Artificial Intelligence and Human Rights: Their Role in the Evolution of AI

*Themis Tzimas**

Abstract

The article analyzes the role of the rule of law and of human rights in relation to the regulation of Artificial Intelligence (AI) and its evolution.

The main argument of the article is that the expanding intellectual autonomy of AI at the level of Artificial Narrow Intelligence (ANI), but even more emphatically in the case of Artificial General and Artificial Super Intelligence (AGI and ASI respectively) transforms social relations and the human centric character of present legal systems, both nationally and internationally.

On such grounds, the article examines the ontological elements of AI and the critical questions which are raised in relation to AGI and ASI, such as the prospect of friendly or unfriendly AGI and ASI.

In the face of such developments the rule of law as the principle and the organizational scheme which institutionalizes justice faces the challenge of a fundamentally new social and legal landscape. In order to identify the role of the rule of law in such a framework, the article examines the rule of law

* Post Doc Researcher at the Faculty of Political Science, Aristotle University of Thessaloniki.

ZaöRV 80 (2020), 533-557

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

from a thick perspective and as a concept not only of national, but also of international dimension. In addition, given its historical evolution and its human-centered character, human rights are viewed as inherent in the rule of law.

The article suggests that human rights can play an important role in terms of machine learning of AI. It also makes an argument against the legitimacy of AGI and ASI because of human rights and of the human-centered nature of existing legal systems.

I. Introduction

Technological, as well as social, economic and political developments are already defined and will be further defined in the future by the rise of Artificial Intelligence.[1] AI gives rise to expectations but also to concerns.[2]

At their core lays the unique ontology of AI, which is built on its growing and expanding autonomy.[3] This expanding intellectual autonomy creates new social relationships, transforms pre-existing ones and subsequently poses unique legal challenges. It is in such a framework that the rule of law and human rights as fundamental, integral part of the rule of law can play a crucial role, as a benchmark regarding what type of AI can be legitimized.

In order to examine the role of the rule of law and human rights, in relation to AI, the article in part 1, articulates – very briefly – the relation between rule of law and human rights; it moves on in part 2 with the ontology of AI and then, in part 3, the role of the human rights, in the development of AI are examined.

II. Rule of Law and Human Rights

In the present article as it is analyzed below, the rule of law is approached from a double perspective: from the so-called, "thick" approach and as a

[1] *D. Ben-Ari/Y. Frish/A. Lazovski/U. Eldan/D. Greenbaum*, "Danger, Will Robinson"? Artificial Intelligence in the Practice of Law: An Analysis and Proof of Concept Experiment Richmond Journal of Law and Technology 23 (2017), 3 et seq., (10).

[2] *S. Yanisky-Ravid/L. A. Velez-Hernandez*, Copyrightability of Artworks Produced by Creative Robots and Originality: The Formality-Objective Model, Minnesota Journal of Law, Science & Technology 19 (2018), 1 et seq. (4 et seq.); *D. A. Larson*, Artificial Intelligence: Robots, Avatars, and the Demise of the Human Mediator, Ohio State Journal on Dispute Resolution 25 (2010), 105 et seq. (106).

[3] *S. Yanisky-Ravid/L. A. Velez-Hernandez* (note 2), 7.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

concept not only of national but of international impact as well. It is also suggested that in order to identify the significance of human rights within the rule of law we need to look into the historical formation of the rule of law as a creation of the era of modernity.

More specifically, the rule of law constitutes both a principle and an organizational "scheme" of the constitutional law[4] and therefore of states, too. It has emerged via different politeiological and constitutional "histories".[5]

The Venice Commission definition of the Rule of Law refers to

> "(1) Legality, including a transparent, accountable and democratic process for enacting law (2) Legal certainty (3) Prohibition of arbitrariness (4) Access to justice before independent and impartial courts, including judicial review of administrative acts (5) Respect for human rights (6) Non-discrimination and equality before the law".[6]

The rule of law is approached in "thin" or "thick" ways; the former comprehend the rule of law mainly on the basis of its formal features, such as the separation of powers and a system of laws which are general, public, prospective, clear, consistent, capable of being followed, stable, and enforced,[7] whereas the thick version combines the former with specific political morals, human and social rights,[8] seeking more "substantive" justice, on top of formal legality.[9]

In the framework of thick approaches, one school of thought comprehends only specific human rights as relevant for the rule of law: namely those which refer to liberty, as well as to the judicial procedures guarantees. However, this school of thought to some extent concedes to the same thin approaches that it attempts to surpass, as it limits the substantial content of the rule of law and the variety of rights that the latter incorporates, to the ones which are necessary for the formalistic definitions of the rule of law.

---

4 *L Pech*, The Rule of Law as a Constitutional Principle of the European Union, Jean Monnet Working Paper Series No. 4/2009, 42.

5 The UK-originated rule of law is based largely on judicial decisions, the "Rechtsstaat" from written constitutions and the nature of the state, the French "État de droit" comprehends the state as guarantor of fundamental rights. <https://www.venice.coe.int>.

6 <https://www.venice.coe.int>.

7 *J. C. S. Ochoa*, Towards a Holistic Approach, in International Practice, to the Design and Implementation of Initiatives to Promote the Rule of Law at the National Level, International Journal of Law in Context 11 (2015), 78 (81); *J. Jowell*, The Rule of Law, in: J. Jowell/D. Oliver (eds.), The Changing Constitution, 2015, 13 et seq.; *L. L. Fuller* The Morality of Law: Storrs Lectures on Jurisprudence, 1977, 1; *J. Raz*, The Rule of Law and Its Virtue, in: J. Raz, The Authority of Law: Essays on Law and Morality, 1979, 210.

8 *M. Zuřn/A. Nollkaemper/R. Peerenboom,* Introduction, in: M. Zuřn/A. Nollkaemper/R. Peerenboom (eds.), Rule of Law Dynamics, 2012, 1.

9 *R. Dworkin*, A Matter of Principle, 1985, 11 et seq.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

Instead if we comprehend the rule of law as a principle and as an organizational scheme, which institutionalizes the concept of justice – *beginning from but not limited to the organization of the state and the form of legislating* – then the whole range of human rights, including civil and political rights must be endorsed within the rule of law.[10] In this sense, the rule of law surpasses the mere level of governance according to the law.

Under such a view however, the rule of law exceeds not only the formalistic, thin approaches within the national states but – *given that the concept of justice transcends the international community as well* – the rule of law exceeds national states, too. If the concept of justice is perceived to be of universal implementation and a foundation of the international community – *as it must be* – then the rule of law, as the principle and the scheme which institutionalizes justice surpasses national states, is applying to the international community as well.[11]

This approach to the rule of law is not unanimously accepted; characteristically, *James Crawford* argues that only a clear International Court of Justice (ICJ) jurisdiction to review judicially the actions of all United Nations (UN) political agencies could establish the rule of law in international political life.[12] The argument implies that since the rule of law at the international level cannot duplicate the characteristics that it has at the national level, it does not exist. It is based on the inaccurate assumption that the rule of law at the national and at the international level must be identical.[13]

It is obvious that the rule of law as an organizational scheme at the national level cannot be replicated at the largely state-centered and mainly horizontally constructed international level. Nevertheless, if the rule of law is comprehended holistically – *as it must be* – with the emphasis not solely on its formal(-istic) characteristics as they are developed in the framework of national states but upon its substance and goals – *i.e., the institutionalization of the concept of justice and the inclusion in this institutionalization of a variety of rights including human rights, too* – then it can be identified at the level of the international community as well.[14]

---

[10] *A. Bedner*, An Elementary Approach to the Rule of Law, Hague Journal on the Rule of Law 2 (2010), 48.

[11] *N. Barber*, The Rechtsstaat and the Rule of Law, U. Toronto L. J. 53 (2003), 443 (452).

[12] *J. Crawford/S. Marks*, The Global Democracy Deficit: An Essay in International Law and Its Limits, in: D. Archibugi/D. Held/M. Kohler (eds.), Re-Imagining Political Community, 1998, 84.

[13] *I. Hurd*, The International Rule of Law: Law and the Limit of Politics, Ethics & Int'l Aff. 28 (2014), 39 et seq. (39).

[14] *A. Watts*, The International Rule of Law, GYIL 36 (1993), 15 et seq. (25, 41).

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

In other words, the role of the rule of law – *meaning the institutionalization of the concept of justice* – and a spectrum of procedural guarantees and rights which "substantiate" this role, albeit in transformed ways when applied at the international level – *i.e., independent judiciary authorities, legal certainty, equality against the law, human, civil and political rights* – establish the rule of law as a foundation of international law, too.[15]

Even the violations of some of these rights or of procedural guarantees, by states or other actors do not *per se* nullify the rule of law; on the contrary they ascertain its existence as well as its necessity.[16]

The emergence of the rule of law at the level of the international community can take several forms:[17] *Simon Chesterman* talks about three possible levels, with the first one being the application of rule of law between states and other subjects of international law, the second referring to the international law supremacy principle, engulfing the human rights norms and standards over domestic legal systems, and the third to a global rule of law which is exercised on individuals directly without the mediation of national law.[18]

Simon Chesterman advocates the first approach as the one providing legal certainty. However, this approach fails to capture the evolution of the international legal order, at least since World War II. Several developments and documents in the post-World War II period prove that international law and the rule of law are built on all three forms in a combined way.

Amongst these documents are the Declaration on the Rule of Law, which combines the rule of law with the concept of justice,[19] the Declaration on Principles of International Law Friendly Relations and Cooperation among States in accordance with the Charter of the United Nations, which mentions the "paramount importance of the Charter of the United Nations in the promotion of the rule of law among nations",[20] and the Universal Declaration of Human Rights, which foresees that

---

[15] *J. Waldron*, Are Sovereigns Entitled to the Benefit of the International Rule of Law?, EJIL 22 (2011), 315 et seq. (316 et seq.); *R. McCorquodale*, Defining the International Rule of Law: Defying Gravity?, ICLQ 65 (2016), 277 et seq. (292).

[16] *B. Tamanaha*, On the Rule of Law: History, Politics, Theory, 2004, 131.

[17] *J. C. S. Ochoa* (note 7), 78.

[18] *S. Chesterman*, "I'll Take Manhattan": The International Rule of Law and the United Nations Security Council, Hague Journal on the Rule of Law 1 (2009), 67 et seq. (68 et seq.); *J. Gathii*, Good Governance as a Counter Insurgency Agenda to Oppositional and Transformative Social Projects in International Law, Buffalo Human Rights Law Review 5 (1999), 107 et seq. (121 et seq.).

[19] Para. 2, Declaration on the Rule of Law.

[20] Declaration on Principles of International Law Friendly Relations and Cooperation among States in Accordance with the Charter of the United Nations, UN GAOR Res. 2625

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

> "it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by the rule of law",

establishing a direct relationship between the rule of law in international law and the people, without the need for national states' mediation.[21]

According to the UN Guidance Note of the UN Secretary-General on the UN Approach to Rule of Law Assistance, which was published in 2008,

> "the rule of law is a principle of governance in which all persons, institutions and entities, public and private, including the State itself, are accountable to laws [...]".[22]

All of the above documents endorse the thick, "inclusive" and holistic approach, to the rule of law,[23] as well as its comprehension as an international concept, with implementation directly on peoples, too, potentially bypassing the will of the states. Support for this approach can be found in the UN Charter, in Art. 1, para. 3, as well as in Arts. 55 and 56.[24]

Summing up, the rule of law emerges at the international level, too, in the sense of the institutionalization of the concept of justice in the international community, comprised of principles and rights that provide it a substantial content, including human rights,[25] and applying directly both to states and to non-state actors.

In order to understand why human rights are essential in the rule of law, we need to think in pre-legal, philosophical and historical terms about the rule of law, namely that it is the era of modernity which produced the concept of the rule of law, in conjunction with the prevalence of human-centered legal systems and a human-centered concept of justice.

On such grounds human rights emerge as integral part of the rule of law because they constitute the preeminent set of rights which fortify the focus on humans of legal systems. Therefore, if the rule of law is to maintain its complete character and nature, it must include human rights.

(XXV) (1970). Para. 7, UN Millennium Declaration 2000, UN Doc. A/Res/55/2. UN Sustainable Development Goals 2015, included as the 2030 Agenda for Sustainable Development, Arts. 8, 9 and 35, UN Doc. A/69/L.85. Access to justice is also included as Goal 16 and Targets. UN Commission on Legal Empowerment of the Poor, Making the Law Work for Everyone, UNDP 2008.

21 Preamble, Universal Declaration of Human Rights 1948.

22 Guidance Note of the Secretary-General, UN Approach to Rule of Law Assistance, (April 2008).

23 *R. Kleinfeld*, Advancing the Rule of Law Abroad: Next Generation Reform, 2012, 16 (92 et. seq.).

24 Arts. 1(3), 55 and 56 UN Charter.

25 Preamble and Art. 1 UN Charter.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

The normative interlinkage between the rule of law, human rights and the human focus of legal systems is central to the significance of AI evolution – both present and potential – as the latter heralds potentially far-reaching transformations of the human focus of the systems. In order to substantiate this analysis, the ontology of AI is examined below.

III. The Ontology of AI

1. Defining AI

What is eventually AI? What is its ontology? Although much ink has been spilled over the issue and there are several definitions and descriptions, the truth is that ambiguity surrounds AI to a significant extent. "[I]n spite of what I regard as AI's significant achievements […] the not so well-kept secret is that AI is internally in a paradigmatic mess", *Chandrasekaran* comments.[26]

The definition of AI constitutes a matter of controversy, too,[27] which results from the different levels of AI development as well as from the extent of anthropomorphism through which we approach AI. Several definitions emphasize the algorithms and technology, which confer upon AI human-like functions or functions which, if conducted by humans, would be considered as the outcome of intelligent activities.[28]

The focus on the "human-like" intelligence of machines,[29] despite being descriptively helpful,[30] can be also deceiving, either because an entity mimicking human intelligence does not necessarily "understand" or share the patterns of human intellect, or because AI entities may develop intelligence of equal or superior level to human intelligence, but not identical to it.[31]

This is why other definitions attempt to capture more thoroughly the significant differences between human and artificial intelligence – with the latter remaining up to a large extent a "black box" for us until now – as well

---

[26] *B. Chandrasekaran*, What Kind of Information Processing Is Intelligence?, in: D. Partridge/Y. Wilks (eds.), The Foundations of Artificial Intelligence, 1990, 14.

[27] *S. J. Russell/P. Norvig*, Artificial Intelligence: A Modern Approach, 2013, 2; *T. Winograd*, Thinking Machines: Can There Be? Are We?, in: D. Partridge/Y. Wilks (note 26), 167.

[28] *M. U. Scherer*, Regulating Artificial Intelligent Systems: Risks, Challenges, Competences, and Strategies, Harvard Journal of Law & Technology 29 (2016), 353 et seq. (363 et seq.).

[29] *D. Laton*, Manhattan_Project.Exe: A Nuclear Option for the Digital Age, Catholic University Journal of Law & Technology 25 (2016), 94 et seq.

[30] *S. J. Russell/P. Norvig* (note 27), 3.

[31] *J. McCarthy*, What Is Artificial Intelligence?, 2007, <http://www.formal.stanford.edu>.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

as the transformations that concepts of intelligence undergo when projected from human on AI intelligence.

These definitions emphasize goal-oriented functions and machine-learning capacities as well as specific intellectual characteristics which evolve in the course of goal oriented behaviors and machine learning.[32] Amongst these characteristics can be "consciousness, self-awareness, language use, the ability to learn, the ability to abstract, the ability to adapt, and the ability to reason".[33]

Within this framework, *Stuart Russell* and *Peter Norvig* among eight definitions of AI prefer the "rational agent" definition, according to which AI agents

> "operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue [the best expected outcome]".[34]

For the purpose of this article, *Russell's* and *Norvig's* definition seems the most helpful, as it can be used both for ANI and for ASI and AGI, as it avoids anthropomorphism and, instead of drifting into lengthy analyses about the transfer of certain intellectual characteristics from human to artificial intelligence, describes a basic set of functions which can be objectively traced in AI.

Therefore, given the mainly legal perspective of the present article, the above-mentioned definition can provide clarity and consistency.

2. Weak and Strong AI

An important distinction in the field of AI is between weak AI, where "the computer is merely an instrument for investigating cognitive processes" and strong AI, where "[t]he processes in the computer are intellectual, self-learning processes".[35] Weak AI is labeled as Artificial Narrow Intelli-

[32] *S. M. Omohundro*, The Basic AI Drives, in: P. Wang/B. Goertzel/S. Franklin (eds.), Artificial General Intelligence 2008. Proceedings of The First AGI Conference, 483 et seq.

[33] *M. U. Scherer* (note 28), 360.

[34] *S. J. Russell/P. Norvig* (note 27), 2 et seq.

[35] *G. Wisskirchen/B. Thibault Biacabe/U. Bormann/A. Muntz/G. Niehaus/G. Jiménez Soler/B. von Brauchitsch*, Artificial Intelligence and Robotics and Their Impact on the Workplace, 10.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

gence while strong AI is further distinguished between Artificial General Intelligence and Artificial Super Intelligence.[36]

The applications of ANI are all around us already, via computers where "intelligent systems have been taught or learned how to carry out specific tasks without being explicitly programmed how to do so", and include both peaceful and military uses.[37]

ANI applications are built on deep learning – *which is based on algorithms and mimics human cognitive functions, having the potential of learning from mistakes* – robotization, with innovations such as 3-D printers and self-learning capacities, de-materialization, in the sense that autonomous software will be collecting data as well as that physical products will become software, gig economy – *which expands self-employment in the sense of crowd working and working on apps* – and autonomous driving.[38]

An element regarding ANI expanding autonomy becomes obvious in applications which "guess" our choices about several products, write articles in newspapers,[39] create novel art[40] and intrude private corporations' business cycle, via the automation of back-office processes.

The fourth industrial revolution constitutes one of the most characteristic areas of ANI implementation with the so-called cyber physical systems – CPS – which refer to "the network connections between humans, machines, products, objects and ICT (information and communication technology) systems",[41] and the emergence of fully automated or so-called "smart" factories and services, with the capacity to provide individualized responses to customers' supposed needs.

ANI applications are present not only in peaceful uses, but in military ones, too, as the debate about killer robots shows. While the automation of weapons is not a completely novel issue,[42] the implementation of AI for military use is not restrained merely in automation, but expands to the field

---

[36] *T. Urban*, The AI Revolution: The Road to Superintelligence, Wait But Why, 2015, <www.waitbutwhy.com>.

[37] *N. Heath*, What is AI? Everything You Need to Know about Artificial Intelligence, 2018, <https://www.zdnet.com>.

[38] *G. Wisskirchen/B. Thibault Biacabe/U. Bormann/A. Muntz/G. Niehaus/G. Jiménez Soler/B. von Brauchitsch* (note 35), 10 et seq.

[39] *N. Sahota*, A.I. May Have Written This Article. But Is That Such a Bad Thing?, 2018, <https://www.forbes.com>.

[40] *A. Elgammal*, Meet AICAN, A Machine that Operates as an Autonomous Artist, The Conversation, 2018, <http://theconversation.com>.

[41] *G. Wisskirchen/B. Thibault Biacabe/U. Bormann/A. Muntz/G. Niehaus/G. Jiménez Soler/B. von Brauchitsch* (note 35), 12.

[42] *M. Ryder*, Killer Robots Already Exist, and They've Been Here a Very Long Time, The Conversation, 2019, <https://theconversation.com>.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

of combat as well as to the decision-making process, with the goal of minimizing human decision-making and political cost, as well as maximizing military efficacy.[43]

However, although ANI has already "outsmarted" humans in certain narrow areas and tasks, it cannot yet compete with humans in terms of adaptable and general intelligence, which eventually could raise general artificial intelligence to a level at least equal to that of humans.

AGI is expected to consist of the

> "type of adaptable intellect found in humans, a flexible form of intelligence capable of learning how to carry out vastly different tasks, anything from haircutting to building spreadsheets or to reasoning about a wide variety of topics, based on its accumulated experience".[44]

A crucial development is therefore the adaptability, the flexibility which allows the entity to choose on its own, where and how to apply its intelligence. The demonstration of

> "a reasonable degree of self-understanding and autonomous self-control, the ability to solve a variety of complex problems in a variety of contexts, and [the ability to] learn to solve new problems that it didn't know about at the time of [the entity's] creation".[45]

The "when" of AGI is debatable, although most analysts agree that it will emerge before the end of the century.[46] In principle, the lesser AI is based on programming and the more it is based on experience and learning, the closer it gets to AGI.[47]

There is extended literature already about the potential evolutionary patterns for designing AGI, as well as the difficulties ahead. The main concept is condensed in the effort to reproduce, either through genetic algorithms or via other, evolutionary algorithms and means, the natural evolutionary pattern, albeit without the "failures" of the natural environment and therefore in a short period and – up to some extent – in a protected environment. The

---

[43] *J. Rohrlich*, The US Army Wants to Turn Tanks Into AI-Powered Killing Machines, Quartz, 2019, <https://qz.com>.

[44] *N. Heath* (note 37).

[45] *C. Pennachin/B. Goertzel*, Preface, in: B. Goertzel/C. Pennachin (eds.), Artificial General Intelligence, 2007, vi.

[46] *D. Tal*, Forecast, How the First Artificial General Intelligence Will Change Society: Future of Artificial Intelligence P2, Quantumrun Special Series, 2018, <https://www.quantumrun.com>.

[47] *D. Tal* (note 46).

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

achievement of such a goal will be the outcome of different components, including algorithms, software and hardware technologies.[48]

Super intelligence moves one step further and refers to the exceeding of human intelligence in the sense of "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest".[49] Or as *Bostrom* had suggested:

> "By a 'superintelligence' we mean an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills."[50]

While the point in time where super intelligence will be attained remains open, its achievability is foreseen with some certainty.[51] After all, since human brain performs computation, in the sense that it "deals in information, converting a pattern of input nerve signals into output nerve signals", a different, non-biological computational entity could perform like the human brain, and eventually out-perform it.[52]

ASI is expected to possess two characteristics regarding its potentially exponential growth: the continuous invention of new machines - *referring both to software and eventually hardware* - by the ASI entities on the one hand and the subsequent acceleration of this procedure, up to explosive levels, on the other.[53] This will eventually constitute the moment of AI achieving "singularity".[54]

---

[48] *C. Shulman/N. Bostrom*, How Hard Is Artificial Intelligence? Evolutionary Arguments and Selection Effects, Journal of Consciousness Studies 19 (2012), 103 et seq.

[49] *N. Bostrom*, Superintelligence, Paths, Dangers, Strategies, 2014.

[50] *N. Bostrom*, How Long Before Superintelligence?, <https://nickbostrom.com>.

[51] *S. Hawking/M. Tegmark/S. Russell/F. Wilczek*, Transcending Complacency on Superintelligent Machines, The Huffington Post, <https://www.huffingtonpost.com>.

[52] What is still missing is the "raw computing power" but there are several other ways that the gap in these regards could close. *A. Snyder-Beattie/D. Dewey*, Explainer: What Is Superintelligence?, The Conversation, 2014, <https://theconversation.com>.

[53] *E. Yudkowsky*, Staring at the Singularity, 1996, <http://yudkowsky.net>.

[54] *D. Chalmers*, The Singularity, A Philosophical Analysis, Journal of Consciousness Studies 17 (2010), 7 et seq. (9); *E. Yudkowsky*, Three Major Singularity Schools, <http://yudkowsky.net>.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

3. AI, Machine Learning and Neural Networks

In order to understand how AI moves from ANI to AGI and ASI, as well as how the developing intellectual autonomy of AI functions,[55] one has to look to the "machine-learning" procedure, which is comprised of a performance and of a learning element. The first one "senses the environment", while the latter employs feedback from the system and amends the performance element.[56]

There are several definitions for machine learning. A comprehensive one defines machine learning as

> "the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding data and information in the form of observations and real-world interactions".[57]

Other definitions of machine learning refer to "the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world".[58]

Machine learning thus resembles more to "coaching" than programming[59] and attempts to mimic human learning procedure.[60] It can also be described through the cumulative contribution of three abilities: to compute information, to learn and to reason.[61]

Machine learning is already giving way – at least to some extent – to neural networks and deep learning. Roughly speaking, neural networks are inspired by the human brain and the synapses between neurons functioning at different layers, through which massive data run, in order to train the system.[62]

---

55 *F. MacDonald*, Harvard Scientists Think They've Pinpointed the Physical Source of Consciousness, Science Alert, 2018, <www.sciencealert.com>.

56 *W. C. Marra/S. K. McNeil*, Understanding "The Loop": Regulating the Next Generation of War Machines, Harv. J. L. & Pub. Pol'y 36 (2013), 1139 et seq. (1145).

57 *D. Faggella*, What is Machine Learning?, Emerj, <https://emerj.com>.

58 *M. Copeland*, What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?, nvidia, 2016, <https://blogs.nvidia.com>.

59 *W. Kowert*, The Foreseeability of Human-Artificial Intelligence Interactions, Tex. L. Rev. 96 (2017), 182 (183); *M. U. Scherer* (note 28), 365.

60 *A. Schuller,* At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law, Harvard National Security Journal 8 (2017), 379 (396).

61 *A. Khoury*, Intellectual Property Rights for "Hubots": On the Legal Implications of Human-Like Robots as Innovators and Creators, Cardozo Arts and Entertainment Law Review 35 (2017), 635 et seq. (640).

62 <https://developer.nvidia.com>.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

Neural networks sustain and enhance machine learning, promoting and accelerating AGI. Further innovations in this area include the deep belief networks (DBN) which are "composed of multiple layers of latent variables ('hidden units'), with connections between the layers but not between units within each layer",[63] and "seed AI", meaning AI with the ability to understand and improve its architecture.[64]

Summing up, machine-learning and the developments or methods which follow are concretely interlinked with the aspiration of mimicking and reproducing human-brain activities and functions, resembling the process by which a child's brain matures and learns.[65]

In the framework of such procedures, AI is expected to develop various components, such as logic[66] – "as a tool of analysis, as a basis for knowledge representation, and as a programming language"[67] – creativity – combined with skills such as problem solving, pattern recognition, classification, learning, induction, deduction, drawing analogies, optimization, surviving in an environment and language processing[68] – communicative capacities, external knowledge, "cognitive autonomy" – in the sense of working "independently without human intervention beyond defining goals" – intuition and strategic thinking.[69]

Therefore the level of "rules-based programming" has already been surpassed.[70] That means that AI possesses already the capacity to function autonomously from the human programmer, to exceed by far human intelligence, currently in narrow, pre-determined areas, but potentially on a much larger scale, to evolve and potentially to even re-program itself.

---

[63] *Artificial Intelligence Blog*, DL Algorithms: Deep Belief Networks (DBN), <https://www.artificial-intelligence.blog>.

[64] *N. Bostrom* (note 49), 29.

[65] *A. M. Turing*, Computing Machinery and Intelligence, Mind, Vol. LIX, No. 236 (1950), 433 et seq. (456).

[66] *R. Thomason*, Logic and Artificial Intelligence, Stanford Encyclopedia of Philosophy, <plato.stanford.edu >.

[67] *R. Thomason* (note 66).

[68] *M. Hutter*, Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability, 2010, 125 et seq. (231).

[69] *S. Yanisky-Ravid/X. Liu,* When Artificial Intelligence Systems Produce Inventions: The 3A Era and an Alternative Model for Patent Law, Cardozo L. Rev. 39 (2018), 2217 (2224); *L. Suchman/J. Weber*, Human-Machine Autonomies, in: N. Bhuta/S. Beck/R. Geib/H. Yan Liu/C. Kreb (eds.), Autonomous Weapon Systems: Law, Ethics, Policy, 2016, 39 et seq.; *M. Tegmark*, Life 3.0, Being Human in the Age of Artificial Intelligence, 2017, 140 et seq.

[70] *D. Pyle/C. San Jose*, An Executive's Guide to Machine Learning, McKinsley Quarterly, 2015, <https://www.mckinsey.com>.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

4. AI Consciousness and Its Nature: The Unpredictability of "Friendliness"

Given the above-mentioned characteristics of AI, AGI and ASI are expected to be gradually endowed with self-awareness,[71] in the sense of being aware of their own existence and of placing themselves in the broader world, with – as mentioned above – adaptable intelligence. That may lead to AI choices not only in terms of means, but also in terms of goals.

Such conception of self-awareness implies a unity of subjective, mental activities, such as imaginative thinking, self-decision, creativity, self-representation and self-discovery, sentience, wakefulness, all of which tend to re-inventing one's own presence in the world.

These elements describe aspects of consciousness – although the latter is difficult, if not completely impossible – and controversial to define –[72] in the sense of self-reflectiveness, of the perception "[…] of perception and the awareness of awareness".[73]

What we do know about consciousness, though, is that it is a multi-layered and multi-paragon function and situation, behind which there is a certain structure and function of matter as well as neurophysiological functions.[74] It necessitates and prerequisites a comprehension of the idea of the "self" as part of the world, but also as distinct from it.[75]

All these eventually lead to subjective experience.[76] Among the several critical questions that subjective experience raises, the most critical regarding AI evolution most likely is where subjective experience will lead an intellect entity, for which the physical world as well as the concept of the "self" is inherently different compared to what they mean for humans, regarding its position towards the latter;[77] whether an entity with subjective experience and with a level of intelligence equal or superior to that of hu-

---

71 *C. Chong*, This Robot Passed a "Self-Awareness" Test That Only Humans Could Handle Until Now, Tech Insider, 2015, <www.businessinsider.com>.

72 *N. Herbert*, Quantum Reality: Beyond the New Physics, 1985, 249.

73 *J. C. Smith*, Machine Intelligence And Legal Reasoning, Chicago-Kent Law Review 73 (1998), 277 et seq. (281).

74 *C. Koch*, What Is Consciousness?, Scientific American, 2018, <www.scientificamerican.com>.; *M. Tegmark* (note 69), 428 et seq.

75 *S. Armstrong/K. Sotala*, How We're Predicting AI – Or Failing To, in: J. Romportl/P. Ircing/E. Zackova/M. Polak/R. Schuster (eds.), Beyond AI: Artificial Dreams, 2012, 52.

76 *M. Tegmark* (note 69), 431.

77 *A. R. Damasio*, Descartes' Error: Emotion, Reason, and the Human Brain, 1994, 247 et seq.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

mans will be "friendly" or not; this is where the optimistic and the apocalyptic view about AI and humanity meet and diverge.[78]

The ambiguity surrounding the concept of "friendliness" of AI derives not only from the technical issues but also from the ambiguity what the concept means for humans and even more how it would be perceived from an AI perspective, once the latter has reached a certain level of subjective experience and some type of consciousness.

In order to reduce the level of ambiguity the goal-oriented approach focuses on the distinctions between the primary goals – *i.e., being friendly towards humans* – and the secondary goals, or in other words, the means in furtherance of the primary goals – for example the maintenance of the security or of the well-being of humans, as well as on the AI learning procedure which is supposed to guide AI towards serving the primary goals.

Therefore, a "Friendly AI" is not an AI duplicating the human friendship instincts, but an AI entity which demonstrates "[…] a set of external behaviors that a human would roughly call 'friendly.'"[79]

Although this approach provides some clarity and constitutes a necessary starting point order to de-codify what "friendliness" on behalf of AI may mean, it still fails to completely answer what may be the impact of machine learning, combined with the subjective AI experience regarding the potential "re-writing" and the interpretation by AI of the primary goal of "friendliness"; it also cannot completely enlighten us on what would happen in the case of non-alignment of primary with secondary goals, nor on who can provide an analytical list of friendly behavior standards to be followed by an AI entity.

These yet unanswered questions would require perfect software on behalf of the human initial programmer on the one hand, as well as the capacity of the AI to improvise in the face of different tasks or even to re-design itself so that it can fulfill the goal of "friendliness", on the other. It would need to develop its own sense of "morality" and cognitive functions whose products could be described as "morality" when projected on an AI entity.[80]

---

[78] *D. Ben-Ari/Y. Frish/A. Lazovski/U. Eldan/D. Greenbaum* (note 1), 17; *A. Eden/E. Steinhart/D. Pearce/J. Moor*, Chapter I; Singularity Hypotheses: An Overview, Introduction: Singularity Hypotheses: A Scientific and Philosophical Assessment, in: A. Eden/J. Moor/J. Soraker/E. Steinhart, Singularity Hypotheses, A Scientific and Philosophical Assessment, 2012, 28.

[79] *E. Yudkowsky*, Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures, 2001, 3.

[80] *E. Yudkowsky* (note 79), 5, 13.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

Given the complexities of the afore-mentioned prerequisites for "friendly" AI, the lack of "trust" towards AGI and ASI becomes understandable. The reason of concern is partly rooted in the human experience which indicates that amassing of power tend to undermine moral boundaries.[81]

On the basis of the analysis above, the legal regulation of AI in both its current weak forms as well as in its potential strong ones appears as absolutely crucial.[82] After all, especially concerning AGI and ASI, once they have realized their full potential, the attempt to regulate them may well come too late.

With regard to the human focus of legal systems as evidenced by the central role of human rights, the critical question thus becomes what level of non-human intelligence and what forms of its subsequent applications could be acceptable from the perspective of human rights and the international rule of law.

IV. A Regulatory Framework for AI – The Role of Human Rights

1. The Role of Human Rights

On the basis of the above-mentioned ontological elements, their potential evolution and the prospect they raise of fundamentally altering human conduct or even inaugurating an era of new, non-human "beings" and legal subjects, the need for a legal framework, capable of regulating present and future developments arises. The goals of such a legal framework must be to minimize the risks to humans and the human focus of present societies and legal systems, as well as to maximize the potential benefit of such a development, in other words to ensure a "friendly" and therefore secure and beneficial (for humans) AI.

What must be stressed is that necessarily any legal framework will be based on assumptions about the development of AI. That means that we

[81] *E. Yudkowsky* (note 79), 42.

[82] *B. M. Hutter*, A Risk Regulation Perspective on Regulatory Excellence, in: C. Coglianese (ed.), Achieving Regulatory Excellence, 2017, 101 et seq.; *N. Bostrom* (note 49), 26, 29, 140, 155; *B. Goertzel*, Response, Human-Level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's Critique of Kurzweil, Artificial Intelligence 171 (2007), 1161 (1162); *M. B. A. van Asselt/O. Renn*, Risk Governance, Journal of Risk Research 14 (2011), 431 et seq. (436 et seq.).

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

have to take into account variables which are still unknown, setting this debate apart from reform discussions in other areas of law, where the debate takes place not on speculative grounds, but *ex post facto*.

Until now, there are only primary efforts for the establishment of a legal framework, as well as declaratory documents by private entities. Indicatively, the European Union (EU) Parliament adopted a resolution about civil law rules on robotics,[83] endorsing *Asimov's* rules for autonomous AI and robotics.[84]

Other states, such as the United States,[85] China[86] and the United Kingdom[87] are also working on regulatory frameworks, though without having produced coherent legal frameworks so far. Private institutions have contributed to the gradual formation of more de-centralized regulatory schemes, although they cannot be substitutes for fully elaborated, legal schemes.[88]

The impact of AI on the one hand and the lack of legal regulation on the other hand bring the need for coherent legal regulation to the forefront, with the rule of law and human rights playing a crucial role.

Given their importance for institutionalizing justice and expressing as well as preserving the human focus of the rule of law, human rights can set the ultimate checks and balances regarding AI development, to the extent that the latter fractures or raises the risk of fracturing this focus.[89]

More specifically, the suggestion is that human rights can and must contribute to a regulatory framework promoting "friendly" AI and prohibiting undesirable as well as enabling desirable AI developments and applications.

---

[83] European Parliament Res. of 16.2.2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[84] "(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws (see *I. Asimov*, Runaround, 1943) and (0) A robot may not harm humanity, or, by inaction, allow humanity to come to harm." European Parliament Res. of 16.2.2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[85] Art. 3 (a)(1) S. 2217, 115th Congress.

[86] *P. Triolo/E. Kania/G. Webster*, Translation: Chinese Government Outlines AI Ambitions Through 2020, New America, 2018, <https://www.newamerica.org>.

[87] House of Lords, Select Committee on Artificial Intelligence, AI in the UK: ready, willing and able? Report of Session 2017-2019, <https://publications.parliament.uk>.

[88] *J. Black*, Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a "Post-Regulatory" World, Current Legal Probs. 54 (2001), 103 et seq.; Future of Life Institute, Asilomar AI Principles, <futureoflife.org>.

[89] *P. Alston*, Conjuring Up New Human Rights: A Proposal for Quality Control, AJIL 78 (1984), 607 et seq.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

In order for such legal regulation to be adequate, different means and phases of regulation must be distinguished.

2. Human Rights and Proactive Training

A first element of regulation should be the obligation of programmers, manufacturers and owners of AI to "train" AI systems so that they endorse the overall goals and the respect for human rights.

This will necessitate that part of big-data with which AI is trained, will be comprised of and devoted to human rights, so that the latter are taken in by AI as a fundamental element of machine-learning and actions.

The training of AI systems in accordance with human rights' treaties will take the form of "exposing" AI systems to legal documents, judicial decisions, legal theory and practice to teach AI the significance, the protection and the implementation of human rights in different environments. This requires a vast collection of big data attuned on the one hand to human rights and on the other hand to different scenarios of implementation of human rights in varying circumstances, so that AI systems will "learn" how to adapt to unpredictable environments. The idea is that in furtherance of friendliness, AI will be able to implement human rights, such as the right to life, liberty, security, prohibition of torture, equal treatment in front of law, non-discrimination etc.

Such training is crucial for a wide variety of AI applications; in the judiciary, in policing, in health, in education, in cyber-defenses or in the military, among other areas. In addition, biased AI constitutes a profound and already existing risk which demonstrates the urgency of human rights oriented machine learning.

Such machine-learning is expected to sustain AI friendliness, through a selection of secondary goals or means in furtherance of the primary goal of "friendliness" – by AI.[90]

Let us take the example of AI projecting the primary goal of "friendliness" on policing; the secondary goal in particular, the means for sustaining friendly AI in the specific task, will be to safeguard security in a given area, in the sense of reducing crime rate; such an outcome however can be achieved in various ways, some of which could very well infringe human rights. Therefore, the AI system must be trained in ways which make sure

[90] *M. R. Waser*, Discovering The Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence, <https://www.aaai.org>.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

that the protection of human rights for AI will be a "*sine qua non*" of the policies that it implements in in order to reduce the crime rate.

Other cases of implementation of friendliness in specific tasks requiring the inclusion of human rights in AI training will continue to arise as AI applications expand. The goal will be to assess all potential tactics which may be employed by AI in the light of human rights and for AI to balance between different tactics accordingly.

Of course, the actual implementation of human rights' guided and trained AI raises complex issues: first of all, the growing autonomy of AI means that although we can try to create AI that will be guided by human rights, we can never be absolutely certain that such guarantees will prove efficient. The expanding autonomy of AI leads to greater unpredictability, which means that essentially AI will be enjoying wider autonomy in terms of the selection of the means it will employ and even in its own, machine learning procedure. On the basis of data that AI may collect or that may be provided to it by humans it will be essential for it to make its own determinations.

In addition, the training of AI on human rights will be made more difficult by a lack of consensus about the meaning and implementation of human rights. In this sense, it will be supremely important to distinguish the good from the bad paradigms in the training of AI.

Most importantly, the implementation of complicated legal norms necessitates intuition, imaginative and creative thinking as well as interpretative approaches which are delicate even for well-trained humans.[91] The employment of such capacities by AI is unpredictable, as are the outcomes of the function of algorithms in general and even more so as their autonomy increases. In addition, we cannot yet predict how AI systems that are autonomous in this sense will comprehend human rights.[92]

Nevertheless and regardless of all these complexities we have to rely in principle on the human rights-learning approach. Despite its potential failures and inherent unpredictability, such learning approach, combined with well-designed software constitutes crucial means in furtherance of achieving friendly AI, especially when the human programmer impact will have been completely surpassed. The machine-learning approach after all mimics the way that humans are trained: the fact that unpredictability is omnipresent

[91] *D. L. Chen*, Machine Learning and the Rule of Law, in: M. Livermore/D. Rockmore (eds.), Computational Analysis of Law, forthcoming, <https://ssrn.com/abstract=3302507>. *A. Liptak*, Sent to Prison by a Software Program's Secret Algorithms, The New York Times, 2017.

[92] *M. Tegmark* (note 69).

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

does not mean that it is not crucial in order to form our ideas and guide our actions.

3. Potential Prohibition of AI Advances

As noted above however, the guarantees from human rights-oriented, machine-learning do not offer complete reassurances about "friendly" AI. A second function of human rights therefore could be to slow down or even prohibit certain technological advances which lead to AGI and ASI, or to potential applications of ANI that may threaten the superiority of human intelligence or the goals of the international community.[93]

There are several areas where ASI is expected to be profoundly threatening to humans: for instance, fully autonomous weapons, based on intelligence which will be superior to the one of humans may lead to a situation resembling science-fiction movies, where our most advanced weapons will not be "ours" anymore. Similarly, ASI systems could very well decide to restructure our political and social systems, eventually even suppressing humans, with the best of intentions, as for example to reverse climate change or secure human welfare.

The source of insecurity is however structural and general; we know from our own history that more intelligent "species" tend to respect less the will of less intelligent ones. ASI is also expected to choose and implement its own goals over the pre-defined ones, given that it will be able to match and supersede human intelligence in all aspects of human intelligence and cognition.

The looming threat therefore is that AI will surpass human intelligence in all its aspects, demonstrate new and more effective types of collective intelligence, up to the point of performing as a "single mind",[94] reproducing itself and evolving further at an unprecedented rate and speed, until finally displaying skills unknown to humans.[95]

As an approach, the potential slowing down or prohibition of AI evolution towards AGI and ASI suffers from the difficulty of attempting to balance between both the beneficial and the possibly harmful aspect of AI. It is therefore necessary to imagine an elaborate and sophisticated legal system that will be able to balance between the two.

---

93 *E. Yudkowsky* (note 79); *H. de Garis*, The Artilect War: Cosmists vs. Terrans, 2005.

94 *M. Shanahan*, Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds, 2010.

95 *N. Bostrom* (note 49), 40 et seq.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

Therefore an approach which combines training, including in human rights, with the distinction between various AI applications and technological advances, depending on a risk impact assessment, is required. Certain applications or stages of evolution may be found for example as too risky for human rights and therefore as illegal under human rights.

On the basis of such evaluation it must also be determined whether AGI and ASI are acceptable from a human rights perspective.

The answer to the question is determined by the relationship between the nature of the existent human rights-based legal systems and the potential nature of AGI and ASI.

Existing legal systems as well as our whole perception about the law are fundamentally human-centered, in the sense that they take for granted that humans are the dominant and most developed form of "being" – intellectually speaking – and that the welfare of humans constitutes the ultimate goal.

The development of AGI or ASI will challenge such common assumptions.[96] In addition, if the prevalent assumption is that AGI and ASI possibly or even likely will become hostile or non-friendly towards humans – posing a medium or high risk threat – then human rights, embodying the human-centered perception of our legal systems, impose the obligation to terminate research moving in this direction, at least "one step" before AI reaches such levels.

In case the assumptions about the risks resulting from AI evolution towards AGI and ASI are more benign, concluding that there is only a low or no risk at all of non-friendly AI, we must focus upon the checks and balances in accordance with human rights, so that AGI and ASI remain safe and beneficial for humans. This constitutes an academic area which is mainly non-legal and a matter of political decision. Law will follow the other scientific areas and the political decisions.

But we need to think not only about the potential risk of directly, non-friendly AI, but also about the internal capacity of human rights, and through them of the human-centered rule of law to adapt to a *fundamentally* transformed legal order, meaning an order where non-biological entities in terms of intelligence will be equal or even superior to humans.

Can such a development be acceptable from a human rights perspective? In order to provide an answer, we need to think that human rights have been developed on the basis of the self-evident truth that there are no intelligent entities which in the last instance can act autonomously from human will and liability. This human-centeredness, which is the basis of and provides the fundamental content to human rights, characterizes all legal sys-

96 *M. Anderson/S. L. Anderson*, Machine Ethics, 2011, 7 et seq.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

tems. Even in largely state-centered legal systems, such as international law, the human focus is still highly present, either explicitly as in the UN Charter and other international treaties, or implicitly, given that states are governed by humans.[97]

However, while human rights constitute a significant element of legal systems and of international law, it is not unanimously accepted that they sit at the top of legal hierarchy.

In this sense, an answer could be that AGI and ASI can be legitimate not under but in parallel to human rights, given that the latter constitute a significant part, albeit only part of a wider variety of sets of rights. Therefore, human rights could co-exist with a type of existential rights of AI entities, which together would comprise the whole of the legal order.

This is a problematic and incomplete approach, however. If human rights are understood as the rights which par excellence embody the human-centeredness of the legal system, then the important issue is not what human rights prescribe *per se* but what they express – namely human-centeredness – as the ultimate foundation of legal orders. Therefore, the real question is not if human rights *per se*, but if the human focus of our legal systems, as it is expressed through human rights, can be preserved after the arrival of AGI and ASI.

The answer to this question is negative. Human-centeredness as the explicit and implicit truth of our legal systems, as the ultimate and superior goal, factor of legitimacy as well as foundation of legal orders, which is endorsed and legally expressed through human rights, – could not survive the emergence of entities with an intelligence equal or superior to that of humans, which would lead to the establishment of a new foundation of legal systems.

After all, no legal system can legitimize the destruction of its foundations. Consequently, human rights cannot legitimize the emergence of AGI and ASI and thus can justify certain restrictions upon AI technological advances and applications which make sure that AI does not reach the stage of AGI and ASI.

V. Post-Human Legal System?

One last critical question raised by the preceding reflections is what may be the role of human rights in a non–human–centered or, in other words, in

97 UN Charter, Preamble.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

a post-human legal system. The post-human character of the legal system will be the consequence of the emergence of entities with an intelligence equal or superior to that of humans which are attributed legal personhood of some type. The discussion so far has largely been based on speculation and imaginative thinking. Still we can consider as a given fact that the emergence of ASI will create new types of legal persons and new realities which will fundamentally alter the structure of legal systems.

In such a framework, humans will have to co-exist with legal persons which will be neither human nor directed or run by humans but have reached or surpassed the level of human intelligence. Therefore, one critical element will be the huge inequality between humans and ASI in terms of the capacities of each side, given that ASI is expected to have exponential self-development. This constitutes a risk factor for the coherence of legal systems; they necessitate – as is the case currently – an extent of relative equality among its component units.

A second critical element is that defining factors of human personality and intelligence which fundamentally shape legal subjectivity and therefore legal systems, too – such as death or the way we comprehend life, physical harm and danger, empathy, relative cultural homogeneity among humans – may be irrelevant or at least will have to be adjusted fundamentally when applied to AI entities.[98] The lack of fear of sanction and the ability of AI to replicate themselves as well as realities which are completely different from the ones upon which legal systems have been until now must also be taken into account.[99]

A third element is the completely different meaning of space and time for AI in relation to humans. The capacity of AI to live in and through the cyberspace, its speed matching that of electrons or of quanta and the duration of its existence which either through each single unit or through self-replication may be indefinite defy all the self-evident norms underlying our legal systems.

Fourth, ASI may demonstrate completely different forms of collective organization compared to the existing ones, such as for example a type of collective conscience, introducing us into an era of post-individualism.

Under such conditions the legal systems will most likely have to adjust to the fact that the existing system of rights and penalties may not be effective for entities which will most likely be indifferent to such types of rights and defiant of such penalties as recognized and cherished by humans – for example patent recognition and deprivation of freedom. We may find our-

[98] *A. Khoury* (note 61), 646.
[99] *M. U. Scherer* (note 28), 367.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

selves in need of coming up with law enforcement mechanisms suited to the cyberspace, matching its space and time characteristics.

Human rights may give way or become part of a wider category of "(co-) existential rights", regulating the basic rights of ASI entities' meaning the rights, which flow out of their conscious existence and from their creations, as well as their interaction with human and human-administered, legal persons.

The co-existence of humans and ASI will have to be regulated so that to the extent possible ASI will not become unfriendly towards humans. Obviously this is mainly an ontological discussion and needs to be regulated before the emergence of ASI. Nevertheless, even following the emergence of ASI and in certain ways especially then, human rights as part of the "existential rights" will have to be further elaborated and evolve so that novel threats will be confronted.

The legal framework will have to design a nexus of norms safeguarding the delicate co-existence of humans and AI while at the same time fortifying human rights. The form that AI evolution will take will eventually determine whether human rights will impose some type of "Segregation" or a meddling of the two intelligences. It is very much likely that the evolution of AI will move towards the second, through the human-machine connection that is already occurring.

Last but not least, the evolution of ASI will determine whether some type of distinct sovereignty will be recognized to collective formations functioning in a different space and time environment and therefore if new political and social rights, as integral part of the set of existential rights will have to be created. Public international law will also have to change, given that new threats to international peace and security will arise following the creation of new types of actors with different formations and collective organization. The norms and organs sustained by the UN Charter and several other treaties will have to be adjusted accordingly. A new legal system will eventually be created, following the emergence of a new world.

VI. Conclusions

The present article has addressed AI from the perspective of human rights, with the latter being conceived as a fundamental element of the rule of law. The legal debate has been sketched on the basis of the evolving autonomy and intellectual capacity of AI. The potential emergence of non-biological intelligence equal or superior to that of humans presents a unique

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

challenge to our societies, as well as to the existing legal systems and to the rule of law.

This is why human rights must be activated, in order to impose checks and balances upon AI development and applications, in order to preserve the human-focus of our legal systems. First, human rights can contribute to an effective machine-learning procedure and the setting of standards that promote or discourage certain AI technological research and applications, on the basis of their compatibility or lack of compatibility with human rights and the human focus of the legal systems of which they form a central part.

It is in this framework that this article proposes to use human rights to discourage the emergence of AGI and ASI: on the one hand, the guarantees for friendly AGI and ASI are not sufficient, at least not yet; on the other hand, even if at some point we can be certain of the emergence of solely friendly AGI and ASI, human rights cannot abdicate before an essentially non-human-centered legal system, as that would be contradictory to their nature and role.

If however AGI and ASI eventually do emerge, then a new, post-human legal system will have to be created. Human rights will be succeeded by a set of rules built around existential rights regulating the legal personhood of AGI and ASI, as well as the interaction of the former with humans and issues which arise from the different relationship of ASI with space and time, a difference likely to affect the very concept of sovereignty itself.

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht

http://www.zaoerv.de
© 2017, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht