United Nations A/78/310

General Assembly

Distr.: General
30 August 2023
English
Original: Spanish

Seventy-eighth session
Item 73 (b) of the provisional agenda*
Promotion and protection of human rights: human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms

Right to privacy

Note by the Secretary-General

The Secretary-General has the honour to transmit to the General Assembly the report prepared by the Special Rapporteur on the right to privacy, Ana Brian Nougrères, submitted in accordance with Human Rights Council resolution 28/16.

* A/78/150.

23-15851 (E) 070923 110923

Report of the Special Rapporteur on the right to privacy, Ana Brian Nougrères

Principles of transparency and explainability in the processing of personal data in artificial intelligence

Summary

In the present report, the Special Rapporteur on the right to privacy, Ana Brian Nougrères, stresses the importance of the principles of transparency and explainability in the processing of personal data using artificial intelligence. The omnipresence of artificial intelligence in all activities and decision-making about people using artificial intelligence demand that the issue be examined and that measures be taken to ensure that the use of artificial intelligence is ethical, responsible and human rights-compliant.

This is important because transparency and explainability not only help to build trust and reliability in artificial intelligence, but also contribute to the protection of human rights. These principles allow individuals affected by artificial intelligence to be informed in a timely, comprehensive, simple and clear manner about basic issues concerning the use of their personal information in artificial intelligence processes or projects and the consequences thereof, and about the specific reasons behind such use. This makes it possible for them to exercise their rights, such as the right to due process and to a defence when faced with decisions made using artificial intelligence tools or technologies.

I. Introduction

1. The High-Level Expert Group on Artificial Intelligence [1] of the European Commission has noted that the principles of transparency and explainability are important components for the promotion of reliable artificial intelligence. To that end, artificial intelligence must be lawful, ethical and robust, "both from a technical and social perspective since, even with good intentions, artificial intelligence systems can cause unintentional harm".[2]

2. In the same vein, the United Nations Educational, Scientific and Cultural Organization (UNESCO) has noted that "transparency and explainability relate closely to adequate responsibility and accountability measures, as well as to the trustworthiness of artificial intelligence systems,"[3] and that "the transparency and explainability of artificial intelligence systems are often essential preconditions to ensure the respect, protection and promotion of human rights, fundamental freedoms and ethical principles."[4]

3. Artificial intelligence is very present on the global agenda. Towards the end of December 2022, for example, the Organisation for Economic Co-operation and Development (OECD) issued a statement on a trusted, sustainable and inclusive digital future,[5] in which it committed to work towards, among other things, advancing a human-centric and rights-oriented digital transformation that includes promoting the enjoyment of human rights, both offline and online, strong protections for personal data, laws and regulations fit for the digital age, and trustworthy, secure, responsible and sustainable use of emerging digital technologies and artificial intelligence.[6] With regard to artificial intelligence, OECD member States have called on the organization to support the development of forward-looking, coherent and implementable policy and legal frameworks for governing artificial intelligence and managing its risks effectively, and to provide evidence, foresight, tools and incident monitoring for effective policy planning and execution to implement trustworthy artificial intelligence.[7]

4. On 23 January 2023, the European Parliament, the Council of Europe and the European Commission adopted the European Declaration on Digital Rights and Principles, in which they committed to:

(a) Promoting human-centric, trustworthy and ethical artificial intelligence systems throughout their development, deployment and use, in line with European Union values;

(b) Ensuring an adequate level of transparency about the use of algorithms and artificial intelligence, and that people are empowered to use them and are informed when interacting with them;

[1] A group of independent experts formed by the European Commission in June 2018.

[2] High-Level Expert Group on Artificial Intelligence, *Ethical guidelines for trustworthy artificial intelligence*, (2019), p. 2. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[3] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, p. 22. Available at https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[4] Ibid.

[5] OECD, *Declaration on a Trusted, Sustainable and Inclusive Digital Future*, 2022. The declaration was the outcome of the meeting held on the island of Gran Canaria, Spain, on 14 and 15 December 2022. Available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0488.

[6] Ibid.

[7] Ibid.

(c) Ensuring that algorithmic systems are based on adequate datasets to avoid discrimination and enable human supervision of all outcomes affecting people's safety and fundamental rights;

(d) Ensuring that technologies such as artificial intelligence are not used to pre-empt people's choices, for example regarding health, education, employment, and their private life;

(e) Providing for safeguards and taking appropriate action, including by promoting trustworthy standards, to ensure that artificial intelligence and digital systems are, at all times, safe and used in full respect for fundamental rights;

(f) Taking measures to ensure that research in artificial intelligence respects the highest ethical standards and relevant European Union law.[8]

5. In view of the above, some considerations on artificial intelligence are set out below, with a brief reference to the following issues that are meant to clarify the content of the principles of transparency and explainability in the context of the processing of personal data in artificial intelligence processes or projects.

II. Artificial intelligence and the processing of personal data

6. Artificial intelligence is now omnipresent in almost every aspect of our society, from the mobile devices that citizens use all the time to the most complex business management systems. This growing presence of artificial intelligence has opened up a wide range of opportunities in various activities and sectors. However, along with these opportunities also come challenges and dangers that must be addressed responsibly so that, among other things, the full potential of artificial intelligence can be harnessed in a safe, ethical and human rights-compliant manner.

7. There is no consensus on the definition of artificial intelligence, but some of its constituent elements have been identified. In a reference text on the subject, the following taxonomy has been proposed:[9]

- Systems that think like humans (e.g., cognitive architectures and neural networks).
- Systems that act like human beings (e.g., automated reasoning and learning).
- Systems that think rationally (e.g., inferences).
- Systems that act rationally (e.g., intelligent software agents and embedded robots that achieve goals through perception, planning, reasoning, learning, communication, decision-making, and acting).

8. All these systems process information to generate results and that information contains, inter alia, personal data. In that regard, the European Commission has stated the following:

> For the purposes of this White Paper, as well as of any possible future discussions on policy initiatives, it seems important to clarify the main elements that compose artificial intelligence, which are "data" and "algorithms". Artificial intelligence can be integrated in hardware. In case of machine learning

[8] European Parliament, Council of Europe and European Commission, "European Declaration on Digital Rights and Principles for the Digital Decade", *Official Journal of the European Union*, 2023/C 23/01, 23 January 2023. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AJOC_2023_023_R_0001.

[9] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Essex, England, Pearson, 2009).

techniques, which constitute a subset of artificial intelligence, algorithms are trained to infer certain patterns based on a set of data in order to determine the actions needed to achieve a given goal.[10]

9. In other words, to develop artificial intelligence, enormous amounts of information are collected, stored, analysed, processed and used to generate various results, actions or behaviours by machines or users of such machines. However, as UNESCO states in its aforementioned recommendation, "privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of [artificial intelligence] systems."[11]

10. With the development of artificial intelligence, the proper processing of personal data is essential to prevent harm or threats to human rights, as the case may be. There are several initiatives and organizations that have worked to demand the development of human rights-compliant artificial intelligence. Some examples are provided below.

11. First, in October 2020, the Global Privacy Assembly adopted its resolution on accountability in the development and use of artificial intelligence,[12] in which it urged organizations that develop or use artificial intelligence systems to consider implementing the following accountability measures:

- Assess the potential impact to human rights (including data protection and privacy rights) before the development and/or use of artificial intelligence;
- Test the robustness, reliability, accuracy and data security of artificial intelligence before putting it into use, including identifying and addressing bias in the systems and the data they use that may lead to unfair outcomes;
- Implement accountability measures which are appropriate regarding the risks of interference with human rights.

12. Along the same lines, UNESCO, in its recommendation, stated that:

> Algorithmic systems require adequate privacy impact assessments, which also include societal and ethical considerations of their use and an innovative use of the privacy by design approach. Artificial intelligence actors need to ensure that they are accountable for the design and implementation of artificial intelligence systems in such a way as to ensure that personal information is protected throughout the life cycle of the [artificial intelligence] system.[13]

13. In June 2019, the Ibero-American Data Protection Network published a document entitled "General recommendations for the treatment of personal data in artificial intelligence",[14] in which it made some suggestions to developers of artificial intelligence products to guide them so that they can take into account the requirements

[10] European Commission, *White Paper on Artificial Intelligence – a European approach to excellence and trust*, COM (2020) 65 final. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1603192201335&uri=CELEX%3A52020DC0065.

[11] See https://unesdoc.unesco.org/ark:/48223/pf0000381137, p. 21.

[12] See https://globalprivacyassembly.org/wp-content/uploads/2020/11/GPA-Resolution-on-Accountability-in-the-Development-and-Use-of-AI-EN.pdf, p. 3.

[13] See https://unesdoc.unesco.org/ark:/48223/pf0000381137, pp. 21–22.

[14] Ibero-American Data Protection Network, "General recommendations for the treatment of personal data in artificial intelligence", (2019). Text adopted by the members of the Network at the session of 21 June 2019, held in Naucalpan de Juárez, Mexico. Available at https://www.redipd.org/sites/default/files/2020-02/guia-recomendaciones-generales-tratamiento-datos-ia.pdf.

of the regulations on personal data processing right from the product design stage. The recommendations are as follows:

- Comply with local regulations on the processing of personal data;
- Conduct privacy impact assessments;
- Embed privacy, ethics and security by design and by default;
- Implement the principle of accountability;
- Design appropriate governance schemes on the processing of personal data in organizations that develop artificial intelligence products;
- Adopt measures to ensure the implementation of the principles on the processing of personal data in artificial intelligence projects;
- Respect the rights of data owners and implement effective mechanisms for the exercise of such rights;
- Ensure the quality of personal data;
- Use anonymization tools;
- Increase trust and transparency with personal data owners.

14. For details on the implementation of some of these recommendations, the Ibero-American Data Protection Network has prepared additional and more detailed guidelines, contained in the document entitled "Specific Guidelines for Compliance with the Principles and Rights that Govern the Protection of Personal Data in Artificial Intelligence Projects".[15] The principle of transparency, which will be referred to later, is discussed in more detail in the present report.

III. Risks inherent to artificial intelligence

15. Society and its digital transformation are being shaped by artificial intelligence, which is present in several aspects of daily life, the economy, science, education, health and many other sectors and activities.

16. Though artificial intelligence offers society undeniable benefits and opportunities, it might also come with intrinsic challenges, risks and threats, which could include its unethical development or use and the making of biased, non-transparent or incorrect decisions about human beings.

17. The risk levels depend on each specific situation.

> The European Commission is of the opinion that a given [artificial intelligence] application should generally be considered high-risk in light of what is at stake, considering whether both the sector and the intended use involve significant risks, in particular from the viewpoint of protection of safety, consumer rights and fundamental rights. More specifically, an [artificial intelligence] application should be considered high-risk when it meets the following two cumulative criteria:
>
> (a) First, the [artificial intelligence] application is employed in a sector where, given the characteristics of the activities typically undertaken,

[15] Ibero-American Data Protection Network, "Specific Guidelines for Compliance with the Principles and Rights that Govern the Protection of Personal Data in Artificial Intelligence Projects", (2019). Available at: https://www.redipd.org/sites/default/files/2020-02/guide-specific-guidelines-ai-projects.pdf.

significant risks can be expected to occur. […]. For instance, health care, transport, energy and parts of the public sector […];

(b) Second, the [artificial intelligence] application in the sector in question is, in addition, used in such a manner that significant risks are likely to arise. […]. The assessment of the level of risk of a given use could be based on the impact on the affected parties. For instance, uses of [artificial intelligence] applications that produce legal or similarly significant effects for the rights of an individual or company; that pose risk of injury, death or significant material or immaterial damage; that produce effects that cannot reasonably be avoided by individuals or legal entities.[16]

18. Artificial intelligence involves different types of risk. The contingencies that should be considered include the inherent risks of operating with algorithms (human bias, technical flaws, security vulnerabilities and failures in their implementation) and their faulty design. Certain issues affect the management and performance of algorithms, as shown in the following graphic:[17]

19. As explained in the literature:

Data input is affected mainly by two variables: bias (incorporation of partial, insufficient, manipulated or outdated data) and pertinence (relevance, inconsistency or completeness of the data). On the other hand, the development of algorithms can be affected by patterns (programming logic bias, including unforeseen functions and inherent failures of the functions used for their codification), and errors (operating conditions that reflect a method of operation that differs from the one planned and goes against the premise of the proposed design). Lastly, risks in output decisions are related to the pertinence and

[16] See https://eur-lex.europa.eu/legal-content/ES/TXT/?qid=1603192201335&uri=CELEX%3A52020DC0065.

[17] See https://www.redipd.org/sites/default/files/2020-02/guia-recomendaciones-generales-tratamiento-datos-ia.pdf, p. 18.

precision of the execution of the algorithms as a direct response to the analysis of the data input.[18]

IV. Principle of transparency in the processing of personal data

20. Transparency is a concept used in several disciplines, including computer science, access to information, law and the processing of personal data. According to UNESCO, "transparency aims at providing appropriate information to the respective addressees to enable their understanding and foster trust".[19]

21. There is no consensus on the scope of transparency in each case and the term has different meanings within each case. For example, the principle of transparency means one thing when used in relation to the processing of personal data in general, and means something else when used in the context of artificial intelligence. In the present report, reference is made to transparency in the processing of personal data in general and specifically in artificial intelligence.

22. The principle of transparency is discussed in several documents of organizations from different parts of the world.[20] The Special Rapporteur had previously noted that, in accordance with the principle of transparency, controllers must inform data subjects of the processing conditions to which their personal information will be subject from the time of collection, so that subjects are in a position to exercise due control over the data.[21]

23. In the cited report, the Special Rapporteur had analysed the principle of transparency based on the following international documents on privacy and the processing of personal data: (a) General Data Protection Regulation of the European Union; (b) Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data; (c) Standards for Personal Data Protection for Ibero-American States, adopted by the Ibero-American Data Protection Network; (d) Recommendations of the Council concerning the Guidelines on the Protection of Privacy and Transborder Flows of Personal Data of the Organisation for Economic

[18] Alejandro Useche and Jeimy Cano, *Robo-Advisors: Asesoría automatizada en el mercado de valores*, Universidad del Rosario and Autorregulador del Mercado de Valores de Colombia (2019), pp. 9–10. Available at: https://www.researchgate.net/publication/331358231_Robo-Advisors_Asesoria_automatizada_en_el_mercado_de_valores.

[19] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, p. 22.

[20] Organisation for Economic Co-operation and Development (OECD), *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, 23 September 1980 and the updated guidelines from July 2013; Council of Europe, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, No. 108, 28 January 1981; United Nations, Guidelines for the regulation of computerized personal data files, 14 December 1990; Council of Europe, Additional Protocol to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, regarding supervisory authorities and transborder data flows, 8 November 2001; Asia-Pacific Economic Cooperation Forum, *Asia-Pacific Economic Cooperation Forum Privacy Framework*, 2004; Spanish Data Protection Agency, *Joint Proposal for a Draft of International Standards on the Protection of Privacy with regard to the Processing of Personal Data*, Madrid, 5 November 2009; Regulation (European Union) 2016/679 of the European Parliament and of the Council 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), 27 April 2016; Ibero-American Data Protection Network, Guidelines for Harmonization of Data Protection in the Ibero-American Community, 2017; Council of Europe, Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, October 2018, and Organization of American States, Inter-American Juridical Committee, *Updated Principles on Privacy and Personal Data Protection*, 2021.

[21] A/77/196, para. 45.

Co-operation and Development; (e) Asia-Pacific Economic Cooperation Forum Privacy Framework, and (f) Updated Principles on Privacy and Personal Data Protection, with annotations, of the Organization of American States.

24. From the analysis, she had concluded that, as a general rule, the following information must be disclosed:

- The identities and addresses of controllers or of their representatives, and the aims or purposes of the processing […] These data are the basic foundations of transparency;
- The rights of the data subject and the ways in which they may be exercised, as well as the recipients of the data or category of recipients;
- The legal foundation or basis for the processing, as well as the existence and/or main characteristics of the processing;
- The category of the data processed and the origin of the data when not obtained directly from the subject.

25. It is worth noting that to implement the principle of transparency, the information provided to the data subject must be in simple, clear, intelligible and easily accessible and understandable language. That mandate must also be upheld in cases involving children and adolescents, with the necessary adjustments being made.

26. Not all the regulatory instruments mentioned above require that the same information be disclosed, since some have more extensive lists of the types of information that must be disclosed. In the particular case of the General Data Protection Regulation of the European Union, the information that must be disclosed includes:[22] the contact details of the data protection officer; the period for which the personal data will be stored or the criteria used to determine that period; whether the controller plans to carry out communications or transfers and the regulations authorizing such communications or transfers; the right to lodge a complaint with a supervisory authority; whether communication is a statutory or contractual requirement, or is necessary to enter into a contract, and whether subjects are required to provide their personal data and the consequences of a failure to do so; the existence of automated decision-making, including profiling, in which case meaningful information about the logic involved must be provided, as well as the significance and envisaged consequences of such processing, and information on the purpose in cases where the controller intends to further process the data for a purpose other than that for which the data were collected.

V. Principle of transparency in the processing of personal data in the field of artificial intelligence

27. It is essential to ensure transparency in artificial intelligence, as a lack of awareness or omission in that connection may generate negative effects. As the European Commission has noted that:

> The lack of transparency (opaqueness of [artificial intelligence]) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights, attribute liability and meet the conditions to claim compensation.[23]

[22] See https://eur-lex.europa.eu/legal-content/ES/TXT/?qid=1532348683434&uri=CELEX%3A02016R0679-20160504.

[23] European Commission, *White Paper on Artificial Intelligence – a European approach to excellence and trust*, 2020, p. 14.

28. The potential opacity of artificial intelligence can be mitigated by requiring compliance with minimum transparency standards. Accordingly, the Commission has identified the following requirements:

> Ensuring clear information to be provided as to the [artificial intelligence] system's capabilities and limitations, in particular the purpose for which the systems are intended, the conditions under which they can be expected to function as intended and the expected level of accuracy in achieving the specified purpose [...] Separately, citizens should be clearly informed when they are interacting with an [artificial intelligence] system and not a human being [...]. It is furthermore important that the information provided is objective, concise and easily understandable.[24]

29. In its recommendation, UNESCO stated that:

> Specific to the [artificial intelligence] system, transparency can enable people to understand how each stage of an [artificial intelligence] system is put in place, appropriate to the context and sensitivity of the [artificial intelligence] system. It may also include insight into factors that affect a specific prediction or decision, and whether or not appropriate assurances (such as safety or fairness measures) are in place.[25]

30. The High-Level Expert Group on Artificial Intelligence has noted that certain requirements, including transparency, must be met in order to achieve trustworthy artificial intelligence. As regards transparency, it is necessary to:

> Communicate, in a clear and proactive manner, information to stakeholders about the [artificial intelligence] system's capabilities and limitations, enabling realistic expectation setting, and about the manner in which the requirements are implemented[; and] be transparent about the fact that they are dealing with an [artificial intelligence] system.[26]

UNESCO has also recommended that "[artificial intelligence] actors should inform users when a product or service is provided directly or with the assistance of [artificial intelligence] systems in a proper and timely manner."[27]

31. According to the UNESCO recommendation:

> Explainability is closely related to transparency, as outcomes and sub-processes leading to outcomes should aim to be understandable and traceable, appropriate to the context. [Artificial intelligence] actors should commit to ensuring that the algorithms developed are explainable. In the case of [artificial intelligence] applications that impact the end user in a way that is not temporary, easily reversible or otherwise low risk, it should be ensured that the meaningful explanation is provided with any decision that resulted in the action taken in order for the outcome to be considered transparent.[28]

32. The High-Level Expert Group on Artificial Intelligence has explained that transparency "is closely linked with the principle of explicability and encompasses transparency of elements relevant to an [artificial intelligence] system: the data, the system and the business models."[29] It has also highlighted the relevance of traceability, explainability and communication in the following terms:

---

[24] Ibid., pp. 23–24.
[25] See https://unesdoc.unesco.org/ark:/48223/pf0000381137, p. 22.
[26] See https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, pp. 2 and 3.
[27] See https://unesdoc.unesco.org/ark:/48223/pf0000381137, p. 22.
[28] Ibid., p. 22.
[29] See https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, p. 18.

- Traceability: The data sets and the processes that yield the [artificial intelligence] system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the [artificial intelligence] system. This enables identification of the reasons why an [artificial intelligence]-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.

- Explainability: Explainability concerns the ability to explain both the technical processes of an [artificial intelligence] system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an [artificial intelligence] system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an [artificial intelligence] system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the [artificial intelligence] system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an [artificial intelligence] system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

- Communication. [Artificial intelligence] systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an [artificial intelligence] system. This entails that [artificial intelligence] systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the [artificial intelligence] system's capabilities and limitations should be communicated to [artificial intelligence] practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the [artificial intelligence] system's level of accuracy, as well as its limitations.[30]

33. The European Data Protection Board and the European Data Protection Supervisor have issued a joint opinion in which they stated that:

> Data subjects should always be informed when their data is used for [artificial intelligence] training and/or prediction, of the legal basis for such processing, general explanation of the logic (procedure) and scope of the [artificial intelligence] system. In that regard, the individuals' right of restriction of processing (article 18 GDPR and article 20 EUDPR as well as of deletion/erasure of data (article 16 GDPR and article 19 EUDPR should always be guaranteed in those cases. Furthermore, the controller should have the explicit obligation to inform the data subject of the applicable periods for objection, restriction, deletion of data, etc. The [artificial intelligence] system must be able to meet all data protection requirements through adequate technical

[30] Ibid.

and organizational measures. A right to explanation should provide for additional transparency.[31]

34. In the report mentioned earlier,[32] the Special Rapporteur had noted that, in cases in which data subjects are subjected to automated decision-making or profiling, they should be able to understand the way in which the information concerning them will be processed (whether artificial intelligence is involved, for example) and with meaningful information about the logic involved and the significance and the envisaged consequences.

35. On that point, the Spanish Data Protection Agency has pointed out that "[t]he word 'meaningful' […] must be understood as information which, once provided to the data subject[s], makes them aware of the type of processing that their data is undergoing and provides certainty and trust as to the associated results".[33]

36. The Agency has also stated that:

> Complying with this obligation by offering technical references on the implementation of the algorithm may be obscure, confusing or lead to information fatigue. Sufficient information should be provided to enable the subjects to understand the behaviour of the processing. Although it will depend on the type of [artificial intelligence] component used, an example of the types of information that may be relevant to the data subject would be:

- Details about the data used for decision-making beyond just the category, especially information regarding the duration of use of the data (how old the data are).
- Relative importance or weight given to each of the data in the decision-making.
- Quality of the training data and the type of models used.
- Profiling activities conducted and their implications.
- Error or precision values according to the specific metrics used to measure the validity of the inference.
- Existence or non-existence of qualified human supervision.
- References to audits, especially audits on possible deviations of inference results, as well as the certification or certifications of the [artificial intelligence] system. For adaptive or evolutionary systems, the last audit conducted.
- If the [artificial intelligence] system includes information referring to identifiable third parties, the prohibition of processing such information without legitimization and the consequences of doing so.[34]

37. The European Data Protection Supervisor has issued an opinion suggesting that if the Commission were to put forward a new artificial intelligence-specific regulatory framework, a certain number of reasonable safeguards should apply to all artificial intelligence applications, regardless of the level of risk, such as having technical and

[31] European Data Protection Board and the European Data Protection Supervisor, 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), 18 June 2021, p. 17. Available at https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf.

[32] A/77/196, para. 55.

[33] Spanish Data Protection Agency, *Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción*, February 2020. p. 24. Available at: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf.

[34] Ibid.

organizational measures in place (including documentation); being fully transparent about the goals, use and design of the algorithmic systems implemented; ensuring the robustness of the artificial intelligence system and implementing and being transparent about the available mechanisms of accountability, redress and independent oversight.[35]

38. The European Data Protection Board and the European Data Protection Supervisor have also noted the need to promote:

> [N]ew, more proactive and timely ways to inform users of [artificial intelligence] systems on the (decision-making) status where the system lays at any time, providing early warning of potential harmful outcomes, so that individuals whose rights and freedoms may be impaired by the machine's autonomous decisions may react or redress the decision.[36]

39. The Ibero-American Data Protection Network is of the opinion that the following actions must be taken to implement the principle of transparency:[37]

- "Communicate to data subjects the main characteristics of the processing to which their personal information will be submitted";
- "Expressly inform data subjects that automation processes will be used in the processing of their personal data";
- "Include all purposes for which the data subjects' data will be processed in the method chosen by the controllers to implement the principle of transparency";
- "Disclose the origin of personal data when such data are obtained through a transfer and, in cases in which the intention is to use artificial intelligence, confirm that the data subjects were notified of this purpose by the first controller who obtained the data to make use of them for that purpose";
- "Develop innovative ways to inform data subjects of the main characteristics of the processing and the level of risk in terms of an increase or decrease in privacy expectations";
- "Safeguard the right to informational self-determination by ensuring that data subjects are always informed in an adequate and timely manner that they will be interacting directly with an artificial intelligence system or when their information will be processed by one";
- "Provide meaningful information on the purpose and effects of artificial intelligence systems to verify continuous alignment with the privacy expectations of data subjects, allowing them to exercise control over the processing of their personal data at all times";
- "Identify and define commonly used terms and create a database so those terms can be reused in different contexts, with standard icons to make information known to the data subjects";

[35] European Data Protection Supervisor, *Opinion 4/2020, European Data Protection Supervisor Opinion on the European Commission's White Paper on Artificial Intelligence – a European approach to excellence and trust*, 29 June 2020, p. 14. Available at: https://edps.europa.eu/sites/edp/files/publication/20-06-19_opinion_ai_white_paper_en.pdf.

[36] European Data Protection Board and the European Data Protection Supervisor, "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)", 18 June 2021, p. 22.

[37] See https://www.redipd.org/sites/default/files/2020–02/guia-orientaciones-espec%C3%ADficas-proteccion-datos-ia.pdf, pp. 17–19.

- “Continuously inform data subjects so that they know how automated decision-making can affect them and how to request human intervention when needed, so they can make an informed decision as to whether or not to consent to the processing”.

40. The Ibero-American Data Protection Network has noted that:

> The information provided regarding the logic of the [artificial intelligence] model should include at least the basic aspects of its operation, as well as the weighting and correlation of the data, written in clear, simple and easily understood language. It will not be necessary to provide a complete explanation of the algorithms used or even to include them.[38]

41. The Ibero-American Data Protection Network has called on those responsible for the processing of data by artificial intelligence to be innovative in order to convey information in a simple and concise manner, indicating that “[t]here are several innovative approaches to providing privacy notices, including the use of videos, cartoons and standardized icons. The use of a combination of approaches can help make complex information on [artificial intelligence] easier for data subjects to understand”.[39]

42. The following paragraphs contain an enunciative and non-exhaustive set of examples of countries that have explicitly or implicitly addressed in their local laws the principle of transparency in the processing of personal data using artificial intelligence.

43. In Ecuador, the Organic Data Protection Act, adopted in 2021, establishes in its article 12, paragraphs 14 and 17, the right to be informed about the existence of the right to not be subject to a decision based solely on automated evaluations, the manner in which that right can be exercised and the existence of automated assessments and decisions, including profiling.

44. The Act also stipulates that in cases in which data are obtained directly from data subjects, the information shall be communicated in advance (at the time the personal data are collected). Article 12 further states that:

> When personal data are not obtained directly from the data subjects or when they have been collected from sources accessible to the public, the data subjects shall be informed within thirty (30) days or in the first communication they receive, whichever occurs first. The data subjects shall be given clear, unambiguous, transparent, understandable, concise and accurate information with no technical hurdles.

45. In Peru, article 72 of the Implementing Regulations of Act No. 29733, the Personal Data Protection Act, addresses the right to the objective processing of personal data, stating as follows:

> To uphold the right to objective processing pursuant to article 23 of the Act,[40] when personal data are processed as part of a decision-making process that does not involve the data subject, the controller of the personal data database or the

[38] See https://www.redipd.org/es/documentos/guia, pp. 17–19.

[39] Ibid.

[40] “Article 23. Right to objective processing. Data subjects have the right to not be subjected to a decision that has legal effects on them or affects them significantly and is supported only by the processing of personal data intended to evaluate certain aspects of their personalities or behaviour, unless it occurs during the negotiation, execution or performance of a contract or in cases of an evaluation for the purposes of taking a position at a public entity, pursuant to the law, without prejudice to the possibility of defending their point of view for the protection of their legitimate interests.”

> controller of the processing shall inform the data subject without delay, except as otherwise provided in the Regulations on the exercise of the other rights set out in the Act and its […] Regulations.

46. In Sao Tome and Principe, Act No. 3/2016 of 2 May 2016, the Individual Personal Data Protection Act, is unique in that it stipulates in its article 21 that controllers or their representatives shall notify the National Personal Data Protection Agency, in writing and no more than eight days before the processing is to begin, that they will begin fully or partially automated processing or batch processing to achieve one or more interrelated ends, with some exceptions. Article 11 of the Act also provides that data subjects, when exercising their right to access, have the right to be informed by the controller of the reasons behind the automated processing of data concerning them.

47. In Uruguay, article 13 of Act No. 18831 of 11 August 2008, the Personal Data Protection Act, establishes that data subjects have the right to be informed, in an express, clear and unmistakable manner, prior to data collection, about the assessment criteria, the processes applied and the technological solution or software utilized in cases in which automated data processing is used to evaluate certain aspects of their personality, such as job performance, creditworthiness, reliability and conduct, to make decisions with legal effects that could significantly affect the data subjects. The Act also states that "when personal data are not collected directly from the data subjects, the information […] shall be provided to them within a period of five business days from the date on which the request is received by the controllers".

VI. Principle of explainability in the processing of personal data in artificial intelligence projects

48. The creation of "virtual profiles" on individuals based on existing information is becoming increasingly common and decisions are often made about them based on the automated processing of their data using various technological tools.

49. Human beings can be positively or negatively affected by the decisions made about them based on the use and processing of data in artificial intelligence projects. There are concerns about how to protect the rights of individuals affected by decisions made about them with artificial intelligence tools or technologies. In the White Paper on artificial intelligence, for example, it is noted that: "as with any new technology, the use of [artificial intelligence] brings both opportunities and risks. Citizens fear being left powerless in defending their rights and safety when facing the information asymmetries of algorithmic decision-making".[41]

50. Given the above, people need to be aware of which data were used to make a decision that affects them, as well as the logic used to reach such decision. Having access to this information will, inter alia, enable the affected person to know whether the decision made about them is correct and, if not, to defend themselves. In other words, such information is necessary to ensure due process, as it will serve as evidence of possible inaccuracies or injustices generated against people during the processing of their personal data in artificial intelligence processes. In this regard, the aforementioned High-Level Expert Group on Artificial Intelligence has emphasized that the principle of explicability:

> is crucial for building and maintaining users' trust in [artificial intelligence] systems. This means that processes need to be transparent, the capabilities and

[41] See https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1603192201335&uri=CELEX%3A52020DC0065, p. 9.

> purpose of [artificial intelligence] 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 [...]. 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.[42]

51. All this explains why transparency in artificial intelligence is important, since such intelligence should not be obscure, secretive or misleading. For this reason, in the aforementioned European Declaration it was stated that:

> Everyone should be empowered to benefit from the advantages of algorithmic and artificial intelligence systems including by making their own, informed choices in the digital environment, while being protected against risks and harm to one's health, safety and fundamental rights.[43]

52. In line with the above, the Ibero-American Data Protection Network recommended in 2019 increasing transparency with personal data subjects.[44]

53. Subsequently, and also related to the above, in its aforementioned 2020 resolution, the Global Privacy Assembly stressed that organizations developing or using artificial intelligence systems should take the following measures into consideration: (a) ensuring transparency and openness by disclosing the use of artificial intelligence, the data being used and the logic involved in the artificial intelligence; (b) ensuring an accountable human actor is identified with whom concerns related to automated decisions can be raised and rights can be exercised, and who can trigger evaluation of the decision process and human intervention; (c) providing explanations in clear and understandable language for the automated decisions made by artificial intelligence upon request; and (d) ensuring human intervention in the automated decision made by artificial intelligence upon request.[45]

54. All of the above is partially aligned with the provisions of the General Data Protection Regulation, which states for example that:

> Where personal data have not been obtained from the data subject, the controller shall provide the data subject with the following information: [...] 2. (g) the existence of automated decision-making, including profiling, referred to in article 22 (1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.[46]

Additionally, the data subject or data owner has the right to:

> obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information: [...] (h) the existence of automated decision-making, including profiling, referred to in article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved,

[42] See https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, p. 13.
[43] Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AJOC_2023_023_R_0001.
[44] See https://www.redipd.org/sites/default/files/2020-02/guia-recomendaciones-generales-tratamiento-datos-ia.pdf, pp 23 and 24.
[45] See https://globalprivacyassembly.org/document-archive/adopted-resolutions/, p. 3.
[46] See http://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679, art. 14, para. 2 (g).

as well as the significance and the envisaged consequences of such processing for the data subject.[47]

55. The National Institute of Standards and Technology summarizes the scope of this principle in the following chart:[48]

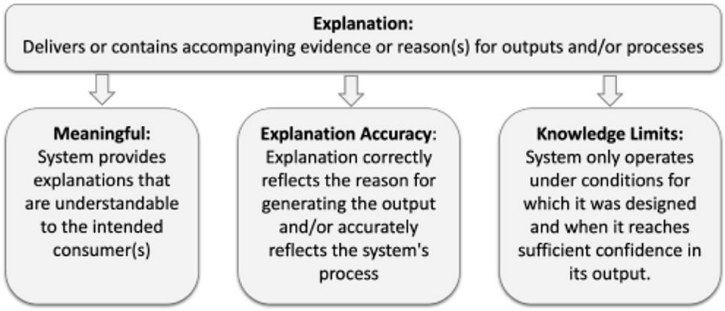

Fig. 1. Illustration of the four principles of explainable artificial intelligence. Arrows indicate that for a system to be explainable, it must provide an explanation. The remaining three principles are the fundamental properties of those explanations.

56. The following table explains the most relevant aspects of each principle according to the National Institute of Standards and Technology document:[49]

Principle	*Meaning or scope*
Explanation	Evidence, support or reasoning related to outputs and/or processes of an [artificial intelligence] system.
Meaningful	Explanation in terms that are understandable to the intended consumer. In other words, this principle seeks to make the explanation comprehensible for a given audience. Many factors affect a good explanation, which is why the target audience or audience to which the explanation is addressed must be taken into account.
Explanation Accuracy	This principle requires the technical explanation to be rigorous, accurate and comprehensive.
Knowledge Limits	Identifying and declaring the limits of knowledge implies making it clear that the system is neither perfect nor infallible because the [artificial intelligence] operates within certain limits and constraints within which it has been programmed. They also depend on the quality and quantity of information processed, among other factors.

57. It has been argued that the explanation must: (a) “be understandable and convincing to the user”; (b) “accurately reflect the system’s reasoning”; (c) “be comprehensive”, and (d) “be specific in the sense that different users with different

[47] Ibid., art. 15 (1).

[48] National Institute of Standards and Technology, *Four Principles of Explainable Artificial Intelligence*, - NISTIR 8312 (2021), p. 3. Available at https://doi.org/10.1017/bhj.2023.10.

[49] The explanation in the table is an adaptation and summary of the original English text cited and available at https://doi.org/10.6028/NIST.IR.8312.

circumstances or different results should obtain different types of explanations".[50] Additionally, it has been noted that:

> the explainability of artificial intelligence is an aspiration that is understandable from an ethical and even legal point of view, but it has profound technical difficulties that are worth knowing and, probably, a large part of the solution will also be technical, to the extent that it is possible to redesign algorithms or identify new ones that satisfy ethical and regulatory aspirations.[51]

UNESCO, for its part, has indicated that:

> Explainability refers to making intelligible and providing insight into the outcome of [artificial intelligence] systems. The explainability of [artificial intelligence] systems also refers to the understandability of the input, output, and functioning of each algorithmic building block and how it contributes to the outcome of the systems.[52]

58. In order to determine the scope of the principle of explainability, it is necessary to bear in mind its objective and, on that basis, to establish what is needed to achieve it. In line with the above, it has been pointed out that:

> if the principle of explainability is intended for any human being to know why a decision is made based on the processing of his or her data with [artificial intelligence] tools, then the explanation should at least be clear, simple, complete, truthful and easily understood by the person requesting the explanation. It is not enough to report on the data used as inputs to generate the decision, rather, the logic or methodology used to reach the decision should be provided. The challenge is not minor, but it is achievable if there is a willingness to easily explain to people why a decision was generated based on the processing of their personal data.[53]

59. Below are some examples of local laws in countries that have tacitly or explicitly incorporated the principle of explainability into their legal frameworks.

60. In Colombia, the law prohibits the processing of data that "misleads"[54] and, in the specific case of decisions made with respect to loan applications, requires those who reject such applications to inform the person concerned in writing, if so required, of "the objective reasons for the rejection".[55]

61. In Ecuador, article 20 of the Data Protection Organic Act establishes that data owners, faced with a decision based solely or partially on assessments resulting from automated processes, including profiling, that produce legal effects on them or that violate their fundamental rights and freedoms, may demand a reasoned explanation of the decision, obtain the assessment criteria on the automated program, submit observations, request information on the types of data used and the source from which

---

[50] Gavilán, Ignacio, "Cuatro principios para una buena explicabilidad de los algoritmos" (2022). Available at: https://ignaciogavilan.com/cuatro-principios-para-una-buena-explicabilidad-de-los-algoritmos/.

[51] Ibid.

[52] See https://unesdoc.unesco.org/ark:/48223/pf0000381137 p. 23.

[53] Nelson Remolina Angarita, "Del principio de explicabilidad en la inteligencia artificial (notas preliminares)", in *Protección de datos personales: doctrina y jurisprudencia*, Pablo Palazzi, ed., vol. III (Centre for Technology and Society, University of San Andrés, Buenos Aires, 2023).

[54] Statutory Act No. 1581 of 2012, which establishes general provisions for the protection of personal data, art. 4 d).

[55] Statutory Act No. 2157 of 2021, which amends and supplements Statutory Act No. 1266 of 2008 and establishes general provisions on habeas data in relation to financial, credit, commercial, service and third-country information and other provisions, art. 5, para. 1.

they were obtained, and contest the decision before the person responsible or in charge (with certain exceptions).

62. In Uruguay, article 16 of Act No. 18331 establishes that:

> individuals have the right not to be subjected to a decision with legal effects that significantly affects them that is based on automated data processing intended to evaluate certain aspects of their personality, such as their job performance, creditworthiness, reliability and conduct. Whoever is affected shall have the right to obtain information from the person responsible for the database both on the evaluation criteria and on the program used in the processing that was used to reach the decision set out in the act.

VII. Conclusions

63. **The following conclusions can be drawn from the foregoing:**

(a) **Transparency and explainability help to build trust in artificial intelligence and to respect human rights;**

(b) **Developers of artificial intelligence must be transparent about how data are processed (how they are collected, stored and used), and about how decisions based on artificial intelligence are made, the reliability of such decisions and the security of the information;**

(c) **Persons affected by decisions made on the basis of artificial intelligence deserve a clear, simple, complete, truthful and understandable explanation of the reasons for that decision. In that regard, the principle of explainability is of cardinal importance not only because it aligns with the principle of transparency, but also because it will make it possible to uphold such persons' right to a defence and due process;**

(d) **Explainability and transparency demand clarity, completeness, truthfulness, impartiality and publicity of the decisions made using artificial intelligence and of the logic, method or reasoning for making decisions about human beings based on information, particularly personal data. Explainability and transparency are, of course, the opposite of opacity, obscurity, deceit, lies and abuse of computing power, which are some of the symptoms of illegal and unethical processing that reflects a lack of respect for human beings and their dignity.**

VIII. Recommendations

64. **In the light of the above, the Special Rapporteur urges States to:**

(a) **Promote transparency in artificial intelligence in order to mitigate the risks that opacity may generate in society, especially with respect to the protection of human rights;**

(b) **Incorporate into their laws the principle of explainability, not only to enable people to understand how the decisions that affect them were made, but also to provide them with the tools to defend their human rights in the face of artificial intelligence;**

(c) **Promote ethical practices that ensure transparency and explainability in the processing of personal data in artificial intelligence projects or processes;**

(d) **Foster, support and facilitate education and digital literacy to enable citizens to better understand the concepts relating to artificial intelligence, transparency and explainability, in order to be able to demand that their rights be respected.**