F. Kittrie considers the public opinion that links the concept of lawfare to the work of Charles Dunlap Jr. in 2001 as an overture; nevertheless, he attributes lawfare to Grotius. He also points out that the application of the law as a strategic tool is also present in the concept of 'warfare without barriers' published in China at the turn of the millennium, well before the Gerasimov doctrine,[10] and in various approaches before that. However, based on the semantic origin of the concept of lawfare, i.e. the combination of law and warfare, Orde F. Kittrie's invaluable analysis relies heavily on the war/military approach and provides its typology and case studies. However, this approach links the use of law as a tool to military strategies rather than a large strategic vision that fits into the diversity of complex security. This approach is also reflected in the author's typology, which sees lawfare as a means intending and able to replace military force and, in the context of acts of war, as a

6 Farkas 2018a, 2019; Kelemen 2017.
7 Dunlap 2001; Bachmann–Munoz Mosquera 2015; Ansah 2010; Kearney 2010; Sari 2017, 2019; Hódos 2021.
8 Hasian 2016; Spitzer 2019, 2020; Kis Kelemen 2018a–b.
9 Sari 2018, 2019; Hódos 2020; Vikman 2021; Farkas–Resperger 2020; Farkas 2020; Kelemen 2021.
10 Kittrie 2016. 4–8.

means of exerting pressures through publishing and promoting violations of the law, typically of military law.[11]

However, due to the complexity of security and the various state and non-state modes of hybrid threats regarding the new comprehensive 21st-century pressure–influence–attack concept, it is important to pay closer attention to and analyse in detail the idea that gaps, deficiencies, contradictions, or inconsistencies of security relevance in legal regulation can pose an extreme risk not only in the case of specific confrontations or in military strategy but also in a larger strategic framework that uses a much wider range of military tools.[12] This is an excellent tool for amplifying various acts of pressure, influence, and destabilization against modern states and for delegitimizing state actions. Thus, in addition to the fact that the legal and military aspects of the issue of lawfare should continue to be the subject-matter of in-depth analyses from a military and strategic perspective, the interpretation of legal vulnerabilities as security risks should also be fine-tuned.

In this respect, it should also be noted that as regards the identification of regulatory failures as a security risk, attention should not be limited to the regulation of defence and security functions but rather a boarder interpretation is needed, including the security aspects of different strategic regulatory areas. It is, therefore, essential that the regulation of the defence and security functions of a state be coherent, up to date, and effective; in addition, in order for the number of channels of influencing and covert operation to be reduced, gaps in the regulation of transport, communications, financial markets, food, pharmaceutical safety, data protection, and migration must be identified, analysed, and bridged. Defective regulation can provide a possibility of making preparations for external interventions, a kind of infiltration that may be disguised by business transactions, acts of organized crime, or lax or circumventable requirements of settlement or setting up businesses. For these reasons, a shift from the concept of lawfare (a combination of law and warfare) linked notionally to warfare towards the concept of legal vulnerability or law as security vulnerability should occur.

With the conceptual issues discussed, revisiting the importance of the modernity and effectiveness of defence and security regulations, we cannot but realize that the operation of the state is extremely widely regulated in its internal and international relations, and this legislation is in the public domain. This serves both law enforcement and legal certainty and also provides an opportunity for specific protective measures to be called into doubt openly if our rules are inadequate. Thus, a specific protective measure combined with an outdated rule can be easily subverted in respect of both the domestic and the international public, for which the World Wide Web is an excellent platform, as it can reach the population directly. There are many examples of this phenomenon, including

11 Id. 11–24.
12 Sari 2017.

cyber attacks, targeted drone strikes, the annexation of Crimea, or disputes in the Far East affecting certain maritime areas.

5. Concluding Remarks

In the foregoing, I have sought to highlight some aspects of the role of law in the 21st-century hybrid security environment by considering the socio-state-regulatory dimensions together. In my view, the idea of lawfare, which arises from the intersection of the concepts of warfare and law, needs to be further developed in a complex and security-oriented analysis of state and law. Just as the concept of security has transcended the dominance of the military, so too should this issue be further developed towards a complex approach and a concept of complex security. It is clear that the exploitation of possible conflicts, gaps, and uncertainties in national and international law can play a prominent role even in the non-military preparatory phase, which is more typical of hybrid threats. It could be said that the information age has also made it much easier to assess and, where appropriate, challenge the law for purposes of influence. And this is an excellent tool for a hybrid narrative since, as I have discussed in relation to the three aspects, the exploitation of such vulnerabilities in the law has a negative impact on the rule of law and the social legitimacy of the rule of law, on the effectiveness of defence organizations, and, where appropriate, on the outcome of a conflict involving hard power through lawfare.

Taken together, the role of the rule of law in maintaining and strengthening security is therefore crucial. It can be analysed, developed, and applied in a modern way if it is able to develop a continuous interaction between security and defence expertise, legal thinking on security and defence issues, and the broad or traditional legal discipline. However, this cooperation is 'only' the professional basis for effective defence against the strategic use of law in a hybrid environment.

Building on these professional foundations, it is also necessary to ensure that society as the legitimacy base for state action and, with it, for the provision of defence, is able to make proper sense of this issue by means of appropriate, balanced, credible, and realistic information and training programmes. It is equally important that, in addition to the social element, political decision-makers, as the determinants of legislation and state decision-making, recognize the hybrid security aspects of the law and show openness to addressing shortcomings, independent of day-to-day political battles, on the one hand, and to establishing and strengthening a defence and security approach in the preparation of various regulations, on the other, in order to prevent future shortcomings and entry points.

References

ANSAH, T. 2010. Lawfare. A Rhetorical Analysis. *Case Western Reserve Journal of International Law* 43: 87–119.

BACHMANN, S. D.–MUNOZ MOSQUERA, A. B. 2015. Lawfare and Hybrid Warfare – How Russia Is Using the Law as a Weapon. *Amicus Curiae, Journal of the Society for Advanced Legal Studies* 2015/Summer: 25–28.

BELLAMY, C. 2016. *The Evolution of Modern Land Warfare.* Abingdon.

DUNLAP, C. J. Jr. 2001. *Law and Military Interventions: Preserving Humanitarian Values in 21st Conflicts.* https://people.duke.edu/~pfeaver/dunlap.pdf (accessed on: 10.06.2021).

FARKAS, Á. 2018a. A honvédelmi jog polgári szabályozási előzményei Magyarországon [A Civil State Regulation Background to National Defence Legislation in Hungary]. In: *A honvédelem jogának elméleti, történeti és kortárs kérdései* [Theoretical, Historical, and Contemporary Issues of National Defence Legislation]. Budapest. 31–57.

2018b. Adalékok a védelmi alkotmány és a védelmi alkotmányjog hazai értelmezéséhez és történetiségéhez [Contributions to the Interpretation of a Defence Constitution and Defence Constitution Law in Hungary]. *Hadtudományi Szemle* 4: 227–255.

2019. *A [hon]védelmi alkotmány polgári evolúciója Magyarországon 1867*–1944 [The Civil State Evolution of the Defence Constitution in Hungary 1867–1944]. Budapest.

2020. Komplex biztonság, hibrid konfliktusok, összetett válaszok [Complex Security, Hybrid Conflicts, and Complex Responses]. *Honvédségi Szemle* 4: 11–23.

FARKAS, Á.–RESPERGER, I. 2020. Az úgynevezett „hibrid hadviselés" kihívásainak kezelése és a nemzetközi jog mai korlátai [Management of the Challenges Posed by 'Hybrid Warfare' and the Current Limitations of International Law]. In: *Új típusú hadviselés a 21. század második évtizedében és azon túl. Intézményi és jogi kihívások* [New Type of Warfare in and beyond the Second Decade of the 21st Century. Institutional and Legal Challenges]. Budapest. 132–149.

FARKAS, Á.–TILL, SZ. 2016. A honvédelmi alkotmány és alkotmányosság alapkérdései Magyarországon [Fundamental Issues of a National Defence Constitution and Constitutionality]. In: *Magyarország katonai védelmének közjogi alapjai* [Public Law Bases of the Military Defence of Hungary]. Budapest. 40–71.

FORGÁCS, B. 2017. *Hadelmélet. A magyar katonai gondolkodás története és a hadikultúrák* [Theory of the Military. A History of Military Theories in Hungary and Military Cultures]. Budapest.

2020. *Gerillák, partizánok, felkelők – Az irreguláris hadviselés elméletének története – korunk kihívásai* [Guerillas, Partisans, and Revolutionaries – A History of Irregular Warfare]. Budapest.

HASIAN, M. Jr. 2016. *Drone Warfare and Lawfare in a Post-Heroic Age*. Tuscaloosa (USA).

HÓDOS, L. 2020. A hibrid konfliktusok felívelési szakasza, avagy a fenyegetés észlelésének, megelőzésének és kezelésének nemzetbiztonsági aspektusai [An Upswing in Hybrid Conflicts, or the National Security Aspects of the Detection, Prevention, and Management of Threats]. *Honvédségi Szemle* 4: 49–64.

2021. A nemzetbiztonsági szolgálatok közelmúltbeli tevékenységét befolyásoló mérföldkövek, avagy az új típusú biztonsági kihívások jelentette veszélyek és az azokra adott kormányzati, illetve jogalkotói válaszok 2010 és 2020 között [Milestones Affecting the Recent Activities of the National Security Services, or the Threats Posed by New Types of Security Challenges and the Governmental and Legislative Responses to Them between 2010 and 2020]. *Szakmai Szemle* 1: 134–149.

KEARNEY, M. 2010. Lawfare, Legitimacy and Resistance: The Weak and the Law. *The Palestine Yearbook of International Law*. Nicosia. 79–129.

KELEMEN, R. 2017. A katonai jog, a katonai büntetőjog helye a dualizmus kori magyar államban [Military Law and the Status of Military Criminal Law in Hungary in the Era of Dualism]. In: *Ünnepi tanulmányok Máthé Gábor oktatói pályafutásának 50. jubileumára* [Studies in Celebration of Gábor Máté's 50-Year Career as a Lecturer]. *Studia sollemnia scientiarum politico-cameralium*. Budapest. 203–210.

2021. A nem állami kibertéri műveletek egyes szereplőinek jelentősége a hibrid konfliktusokban [The Importance of Certain Actors of Non-governmental Cyberspace Operations in Hybrid Conflicts]. *SmartLaw Research Group Working Paper* 2: https://www.academia.edu/50320075/A_nem_%C3%A1llami_kibert%C3%A9ri_szerepl%C5%91k_jelent%C5%91s%C3%A9ge_a_hibrid_konfliktusokban (accessed on: 10.06.2021).

KIS KELEMEN, B. 2018a. Drónok háborúja (1.) [The War of Drones (1)]. *Honvédségi Szemle* 1: 70–82.

2018b. Drónok háborúja (2.) [The War of Drones (2)]. *Honvédségi Szemle* 2: 16–29.

KITTRIE, R. F. 2016. *Lawfare. Law as a Weapon of War*. Oxford.

MACHIAVELLI, N. 2001. *A háború művészete* [The Art of War]. Szeged.

PATYI, A. 2015. Demokratikus legitimáció, választási felhatalmazás és alkotmány a haderő mögött [Democratic Legitimacy, Elections, and the Constitution Empowering the Armed Forces]. *Hadtudomány* 1–2: 72–75.

2016. A védelmi alkotmány alapkérdése: a fegyveres erő rendeltetése [The Fundamental Issue of a Defence Constitution Is the Intended Purpose of the

Armed Forces]. In: *Közjog és jogállam. Tanulmányok Kiss László professzor 65. születésnapjára* [Public Law and the Rule of Law: Studies for the 65th Birthday of Professor László Kiss]. Pécs. 233–249.

SARI, A. 2017. *Hybrid Warfare, Law and the Fulda Gap.* https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2927773_code957129.pdf?abstractid=2927773.&mirid=1 (accessed on: 30.07.2021).

2018. *Blurred Lines: Hybrid Threats and the Politics of International Law.* https://www.hybridcoe.fi/wp-content/uploads/2020/07/Strategic-Analysis-4-Sari.pdf (accessed on: 30.07.2021).

2019. *Legal Resilience in an Era of Gray Zone Conflicts and Hybrid Threats.* Exeter.

SPITZER, J. 2019. *Önvédelem versus terrorizmus: Az erőszak tilalma és az önvédelem joga a nemzetközi jogban, különös tekintettel az Iszlám Állam elleni nemzetközi fellépés lehetőségeire* [Self-Defence Versus Terrorism. The Prohibition of Violence and the Right to Self-Defence in International Law, with Special Regard to Potential International Actions against the Islamic State]. Budapest.

2020. A felfegyverzett drónok alkalmazásának egyes nemzetközi jogi kérdései [Certain International Law Implications of the Use of Armed Drones]. In: *Új típusú hadviselés a 21. század második évtizedében és azon túl. Intézményi és jogi kihívások* [New Type of Warfare in and beyond the Second Decade of the 21st Century. Institutional and Legal Challenges]. Buadpest. 172–190.

SZENDY, I. 2017. A hadviselés mint tudományelméleti és tudományrendszertani kategória [Warfare as a Category of Theory of Sciences and Scientific Systems]. *Hadtudomány* 3–4: 106–129.

VIKMAN, L. 2021. A műveletтervezés jogi feladatai [Legal Tasks Related to Operational Planning]. *Honvédségi Szemle* 2: 44–56.

Acta Univ. Sapientiae, Legal Studies, **11**, 2 (2022) 125–138

DOI: 10.47745/AUSLEG.2022.11.2.08

The 'Fertile Source' of Hungarian Constitutional Law: Thoughts on the 800-Year-Old Golden Bull

József Zoltán FAZAKAS
PhD, Associate Professor
Károli Gáspár University of the Reformed Church in Hungary (Budapest),
Faculty of Law
e-mail: fazakas.zoltan.jozsef@kre.hu

Abstract. In the year 2022, Hungary had the opportunity for a double celebration on the occasion of the tenth anniversary of the entry into force of the Fundamental Law and the eight hundredth anniversary of the Golden Bull issued by Andrew II. Eight hundred years ago, the Golden Bull, as one of the roots of the Hungarian historical constitution, formulated answers to certain questions of constitutional importance, which later proved to be suitable for the interpretation of power in Hungary as bound by law. In the context of constitutional law, the Golden Bull was one of the most important fundamental bills of noble liberties. The revolution of 1848, which laid the foundations of the modern constitutional state, was based precisely on the extension of these noble liberties, and thus some of their theses were also applied later. Because of the social change of the 19th century, which was partly inspired by it, the Golden Bull was a cardinal law, the basis of the modern Hungarian rule of law, which was valid as part of the historical constitution. Today, through the provisions of the Fundamental Law that name the historical constitution, it is not only a historical monument but the root of living law, and thus it retains its critical and interpretative significance. For these reasons, the present study outlines the relationship between the current Fundamental Law and the historical Hungarian constitution, cited in several provisions of the Fundamental Law, and then analyses the place of the Golden Bull as a cardinal law, the constitutional context of its origins in the Hungarian unwritten constitution, and its direct relationship with the constitutional revolution of 1848. The next part of the study explores the roots of those constitutional institutions that are still in force today, which can be derived directly or indirectly from the Golden Bull, thus paying homage to the eight-hundred-year-old source of constitutional law of the more than one-thousand-year-old Hungarian statehood, as well as to the current Fundamental Law and its promulgator, King Andrew II, who is often misunderstood by the public.

Keywords: Golden Bull, historical constitution, Fundamental Law, Hungarian constitutional law, monarchy

1. Introduction

In the year 2022, Hungary could celebrate a double constitutional anniversary: the tenth anniversary of the entry into force of the Fundamental Law, which is the basis of the country's current constitutional order, and also the eight hundredth anniversary of the Golden Bull issued by King Andrew II of Hungary (d. 1205–1235) on 24 April 1222 – if we accept the date around which the historical consensus has formed,[1] despite the difficulty of reconstructing the circumstances. The moment in relation to the historical horizon and the anniversaries of eight centuries linked to the millennium of statehood thus justify the commemoration of the Hungarian Golden Bull as 'this fertile source of our common law'.[2]

The relationship between the two highly significant legal sources is by no means a field for abstract speculation since the Preamble of Hungary's Fundamental Law, the National Avowal, also emphasizes that 'We honour the achievements of our historic constitution and we honour the Holy Crown, which embodies the constitutional continuity of Hungary's statehood and the unity of the nation. We hold that the protection of our identity rooted in our historic constitution is a fundamental obligation of the State.' Paragraph (3) of Article (R) specifically underscores this relationship by stating[3] that 'The provisions of the Fundamental Law shall be interpreted in accordance with their purposes, the National Avowal contained therein and the achievements of our historic constitution.' Consequently, the provisions above are not just a mere homage to the traditions of the Hungarian constitutional heritage; on the contrary, according to some authors, the unwritten Hungarian historical constitution is part of the Hungarian national and constitutional identity through its designation in the Fundamental Law,[4] and its achievements serve to reveal the correct content of the historical constitution; thus, the historical constitution not only lives as a set of cardinal laws and customs[5] but is also indispensably linked to the interpretation of the constitution today.

The Hungarian historical constitution is a set of specific laws, rules of customary law, and principles laid down in legal literature,[6] whose elements and results, i.e. achievements, provide fundamental assistance and guidance[7] to the constitutional institutions for the interpretation of the present-day state system[8] and the Fundamental Law. Consequently, especially the laws constituting the civic transformation of the nineteenth century that was partly inspired by the Golden

1 Zsoldos 2011. 4–5, 31–32.
2 Fest 1934. 273.
3 Árva 2013. 13–15, 66–67.
4 Sulyok 2016. 351–352.
5 Balogh 2016. 543.
6 Horváth 2022. 227.
7 Trócsányi 2014. 62–66.
8 Szabó 2016. 21.

Bull are institutions[9] that can be interpreted as part of the historical constitution as the foundations of the modern Hungarian rule of law, and thus are of critical and interpretative importance today.[10] It is precisely with this constitutional law perspective in mind that the following is an attempt to partially elicit the individual elements of the Golden Bull, which, as part of the historical constitution, as an achievement[11] of the organic development of the Hungarian state, were passed down[12] through the millennia into the acts of April 1848,[13] which form one of the foundations of modern Hungarian statehood and, as a result of these laws, are still part of living law today, even if by explicit codification.

2. The Golden Bull as a Source of Law and as a Cardinal Law of the Historical Constitution

The publication of the Golden Bull as a legal source is linked to the period of the dissolution of the patrimonial monarchy and the initial period of the development of the Hungarian estates.[14] In this period, Hungary, like many other European countries, was undergoing social, economic, political, and legal processes that foreshadowed a slow but inexorable transformation of the former state system. In the centuries-long transition from patrimonial monarchy to the monarchy of the estates, the significance and central role of the Golden Bull is inescapable despite the fact that in the course of later times not only Andrew II,[15] who granted the charter, but also his successors[16] and the kings from other dynasties who came later tried to relegate it to the background.[17]

At the same time, these circumstances and events must be interpreted in the context of the times, accepting the view that 'The past cannot be modified in retrospect according to wishful thinking, daydreams, or even tactical tricks.'[18] Accordingly, it should be noted that the Golden Bull was also the result of a sovereign decision and legislation of the monarchy,[19] and although it was issued at a particularly

9 Constitutional Court ruling 33/2012 (VII.17).
10 Horváth 2022. 228.
11 Id. 228–229.
12 Szabó 2020. 83–122.
13 The package of 31 articles of law sanctioned and issued by King Ferdinand V (r. 1830–1848) on 11 April 1848. See Toldy 1866. X, 279–307.
14 Hajnik 1867. 60–78.
15 Karácsonyi 1899. 4.
16 Hajnik 1867. 63, 71–73.
17 Ferdinandy 1899. 53–167.
18 Kosáry 1987. 5. Translation by the author.
19 Timon 1903. 100–104.

charged time of transition from patrimonial rule to monarchy-limited law,[20] its issuance was ultimately at the will of the monarch. Accordingly, it should be stressed that although the influence of political interest groups should not be understated in this period, the right to legislate was basically vested in the king until the establishment of the monarchical parliament or, more precisely, the monarchical dualism.[21] Another aspect of the issue is the fact that at the time of the publication of the Golden Bull, over the specific laws issued by the ruler,[22] the role of customary law was dominant,[23] which explains why the Golden Bull was subsequently confirmed and amended by separate laws.[24] These ratifications and amendments accompanied the Hungarian historical constitution throughout the period up to the beginning of the 20th century. Of these, the ratification of King Louis I (the Great) (r. 1342–1382) in 1351[25] was the most important.[26] It confirmed the effect of the provisions[27] laid down in the document; its constitutional nature[28] became undoubted centuries later, the last time it was part of the royal coronation oath[29] being[30] in 1916.[31]

1222 was a particularly turbulent year for the transition from patrimonial monarchy[32] to the monarchy of the estates, and recent research has shed new light on this year and the changes that took place, in contrast to the – in many cases trite – findings of earlier research.[33] Instead of the figure of a powerless king leading his country to ruin,[34] the exact content and circumstances of the Golden Bull's issuance unfold the image of a responsible statesman who wishes to set the Hungarian state on a new foundation.[35] This picture can basically be seen from two angles: firstly in the historical context of the processes of the time and secondly in its jurisprudential legacy.

As a result of recent research, the economic, historical, and political contexts of the first viewpoint have nuanced the earlier picture of the disintegration of

20 Zsoldos 2022. 11–13.
21 Timon 1903. 187.
22 Béli 2022. 122–126.
23 Hajnik 1872. 275–276.
24 Knauz 1869. 9–19.
25 Csukovits 2022. 190–200.
26 Béli 2022. 143.
27 Wenzel 1873. 4–5, 9–13.
28 Toldy 1866. VIII, 23–36.
29 Karácsonyi 1900. 68–69.
30 Ferdinandy 1899. 161–167.
31 Act III of 1917 on the enactment into the law of the land of the royal charter issued by His Majesty the King before his successful investiture and coronation and the royal oath taken at his coronation.
32 Timon 1903. 110–118.
33 Zsoldos 2022. 11–13.
34 Knauz 1869. 7. 77–79.
35 Zsoldos 2022. 14–31.

the patrimonial monarchy, focusing on the role of Andrew II in this process in a way that reveals the much more complex motivations for the monarch's actions. Accordingly, the reasons for the issuance of the Golden Bull are placed, from the monarch's point of view, in a complex reform process, the essential element of which was the introduction of modern monetary management, tax and customs policies, instead of the previous accumulation of royal wealth in kind, and thus aimed at reducing the political influence of the royal county officials.[36] In this context, the role of the royal servants (Lat. servientes) is also cast in a new light: far from being a vulnerable and subjected class, they are the natural political allies of the ruler against the former ruling class, and with their help and the redemptions they were given, the foundations of the new economic and political system were laid. The system of the noble estates supplanted the patrimonial royal county system.[37] In this light, the provisions of the Golden Bull, especially in the areas of property policy, financial management, and military organization, were in fact conscious steps to strengthen the servile class and to diminish the rights of the former county leaders so that the Hungarian state could embark on a more modern state organization and economic path in keeping with the times,[38] i.e. to become a monarchy complete with the estates.

The significance of the Golden Bull can be found precisely in this programme and in its development over centuries and its results: by articulating the need to subject power[39] previously considered unlimited, to constitutional limits, the Golden Bull carried with it a partial programme of the rule of law in our modern sense. In other words, the royal programme could only succeed in putting the Hungarian state on a new economic and political footing if the cooperation between the monarch and the supporting classes was real, and consequently the Golden Bull necessarily already represented the king's obligations[40] and the guarantee elements for enforcing his promises.[41]

The structure of the Golden Bull as a piece of legislation can also be judged by the above context and the characteristics of the time. The thirty-one articles of the Golden Bull are therefore far from being the result of codification in the modern sense but rather a collection of specific responses to the conditions of the time, to social, economic, and political needs, and a law of exceptional importance despite its mixed provisions that formulate the legal nature of power.[42]

On the other hand, with regard to the provisions of the Golden Bull, it is also worth noting that not all of its articles contained or could contain the roots of legal

36 Id. 15–16.
37 Id. 15–18.
38 Id. 18–28.
39 Timon 1903. 166.
40 Béli 2022. 140.
41 Zsoldos 2022. 25–28.
42 Horváth 2022. 232–233.

institutions of future relevance. Regardless of this, the later achievements of the historical constitution can be clearly traced back to the Golden Bull in certain areas, following the establishment of the monarchy and as a result of the peculiarities of the organic development of the Hungarian state.[43] Due to the characteristics of the development of the Hungarian state, some articles have inevitably become obsolete over the centuries, while others have grown in importance and thus have become the cornerstones of the historical constitution.[44] The latter articles as a framework legislation, providing ample scope for subsequent interpretation of the law and explanatory customary law,[45] can be seen essentially in the context of personal liberty and the fundamental rights of the nobility, legislation, the judiciary, and public administration. As a result of more than six centuries of legal development, these provisions of the Golden Bull became the basic tenets of the modern Hungarian state with the extension of the law in 1848.

The Golden Bull as the cardinal law of noble rights led to the unquestionable basic tenet of the Hungarian constitutional system – after the establishment of the estates until the laws of April 1848 passed by the last Diet of the estates of 1847/48 – that the nobility became the holders of political rights as equal members of the Hungarian nation meaning the totality of the estates.[46] The members of the Hungarian nobility – roughly 3.5–4%[47] of the country's population – enjoyed rights regardless of their language,[48] which gave rise to the legislative position in 1848 that the Hungarian nation was a Hungarian nation of the estates, i.e. the populus,[49] by abolishing feudal privileges and extending its rights to the plebians[50] excluded from political rights – and as a result creating a uniform law;[51] it goes without saying that the population of the country, including not only the nationalities but also the Hungarian-speaking population not considered as the Hungarian nobility, would be granted some rights.[52] The results expected from the legislative concepts thus also implied the joint rise of the Hungarian- and non-Hungarian-speaking social strata living under legal restrictions in order to achieve real equality of rights. The above programme was also fully in line with the social phenomenon of the acquisition of nobility and the attainment of noble freedom by individuals in order to enjoy full rights.[53]

43 Zsoldos 2022. 30–31.
44 Ferdinandy 1899. 168.
45 Horváth 2022. 236–237.
46 Szabó 1848. 60–61.
47 Kósa 2003. 33.
48 Szabó 1848. 60–61.
49 Timon 1903. 552–554.
50 Eckhart 1935. 242–244.
51 Szabó 1848. 110–120.
52 Szemere 1941. 39.
53 Hermann 2001. 147.

Accordingly, the starting point of the new Hungarian constitutional system established by the Fundamental Acts of April 1848[54] was the equality of rights based on the extension of the concept of the political nation of the estates and their rights,[55] which had its origins in the Golden Bull. The last Diet of the estates laid the foundations of the modern Hungarian state on the basis of the above thesis by abolishing the estates themselves with the Acts of April 1848, by which the principle of general equality of rights of the civic transformation was achieved without separate codification and without the formulation of a charter constitution,[56] but by abolishing the estates, abolishing the relations of subservience to a lord, introducing public taxation and codifying the most important rights,[57] the Hungarian State was given actual content and legal recognition and was transformed into a constitutional monarchy.[58]

Thus, the cornerstones of the Golden Bull concerning the freedom of the nobility, being directly related to the Acts of April 1848, carry the heritage that is also formulated in the current Fundamental Law or, as the heritage of the historical constitution, provide fundamental help and guidance for the interpretation of the current Fundamental Law,[59] pointing out its survival through the centuries.

3. The Relationship between the Golden Bull and the Current Fundamental Law in the Light of Certain Provisions of the Latter

In reviewing the provisions of the Golden Bull, it can be observed that both individual and collective perspectives on the limitation of power are prevalent and gain ground in the document. There is no difference in the constitutionalism of our time, where individual rights and individual freedom and community rights and freedom to exercise them in the community are also found together in the case of the Fundamental Law.[60] A crucial element of the individual perspective is the question of personal freedom, the guarantee of which, from the point of view of power, leads to the conclusion that individual freedom can be restricted, but that this should not lead to vulnerability, but that the restrictions must be justified from a constitutional point of view and essentially in the community interest.[61]

54 Toldy 1866. X. 279–307.
55 Szabó 1848. 110–112, 117–118, 119–120, 122–123.
56 Szabó 2015. 176–177, 182.
57 Csizmadia 1998. 295–297.
58 Sólyom 2019. 508–510.
59 Trócsányi 2014. 62–66.
60 Árva 2013. 93–97.
61 Ferdinandy 1899. 168–169.

The validity of the above statement from the end of the 19th century is still evident today; it means in fact that public and private interests in this context cannot lead to hierarchization, or, more precisely, 'neither the individual is for the public nor the public for the individual, but both are mutually for each other'.[62]

In accordance with the conditions of the times and the emerging order, the full enjoyment of the rights of individual liberty was enshrined as a right of the nobility, but the granting of these noble rights gave the opportunity to all the persons with these rights to become the counter-pole to royal power.[63] Consequently, the royal power, or, more precisely, the power of the state, was limited by individual liberty, which was further guaranteed by the fact that once recognized as an acquired right, liberty could no longer be challenged, withdrawn, or annulled by the king.[64]

A prominent element of the right to personal liberty is its individual and case-by-case restriction, which is essentially a feature of criminal law. In this context, Article II of the Golden Bull[65] sets out the basic conditions for lawful summons[66] and arrest.[67] The culmination of the centuries-long development of the cited provision of the Golden Bull was Act XXXIII of 1896 on the Code of Criminal Procedure, which exhaustively set forth the powers granted to the state, thereby limiting them similarly to the Dualism-era regulations that are also reflected in our current law.[68] Likewise today, Article XII,[69] which lays down the principle of individual criminal responsibility,[70] and Article XXVIII, which essentially contains the basic principle of public justice, are fundamental starting points of modern regulation.[71]

At first reading, the provision of Article IV of the Golden Bull seems to be a right of free testamentary disposition, but this norm was much more complex and contained rules that were only valid in medieval private law, referring to the rights of daughters and later those of entailment (*aviticitas*), so it was only valid in an orderly framework.[72] This article was effectively repealed in 1848 with the abolition of primogeniture,[73] but indirectly we find an institution that continues to this day in the declaration of the necessary succession of the king or, more precisely, of the state, in cases of vacant succession.[74]

62 Ferdinandy 1899. 169. Translation by the author.
63 Timon 129–131.
64 Ferdinandy 1899. 169.
65 Timon 1903. 130.
66 Horváth 2022. 240–242.
67 Ferdinandy 1899. 170–173.
68 Horváth 2022. 242.
69 Árva 2013. 113–116.
70 Ferdinandy 1899. 177.
71 Id. 179.
72 Horváth 2022. 243.
73 Ferdinandy 1899. 174.
74 Act V of 2013 on the Civil Code § 7:74.

The constitutionality of our times has been only indirectly influenced by Article XVII of the Golden Bull, which prohibits the repossession of land acquired in return for just services, and Article XXII, which states that nobles are not obliged to tolerate the king's pigs grazing on their property. In fact, the essence of the provisions is the undisturbed enjoyment and protection of possession or property,[75] and thus, along with the necessary distinction between constitutional and civil property,[76] they can be seen as one of the roots of the law of property today.

Summarizing the question of individual rights, it can be concluded that in the case of Hungary, the above-quoted provisions of the Golden Bull became a specific right and part of the historical constitution as a result of the organic development of the state through the constitutional extension[77] of 1848.[78] 'Individual freedom was guaranteed by the Golden Bull (…). That is, the nobility of the political nation of that time had all the rights that the whole nation has today.'[79]

The importance of the Golden Bull beyond individual rights is also confirmed by its provisions on the organization of the state. Some of its rules are still reflected in the constitutional structure of our times. The first of these articles is Article I of the Golden Bull, which deals with the celebration of St Stephen's Day and the king's personal jurisdiction or, in his absence, the justice served by the palatine and the right to lodge a complaint.[80] Also because of the complexity of its provisions, Article I of the Golden Bull is in fact the root of several legal institutions. These can be clearly identified and named, so in addition to the national or, more precisely, the state holiday,[81] they are the basic sources of justice and the right of citizens to complaint and redress. Although at first reading the article seems to be based on the right of recourse to the courts, the institution of the 'days when the law is laid down at Székesfehérvár' (the set periods when the king or the palatine serves justice) can be seen in fact as a precursor of the regularly convened Parliament, and the public law literature of the 19th century traces the institution of ministerial responsibility back to this norm.[82] The breakthrough and subordination of royal power by the Golden Bull, which had previously been considered unlimited, led to the later formulation of, among other things, the dualism of the estates, establishing the division of legislative power between the King and the Diet and the joint right of the latter.[83] Based on the traditional historical constitution and the doctrine of the Holy Crown,

75 Ferdinandy 1899. 177.
76 Árva 2013. 157–158.
77 Máriássy 1896. 194-195.
78 Eckhart 1935. 83–84.
79 Máriássy 1896. 13.
80 Ferdinandy 1899. 169–170.
81 Árva 2013. 49–50.
82 Horváth 2022. 237–240.
83 Ferdinandy 1906. 55–81, 106–126.

the above legislative power of the Diet was clearly established in the customary law in the Tripartitum and then as a specific norm of law in Article XVIII of the Act of 1635. The right of recourse to the courts,[84] the right of appeal to the authorities,[85] the right of complaint,[86] and the right of the supreme representative body of the people,[87] the National Assembly,[88] are all fundamental institutions of the modern Hungarian state.

Beyond the right of noble tax exemption,[89] Article III of the Golden Bull can in fact be interpreted as the basis for codifying legislation[90] that developed in later times, containing the right of Parliament to offer taxes, which in the modern constitutional framework is embodied in budgetary law.[91]

The rules of articles V, VI, VIII, and IX of the Golden Bull essentially contain the regulatory roots of the division of jurisdiction and powers in the administration of justice.[92] From these rules, it is a fundamental principle that no one may be deprived of the authority of a competent judge and that justice must be served without distinction of person.[93] The wording reflects today's expectations of equality before the law and fair trial.[94]

Article VII of the Golden Bull is the foundation of the constitutional institutions that continue to live on in terms of the military obligations of the nobles[95] and the conditions for the use of military forces abroad and at home.[96] In the course of the development of the state, this article was essentially recodified[97] during the era of dualism, and some of the basic principles were laid down that are still valid today and that are also valid for the current Fundamental Law in the context of the decisions of the National Assembly that created both the obligation to participate in the national defence and its material basis.[98] Article X of the Golden Bull, which ordered the rewarding of the sons of those who died heroically in war, can also be linked to this item, and this institution is still in force today in the rules of honour of the armed forces and law enforcement and disaster management agencies.

84 Árva 2013. 200–206.
85. Id 193–194
86. Id 195–196.
87. Id 188–192.
88. Id 222–258.
89 Timon 1903. 130.
90 Ferdinandy 1899. 173–174.
91 Árva 2013. 213–214.
92 Horváth 2022. 243.
93 Ferdinandy 1899. 174.
94 Árva 2013. 162–167, 201–206.
95 Timon 1903. 128.
96 Ferdinandy 1899. 174–176.
97 Horváth 2022. 243.
98 Árva 2013. 215–219, 228–229.

Article XI and Article XXIV of the Golden Bull[99] forbade foreigners, non-citizens in today's terms, to hold office. The right to hold office[100] without distinction of order was guaranteed by Article V of 1844,[101] and then the laws of denominational equality abolished the last vestiges of religious distinction. Nowadays, people are also entitled to hold various offices and positions without distinction, but Hungarian citizenship is still a basic requirement.[102]

Articles XIII, XIV, and XV of the Golden Bull provided protection against the excesses of public authority as we understand them today, and at the same time created the legal basis for the liability of public officials,[103] which can be found in the current Fundamental Law and enforced under the provisions of the relevant sectoral legislation.[104] By prohibiting the accumulation of offices, Article XXX of the Golden Bull also established an early conflict-of-interest rule, and it can be interpreted as its root,[105] these conflict-of-interest rules – in line with the principle of separation of powers[106] – still being very much in force today.

Various provisions of the other articles of the Golden Bull, such as Article XVI, Article XVIII, Article XIX, Article XX, Article XXI, Article XXIII, Article XXV, Article XXVII, and, finally, Article XXIX, were partially invalidated after their adoption and before the fall of the medieval Kingdom of Hungary in the period preceding the Battle of Mohács, or at the latest with the civic transformation of 1848,[107] so they have not had a demonstrable impact on the constitutionality of our times and can only be mentioned as monuments of legal history.

Last but not least, it is necessary to mention Article XXXI of the Golden Bull, which contained the famous resistance clause. The resistance clause was repealed by Article 1 of Act IV of 1687,[108] which confirmed the other provisions of the Golden Bull[109] even before the civic transformation, without ever having been applied,[110] with the result that some late-19th-century constitutional law scholars considered the elements of the constitutional guarantees and safeguards contained in the resistance clause to be valid unchanged – as contained in other legislation.[111]

The naming of individual rights and the rights of the state organization in the Golden Bull laid the foundations for the further development of the Hungarian

99 Ferdinandy 1899. 176–177.
100 Timon 1903. 130–131.
101 Eötvös 1903. 165.
102 Árva 2013. 188–192.
103 Ferdinandy 1899. 177–178.
104 Árva 2013. 193–194.
105 Ferdinandy 1899. 179.
106 Árva 2013. 29–31.
107 Ferdinandy 1899. 178–179.
108 Béli 2022. 143–144.
109 Article IV of the Act of 1687 about Article 31 of the Act of 1222 by King Andrew II (of Jerusalem) is explained in some parts.
110 Horváth 2022. 244–251.
111 Ferdinandy 1899. 180–181.

historical constitution. The renowned Hungarian public lawyer, Géza Ferdinandy, summarized the significance of the Golden Bull and its fundamental constitutional legacy in ten points, stating that these items should be considered the basis of the rights and constitutional status of all Hungarian citizens by the extension of the law. According to Ferdinandy, the Golden Bull: (1) fixed the limits of royal power and its subordination to public law; (2) established the right of citizens to petition and complaint; (3) ensured personal liberty and the inviolability of property; (4) established the Parliament and its rights to levy taxes and decide on national defence – through the new enactment; (5) established the exclusive right of Hungarian citizens to hold office; (6) established the criminal and property liability of civil servants; (7) established the territorial integrity of the country as a fundamental principle; (8) established the conditions for the legitimate exercise of judicial power and the principles of justice, including the right of access to the courts; (9) established the preservation of the Constitution as a fundamental principle; (10) established the right of passive resistance of the nation and its citizens.[112]

4. Conclusions

The Golden Bull, as one of the roots of the Hungarian historical constitution, could not have prevailed for centuries if the principles and provisions it contained had not been effective, sufficiently flexible, and capable of dealing with the challenges that have arisen over the centuries. The significance of the royal document, therefore, lies precisely in the fact that eight hundred years ago it formulated answers to certain questions of constitutional significance that later proved to be suitable for the interpretation of power in Hungary as bound by law. Of course, the Golden Bull was born in an era in which constitutionalism in the modern sense could only be understood by a narrow stratum, but in the context of constitutional law, the Golden Bull ultimately laid the foundations for the legal binding of power, and by formulating these principles, Hungarian constitutional law was able to extend the principles it contained in 1848 and lay the foundations of the modern rule of law. After the establishment of the monarchy, when the nobility became the holders of political power, all the rights of the nobility were traced back to the Golden Bull, and the Hungarian nobility imagined equality of rights by extending these rights. Accordingly, during the period of the constitutional monarchy and under the historical constitution until the end of the Second World War, the Golden Bull and the historical constitution could be directly enforced and, on the whole, it resulted in a constitutional state governed by the rule of law.

112 Id. 181.

The Golden Bull was therefore a cardinal piece of legislation as part of the historical constitution because of the civic transformation of the 19th century, which was partly inspired by it, and because of the foundation of the modern Hungarian constitutional state. Today, the provisions of the Fundamental Law that name the historical constitution make it not only a historical monument but also the root of living law, and thus it is still of critical and interpretative importance.

A brief review of the provisions of the Golden Bull, the circumstances of its publication, and its legacy over the centuries and the work of historians and public lawyers in the 19th century, as well as the results of new research, should help to clarify the public image of Andrew II as a powerless king.[113] It is precisely because of the Golden Bull and its constitutional legacy that, instead of a king who was powerless and led his country to ruin, we can honour a responsible statesman who put the Hungarian state on a new footing in the person of King Andrew II.

References

ÁRVA, Zs. 2013. *Kommentár Magyarország Alaptörvényéhez*. Budapest.

BALOGH, E. 2016. Alkotmányunk történetisége, kitekintéssel az Alkotmánybíróság judikatúrájára. In: *Számadás az Alaptörvényről*. Budapest.

BÉLI, G. 2022. II. András korabeli jogforrások különös tekintettel az Aranybullára. In: *Aranybulla 800*. Budapest.

CSIZMADIA, A. 1998. Az állampolgári jogegyenlőség, a földesúr–jobbágy viszony felszámolása és a szabadságjogok a forradalom és a szabadságharc alatt. In: *Magyar állam- és jogtörténet*. Budapest.

CSUKOVITS, E. 2022. Az Aranybullát átíró 1351. évi törvény. In: *Aranybulla 800*. Budapest.

ECKHART, F. 1935. *Magyarország története*. Budapest.

EÖTVÖS, J. 1903. Az 1843-44-iki országgyűlésről. In: *A nemzetiségi kérdés*. Budapest.

FERDINANDY, G. 1899. *Az arany bulla*. Budapest.

1906. *A magyar alkotmány történelmi fejlődése*. Budapest.

FEST, S. 1934. Magna Carta – Aranybulla. *Budapesti Szemle* 1934/682.

HAJNIK, I. 1867. *Magyarország az Árpád-királyoktól az ősiségnek megállapitásáig és a hűbéri Európa*. Pest.

1872. *Magyar alkotmány és jog az Árpádok alatt*. Pest.

HERMANN, G. M. 2001. A reformkori nemesi liberalizmus székelyföldi lecsapódása a korabeli sajtó tükrében. In: *A székelység története a 17–19. században*. Mierscurea Ciuc.

113 Knauz 1869. 7, 77–79.

HORVÁTH, A. 2022. Az 1222. évi Aranybulla mint történeti alkotmányunk sarkalatos törvénye a „hosszú" 19. század közjogi irodalmában. In: *Aranybulla 800*. Budapest.

KARÁCSONYI, J. 1899. *Az Aranybulla keletkezése és első sorsa*. Budapest.
1900. Ismertetés: Az arany bulla. Közjogi tanulmány. Irta dr. Ferdinandy Gejza. *Századok* 1900/1.

KNAUZ, N. 1869. II. Endre szabadságlevelei. *Értekezések a Történeti Tudományok köréből* 1. köt.

KÓSA, L. 2003. *Nemesek, polgárok, parasztok*. Budapest. 2003.

KOSÁRY, D. 1987. Előszó. In: *A történelem veszedelmei*. Budapest.

MÁRIÁSSY, B. 1896. *A szabadelvüség multja, jelene és jövője*. Győr.

SÓLYOM, L. 2019. Az Alkotmány emberi jogi generálklauzulájához vezető út. In: *Documenta–Alkotmányjog*. Budapest.

SULYOK, M. 2016. *Kettő az egyben? Alkotmány és identitás*. In: *Számadás az Alaptörvényről*. Budapest.

SZABÓ, B. 1848. *A magyar korona országainak státusjogi és monarchiai állása a Pragmatica Sanctio szerint*. Pozsony.

SZABÓ, I. 2015. Történeti alkotmány a polgári a korban. In: *A Hármaskönyv 500. évfordulóján. A boldogságos Szent Erzsébet özvegy ünnepén*. Budapest.
2016. Közjogi hagyományok és jogtörténet. *Jogtörténeti Szemle* 2016/3.
2020. Az ősi alkotmány. In: *A magyar közjog alapintézményei*. Budapest.

SZEMERE, B. 1941. *Szemere Bertalan miniszterelnök emlékiratai az 1848/49-i magyar kormányzat nemzetiségi politikájáról*. Budapest.

TIMON, Á. 1903. *Magyar Alkotmány- és jogtörténet különös tekintettel a nyugati államok jogfejlődésére*. Budapest.

TOLDY, F. 1866. *A Magyar Birodalom Alaptörvényei*. Pest.

TRÓCSÁNYI, L. 2014. *Az alkotmányozás dilemmái*. Budapest.

WENZEL, G. 1873. Adalék 1352-ből az Aranybulla néhány czikkének alkalmazásához és magyarázatához. *Értekezések a Történeti Tudományok köréből* (III)2.

ZSOLDOS, A. 2011. II. András Aranybullája. *Történelmi Szemle* 2011/1.
2022. Az Aranybulla és története. In: *Aranybulla 800*. Budapest.