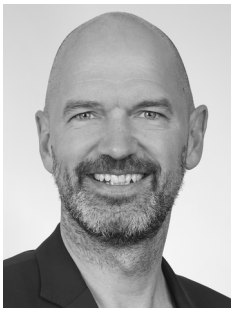
Received: 17.12.2025 | Revised: 22.01.2026 | Accepted: 29.01.2026

Steps Towards "Responsible" and Human-Centered "AI" – Some Ethical Considerations



Peter G. Kirchschlaeger

Abstract: So-called "artificial intelligence (AI)" – more adequately referred to as "data-based systems (DS)" – opens ethical downsides and upsides. There is a necessity to identify precisely the ethical opportunities and risks of DS in order to promote the former and in order to avoid the latter – for the benefit of all people and the planet earth. Companies can contribute to the realization of DS with ethics by, first, living up to the exclusive human responsibility for machines; second, while running innovation- and research-processes, by implementing always right from the start an interaction between ethics and technologies; third, by promoting global human rights-based regulation of DS as well as the establishment of an International Data-Based Systems Agency (IDA) at the UN enforcing this global regulation of DS.

Keywords: Artificial Intelligence (AI), Data-Based Systems (DS), Ethics, Human Rights, International Data-Based Systems Agency (IDA), UN

Schritte zu einer "verantwortungsvollen" und menschenzentrierten "KI" – ein paar ethische Überlegungen

Zusammenfassung: Die sogenannte „künstliche Intelligenz (KI)" – besser bezeichnet als „datenbasierte Systeme (DS)" – birgt ethische Vor- und Nachteile. Es ist notwendig, die ethischen Chancen und Risiken von DS genau zu identifizieren, damit die ethischen Vorteile gefördert und die ethischen Nachteile der DS vermieden werden können – zugunsten aller Menschen und des Planeten Erde. Unternehmen können zur Verwirklichung ethischer DS beitragen, indem sie erstens der ausschliesslichen Verantwortung des Menschen für Maschinen gerecht werden, zweitens bei der Durchführung von Innovations- und Forschungsprozessen von Anfang an eine Interaktion zwischen Ethik und Technologien implementieren und drittens eine globale, auf Menschenrechten basierende Regulierung von DS sowie die Einrichtung einer Internationalen Agentur für datenbasierte Systeme (IDA) bei der UNO fördern, die diese globale Regulierung von DS durchsetzt.

Schlüsselwörter: Künstliche Intelligenz (KI), Datenbasierte Systeme (DS), Ethik, Menschenrechte, Internationale Agentur für datenbasierte Systeme (IDA), UNO

1 So-Called "AI"? Data-Based Systems (DS)!

The ethical analysis of so-called "Artificial Intelligence (AI)" starts with a critical examination of the term itself from an ethical standpoint. So-called "AI" can be defined as striving by technical means to imitate or fulfill cognitive functions of human thought. During an ethical critique of so-called "AI", it becomes clear that so-called "AI" does not comprise

the sum of human knowledge, nor is it objective, fair and neutral. It is not trained by or does not represent reality, facts, or scientific evidence, but is trained by and represents the online past. It is only based on certain data. Looking at the above-mentioned definition of so-called "AI" from an ethical perspective, the term "artificial" is not questioned because this technology is created by humans.

A first criticism arises regarding "intelligence" because intelligence does not just consist of the solution of a cognitive task but also in the way it is pursued (Misselhorn, 2018). A second criticism highlights that so-called "AI" is limited to certain areas of intelligence (e.g., certain cognitive capacities). Among others, in the domain of emotional and social intelligence, machines are only able to simulate emotions, personal interaction, and relationships and lack authenticity. For instance, a health care robot can be trained to cry when the patient is crying, but no one would argue that the robot feels real emotions and cries due to them. The robot does not even care nor not care about it. On the contrary, one could train the exact same robot to slap the patient's face when the patient is crying, and the robot would perform this function in the same perfect way. Again, the robot does not even care nor not care about it. Machines cannot reach emotional and social intelligence – neither today nor tomorrow as the expectable further increase of compute of machines in the future can indeed improve the simulation of emotions by machines but does not create emotional and social intelligence.

Beyond that area of human intelligence, in the domain of moral capability, one cannot ascribe machines with moral capability because they are presupposed to follow ethical rules given by humans. Technologies are primarily made for their suitability and may set rules as a self-learning system, for example, to increase their efficiency, but these rules do not contain any ethical quality. E.g., a self-driving car could set the rules for itself, but it is not aware of the ethical quality of these rules. It could give itself the rule to get from A to B as fast as possible including harming humans and nature, to optimally fulfill the task of reaching B in the shortest time possible, without being able to recognize ethical rules for itself, which would allow the machine to perceive the illegitimacy of its rules and actions. A human driver instead possesses the potential to recognize for himself or herself binding ethical rules, which empower him or her to see that harming humans and nature might be more efficient but illegitimate. While humans are able to recognize by themselves ethical rules for themselves (Kirchschlaeger, 2023), machines cannot. *They don't even do not care* if they fulfill a legitimate or illegitimate task. The potential that DS possess in relation to ethical actions is nowhere close to moral capability because DS lack not only autonomy but also vulnerability, conscience, freedom, and responsibility, which are all essential for human morality (Kirchschlaeger, 2021).

The term "data-based systems (DS)" (Kirchschlaeger, 2021; 2022) would be more appropriate than "AI" because this term describes what actually constitutes "AI": generation, collection, and evaluation of data; data-based perception (sensory, linguistic); data-based predictions; data-based decisions. The mastery of an enormous quantity of data depicts the key asset of these technologies – of *data-based systems (DS)*.

The above reflection leads to the main conclusion that DS can not be responsible. Humans are and remain exclusively responsible for DS (Johnson, 2006; Yampolski, 2013). Companies need to live up to this exclusive responsibility for DS ensuring that DS are human-centered.

2 Implementing Interaction Between Ethics and Technologies

Beyond that, companies should avoid understanding ethics as an afterthought in ventures and innovation- and research-processes (Kirchschlaeger, 2024a; 2024b). Instead, the relationship of ethics and technology should be understood as an interaction, with each contributing to the other (Kirchschlaeger, 2021). Applications of groundbreaking technologies often reshape the ethical environment by creating new solutions to societal challenges and new values. At the same time, scientists and technologists all perform their work within an ethically informed context. Meanwhile, ethics contributes to technology by stimulating technological innovation, by recognizing technological inventions, and by providing ethical guidance.

Moreover, ethics belongs to technology. Horizons of meaning and ethical ends inform technology in an ethical sense. Ethics should be considered right from the start because of the very nature of technology as a human creation. Ethics can provide ethical guidance to the agenda-setting for innovation and research.

Finally, ethical principles and norms inform legal principles and norms guaranteeing freedom and independence of research. Only this freedom and this independence enable new explorations and insights as well as innovation.

3 Promoting Human Rights-Based Global Regulation and the Establishment of an International Data-Based Systems Agency (IDA) at the UN

Humans need to become active so that DS do not simply happen, but that humans shape them. This is necessary so that DS will not be reduced to an instrument serving pure efficiency but can rise to their ethically positive potential. More importantly, there is a need for ethical guidance to review the economic self-interests that run DS so far almost exclusively.

Beyond the so-far elaborated two concrete measures on an organizational level, living up to the exclusive human responsibility for DS as well as the promotion of an interaction between ethics and technologies right from the start of a venture or a research- and innovation-process (meso-level), companies should – on a macro-level – join the efforts striving for human rights-based DS (a global regulatory framework encompassing the respect and implementation of human rights in the entire life-circle of DS) (Kirchschlaeger, 2021; 2024c; 2025) and support the establishment of an International Data-Based Systems Agency (IDA) at the UN (Kirchschlaeger, 2021). IDA at the UN should fulfill the following three key functions:

1. Providing an access to market approval-process which several other industries know since decades (e.g., the pharmaceutical industry) in order to avoid harm of humans and the environment; the access to market approval-process orchestrated by IDA should ensure that human rights-violating DS including so-called "frontier models" as well as applications and products (like, e.g., an app sexualizing pictures of children) (Heikkilae, 2022; Snow, 2022; Lensa, 2025) do not even end up on the market. By this, a preventive impact is caused by IDA that the private sector does not even design and develop such human rights-violating DS knowing that they will not pass the access to market approval-process;
2. Monitoring that human rights are not violated with or by DS;

3. Fostering international technical collaboration in the sphere of DS in order to enable humanity to reach faster and better the positive potential of DS.

The aim of IDA at the UN is to ensure and to promote the development and deployment of HRBDS as a regulatory framework guaranteeing the use of the ethical positive potential of DS for the benefit of all humans and the planet as well as the handling of its ethical negative potential, including the destruction of humankind and the planet. Serving this aim is the establishment of robust governance mechanisms.

IDA should be built following the model of the International Atomic Energy Agency (IAEA) at the UN as an "institution with teeth" because, thanks to its legal powers, functions, enforcement mechanisms, and instruments, the IAEA was able to foster innovation and ethical opportunities while at the same time protecting humanity and the planet from the existential risks in the domain of nuclear technologies, which also embrace the same dual nature as DS, covering both ethical upsides and downsides. Leveraging the lessons learned from nuclear technologies and the establishment of the IAEA, the establishment of IDA presents a viable pathway towards effective global governance of existential AI risks, ensuring the responsible and ethical development of DS for the betterment of humanity and the planet.

What makes the establishment of an IDA realistic is not only its essential and minimum normative framework, its practice-oriented and participatory governance-structure, , as well as its striving for legitimacy combined with fostering innovation but also that in the past, humanity has shown that when the well-being of people and the planet is at stake, humanity can focus on what is technically feasible rather than blindly pursuing all that is technically possible.

Humanity did pursue nuclear technology, develop the atomic bomb, and even deploy it more than once. But to prevent yet worse events, humanity then massively restricted the research and development of nuclear technology despite overwhelming opposition by state and non-state actors. That nothing worse has happened is largely due to international guidelines, concrete enforcement mechanisms, and the International Atomic Energy Agency (IAEA) of the UN (Kirchschlaeger, 2021).

Beyond that, DS distinguish themselves from nuclear technology especially in three characteristics that increase the realizability of the establishment and the existential impact of IDA for humanity and the planet:

1. In order to function, DS must have power. This means that if a DS is violating human rights, threatening peace, or destroying the planet, it can be stopped by taking it off the power grid or by cutting off the power supply.
2. In order to function, DS must be connected because of its dependence on data flow. This means that if a DS is violating human rights, threatening peace, or destroying the planet, it can be stopped by disconnecting it.
3. While operating, every DS leaves data traces, allowing identification and accountability.

Finally, it also builds an advantage over attempts to ensure that all humans benefit from a previous technology-based innovation and to master the ethical dangers of a previous technology-based innovation (like in the case of nuclear technologies with the establishment of the International Atomic Energy Agency), IDA could also rely on DS-based solutions to implement HRBDS.

Supporting HRBDS and IDA means for companies to join forces with a fast growing international and interdisciplinary network of experts, business leaders, entrepreneurs, and global leaders (IDA, 2025), including, among others, UN Secretary General António Guterres, Pope Francis, His Holiness the Dalai Lama, UN High Commissioner for Human Rights Volker Türk, Sam Altman (Founder of Open AI), Mustafa Suleyman (CEO of Microsoft AI, Co-Founder and former Head of applied AI at DeepMind), and "The Elders" (Kirchschlaeger, 2024b; 2024d; 2024e; 2024f; 2024g; 2025).

References

Heikkilae, M. (2022, December 12). The viral AI avatar app Lensa undressed me – without my consent. MIT Technology Review. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent

IDA. (2025). International Data-Based Systems Agency IDA at the UN: Supporters of IDA. IDA. Retrieved October 9, 2025, from https://idaonline.ch/supporters-of-ida/

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 8(4), 195-204. https://doi.org/10.1007/s10676-006-9111-5

Kirchschlaeger, P. G. (2021). Digital Transformation and Ethics. Ethical Considerations on the robotization and automation of society and the economy and the use of Artificial Intelligence. Nomos.

Kirchschlaeger, P. G. (2022). Ethische KI? Datenbasierte Systeme (DS) mit Ethik. HMD-Praxis der Wirtschaftsinformatik, 59(2), 482-494. https://doi.org/10.1365/s40702-022-00843-2

Kirchschlaeger, P. G. (2023). Ethical Decision-Making. Nomos. https://doi.org/10.5771/9783748918684

Kirchschlaeger, P. G. (2024a, April 11). In an era of digital disruptions, ethics can't be an afterthought – Part 1. Business and Human Rights Journal Blog.

https://www.cambridge.org/core/blog/2024/04/11/in-an-era-of-digital-disruptions-ethics-cant-be-an-afterthought/

Kirchschlaeger, P. G. (2024b, April 12). In an era of digital disruptions, ethics can't be an afterthought – Part 2. Business and Human Rights Journal Blog. https://www.cambridge.org/core/blog/2024/04/12/in-an-era-of-digital-disruptions-ethics-cant-be-an-afterthought-part-2/

Kirchschlaeger, P. G. (2024c). Artificial intelligence and the complexity of ethics. Asian Horizons, 14(3), 375-389. https://dvkjournals.in/index.php/ah/article/view/4590/3752

Kirchschlaeger, P. G. (2024d, December 3). Protecting children from Anti-Social media. *Project Syndicate*. https://www.project-syndicate.org/commentary/australia-ban-on-children-using-social-media-should-be-emulated-by-peter-g-kirchschlager-2024-12

Kirchschlaeger, P. G. (2024e). The need for an International Data-Based Systems Agency (IDA) at the UN: governing "AI" globally by keeping the planet sustainably and protecting the weaker from the powerful. Journal of AI Humanities, 18, 213-248.

Kirchschlaeger, P. G. (2024f). An International Data-Based Systems Agency IDA: striving for a peaceful, sustainable, and Human Rights-Based future. Philosophies, 9(3), 73. https://doi.org/10.3390/philosophies9030073

Kirchschlaeger, P. G. (2024g). Securing a peaceful, sustainable, and humane future through an International Data-based Systems Agency (IDA) at the UN. Data & Policy, 6(78). https://doi.org/10.1017/dap.2024.38

Kirchschlaeger, P.G. (2025). Artificial Intelligence – an Analysis from the Rights of the Child Perspective. Berkley Journal of International Law. https://www.berkeleyjournalofinternationallaw.com/post/artificial-intelligence-an-analysis-from-the-rights-of-the-child-perspective

Lensa. (2025). Lensa AI: Influencers' best kept secret. Lensa App. Retrieved October 9, 2025, from https://lensa.app/

Misselhorn, C. (2018). Grundfragen der Maschinenethik. Reclam.

Snow, O. (2022, December 7). 'Magic Avatar' app Lensa generated nudes from my childhood photos. The dreamy picture-editing AI is a nightmare waiting to happen. Wired. https://www.wired.com/story/lensa-artificial-intelligence-csem/?bxid=5cc9e15efc942d13eb203f10

Yampolskiy, R.V. (2013). Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach. In V. Müller (Ed), Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics (pp. 389-396). Springer. https://doi.org/10.1007/978-3-642-31674-6_2

Peter G. Kirchschlaeger, Prof. Dr., is Ethics-Professor and Director of the Institute for Social Ethics ISE at University of Lucerne, Research Fellow at the University of the Free State, Bloemfontein (South Africa), Visiting Professor at the Chair of Neuronal Learning and Intelligent Systems at ETH Zurich and at the ETH AI Center as well as Visiting Fellow at the University of Tuebingen (Germany). Previously, he was a Visiting Fellow at Yale University (USA).

Address: University of Lucerne, Institute of Social Ethics ISE, Frohburgstrasse 3, Postfach, 6002 Luzern, Switzerland, Tel.: +41 41 229 52 61, E-Mail: peter.kirchschlaeger@unilu.chORCID: https://orcid.org/0000-0001-9528-1228