联合国 A/78/310

大 会

Distr.: General
30 August 2023
Chinese
Original: Spanish

第七十八届会议

临时议程* 项目 73(b)

促进和保护人权：人权问题，包括增进人权和基本自由切实享受的各种途径

隐私权

秘书长的说明

秘书长谨向大会转递隐私权特别报告员安娜·布里安·努格蕾蕾斯根据人权理事会第 28/16 号决议编写并提交的报告。

* A/78/150。

23-15851 (C) 140923 210923

隐私权特别报告员安娜·布里安·努格蕾蕾斯的报告

人工智能处理个人数据的透明度和可解释性原则

摘要

隐私权特别报告员安娜·布里安·努格蕾蕾斯在本报告中强调，在使用人工智能处理个人数据时，必须遵守透明度和可解释性原则。人工智能在所有活动中无所不在，使用人工智能对人作出决策，都要求对这一问题进行审视，并采取措施确保人工智能的使用合乎伦理、负责任、符合人权。

这一点非常重要，因为透明度和可解释性不仅有助于建立对人工智能的信任和建立人工智能的可靠性，也有助于保护人权。这些原则可以让受人工智能影响的个人能够及时、全面、简单、清晰地了解人工智能过程或项目中使用其个人信息的基本问题及其后果，以及这些使用背后的具体原因。这使他们能够行使自己的权利，例如在面对使用人工智能工具或技术作出的决定时享有正当程序权和辩护权。

一. 导言

1. 欧洲联盟委员会人工智能高级别专家组[1] 指出，透明度和可解释性原则是促进可靠人工智能的重要组成部分。为此，人工智能必须是合法、合乎伦理和稳健的，“无论是从技术角度还是从社会角度来看，即使有良好意图，人工智能系统也可能造成无意伤害”。[2]

2. 同样，联合国教育、科学及文化组织(教科文组织)指出，“透明度和可解释 性与适当的责任和问责措施以及人工智能系统的可信度密切相关”，[3] “人工智能系统的透明度和可解释性往往是确保尊重、保护和促进人权、基本自由和道德原则的必要先决条件”。[4]

3. 人工智能已被提上全球议程。例如，临近 2022 年 12 月底，经济合作与发展组织经合组织(经合组织)发表了一份关于可信、可持续和包容的数字未来的声明。[5] 其中承诺，除其他外，努力推进以人为本、以权利为导向的数字化转型，包括促进线上线下人权的享受，对个人数据的强有力保护，适合数字时代的法律和条例，以及可信、安全、负责任和可持续地使用新兴的数字技术和人工智能。[6] 关于人工智能，经合组织成员国呼吁该组织支持制定前瞻性、一致性和可执行的政策和法律框架，以有效治理人工智能并管理其风险，并为有效的政策规划和执行提供证据、远见、工具和事件监测，以实施可信赖的人工智能。[7]

4. 2023 年 1 月 23 日，欧洲议会、欧洲委员会和欧盟委员会通过了《欧洲数字权利和原则宣言》，承诺：

(a) 根据欧洲联盟的价值观，在人工智能系统的整个开发、部署和使用过程中促进以人为本、值得信赖和符合伦理的人工智能系统；

(b) 确保算法和人工智能的使用有足够的透明度，确保人们有权使用算法和人工智能，并在与算法和人工智能互动时了解相关信息；

(c) 确保算法系统基于适当的数据集，避免歧视，并使人类能够对影响人们安全和基本权利的所有结果进行监督；

---

[1] 欧盟委员会于 2018 年 6 月成立的一个独立专家小组。

[2] 人工智能高级别专家组，*Ethical guidelines for trustworthy artificial intelligence*, (2019)，第 2 页。可查阅：https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai。

[3] 教科文组织，《人工智能伦理问题建议书》，2021 年，第 22 页。可查阅 https://unesdoc.unesco.org/ark:/48223/pf0000381137。

[4] 同上。

[5] 经合组织，*Declaration on a Trusted, Sustainable and Inclusive Digital Future*，2022。该宣言是 2022 年 12 月 14 日和 15 日西班牙大加纳利岛举行的会议的成果。可查阅 https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0488。

[6] 同上。

[7] 同上。

(d) 确保人工智能等技术不被用来预先阻止人们的选择，例如在健康、教育、就业及其私人生活方面的选择；

(e) 规定保障措施并采取适当行动，包括推广可信赖的标准，以确保人工智能和数字系统在任何时候都是安全的，并在使用时充分尊重基本权利；

(f) 采取措施确保人工智能研究尊重最高伦理标准和欧洲联盟相关法律。[8]

5. 鉴于上述情况，下文列出有关人工智能的一些考虑因素，并简要提及以下问题，这些问题旨在澄清在人工智能过程或项目中处理个人数据时透明度和可解释性原则的内容。

二. 人工智能与个人数据处理

6. 人工智能现在几乎遍布我们社会的各个方面，从公民经常使用的移动设备到最复杂的商业管理系统。人工智能的日益普及为各种活动和行业带来广泛机遇。然而，伴随这些机遇而来的也有挑战和危险，必须负责任地加以应对，以便除其他外，能够以安全、道德和符合人权的方式充分发挥人工智能的潜力。

7. 对人工智能的定义尚未达成共识，但它的一些构成要素已经确定。该主题的参考文献提出以下分类法：[9]

- 像人类一样思考的系统(如认知架构和神经网络)。
- 像人类一样行动的系统(如自动化推理和学习)。
- 理性思考的系统(如推断)。
- 理性行动的系统(如通过感知、规划、推理、学习、沟通、决策和行动来实现目标的智能软件代理和嵌入式机器人)。

8. 所有这些系统都处理信息以生成结果，而这些信息除其他外包含个人数据。为此，欧盟委员会声明如下：

> 为了本白皮书以及未来可能就政策倡议进行的任何讨论，似乎有必要澄清构成人工智能的主要要素，即“数据”和“算法”。人工智能可以集成在硬件中。机器学习技术是人工智能的一个子集，算法经过训练，可根据一组数据推断出某些模式，从而确定实现给定目标所需的行动。[10]

---

[8] 欧洲议会、欧洲理事会、欧盟委员会，“European Declaration on Digital Rights and Principles for the Digital Decade”, *Official Journal of the European Union*, 2023/C 23/01, 23 January 2023.可查阅 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).

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

9. 换言之，为了开发人工智能，需要收集、储存、分析、处理和使用大量信息，以便机器或机器用户生成各种结果、行动或行为。然而，正如联合国教科文组织在上述建议中指出，“隐私权对于保护人的尊严、自主权和能动性不可或缺，在人工智能系统的整个生命周期内必须予以尊重、保护和促进”。[11]

10. 随着人工智能的发展，适当或按情况处理个人数据对于防止人权受到伤害或威胁至关重要。有几个倡议和组织致力于要求开发符合人权的人工智能。下文提供了一些实例。

11. 首先，2020 年 10 月，全球隐私大会通过了关于开发和使用人工智能的问责制决议，[12] 其中敦促开发或使用人工智能系统的组织考虑执行以下问责措施：

- 在开发和/或使用人工智能之前，评估对人权(包括数据保护和隐私权)的潜在影响；
- 在人工智能投入使用前，测试其稳健性、可靠性、准确性和数据安全性，包括识别和解决系统及其使用的数据中可能导致不公平结果的偏差；
- 针对干预人权的风险，执行适当的问责措施。

12. 教科文组织还在建议中指出：

> 需要对算法系统开展充分的隐私影响评估，其中包括使用算法系统的社会和伦理考量以及通过设计方法对于隐私的创新使用。人工智能行为者需要确保他们对人工智能系统的设计和实施负责，以确保个人信息在人工智能系统的整个生命周期内受到保护。[13]

13. 2019 年 6 月，伊比利亚美洲数据保护网发布了一份题为“人工智能处理个人数据的一般建议”的文件。[14] 其中向人工智能产品开发者提出一些建议，指导他们从产品设计阶段就考虑到个人数据处理权条例的要求。建议如下：

- 遵守当地的个人数据处理条例；
- 进行隐私影响评估；
- 通过设计和默认嵌入隐私、伦理和安全；
- 执行问责制原则；

---

[11] 见 https://unesdoc.unesco.org/ark:/48223/pf0000381137，第 21 页。

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

[13] 见 https://unesdoc.unesco.org/ark:/48223/pf0000381137，第 21-22 页。

[14] 伊比利亚美洲数据保护网，“General recommendations for the treatment of personal data in artificial intelligence”，(2019)。保护网成员 2019 年 6 月 21 日在墨西哥 Naucalpan de Juárez 举行的会议上通过的文本。可查阅 https://www.redipd.org/sites/default/files/2020-02/guia-recomendaciones-generales-tratamiento-datos-ia.pdf。

- 在开发人工智能产品的组织中，就个人数据的处理设计适当的治理方案；
- 采取措施，确保在人工智能项目中执行个人数据处理原则；
- 尊重数据所有者的权利，并实施有效机制来行使这些权利；
- 确保个人数据的质量；
- 使用匿名化工具；
- 对个人数据所有者的信任度和透明度。

14. 就执行其中一些建议的细节，伊比利亚美洲数据保护网还进一步编写了细则，载于题为“遵守人工智能项目中个人数据保护原则和权利的具体准则”的文件。[15] 本报告将更详细地讨论稍后提到的透明度原则。

三. 人工智能的固有风险

15. 人工智能正在塑造社会及其数字化转型，它存在于日常生活、经济、科学、教育、卫生以及许多其他部门和活动的多个方面。

16. 尽管人工智能为社会提供了不可否认的益处和机会，它但也可能带来内在的挑战、风险和威胁，其中可能包括不道德的开发或使用，以及对人类作出有偏见、不透明或不正确的决定。

17. 风险程度取决于每种具体情况。

> 欧盟委员会认为，考虑到利害关系，特别是从保护安全、消费者权利和基本权利的角度，考虑到该部门和预期用途是否涉及重大风险，某一特定[人工智能]应用通常应被视为高风险。更具体地说，当[人工智能]应用满足以下两个累积标准时，应被视为高风险：

(a) 首先，使用[人工智能]应用的部门，鉴于通常开展的活动的特点，预计会出现重大风险。[……]例如，卫生保健、运输、能源和部分公共部门[……]；

(b) 其次，有关部门使用[人工智能]应用的方式有可能产生重大风险。[……]。对特定用途风险程度的评估可基于对受影响方的影响。例如，使用[人工智能]应用对个人或公司的权利产生法律或类似的重大影响；造成伤害、死亡或重大物质或非物质损害的风险；产生个人或法律实体无法合理避免的影响。[16]

[15] 伊比利亚美洲数据保护网，“Specific Guidelines for Compliance with the Principles and Rights that Govern the Protection of Personal Data in Artificial Intelligence Projects”，(2019)。可查阅 https://www.redipd.org/sites/default/files/2020-02/guide-specific-guidelines-ai-projects.pdf。

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

18. 人工智能涉及不同类型的风险。应考虑的意外情况包括算法运行的固有风险(人为偏见、技术缺陷、安全漏洞、算法实施失败)及其错误的设计。如下图所示，某些问题会影响算法的管理和性能：[17]

19. 如文献中所解释的：

> 数据输入主要受两个变量的影响：偏差(纳入部分、不充分、被操纵或过时的数据)和相关性(数据的相关性、不一致性或完整性)。另一方面，算法的开发可能会受到模式(编程逻辑偏差，包括不可预见的功能和用于编码的功能的固有故障)和错误(运行条件反映了不同于计划的运行方法，且违背了拟议设计的前提)的影响。最后，作为对数据输入分析的直接响应，输出决定的风险与算法执行的相关性和精确性有关。[18]

四. 处理个人数据的透明度原则

20. 透明度是计算机科学、信息获取、法律、个人数据处理等多个学科使用的一个概念。教科文组织认为，“透明度的目的是为相关对象提供适当的信息，以便他们理解和增进信任”。[19]

21. 对每种情况下的透明度范围并无共识。该词在每种情况下有着不同含义。例如，透明度原则用于一般个人数据处理时意味着一件事，用于人工智能时则意味着另一件事。本报告提到了一般个人数据处理的透明度，特别是人工智能方面的透明度。

---

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

[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)，第 9-10 页。可查阅 https://www.researchgate.net/publication/331358231_Robo-Advisors_Asesoria_automatizada_en_el_mercado_de_valores。

[19] 教科文组织，《人工智能伦理问题建议书》，2021 年，第 22 页。

22. 世界各地组织的若干文件讨论了透明度原则。[20] 特别报告员此前指出，控制方必须从收集时起就告知数据主体其个人信息将面对的处理条件，以便主体能够对数据行使应有的控制。[21]

23. 在引用的报告中，特别报告员根据以下关于隐私和个人数据处理的国际文件分析了透明度原则：(a) 《欧洲联盟一般数据保护条例》；(b) 《关于在自动处理个人数据方面保护个人的公约》；(c) 伊比利亚美洲数据保护网通过的《伊比利亚美洲国家个人数据保护标准》；(d) 理事会关于《经济合作与发展组织保护隐私和个人数据跨境流动准则》的建议；(e) 亚太经济合作论坛隐私权框架；(f) 美洲国家组织《隐私和个人数据保护最新原则》及说明。

24. 她从分析中得出结论认为，作为一般性规则，必须披露以下信息：

- 控制方或其代表的身份和地址，以及处理的目的或宗旨[……]这些数据是透明度的基本基础；
- 数据主体的权利和行使权利的方式，以及数据的接收者或接收者类别；
- 数据处理的法律基础或依据，以及数据处理的存在和/或主要特点；
- 所处理的数据类别和数据来源(如果不是直接从数据主体处获得)。

25. 值得注意的是，为执行透明度原则，向数据主体提供的信息必须使用简单、清晰、易懂、易及、易理解的语言。在涉及儿童和青少年时也必须坚持这项任务，并作必要调整。

26. 并非所有上述规范性文书都要求披露同样信息，因为有些文书对必须披露的信息类型列出了更详尽的清单。《欧盟一般数据保护条例》必须披露的信息包括：[22] 数据保护官的联系方式；个人数据的保存期限或确定保存期限的标准；控制方是否计划进行通信或转让，条例是否授权此类通信或转让；向监督当局投诉的权利；数据通信是否是法定要求或合同要求，或签订合同的必要条件，主体是否需要提供其个人数据，

[20] 经济合作与发展组织 (经合组织)，“保护隐私和个人数据跨境流通指导原则”，1980 年 9 月 23 日，以及 2013 年 7 月以来更新的指导原则；欧洲理事会，《关于在自动处理个人数据方面保护个人的公约》，第 108 号，1981 年 1 月 28 日；联合国，《电脑个人数据档案的管理准则》，1990 年 12 月 14 日；欧洲理事会，涉及监管机构和跨境数据流动的《关于在自动处理个人数据方面保护个人的公约的附加议定书》，2001 年 11 月 8 日；亚洲太平洋经济合作组织，*Asia-Pacific Economic Cooperation Forum Privacy Framework*, 2004；西班牙数据保护局, *Joint Proposal for a Draft of International Standards on the Protection of Privacy with regard to the Processing of Personal Data*, 2009 年 11 月 5 日，马德里；欧洲议会和欧洲理事会关于在处理个人数据方面保护自然人以及关于此类数据自由流动的(欧洲联盟)条例第 2016/679 号并废除第 95/46/EC 号指令(《一般数据保护条例》)，2016 年 4 月 27 日；伊比利亚美洲数据保护网，《伊比利亚美洲社区统一数据保护准则》, 2017 年；欧洲理事会，《在个人数据自动处理方面保护个人公约修订议定书》，2018 年 10 月；美洲国家组织，美洲法律委员会，*Updated Principles on Privacy and Personal Data Protection*, 2021。

[21] A/77/196，第 45 段。

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

以及不提供的后果；是否存在自动化决策，包括概况分析。在这种情况下，必须提供有关所涉逻辑的实质信息，以及这种处理的意义和预期后果，如果控制方打算进一步处理数据的目的与收集数据的目的不同，还必须提供有关目的的信息。

五. 人工智能领域处理个人数据的透明度原则

27. 必须确保人工智能的透明度，因为缺乏这方面的认识或疏忽可能会产生负面影响。欧盟委员会指出：

> 缺乏透明度([人工智能]的不透明性)使得难以识别和证明可能违反法律得行为，如保护基本权利、确定责任归属、满足索赔条件的法律规定。[23]

28. 可以通过要求遵守最低透明度标准来减轻人工智能潜在的不透明性。因此，委员会确定了以下要求：

> 确保提供明确信息。说明[人工智能]系统的能力和局限性，特别是系统的预期目的、系统可望按预期运行的条件以及实现特定目的的预期准确度[……]。另外，公民在与[人工智能]系统而不是与人类互动时，应被清楚告知[……]。此外，所提供的信息必须客观、简明和易于理解。[24]

29. 教科文组织在建议中指出：

> 具体到人工智能系统，透明度可以帮助人们了解人工智能系统各个阶段是如何按照该系统的具体环境和敏感度设定的。透明度还包括深入了解可以影响特定预测或决定的因素，以及了解是否具备适当的保证(例如安全或公平措施)。[25]

30. 人工智能高级别专家组指出，为实现可信赖的人工智能，必须满足某些要求，包括透明度。关于透明度，必须：

> 明确、主动地向利益攸关方传达有关[人工智能]系统的能力和局限性的信息，使其设定现实的期望值，并传达执行要求的方式[；并且]对他们正在与[人工智能]系统打交道这一事实保持透明。[26]

教科文组织还建议，“对于由人工智能系统直接提供或协助提供的产品或服务，人工智能行为者应以适当和及时的方式告知用户”。[27]

---

[23] 欧洲委员会，*White Paper on Artificial Intelligence-a European approach to excellence and trust*，2020 年，第 14 页。

[24] 同上，第 23-24 页。

[25] 见 https://unesdoc.unesco.org/ark:/48223/pf0000381137，第 22 页。

[26] 见 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai，第 2 和 3 页。

[27] 见 https://unesdoc.unesco.org/ark:/48223/pf0000381137，第 22 页。

31. 根据教科文组织的建议：

> 可解释性与透明度密切相关，结果和导致结果的子过程应以可理解和可追溯为目标，并且应切合具体情况。人工智能行为者应致力于确保开发出的算法是可以解释的。就[人工智能]对终端用户所产生的影响不是暂时的、容易逆转的或低风险的人工智能应用程序而言，应确保为导致所采取行动的任何决定提供有意义的解释，以便使这一结果被认为是透明的。[28]

32. 人工智能高级别专家组解释说，透明度“与可解释性原则密切相关，包括与[人工智能]系统相关的要素：数据、系统和业务模式的透明度”。[29] 它还强调了可追溯性、可解释性和沟通的相关性，具体如下：

- 可追溯性：数据集和产生[人工智能]系统决定的过程，包括数据收集和数据标注以及所使用的算法，应按照尽可能好的标准进行记录，以实现可追溯性并提高透明度。这也适用于[人工智能]系统作出的决定。这样就能找出[人工智能]决定出错的原因，从而有助于防止今后再出错。可追溯性有助于可审计性和可解释性。

- 可解释性：可解释性涉及解释[人工智能]系统的技术过程和相关人为决定(例如系统的应用领域)的能力。技术可解释性要求[人工智能]系统作出的决定能够被人类理解和追踪。此外，可能需要在增强系统的可解释性(这可能降低其准确性)或提高其准确性(以可解释性为代价)之间作出权衡。每当[人工智能]系统对人们生活产生重大影响时，就应可要求对[人工智能]系统的决策过程作出适当解释。这类解释应当及时，并适应相关利益攸关方(例如外行、监管者或研究人员)的专业知识。此外，还应解释[人工智能]系统影响和塑造组织决策过程的程度、系统的设计选择以及部署系统的理由(从而确保业务模式的透明度)。

- 沟通[人工智能]系统不应该向用户表示自己是人类；人类有权被告知他们正在与[人工智能]系统互动。这意味着[人工智能]系统必须是可识别的。此外，必要时还应提供选项，可决定不进行这种互动而进行人际互动，以确保遵守基本权利。除此之外，[人工智能]系统的能力和局限性应当以适合当前使用情况的方式传达给[人工智能]从业人员或终端用户。这可以包括传达[人工智能]系统的准确程度及其局限性。[30]

33. 欧洲数据保护委员会和欧洲数据保护监督员发表一项联合意见，其中指出：

> 当数据主体的数据用于[人工智能]训练和/或预测时，应始终告知数据主体此类处理的法律依据、逻辑(程序)的一般解释和[人工智能]系统的范围。在这种情况下，应始终保障个人限制处理的权(《一般数据保护条例》第 18 条、《欧盟数据

[28] 同上，第 22 页。

[29] 见 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai，第 18 页。

[30] 同上。

保护条例》第 20 条)以及删除/擦除数据权(《一般数据保护条例》第 16 条和《欧盟数据保护条例》第 19 条)。此外，控制方应有明确义务告知数据主体反对、限制、删除数据等的适用期限。[人工智能]系统必须能够通过适足的技术和组织措施满足所有数据保护要求。解释权应增加透明度。[31]

34. 特别报告员在上文提到的报告[32] 中指出，在数据主体要接受自动化决策或概况分析的情况下，他们须了解处理与其相关的信息的方式(例如是否涉及人工智能)、关于所涉逻辑的实质信息、重要意义和预期后果的信息。

35. 关于这一点，西班牙数据保护局指出，“‘实质’ [……]一词必须理解为一旦提供给数据主体，就能使数据主体了解其数据正在经历的处理类型，并对相关结果提供确定性和信任的信息”。[33]

36. 西班牙数据保护局还指出：

通过提供有关算法实施的技术参考来遵守这一义务可能会模糊不清、令人困惑或导致信息疲劳。应提供足够的信息，使主体能够理解处理行为。虽然这取决于所使用的[人工智能]组件的类型，但与数据主体相关的信息类型举例如下：

- 用于决策的数据的详细信息，而不仅限于类别，特别是关于数据使用期限的信息(数据有多长时间)。
- 决策过程中给予每个数据的相对重要性或权重。
- 训练数据的质量和所用模型的类型。
- 开展的概况分析活动及其影响。
- 根据用于衡量推断有效性的特定指标得出的误差或精度值。
- 是否存在合格的人工监督。
- 审计参考，特别是对推断结果可能出现的偏差进行的审计，以及[人工智能]系统的认证。对于适应型或进化型系统，是最近一次进行的审计。
- 如果[人工智能]系统包含涉及可识别第三方的信息，禁止未经合法授权处理此类信息以及这样做的后果。[34]

[31] 欧洲数据保护委员会和欧洲数据保护监督员，关于欧洲议会和欧洲理事会制定人工智能统一规则的条例提案的第 5/2021 号联合意见(人工智能法)，2021 年 6 月 18 日，第 17 页。可查阅 https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf。

[32] A/77/196，第 55 段。

[33] 西班牙数据保护局，*Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial.Una introducción*，2020 年 2 月。第 24 页。可查阅 https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf。

[34] 同上。

37. 欧洲数据保护监督员已发布一项意见，建议如果欧盟委员会要提出一个新的人工智能专用监管框架，那么，一定数量的合理保障措施应适用于所有人工智能应用，无论风险程度如何。这些措施包括，制定技术和组织措施(包括文件)；对所实施的算法系统的目标、使用和设计完全透明；确保人工智能系统的稳健性，执行现有的问责、补救和独立监督机制并保持其透明度。[35]

38. 欧洲数据保护委员会和欧洲数据保护监督员也指出，需要促进：

> [以新的、更积极主动、更及时的方式，随时向[人工智能]系统用户通报系统所处的(决策)状态，对潜在的有害结果发出预警，以便权利和自由可能因机器的自主决定而受损的个人作出反应或纠正决定。[36]

39. 伊比利亚美洲数据保护网认为，为执行透明度原则，必须采取下列行动：[37]

- “向数据主体传达其个人信息将被提交处理的主要特点”；
- “明确告知数据主体，将使用自动化过程处理其个人数据”；
- “在控制者为实施透明度原则而选择的方法中包括处理数据主体数据的所有目的”；
- “在通过转让获得个人数据时，披露这些数据的来源；在有意使用人工智能时，确认为用于该目的而获得数据的第一控制方将该目的通知数据主体”；
- “开发创新方法，告知数据主体处理的主要特点以及隐私预期增加或减少的风险程度。
- “保障信息自决权，确保始终以适当、及时的方式告知数据主体他们将与人工智能系统直接互动，或他们的信息将由人工智能系统处理”。
- “提供关于人工智能系统的目的和效果的实质信息，以验证是否持续符合数据主体的隐私预期，使数据主体能够随时对其个人数据的处理行使控制”；
- “确定和界定常用术语，并创建数据库，以便这些术语可在不同背景下重复使用，并配有标准图标，使数据主体了解信息”；
- “持续告知数据主体，以便他们了解自动化决策对他们的影响，以及在需要时如何请求人工干预，以便他们就是否同意处理作出知情决定”。

---

[35] 欧洲数据保护监督员，*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*，2020 年 6 月 29 日，第 14 页。可查阅 https://edps.europa.eu/sites/edp/files/publication/20-06-19_opinion_ai_white_paper_en.pdf。

[36] 欧洲数据保护委员会、欧洲数据保护监督员，关于欧洲议会和欧洲理事会制定人工智能统一规则条例(人工智能法)提案的第 5/2021 号联合意见，2021 年 6 月 18 日，第 22 页。

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

40. 伊比利亚美洲数据保护网指出：

> 提供的关于[人工智能]模型逻辑的信息应至少包括其运行的基本方面，以及数据的权重和相关性，以清晰、简单和易于理解的语言编写。没有必要对所使用的算法进行完整解释，甚至没有必要包括它们。[38]

41. 伊比利亚美洲数据保护网络呼吁负责人工智能处理数据的人进行创新，以便以简单明了的方式传达信息，并表示“有几种提供隐私通知的创新方法，包括使用视频、卡通和标准化图标。综合使用各种方法有助于数据主体更容易理解复杂的[人工智能]信息”。[39]

42. 一些国家在本国法律中明示或默示使用人工智能处理个人数据的透明度原则，以下各段列举了部分国家实例。

43. 在厄瓜多尔，2021 年通过的《有机数据保护法》第 12 条第 14 款和第 17 款规定，有权获知是否存在不接受仅基于自动化评价所作决定的影响的权利、行使该权利的方式以及是否存在自动化评估和决定，包括概况分析。

44. 该法还规定，在直接从数据主体获得数据的情况下，应事先(在收集个人数据时)告知有关信息。第 12 条进一步规定：

> 如果个人数据不是直接从数据主体处获得，或是从公众可获取的来源收集的，应在三十(30)天内或在数据主体收到的第一次通知中(以先发生者为准)通知数据主体。应向数据主体提供清晰、明确、透明、易懂、简洁和准确的信息，不得有任何技术障碍。

45. 在秘鲁，第 29733 号法《个人数据保护法》执行条例第 72 条对客观处理个人数据的权利作了以下规定：

> 为维护根据本法第 23 条规定的客观处理权，[40] 当个人数据作为决策过程的一部分进行处理而数据主体没有参与时，个人数据数据库的控制方或处理的控制方应立即通知数据主体，除非关于行使本法及其[……]条例中规定的其他权利的条例另有规定。

46. 在圣多美和普林西比，2016 年 5 月 2 日第 3/2016 号法《个人数据保护法》的独特之处在于，该法第 21 条规定，控制方或其代表应在开始处理前不超过八天书面通知国家个人数据保护局，他们将开始完全或部分自动化处理或分批处理，以实现一个或多个相关目的，但有一些例外。该法第 11 条还规定，数据主体在行使其查阅权时，有权要求控制方告知对其相关数据进行自动化处理的原因。

---

[38] 见 https://www.redipd.org/es/documentos/guia，第 17-19 页。

[39] 同上。

[40] “第 23 条。客观处理权。数据主体有权不受制于对其具有法律效力或对其有重大影响，仅借助旨在评估其个性或行为的某些方面的个人数据处理而作出的决定，除非该决定发生在合同的谈判、执行或履行期间，或是为在公共实体任职而依法进行评价的情况下，且不损害数据主体为保护其合法权益而捍卫其观点的可能性”。

47. 在乌拉圭，2008 年 8 月 11 日第 18831 号法律《个人数据保护法》第 13 条规定，在收集数据之前，数据主体有权以明确、清晰和无误的方式获知在使用自动化数据处理评价其个性的某些方面，如工作表现、信誉、可靠性和行为时所采用的评估标准、程序以及技术解决方案或软件，以作出可能对数据主体产生重大影响的具有法律效力的决定。该法还规定，“如果不是直接从数据主体收集个人数据，应在控制方收到请求之日起五个工作日内向其提供[……]信息”。

六. 人工智能项目中处理个人数据的可解释性原则

48. 根据现有信息建立个人“虚拟概况”的做法正变得越来越普遍，而且决定往往是在使用各种技术工具对个人数据进行自动化处理的基础上作出的。

49. 在人工智能项目使用和处理数据的基础上对人类作出的决定可能对人类产生积极或消极影响。如何保护因人工智能工具或技术所作决策而受到影响的个人的权利，是令人关切的问题。例如，人工智能《白皮书》指出：“与任何新技术一样，使用[人工智能]既带来机遇，也带来风险。公民担心在面对算法决策的信息不对称时，他们无力捍卫自己的权利和安全”。[41]

50. 鉴于上述情况，人们需要了解哪些数据被用来作出影响他们的决定，以及作出这些决定的逻辑。除其他外，获得这些信息将使受影响者能够知道对他们作出的决定是否正确，如果不正确，则为自己辩护。换句话说，这些信息对于确保正当程序是必要的，因为它可以作为证据，证明在人工智能过程中处理个人数据时可能产生的不准确或不公正。在这方面，上述人工智能高级别专家组强调，可解释性原则：

> 对于建立和维持用户对[人工智能]系统的信任至关重要。这意味着过程需要透明，[人工智能]系统的能力和目的需要公开沟通，决定需要尽可能向直接和间接受影响者解释。没有这些信息，就无法对决定提出正当的质疑[……]。如果输出是错误或不准确的，对可解释性的需求在很大程度上依赖于背景以及后果的严重性。[42]

51. 所有这些都解释了为什么人工智能的透明度很重要，因为这种智能不应是模糊、秘密或误导性的。为此，上述《欧洲宣言》指出：

> 每个人都应能够受益于算法和人工智能系统的优势，包括在数字环境中作出自己的知情选择，同时保护自己的健康、安全和基本权利免受风险和伤害。[43]

52. 据此，伊比利亚美洲数据保护网在 2019 年建议，提高对个人数据主体的透明度。[44]

---

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

[42] 见 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai，第 13 页。

[43] 可查阅 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AJOC_2023_023_R_0001。

[44] 见 https://www.redipd.org/sites/default/files/2020-02/guia-recomendaciones-generales-tratamiento-datos-ia.pdf，第 23 和 24 页。

53. 随后，也与上述内容相关的是，全球隐私大会在其上述 2020 年决议中强调，开发或使用人工智能系统的组织应考虑以下措施：(a) 通过披露人工智能的使用、正在使用的数据以及人工智能所涉及的逻辑，确保透明度和公开性；(b) 确保确定一个负责任的人类行为者，可以向其提出与自动化决定有关的关切并对其行使权利，可以启动对决定过程的评价和人为干预；(c) 应要求以清晰易懂的语言解释人工智能所作的自动化决定；(d) 确保应要求对人工智能所作的自动化决定进行人为干预。[45]

54. 上述所有部分符合《一般数据保护条例》的规定，例如：

> 在未从数据主体处获得个人数据的情况下，控制方应向数据主体提供以下信息：[……] 2.(g)是否存在第 22 条第(1)款和第(4)款所述的自动化决策，包括概况分析，以及至少在这些情况下，关于所涉逻辑的实质信息，以及这种处理对数据主体的意义和预期后果。[46]
>
> 此外，数据主体或数据所有者有权：
>
> 从控制方处获得是否正在处理其个人数据的确认，如是，则有权查阅其个人数据和以下信息：[……](h)是否存在第 22 条第(1)款和第(4)款所述的自动化决策，包括概况分析，以及至少在这些情况下，关于所涉逻辑的实质信息，以及这种处理对数据主体的意义和预期后果。[47]

55. 美国国家标准和技术研究所在下表中概括了这一原则的范围：[48]

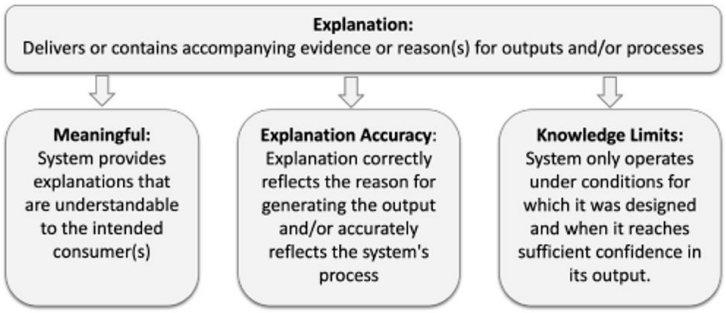

Fig. 1. Illustration of the four principles of explainable artificial intelligence. Arrows indicate that for a system to be explainable, it must provide an explanation. The remaining three principles are the fundamental properties of those explanations.

[45] 见 https://globalprivacyassembly.org/document-archive/adopted-resolutions/，第 3 页。

[46] 见 http://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679，第 14 条，第 2 (g)款。

[47] 同上，第 15 条第(1)款。

[48] 美国国家标准和技术研究所，*Four Principles of Explainable Artificial Intelligence*, - NISTIR 8312 (2021)，第 3 页。可查阅 https://doi.org/10.1017/bhj.2023.10。

56. 下表根据美国国家标准和技术研究所文件解释了每项原则最相关的方面：[49]

原则	含义或范围
解释	与[人工智能]系统的输出和/或过程相关的证据、支持或推理。
实质	用目标消费者可以理解的术语进行解释。换言之，这一原则旨在使给定受众理解所作的解释。许多因素会影响一个好的解释，因此必须考虑到解释的目标受众或受众。
解释的准确性	这一原则要求技术解释要严谨、准确、全面。
知识的局限性	识别和宣布知识的局限性意味着要明确，系统既不完美也不绝对可靠，因为[人工智能]在它被编程的某些限制和约束内运行。除其他因素外，它们还取决于所处理信息的质量和数量。

57. 有人认为，解释必须：(a) “对用户来说是可以理解和令人信服的”；(b) “准确反映系统的推理”；(c) “全面”，以及(d) “具体，即不同用户有不同的情况或不同的结果，应得到不同类型的解释”。[50] 此外，还指出：

> 人工智能的可解释性从伦理甚至法律的角度来看都是可以理解的一种愿望，但它有深层次的技术困难，值得我们去了解，而解决方案的很大一部分也很可能是技术性的，以至于有可能重新设计算法或确定新算法来满足伦理和监管愿望。[51]

教科文组织表示：

> 可解释性是指让人工智能系统的结果可以理解，并提供阐释说明。人工智能系统的可解释性也指各个算法模块的输入、输出和性能的可解释性及其如何促成系统结果。[52]

58. 为确定可解释性原则的范围，必须铭记其目标，并在此基础上确定实现这一目标所需的条件。根据上述情况，有人指出：

> 如果可解释性原则的目的是让任何人知道，为什么一项决定是根据[人工智能]工具对其数据进行处理而作出的，那么，解释至少应清晰、简单、完整、真实，并让请求解释的人容易理解。仅仅报告用于生成决定的输入的数据是不够的，还应提供用于作出决定的逻辑或方法。这个挑战并不小，但是，如果愿意向人们深入浅出地解释为什么一个决定是基于对他们个人数据的处理而产生的，这是可以实现的。[53]

---

[49] 表中的解释是引用的英文原件的改编和摘要，可查阅 https://doi.org/10.6028/NIST.IR.8312。

[50] Gavilán, Ignacio, “Cuatro principios para una buena explicabilidad de los algoritmos”(2022)。可查阅 https://ignaciogavilan.com/cuatro-principios-para-una-buena-explicabilidad-de-los-algoritmos/。

[51] 同上。

[52] 见 https://unesdoc.unesco.org/ark:/48223/pf0000381137，第 23 页。

[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).

59. 以下是一些国家的地方法律的一些例子，这些国家已默认或明示将可解释性原则纳入其法律框架。

60. 哥伦比亚法律禁止处理“误导”的数据，[54] 并在就贷款申请作出决定的具体情况下，要求拒绝此类申请的人在有要求时以书面形式告知有关人员“拒绝的客观理由”。[55]

61. 厄瓜多尔《有机数据保护法》第 20 条规定，数据所有者在面临完全或部分基于自动化过程(包括概况分析)产生的评估作出的决定时，如果该决定对其产生法律效力或侵犯其基本权利和自由，可要求对该决定作出合理解释，获得自动化程序的评估标准，提交意见，要求提供所用数据类型和数据来源的信息，并向责任人或负责人提出对该决定的异议(除某些例外)。

62. 乌拉圭第 18331 号法律第 16 条规定：

> 个人有权不受制于基于自动化数据处理、对其产生重大影响、具有法律效力的决定。该决定旨在评估其个性的某些方面，如工作表现、信誉、可靠性和行为。受影响的任何人都有权从数据库负责人处获得有关评价标准和数据处理所用程序的信息。这些信息是用来作出该法律所述的决定的。

七. 结论

63. 从上文可以得出以下结论：

(a) 透明度和可解释性有助于建立对人工智能的信任和尊重人权；

(b) 人工智能的开发者必须对如何处理数据(如何收集、存储和使用数据)以及如何作出基于人工智能的决定、这些决定的可靠性、信息的安全性保持透明；

(c) 受基于人工智能所作决定影响的人应得到关于该决定原因的清晰、简单、完整、真实和可理解的解释。在这方面，可解释性原则至关重要，不仅因为它与透明度原则相一致，还因为它能够维护这些人的辩护权和正当程序权；

(d) 可解释性和透明度要求使用人工智能作出的决定以及基于信息、特别是个人数据作出关于人类的决定的逻辑、方法或推理清晰、完整、真实、公正和公开。可解释性和透明度与不透明、模糊、欺骗、谎言和滥用计算能力截然相对，后者是非法和不道德处理的一些表现，反映了对人类及其尊严缺乏尊重。

[54] 2012 年第 1581 号法令确立了保护个人数据的一般规定，第 4 d)条。

[55] 2021 年第 2157 号法令修正和补充了 2008 年第 1266 号法令，并确立了关于金融、信贷、商业、服务和第三国信息的个人数据保护的一般规定和其他规定，第 5 条第 1 款。

八. 建议

64. 鉴于上述情况，特别报告员敦促各国：

(a) 促进人工智能的透明度，以减轻不透明可能给社会带来的风险，特别是在保护人权方面；

(b) 在法律中纳入可解释性原则。这不仅使人们了解影响他们的决定是如何作出的，还向他们提供在人工智能面前捍卫人权的工具；

(c) 促进伦理实践，确保人工智能项目或过程中处理个人数据的透明度和可解释性；

(d) 培养、支持和推动教育和数字素养，使公民更好地理解与人工智能、透明度和可解释性有关的概念，从而要求尊重其权利。