INTERNATIONAL REVIEW of the Red Cross

# 利用人工智能的潜力开展人道行动：机遇与风险

安娜·贝杜斯基*著/张灯**译


## 摘要

数据驱动的人工智能（AI）技术正在逐步改变人道领域，但这些技术给冲突和危机局势中的弱势个体和群体的保护带来了重大风险。本文考察了在人道行动中使用人工智能的机遇和风险，并考察了人工智能是否以及在何种情况下可以安全地使用以支持人道参与方在一线开展的工作。本文认为，人工智能有潜力支持人道参与方，因其实现了从反应性到预测性的人道行动方式的范式转变。但与此同时，本文建议，若要使人工智能为人道行动服务，而非以牺牲人道为代价地使用人工智能，就必须优先处理现有风险，包括与算法偏见和数据隐私问题有关的风险。由此，本文将对当前有关是否有可能在人道行动中负责任地利用人工智能潜力的讨论有所助益。


**关键词：**人工智能；“不伤害”；人道行动；人道原则；人权。

* 安娜·贝杜斯基（Ana Beduschi）是埃克塞特大学法学副教授和日内瓦国际人道法和人权研究院高级研究员。
** 复旦大学国际关系与公共事务学院博士研究生。

# 引言

在人道行动中使用数字技术并非一个新现象。几十年来，人道参与方一直在利用数字技术来帮助和保护受冲突和危机影响的群体。[1]然而，当前计算能力的进步，加上海量数据（包括大数据）的可得性，已使在人道背景下更广泛地使用数字技术成为可能。[2]新冠疫情进一步加速了使用数字技术以帮助维持人道行动的趋势。[3]

人工智能就是这样一种正在逐步改变人道领域的数字技术。尽管没有国际公认的定义，但人工智能被广泛理解为“将数据、算法和计算能力结合在一起的技术集合”。[4]这些技术包括：

> 由人类设计的、在给定复杂目标的情况下，通过借助数据采集感知环境，解释收集的结构化或非结构化数据，对知识进行推理或处理从这些数据得出的信息，并决定为实现给定目标而采取的最佳行动的，在物理或数字维度上发挥作用的软件（可能也包括硬件）系统。[5]

这个定义包括两方面主要内容：知识型系统和机器学习系统。知识型系统见于计算机程序，其使用现有知识库来解决通常依赖人类专门技能的问题。[6]机器学习是“对算法和系统的系统性研究，以通过经验提升其知识或性能”。[7]通过机器学习，机器可以被训练以理解数据。例如，人工智能系统可以被训练来执行自然语言处理等任务，利用计算机的能力来解析和解释书面和口语文字。[8]深度学习是机器学习的一个子概念，特别用于执行图像、视频、语音和音频处理等任务。[9]本文的分析对这两类系统都适用。

人工智能系统通常利用大量的数据，包括由人道参与方直接收集的以及其他来源的信息，如大数据等，来学习、发现模式（patterns），对这些模式进行推断，并预测未来的行为。[10]大数据，或曰“大量高速、复杂和可变的数据”，[11]在人道背景下也越来越重要。大数据的一个重要组成部分源于社交媒体和在线平台上用户生成的内容，如文本、图像、音频和视频。[12]社交媒体平台往往为

---

[1] Patrick Meier, “New Information Technologies and Their Impact on the Humanitarian Sector”, *International Review of the Red Cross*, Vol. 93, No. 884, 2011; Anja Kaspersen and Charlotte Lindsey-Curtet, “The Digital Transformation of the Humanitarian Sector”, *Humanitarian Law and Policy Blog*, 5 December 2016, available at: https://blogs.icrc.org/law-and-policy/2016/12/05/digital-transformation-humanitarian-sector/ (all internet references were accessed in April 2022); Dzhennet-Mari Akhmatova and Malika-Sofi Akhmatova, “Promoting Digital Humanitarian Action in Protecting Human Rights: Hope or Hype”, *International Journal of Humanitarian Action*, Vol. 5, 2020.

[2] Ana Beduschi, “The Big Data of International Migration: Opportunities and Challenges for States under International Human Rights Law”, *Georgetown Journal of International Law*, Vol. 49, No. 4, 2018; Michael Pizzi, Mila Romanoff and Tim Engelhardt, “AI for Humanitarian Action: Human Rights and Ethics”, *International Review of the Red Cross*, Vol. 102, No. 913, 2021.

[3] Saman Rejali and Yannick Heiniger, “The Role of Digital Technologies in Humanitarian Law, Policy and Action: Charting a Path Forward”, *International Review of the Red Cross*, Vol. 102, No. 913, 2021; Jo Burton, “‘Doing no Harm’ in the Digital Age: What the Digitalization of Cash Means for Humanitarian Action”, *International Review of the Red Cross*, Vol. 102, No. 913, 2021; John Bryant, Kerrie Holloway, Oliver Lough and Barnaby Willitts-King, *Bridging Humanitarian Digital Divides during Covid-19*, Overseas Development Institute, London, 2020; Theodora Gazi and Alexandros Gazis, “Humanitarian Aid in the Age of COVID-19: A Review of Big Data Crisis Analytics and the General Data Protection Regulation”, *International Review of the Red Cross*, Vol. 102, No. 913, 2021.

[4] European Commission, *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, COM (2020) 65 final, 2020, p. 2.

[5] European Union High Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Scientific Disciplines*, Brussels, 2019, p. 6.

[6] Martin Swain, “Knowledge-Based System”, in Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho and Hiroki Yokota (eds), *Encyclopedia of Systems Biology*, Springer, New York, 2013.

[7] Peter Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, Cambridge, 2012, p. 3.

[8] *Ibid*., Jacob Eisenstein, *Introduction to Natural Language Processing*, MIT Press, Cambridge, MA, 2019.

[9] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, “Deep Learning”, *Nature*, Vol. 521, 2015; Neil Savage, “How AI and neuroscience drive each other forwards”, *Nature*, Vol. 571, No. 7553, 2019.

[10] Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms”, *Big Data & Society*, Vol. 3, No. 1, 2016; Sandra Wachter, Brent Mittelstadt and Chris Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”, *Harvard Journal of Law & Technology*, Vol. 31, No. 2, 2018.

[11] Tech America Foundation, *Demystifying Big Data: A Practical Guide to Transforming the Business of Government*, Washington, DC, 2012.

[12] Amir Gandomi and Murtaza Haider, “Beyond the Hype: Big Data Concepts, Methods, and Analytics”, *International Journal of Information Management*, Vol. 35, No. 2, 2015.

用户在冲突或危机期间提供特定的参与和沟通渠道。[13]例如，脸书已启用安全检查功能，用户可以在自然灾害或其他冲突或紧急情况发生时报告自己的状态。[14]人工智能系统可以在这些不同类型数据的基础上，描绘冲突和危机的演变。

在这方面，人工智能技术具有支持人道参与方的潜力，因其实现冲突或危机中人道行动方式从反应性到预测性的范式转变。[15]例如，2019 年，人工智能支持下的灾害测绘帮助人道工作者在莫桑比克迅速开展了紧急应对行动。[16]数据驱动的人工智能系统还可以建立在预测分析技术的基础上，以寻求识别数据中的模式和关系，从而预测一线的进展情况。[17]例如，由联合国难民署发起的“杰森项目”（Project Jetson）就利用预测分析技术来预判人们被迫流离失所的情况。[18]

然而，学者和社会活动家越来越多地表达了对在人道背景下使用人工智能所带来的风险的担忧。这些担忧包括从“监控人道主义”（surveillance humanitarianism）[19]带来的危险到过度的“技术解决主义”（techno-solutionism）[20]“技术殖民主义”（techno-colonialism）潜在上升有关的问题。[21]这些都是重大的风险，因为它们可能会使已经受到冲突或危机影响的群体面临额外的伤害和人权侵犯。

在此背景下，本文探讨了在人道行动中使用人工智能的机遇和风险。本文运用法律、政策导向和技术方面的学术和专业文献，以评估是否以及在什么情况下可以安全地使用人工智能来支持人道参与方在一线开展的工作。尽管很多学术和专业文献指出了对于在武装冲突中将人工智能用于军事行动的日益增长的兴趣，但这一领域仍不在本文的讨论范围之内。[22]这种选择的理由是，人工智能在军事行动之外的运用越来越多，以支持冲突、灾难和危机局势中的人道援助。

本文的分析分三步进行。首先，本文研究人工智能为支持人道参与方在一线工作所带来的机遇。其次，对这些技术所带来的现有风险进行评估。再次，基于“不伤害”的人道要求，本文提出了在

---

[13] Billy Haworth and Eleanor Bruce, “A Review of Volunteered Geographic Information for Disaster Management”, *Geography Compass*, Vol. 9, No. 5, 2015; A. Beduschi, above note 2; Pankaj Sharma and Ashutosh Joshi, “Challenges of Using Big Data for Humanitarian Relief: Lessons from the Literature”, *Journal of Humanitarian Logistics and Supply Chain Management*, Vol. 10, No. 4, 2020; T. Gazi and A. Gazis, above note 3.

[14] Facebook, “Crisis Response”, available at: www.facebook.com/about/safetycheck/.

[15] Mark Lowcock, “Anticipation Saves Lives: How Data and Innovative Financing Can Help Improve the World’s Response to Humanitarian Crises”, speech delivered at the London School of Economics, 2019, available at: https://reliefweb.int/report/world/mark-lowcock-under-secretary-general-humanitarian-affairs-and-emergency-relief; United Nations Office for the Coordination of Humanitarian Affairs (OCHA), *From Digital Promise to Frontline Practice: New and Emerging Technologies in Humanitarian Action*, New York, 2021; Christopher Chen, “The Future is Now: Artificial Intelligence and Anticipatory Humanitarian Action”, *Humanitarian Law and Policy Blog*, 19 August 2021, available at: https://blogs.icrc.org/law-and-policy/2021/08/19/artificial-intelligence-anticipatory-humanitarian/.

[16] OCHA, above note 15, p. 7.

[17] A. Gandomi and M. Haider, above note 12, p. 143.

[18] 见“杰森项目”网站：https://jetson.unhcr.org。

[19] “监控人道主义”一词是指在没有适当保障措施的情况下，人道组织增加数据收集的做法可能会无意中增加需要人道援助的个体的脆弱性。See Mark Latonero, “Stop Surveillance Humanitarianism”, *New York Times*, 11 July 2019. See also Keren Weitzberg, Margie Cheesman, Aaron Martin and Emrys Schoemaker, “Between Surveillance and Recognition: Rethinking Digital Identity in Aid”, *Big Data & Society*, Vol. 8, No. 1, 2021.

[20] “技术解决主义”一词是指决策者希望利用数字技术来解决那些不仅仅需要技术解决方案的复杂社会问题的意愿。See Mark Duffield, “The Resilience of the Ruins: Towards a Critique of Digital Humanitarianism”, *Resilience*, Vol. 4, No. 3, 2016; Petra Molnar, *Technological Testing Grounds*, EDRi and Refugee Law Lab, Brussels, 2020, available at: https://edri.org/wp-content/uploads/2020/11/Technological-Testing-Grounds.pdf.

[21] “技术殖民主义”一词广义上是指在数字创新中可能导致在世界各地不同群体之间再现依赖和不平等的殖民关系的做法。See Mirca Madianou, “Technocolonialism: Digital Innovation and Data Practices in the Humanitarian Response to Refugee Crises”, *Social Media & Society*, Vol. 5, No. 3, 2019; Nick Couldry and Ulises A. Mejias, “Data Colonialism: Rethinking Big Data’s Relation to the Contemporary Subject”, *Television & New Media*, Vol. 20, No. 4, 2019.

[22] Rain Liivoja, Kobi Leins and Tim McCormack, “Emerging Technologies of Warfare”, in Rain Liivoja and Tim McCormack (eds), *Routledge Handbook of the Law of Armed Conflict*, Routledge, London, 2016; Ronald Alcala and Eric Talbot Jensen, *The Impact of Emerging Technologies on the Law of Armed Conflict*, Oxford University Press, Oxford, 2019; 红十字国际委员会，《武装冲突中的人工智能与机器学习：以人为本的方法》，日内瓦，2019 年；Hitoshi Nasu, “Artificial Intelligence and the Obligation to Respect and to Ensure Respect for IHL”, in Eve Massingham and Annabel McConnachie (eds), *Ensuring Respect for International Humanitarian Law*, Routledge, London, 2020; Jai Galliott, Duncan MacIntosh and Jens David Ohlin, *Lethal Autonomous Weapons: Re-examining the Law and Ethics of Robotic Warfare*, Oxford University Press, Oxford, 2021.

人道背景下使用人工智能的主要建议。最后，本文就是否有可能安全地利用人工智能的益处，同时尽量减少其对人道行动构成的风险得出结论。

## 人工智能支持范式转变：从反应性到预测性的人道行动方式

如前所述，人工智能具有支持人道参与方的潜力，因其使人道行动方式实现了从反应性到预测性的范式转变。[23]这一转变意味着一旦可能预见到危机，就要采取行动，积极主动地减轻对弱势群体的不利影响。[24]在这方面，人工智能技术可以在三个主要的方面进一步扩充执行人道使命的“工具箱”：准备、应对和恢复。

准备是一个持续不断的过程，其目的是了解现有的风险并提出应对这些风险的行动建议，从而支持对危机和紧急情况作出更有效的人道应对。[25]应对的重点是向有需要的人提供援助，[26]而恢复则是指在提供即时人道救济之外进一步的方案。[27]因此，恢复是一个重要因素，因为当代人道危机往往日益复杂且旷日持久，从而超越了人道援助与发展合作之间的界限。[28]

### 准备

人工智能技术可以支持人道准备工作，因为人工智能系统可以用来分析大量数据，从而提供关于受影响群体潜在风险的基本判断。这些判断可以在危机或人道灾难发生之前，让人道工作者了解这些风险。[29]在这方面，基于数据驱动的机器学习和统计模型的预测分析技术可用于计算和预测即将发生的自然灾害、流离失所和难民流动、饥荒和全球卫生突发事件。[30]迄今为止，这种系统在早期预警和短期预测方面表现最好。[31]此外，它们的潜力是巨大的，因为执行预测分析的人工智能系统可以为准备工作提供帮助。

例如，红十字会与红新月会国际联合会通过使用基于预测的筹资方案，能够迅速分配人道资源，以便及早采取行动。[32]该方案使用各种数据来源，如气象数据和市场分析，以确定应在何时何地分配人道资源。[33]

---

[23] M. Lowcock, above note 15; OCHA, above note 15.
[24] M. Lowcock, above note 15.
[25] Inter-Agency Standing Committee, *The Implementation of the Humanitarian Programme Cycle*, Geneva, 2015.
[26] *Ibid.*
[27] International Federation of Red Cross and Red Crescent Societies (IFRC), “Recovery”, available at: www.ifrc.org/recovery.
[28] Executive Board of the United Nations Development Programme and of the United Nations Population Fund, *Role of UNDP in Crisis and Post-Conflict Situations*, UN Doc. DP/2001/4, Geneva, 2000, para. 48; Lucy Earle, “Addressing Urban Crises: Bridging the Humanitarian-Development Divide”, *International Review of the Red Cross* Vol. 98, No. 901, 2016; Atsushi Hanatani, Oscar A. Gómez and Chigumi Kawaguchi, *Crisis Management Beyond the Humanitarian-Development Nexus*, Routledge, London, 2018; Jon Harald Sande Lie, “The Humanitarian-Development Nexus: Humanitarian Principles, Practice, and Pragmatics”, *Journal of International Humanitarian Action*, Vol. 5, 2020.
[29] Kevin Hernandez and Tony Roberts, *Predictive Analytics in Humanitarian Action: A Preliminary Mapping and Analysis*, Institute for Development Studies, London, 2020.
[30] *Ibid.*; Petra Molnar, “Technology on the Margins: AI and Global Migration Management from a Human Rights Perspective”, *Cambridge Journal of International Law*, Vol. 8, No. 2, 2019; Ana Beduschi, “International Migration Management in the Age of Artificial Intelligence”, *Migration Studies*, Vol. 9, No. 3, 2020; Centre for Humanitarian Data, “OCHA-Bucky: A COVID-19 Model to Inform Humanitarian Operations”, The Hague, 2021, available at: https://centre.humdata.org/ocha-bucky-a-covid-19-model-to-inform-humanitarian-operations/; T. Gazi and A. Gazis, above note 3.
[31] K. Hernandez and T. Roberts, above note 29; Jessica Bither and Astrid Ziebarth, *AI, Digital Identities, Biometrics, Blockchain: A Primer on the Use of Technology in Migration Management*, Migration Strategy Group on International Cooperation and Development, Berlin, 2020.
[32] IFRC, “Forecast-based Financing: A New Era for the Humanitarian System”, 2021, available at: www.forecast-based-financing.org/wp-content/uploads/2019/03/DRK_Broschuere_2019_new_era.pdf.
[33] Toke Jeppe Bengtsson, *Forecast-based Financing: Developing Triggers for Drought*, Lund University, Lund, 2018.

另一个例子是联合国难民署的“杰森项目”，该项目利用预测分析技术预测由于索马里暴力和冲突升级所导致的被迫流离失所的情况。[34]“杰森项目”基于各种数据源，包括气候数据（如河流水位和降雨模式）、市场价格、汇款数据和该机构收集的其他数据，以训练其机器学习算法。

在另一个地区，世界粮食计划署开发了一个模型，利用预测分析技术来预测冲突地区的粮食不安全状况，因为在这些地区，传统的数据收集工作面临挑战。[35]该模型提供了一张地图，描绘了世界各地人口中营养不良的发生率。

但是，使用人工智能系统，特别是那些使用预测分析模型的系统，是否会让人道行动做好更充分的准备？这个问题需要细致入微的回答。一方面，在某些情况下，人工智能系统可能有益于人道行动，因为它们可能有助于更好地了解局势和更好地预测应对行动。例如，更好的准备工作有助于及早分配资源，这对一线行动的有效性来说可能是至关重要的。另一方面，对历史数据的分析不应是为未来行动提供信息和框架的唯一途径。基于过去数据分析的模型可能没有考虑到诸如人类行为和环境变化等变量，因此可能提供不完整或错误的预测。例如，在新冠疫情期间，大多数人工智能模型未能为医疗决策提供高效支持，以应对疫情暴发。[36]这在一定程度上是由于数据质量低（与新冠病毒无关的历史数据）以及存在偏见的风险高。[37]此外，侧重于分析过去数据的人工智能系统可能继续重现错误和不准确性，并使历史上的不平等、偏见和不公平现象长期存在。[38]因此，认真考虑即将使用人工智能系统的人道环境背景的特殊性，可能有助于避免不必要诉诸技术的情况，并防止“技术解决主义”的加剧。

“技术解决主义”，或相信技术可以解决大多数社会问题的信念，已被证明在人道领域导致了喜忧参半的结果。例如，研究表明，侧重于大数据分析以预测西非的埃博拉疫情暴发并不总是像投资适当的公共卫生和社会基础设施那样有效。[39]与受影响社区密切合作——例如，通过参与式设计[40]——可有助于根据社区的关键需求开展针对性的预测性干预措施，从而在冲突或危机发生之前更好地为人道行动提供参考并加以准备。如下节所述，这也可以适用于在人道应对中使用的人工智能系统。

## 应对

人工智能系统可用于支持冲突和危机期间的人道应对。例如，在深度学习、自然语言处理和图像处理方面的最新进展，可以使在危机和冲突局势下对社交媒体信息的分类进行得更快、更精确。这可协助人道参与方对紧急情况作出应对。[41]特别是，这些先进的人工智能技术可以帮助确定哪些领域将从向有需要者提供高效援助的工作中受益。

---

[34] 见“杰森项目”网站，前注 18；UNHCR Innovation Service, “Is It Possible to Predict Forced Displacement?”, *Medium*, 2019, available at: https://medium.com/unhcr-innovation-service/is-it-possible-to-predict-forced-displacement-58960afe0ba1.

[35] 见世界粮食计划署“饥饿地图”：https://hungermap.wfp.org/。

[36] Laure Wynants *et al*., “Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal”, *BMJ*, Vol. 369, 2020.

[37] *Ibid*., pp. 5–6.

[38] 见下一节关于“以牺牲人道为代价的人工智能：对受影响群体的风险”的讨论。See also Rashida Richardson, Jason Schultz and Kate Crawford, “Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice”, *New York University Law Review*, Vol. 94, 2019, p. 224.

[39] Dilon Wamsley and Benjamin Chin-Yee, “COVID-19, Digital Health Technology and the Politics of the Unprecedented”, *Big Data & Society*, Vol. 8, No. 1, 2021, p. 3.

[40] 参与式设计是一个从技术设计的早期阶段就包含各种利益相关方的过程。See Peter M. Asaro, “Transforming Society by Transforming Technology: The Science and Politics of Participatory Design”, *Accounting, Management and Information Technologies*, Vol. 10, No. 4, 2000; Elizabeth Rosenzweig, “UX Thinking”, in Elizabeth Rosenzweig (ed.), *Successful User Experience: Strategy and Roadmaps*, Elsevier, Amsterdam, 2015.

[41] Swati Padhee, Tanay Kumar Saha, Joel Tetreault and Alejandro Jaimes, “Clustering of Social Media Messages for Humanitarian Aid Response during Crisis”, 2020, available at: https://arxiv.org/pdf/2007.11756.pdf; Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam and Umair Qazi, “Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response”, *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2020.

例如，“紧急情况观测”（Emergency Situation Awareness）平台能够监测澳大利亚和新西兰的推特上的内容，以向用户提供有关地震、森林火灾和洪水等自然灾害的影响和范围的信息。[42]类似地，“应灾人工智能”（Artificial Intelligence for Disaster Response）是一个利用人工智能对社交媒体内容进行过滤和分类的开放平台，能够提供对灾难演变的判断。[43]诸如此类的平台可以对如社交媒体上发布的显示基础设施受损情况和灾民伤害程度的相关图片等的内容进行分流和分类，这对灾害应对和管理很有帮助。[44]

另一个例子是快速制图服务（Rapid Mapping Service），这是一个由联合国训练研究所（United Nations Institute for Training and Research）、联合国业务卫星应用方案（UN Operational Satellite Applications Programme）和联合国全球脉动（UN Global Pulse）联合开发的项目。[45]该项目将人工智能应用于卫星图像，以便迅速绘制洪灾地区的地图，并评估冲突或自然灾害（如地震和山体滑坡）造成的损害，从而为一线的人道应对提供信息。

这些例子是否就表明，人工智能可以促成在人道背景下开展更有效的应对行动呢？根据其设计和使用情况，人工智能系统或许可以支持冲突和危机中的人道应对行动。然而，这在很大程度上取决于具体环境。

利用人工智能技术绘制受灾地区的地图，似乎产生了令人满意的结果。例如，“人道主义开放街道地图”（Humanitarian OpenStreetMap）项目依靠的是能够绘制受灾地区地图的人工智能系统。[46]这个项目利用众包的社交媒体数据和卫星及无人机图像，提供关于哪些地区受灾害情况影响，且需要需要优先应对其灾情的可靠信息。然而，在武装冲突局势下的人道应对中，这样的项目可能不会产生具有相关性的结果。例如，虚假的信息宣传可能会影响武装冲突期间可靠数据的获得。[47]更为普遍的是，在武装冲突局势下可能很难获得高质量数据的问题，或许还会影响该背景下人工智能系统的设计和开发，从而影响其制图工具的适用性。

因此，人工智能技术尽管可能为支持有效的人道救济应对行动提供机遇，但不应将其理解为人道行动领域内任何情况下的现成的“一刀切”解决方案。

## 恢复

人工智能或可有效用于恢复工作，因为当代危机的复杂性往往导致旷日持久的冲突局势。[48]信息技术可以成为在此类背景下促进人道工作者和受影响社群之间互动协作的一项额外有利条件。[49]

人工智能技术可以支持长期局势中的人道行动。例如，红十字国际委员会开发的“照片寻人”（Trace the Face）工具旨在帮助难民和移民寻找失踪的家庭成员。[50]该工具使用面部识别技术自动

---

[42] 见“紧急情况观测”网站：https://esa.csiro.au/aus/about-public.html。
[43] 见“应灾人工智能”网站：http://aidr.qcri.org/。
[44] Declan Butler, “Crowdsourcing Goes Mainstream in Typhoon Response”, *Nature*, 2013, available at: www.nature.com/articles/nature.2013.14186; Wenjuan Sun, Paolo Bocchini and Brian D. Davison, “Applications of Artificial Intelligence for Disaster Management”, *Nature Hazards*, Vol. 103, No. 3, 2020.
[45] Felicia Vacarelu and Joseph Aylett-Bullock, “Fusing AI into Satellite Image Analysis to Inform Rapid Response to Floods”, United Nations Institute for Training and Research, 2021, available at: https://unitar.org/about/news-stories/news/fusing-ai-satellite-image-analysis-inform-rapid-response-floods.
[46] 见“人道主义开放街道地图”网站：www.hotosm.org。
[47] ICRC, *Harmful Information. Misinformation, Disinformation and Hate Speech in Armed Conflict and Other Situations of Violence*, Geneva, 2021, available at: https://shop.icrc.org/harmful-information-misinformation-disinformation-and-hate-speech-in-armed-conflict-and-other-situations-of-violenceicrc-initial-findings-and-perspectives-on-adapting-protection-approaches-pdf-en.html.
[48] Edwin Odhiambo Abuya, “From Here to Where? Refugees Living in Protracted Situations in Africa”, in Alice Edwards and Carla Ferstman (eds), *Human Security and Non-Citizens: Law, Policy and International Affairs*, Cambridge University Press, Cambridge, 2010; 红十字国际委员会，《旷日持久的冲突与人道行动：红十字国际委员会的近期经验》，日内瓦，2016 年，第 9～11 页；Ellen Policinski and Jovana Kuzmanovic, “Editorial: Protracted Conflicts: The Enduring Legacy of Endless War”, *International Review of the Red Cross*, Vol. 101, No. 912, 2019.
[49] Mirca Madianou, Liezel Longboan and Jonathan Corpus Ong, “Finding a Voice through Humanitarian Technologies? Communication Technologies and Participation in Disaster Recovery”, *International Journal of Communication*, Vol. 9, 2015; ICRC, above note 48, pp. 15, 37.
[50] ICRC, “Rewards and Risks in Humanitarian AI: An Example”, *Inspired: Innovation to Save Lives and Defend Dignity*, 2019, available at: https://blogs.icrc.org/inspired/2019/09/06/humanitarian-artificial-intelligence/.

搜索和匹配，从而简化了整个过程。另一个例子是人工智能驱动的聊天机器人，它可以为受影响的社区成员提供一个接触人道组织并获得相关信息的途径。这些聊天机器人目前正在向移民和难民提供咨询服务。[51]同样，人道组织也可以使用短信聊天机器人与受影响的群体联系。[52]

然而，至关重要的问题是，是否有可能从这些例子中归纳出人工智能有助于更好地采取恢复行动的结论。正如前文在分析准备和应对行动时所指出的那样，运用人工智能的益处在很大程度上取决于使用这些技术的具体环境。对于恢复行动而言同样如此。社区参与和以人为本的方式可能有助于确定在哪些领域，技术可以有效地支持一线的恢复工作，或者反过来，确定那些人工智能系统不会为恢复工作增加价值的领域。这应为在恢复方案中使用人工智能系统的决策提供信息。此外，人工智能技术也可能给受影响的群体带来相当大的风险，如加剧不相称的监视或由于算法偏见而导致不平等的现象长期存在。这些风险将在下一节中进行分析。

## 以牺牲人道为代价的人工智能：对受影响群体的风险

尽管人工智能可能会在人道领域带来潜在的宝贵成果，但使用这些系统并非没有风险。在人道行动的背景下，有三个主要领域是与之特别相关的：数据质量、算法偏见以及对数据隐私的尊重和保护。

### 数据质量

对用于训练人工智能算法的数据质量的担忧并不只限于人道领域，但这个问题可能会对人道行动产生重大影响。一般来说，糟糕的数据质量会导致同样糟糕的结果。[53]例如，在预测性警务和风险评估算法的背景下，情况就是如此。这些算法通常利用历史上的犯罪数据，如每个邮政编码对应的警察逮捕率和以往的犯罪记录，来预测未来的犯罪发生率和累犯风险。[54]如果用于训练这些算法的数据不完整或包含错误，那么算法的结果（即犯罪预测和累犯风险评分）可能同样质量不佳。研究确实发现，历史犯罪数据集可能是不完整的，而且可能包含错误，因为在美国等一些司法管辖区，违警记录中经常存在种族偏见。[55]如果这些算法被用来支持司法决策，就可能会导致基于种族的不公平和歧视。[56]

在人道背景下，数据质量差会导致不良后果，从而可能直接影响到因冲突或危机而本已处于脆弱境况的群体。用不准确、不完整或存在偏见的数据训练出来的人工智能系统有可能会使这些错误长期存在并逐级向前发展。例如，最近的一项研究发现，十个最常用的计算机视觉、自然语言和音

[51] Ana Beduschi and Marie McAuliffe, “AI, Migration and Mobility: Implications for Policy and Practice”, in Marie McAuliffe and Anna Triandafyllidou (eds), *World Migration Report 2022*, International Organization for Migration, Geneva, 2021; Marie McAuliffe, Jenna Blower and Ana Beduschi, “Digitalization and Artificial Intelligence in Migration and Mobility: Transnational Implications of the COVID-19 Pandemic”, *Societies*, Vol. 11, No. 4, 2021.

[52] ICRC, The Engine Room and Block Party, *Humanitarian Futures for Messaging Apps*, Geneva, 2017, available at: www.icrc.org/en/publication/humanitarian-futures-messaging-apps; Joanna Misiura and Andrej Verity, *Chatbots in the Humanitarian Field: Concepts, Uses and Shortfalls*, Digital Humanitarian Network, Geneva, 2019, available at: www.digitalhumanitarians.com/chatbots-in-the-humanitarian-field-concepts-uses-and-shortfalls/.

[53] Thomas Redman, “If Your Data Is Bad, Your Machine Learning Tools Are Useless”, *Harvard Business Review*, 2 April 2018, available at: https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless; R. Richardson, J. Schultz and K. Crawford, above note 38.

[54] Andrew Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, New York University Press, New York, 2017; Sarah Brayne, *Predict and Surveil: Data, Discretion, and the Future of Policing*, Oxford University Press, Oxford, 2020; R. Richardson, J. Schultz and K. Crawford, above note 38.

[55] S. Brayne, above note 54, pp. 33–34, 105; A. Ferguson, above note 54, p. 23; C. Dominik Güss, Ma Teresa Tuason and Alicia Devine, “Problems with Police Reports as Data Sources: A Researchers’ Perspective”, *Frontiers in Psychology*, Vol. 11, 2020.

[56] See, notably, Sonja B. Starr, “Evidence-Based Sentencing and the Scientific Rationalization of Discrimination”, *Stanford Law Review*, Vol. 66, No. 4, 2014; Supreme Court of Wisconsin, *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), 2016, p. 769（要求在量刑时使用算法风险评估之前应当提出警告，并规定风险评分不得用于“决定罪犯是否被监禁”或“决定刑罚的严重程度”）.

频数据集包含重大的标签错误（即对图像、文本或音频的错误识别）。[57]由于这些数据集经常被用于训练人工智能算法，因而错误将持续存在于生成的人工智能系统中。

遗憾的是，由于人道行动受到多方面的限制，故其可能难以获得高质量的数据。[58]例如，由于偏远地区的互联网连通不足，人道工作者在收集数据时可能会遇到问题。包含不同人道参与方所收集信息的不完整和重叠的数据集也可能是一个问题，例如，如果数据集里保留过时的信息，则不准确性可能会继续存在。[59]当使用大数据和众包数据时，也会出现错误和不准确的情况。[60]因此，使用这些数据集的团队尽可能地控制错误，这一点至关重要。然而，数据集和人工智能系统也可能受到算法偏见的影响，这个话题与数据质量有关，但具有更大的社会影响，因此将在下一节中讨论。

## 算法偏见

与数据质量问题相关的是人工智能系统设计和开发中存在的偏见问题。此处，偏见不仅被认为是一种技术或统计误差，而是还被认为是反映在人工智能系统中的、可能导致不公平的结果和歧视的人类观点、偏见和刻板印象。[61]人工智能系统确实可以反映其人类设计者和开发者的偏见。[62]一旦使用了这样的系统，反过来又可能导致非法歧视。

国际人权法禁止基于种族、肤色、性、性别、性取向、语言、宗教、政治或其他见解、民族或社会出身、财产、出生或其他身份的直接和间接形式的歧视。[63]当一个人基于其中一个或多个理由受到不利待遇时，就受到了直接歧视。间接歧视在即使措施表面上看似中立的情况下也会存在，因为相关措施实际上可能导致对个人的基于一个或多个受保护理由的不利待遇。

人工智能系统中的偏见可能加剧不平等，使直接和间接形式的歧视持续存在，特别是基于性别和种族的歧视。[64]例如，针对少数群体的结构性和历史性偏见由于其普遍性，可能反映在人工智能系统中。[65]偏见还通常来自用于训练人工智能算法的数据集中不同群体的代表性差距。[66]例如，研究人员已经证明，市场上可以获得的面部识别算法在识别肤色较深的女性方面不太准确，部分原因

---

[57] 相关实例见于 Curtis G. Northcutt, Anish Athalye and Jonas Mueller, “Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks”, 35th Conference on Neural Information Processing Systems, 2021, available at: https://arxiv.org/pdf/2103.14749.pdf。

[58] Christopher Kuner and Massimo Marelli, *Handbook on Data Protection in Humanitarian Action*, 2nd ed., ICRC, Geneva, 2020, p. 39; OCHA, above note 15, p. 10; ICRC, The Engine Room and Block Party, above note 52, p. 32.

[59] Anne Singleton, *Migration and Asylum Data for Policy-Making in the European Union: The Problem with Numbers*, CEPS Papers in Liberty and Security in Europe No. 89, 2016, available at www.ceps.eu/ceps-publications/migration-and-asylum-data-policy-making-european-union-problem-numbers/; European Union Agency for Fundamental Rights, *Data Quality and Artificial Intelligence: Mitigating Bias and Error to Protect Fundamental Rights*, Vienna, 2019.

[60] B. Haworth and E. Bruce, above note 13; P. Sharma and A. Joshi, above note 13.

[61] Kate Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press, New York, 2021, pp. 133 – 135.

[62] Batya Friedman and Helen Nissenbaum, “Bias in Computer Systems”, *ACM Transactions on Information Systems*, Vol. 14, No. 3, 1996; James Zou and Londa Schiebinger, “AI Can Be Sexist and Racist - It’s Time to Make It Fair”, *Nature*, Vol. 559, 2018; Harini Suresh and John V. Guttag, “A Framework for Understanding Unintended Consequences of Machine Learning”, 2020, available at: https://arxiv.org/pdf/1901.10002.pdf.

[63] 《世界人权宣言》，1948 年 12 月 10 日，第 2 条和第 7 条；《公民权利和政治权利国际公约》，1966 年 12 月 16 日，第 26 条；《欧洲人权公约》，1950 年 11 月 4 日，第 14 条；《美洲人权公约》，1969 年 11 月 22 日，第 1 条；《非洲人权和民族权利宪章》，1981 年 6 月 27 日，第 2 条。See also Rachel Murray and Frans Viljoen, “Towards Non-Discrimination on the Basis of Sexual Orientation: The Normative Basis and Procedural Possibilities before the African Commission on Human and Peoples’ Rights and the African Union”, *Human Rights Quarterly*, Vol. 29, No. 1, 2007; 人权理事会，《基于性取向和性别认同对个人的歧视性法律、做法和暴力行为——联合国人权事务高级专员的报告》，联合国第 A/HRC/19/41 号文件，2011 年 11 月 17 日；人权理事会，《防止基于性取向和性别认同的暴力和歧视》，联合国第 A/HRC/RES/32/2 号文件，2016 年 7 月 15 日。

[64] Noel Sharkey, “The Impact of Gender and Race Bias in AI”, *Humanitarian Law and Policy Blog*, 28 August 2018, available at: https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/; UN Secretary-General’s High-Level Panel on Digital Cooperation, *The Age of Digital Interdependence*, New York, 2019, available at: www.un.org/en/pdfs/DigitalCooperation-report-for%20web.pdf; 联合国大会，《当代形式种族主义、种族歧视、仇外心理和相关不容忍行为特别报告员（滕达依 • 阿丘梅）的报告》，联合国第 A/75/590 号文件，2020 年 11 月 10 日。

[65] H. Suresh and J. V. Guttag, above note 62.

[66] J. Zou and L. Schiebinger, above note 62.

是训练数据集缺乏多样性。[67]同样，研究人员表明，当残疾人使用轮椅等辅助技术时，人工智能算法更难识别这些个体。[68]

在这方面，有偏见的人工智能系统可能不会被检测到，并继续支持可能导致歧视性结果的决定。[69]这在一定程度上是由于某些机器学习和深度学习算法的不透明性，即所谓的“黑箱问题”。[70]此外，基于深度学习技术的人工智能系统的复杂性导致其设计者和开发者往往无法理解和充分解释机器是如何做出某些决定的。这反过来可能会使识别算法中的偏见变得更具挑战性。

在人道背景下，使用有偏见的人工智能系统的后果可能很严重。例如，在面部识别技术是身份识别和身份核验的唯一手段的情况下，这种系统的不准确性可能导致对肤色较深的个体进行错误识别。如果以这些方式进行身份识别和身份核验是获得人道援助的先决条件，那么错误识别可能导致个体无法得到援助。如果用于分类的系统错误地显示某个人已经接受了有关的援助（如紧急食品供应或医疗服务），这种情况就可能发生，从而将对受影响的个体产生严重后果。如果知道人工智能系统的风险而不加以解决，就可能导致非法的种族歧视。这也可能违背人道的原则，根据这一原则，须努力应对人们的疾苦，不论这种痛苦发生在什么地方。[71]

因此，必须制定保障措施，以确保用于支持人道工作的人工智能系统不会变成排斥需要援助的个体或群体的工具。例如，如果网上关于战争中儿童的照片倾向于显示大量有色人种儿童持有武器（即作为儿童兵），而将具有白人种族背景的儿童描述为受害者，那么根据这类数据集训练的人工智能算法可能会继续延续这种区分。这反过来又可能助长人道行动中对有色人种儿童的现有偏见，加剧武装冲突已经造成的痛苦。因此，对这种类型的偏见的认识和控制应贯穿于将在人道背景下使用的人工智能系统的设计和开发中。另一个例子涉及面部识别技术——只要这些技术在识别肤色较深的人方面仍然不准确，它们就不应被用于协助在决定人道援助提供方面至关重要的决策。

## 数据隐私

正如国际上所同意的那样，“人们在互联网下所享有的权利在互联网上同样应该得到保护”。[72]这一点应该适用于人工智能系统。

国际人权法律文件承认隐私权。[73]此外，具体的法律制度，如《通用数据保护条例》（以下简称《条例》），建立了保护个人数据的基本标准。虽然该《条例》是一项欧盟法律制度，并不约束全球所有人道参与方，但它在欧盟以外仍具有重要意义，因为它在全球范围内激发了类似规范的产生。[74]

---

[67] Joy Buolamwini and Timmit Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, *Proceedings of Machine Learning Research: Conference on Fairness, Accountability and Transparency*, Vol. 81, 2018. But see Stewart Baker, “The Flawed Claims about Bias in Facial Recognition”, *Lawfare*, 2 February 2022, available at: www.lawfareblog.com/flawed-claims-about-bias-facial-recognition.

[68] Meredith Whittaker *et al.*, *Disability, Bias and AI*, AI Now Institute, New York University, 2019, pp. 9-10.

[69] M. Pizzi, M. Romanoff and T. Engelhardt, above note 2.

[70] “黑箱问题”出现于用户和第三方无法看到人工智能系统的运行的时候。See Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*, Harvard University Press, Cambridge, MA, 2016.

[71] 联合国大会，第 46/182 号决议，1991 年 12 月 19 日；《国际红十字与红新月运动章程》，第 25 届红十字与红新月国际大会通过，日内瓦，1986 年（于 1995 年和 2006 年进行了修订），序言。

[72] 联合国大会，第 68/167 号决议，2014 年 1 月 21 日，第 2 段；人权理事会，《在互联网上增进、保护和享有人权》，联合国第 A/HRC/20/L.13 号文件，2012 年 6 月 29 日，第 1 段。

[73] 《世界人权宣言》第 12 条；《公民权利和政治权利国际公约》第 17 条；《欧洲人权公约》第 8 条；《美洲人权公约》第 11 条。

[74] 根据《条例》第 45 条，欧盟委员会可以发布充分性决定（adequacy decision），承认第三国的国内法提供了与《条例》基本相同的充分的数据保护水平。这种决定的结果是，数据流可以继续，而不需要进一步的保护措施。迄今为止，欧盟委员会已经发布了关于安道尔、阿根廷、加拿大（商业组织）、法罗群岛、根西岛、以色列、马恩岛、日本、泽西岛、新西兰、韩国、瑞士、英国和乌拉圭的充分性决定。See European Commission, “Adequacy Decisions”, available at: https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en.

《人道行动中的数据保护手册》（*Handbook on Data Protection in Humanitarian Action*）也考虑到了《条例》中规定的原则，[75]该手册被认为是为人道背景下的个人数据处理设定最低标准的主要资源。这些原则包括个人数据处理中的合法性、公平性和透明度（《条例》第 5 条）。

个人数据处理有合法依据是一项法律要求（《条例》第 6 条）。在人道背景下，同意往往被用作个人数据处理的合法依据。根据法律标准，同意必须是充分知情的、具体的、明确的和自由的（《条例》第 4 条第 11 款）。但是，在人道背景下，由于人道组织和人道援助受益人之间所固有的权力不平衡性，同意可能不是完全明确的和自由的。拒绝同意收集和处理个人数据，实际上可能会导致被拒绝给予人道援助。[76]然而，由于语言上的障碍以及行政和制度上的复杂性，人道参与方可能难以确保人道援助的受益人实际上理解了同意的含义。

鉴于人工智能系统经常使用数据来进一步完善和开发其他人工智能解决方案，充分知情、具体、明确和自由的同意也可能很难实现。虽然个人可能同意将其个人信息用于与人道行动相关的特定目的，但他们可能不知道或不同意这些数据日后被用于开发其他人工智能系统。[77]对“监控人道主义”的批评进一步加剧了这种担忧，因为人道工作者收集越来越多的数据和使用越来越多的技术可能无意中增加需要援助者的脆弱性。[78]

由于科技公司和人道组织之间的合作越来越普遍，这些做法需要受到更多的审查。[79]这些公司在这一领域发挥着核心作用，因为它们设计和开发了人道工作者之后将在一线使用的人工智能系统。可以说，科技公司的利益和世界观往往主要反映在人工智能系统的设计和开发上，从而忽视了用户的需求和体验。[80]这尤其关系到在人道背景下使用人工智能系统的问题，因为受冲突或危机影响的群体会面临很大风险。因此，有必要为在人道背景下应用人工智能制定一套明确的指导方针，特别是将“不伤害”的人道要求置于核心位置，这将在下一节中讨论。

## 为人道行动服务的人工智能：“不伤害”的人道要求

如前所述，虽然人工智能可能为促进人道行动带来新的机会，但它在人道背景下使用时也会带来重大风险。本节阐述了“不伤害”的人道要求，并就如何使人工智能支持人道行动而不损害受冲突和危机影响的群体提出了建议。

### 人工智能时代的“不伤害”

面对不断发展的人工智能技术，至关重要的是，人道工作者必须将“不伤害”作为人道行动中所有人工智能系统使用时的重中之重。不伤害原则长期以来被认为是生物伦理学的核心原则之一。

[75] C. Kuner and M. Marelli, above note 58, p. 23.
[76] M. Madianou, above note 21, p. 9; M. Pizzi, M. Romanoff and T. Engelhardt, above note 2, p. 152.
[77] C. Kuner and M. Marelli, above note 58, p. 284; Meg Leta Jones and Elizabeth Edenberg, “Troubleshooting AI and Consent”, in Markus D. Dubber, Frank Pasquale and Sunit Das (eds), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford, 2020, p. 366.
[78] M. Latonero, above note 19； P. Molnar, above note 20; Pierrick Devidal, “Cashless Cash: Financial Inclusion or Surveillance Humanitarianism?”, *Humanitarian Law and Policy Blog*, 2 March 2021, available at: https://blogs.icrc.org/law-and-policy/2021/03/02/cashless-cash/.
[79] M. Pizzi, M. Romanoff and T. Engelhardt, above note 2; Linda Kinstler, “Big Tech Firms Are Racing to Track Climate Refugees”, *MIT Technology Review*, 17 May 2019, available at: www.technologyreview.com/2019/05/17/103059/big-tech-firms-are-racing-to-track-climate-refugees/.
[80] Ziv Carmon, Rom Schrift, Klaus Wertenbroch and Haiyang Yang, “Designing AI Systems that Customers Won’t Hate”, *MIT Sloan Management Review*, 16 December 2019, available at: https://sloanreview.mit.edu/article/designing-ai-systems-that-customers-wont-hate/.

[81]它最初是由玛丽·安德森（Mary Anderson）在人道领域提出的；[82]随后，各人道组织进一步发展了其应用。[83]今天，这一原则在技术伦理学和人工智能领域也被普遍引用。[84]

“不伤害”原则要求人道参与方考虑其作为或不作为可能在无意中给他们计划提供服务的群体造成伤害或带来新风险的潜在方式。[85]例如，人道“创新”可能会给本已处于脆弱境况的群体带来不必要的风险：新引进的系统出现技术故障导致援助派发的延迟、中断或取消，就是其中一种风险。[86]因此，避免或防止伤害和减轻风险是这一人道要求的核心。

风险分析和影响评估可用于落实“不伤害”原则。风险分析有助于查明人道行动所产生的潜在风险，并为减轻风险提供明确的途径。影响评估可以提供确定具体人道方案负面影响的手段，以及避免或预防伤害的最佳方式。当人道组织预想将人工智能技术用于人道行动时，这些程序可能会帮助它们。有时，这些程序甚至可能得出这样的结论：在特定的情况下不应使用任何技术，因为这些技术对其受益人造成的伤害大于好处。在某些情况下，一项技术的存在并不意味着也必须使用它。

人工智能技术存在一些众所周知的风险，在将人工智能系统用于人道行动之前，人道参与方应该解决这些风险。例如，使用数据驱动的人工智能系统的人道组织应识别可能导致其工作人员及其受益人敏感信息泄露的数据安全漏洞风险。他们还应评估使用人工智能系统是否会对受影响群体产生负面影响——例如，在描绘冲突演变的情况时暴露他们的位置，从而在无意中使他们受到迫害。总之，人工智能系统的使用绝不应给受影响群体带来额外的伤害或风险。

因此，人道参与方不得过度依赖人工智能技术，特别是那些在某些情况下仍然不够准确的技术，如面部识别技术。[87]在采用人工智能系统之前，人道参与方应评估是否有必要在一线使用这些技术、这些技术是否增加了有关人道方案的价值以及这是否能够以保护弱势群体免受更多伤害的方式实现。

## 避免和减轻数据隐私损害的机制

在数字时代，避免或减轻伤害也需要保护数据隐私。数据隐私应该在人工智能从设计、开发再到实施的整个生命周期中都得到保护和尊重。

在这方面，“隐私设计”（privacy by design）原则提供了一个很好的起点。[88]它们基于以用户为中心的方法，提供了一套主动的（而非反应性的）和预防性的（而非补救性的）原则。这些都是构建更好的数据隐私保护的宝贵工具。[89]

---

[81] Tom L. Beauchamp and James F. Childress, *Principles of Biomedical Ethics*, 8th ed., Oxford University Press, Oxford, 2019; Luciano Floridi and Josh Cowls, “A Unified Framework of Five Principles for AI in Society”, *Harvard Data Science Review*, Vol. 1, No. 1, 2019.

[82] Mary B. Anderson, *Do No Harm: How Aid Can Support Peace or War*, Lynne Rienner, Boulder, CO, 1999; Mary B. Anderson, *Options for Aid in Conflict: Lessons from Field Experience*, CDA Collaborative Learning Projects, Cambridge, MA, 2000.

[83] 红十字国际委员会，《红十字国际委员会保护政策》，《红十字国际评论》，第 90 卷，第 871 期，2008 年，第 3 页；Sphere Association, *The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response*, 4th ed., Geneva, 2018.

[84] L. Floridi and J. Cowls, above note 81; Luciano Floridi, *The Ethics of Information*, Oxford University Press, Oxford, 2013; C. Kuner and M. Marelli, above note 58.

[85] Sphere Association， above note 83, p. 268.

[86] Kristin Bergtora Sandvik, Katja Lindskov Jacobsen and Sean Martin McDonald, “Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation”, *International Review of the Red Cross*, Vol. 99, No. 1, 2017.

[87] Davide Castelvecchi, “Is Facial Recognition too Biased to Be Let Loose?”, *Nature*, Vol. 587, 2020.

[88] 这些原则是由安·卡沃乌基扬（Ann Cavoukian）在 2010 年提出的，她当时担任加拿大安大略省的信息和隐私专员。See Ann Cavoukian, “Privacy by Design: The 7 Foundational Principles”, Toronto, 2010, available at: www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf. 这些原则后来得到了国际数据保护和隐私专员会议（International Conference of Data Protection and Privacy Commissioners）的认可。See “Resolution on Privacy by Design”, 32nd International Conference of Data Protection and Privacy Commissioners, Jerusalem, 27-29 October 2010, available at: http://globalprivacyassembly.org/wp-content/uploads/2015/02/32-Conference-Israel-resolution-on-Privacy-by-Design.pdf. See also Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, Washington*, DC, 2012, available at: www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

[89] Lina Jasmontaite, Irene Kamara, Gabriela Zanfir-Fortuna and Stefano Leucci, “Data Protection by Design and by Default:

对于受欧盟法律管辖的人道组织，《条例》第 25 条对在设计和默认情况下的数据保护提出了更为全面的要求。[90]该条款要求实施适当的技术和组织措施，以将（《条例》第 5 条中列举的）核心数据保护原则纳入个人数据处理系统的设计和开发中。如前所述，这些核心原则是合法性、公平性和透明，以及目的限制（purpose limitation）、数据最小化（data minimization）、准确性（accuracy）、存储限制（storage limitation）、完整性（integrity）和保密性（confidentiality）以及可归责性（accountability）。这也符合红十字国际委员会提出的基本数据保护原则。[91]

因此，设计人工智能解决方案或从私营部门供应商处采购人工智能系统的人道组织应确保这些人工智能系统是在设计和默认情况下实施数据保护的。例如，它们应确保在处理个人信息时获得同意，或依靠如数据主体或另一人的重大利益、公共利益、合法利益、合同的履行或法律义务遵守等其他法律依据进行处理。[92]同样，数据收集应保持在所需的最低限度，存储应是网络安全的（cyber-secure），个人数据一旦不再需要，就应将其销毁，而且个人信息应仅用于最初收集时的目的。

此外，开展数据保护影响评估也可能有助于人道参与方了解在人道方案中使用人工智能技术的潜在负面影响。影响评估是一种识别个人数据隐私保护风险以及减轻这些风险的方法的程序。[93]如果损害个人权利和自由的风险很高，受《条例》管辖的人道组织有义务在处理数据之前进行一次影响评估（《条例》第 35 条第 1 款）。即使该组织在法律上没有义务开展评估，这一程序仍可增加人道项目的价值。评估可以有助于提供一个清晰的路线图，以确定有关数据驱动人工智能系统的风险、解决方案和建议。

例如，影响评估可以用来确定被用于训练人工智能算法的匿名数据可能被重新识别从而重新成为个人信息并引起数据保护法律制度适用的情况。重新识别（re-identification）发生于最初匿名化的数据被去匿名化的时候。当不同来源的信息被用来匹配以识别原本匿名的数据集中的个体时，这种情况就会发生。例如，一项研究发现，有可能通过信息匹配从一份包含 50 万名网飞订阅用户的匿名电影评分列表中识别出个人，并同时了解他们明显的政治倾向和其他潜在敏感信息。[94]总之，研究表明，在某些情况下，即使数据集最初是匿名的，个体仍有超过 99%的机会被重新识别。[95]

在人道背景下，匿名化可能不足以防止对弱势群体身份的重新识别，并且如果不能以网络安全的方式保留信息，这些群体就有可能遭受迫害和伤害。影响评估则可有助于确定其他解决方案和组织措施以防止重新识别的发生。

### 透明度、可归责性和补救措施

“不伤害”的原则还意味着，人道参与方应考虑建立一个总体框架，以确保人道行动中人工智能使用所亟需的透明度和可归责性。

这里使用“透明度”一词是为了表明人道参与方应该就他们是否以及如何在人道行动中使用人工智能系统进行沟通。他们应该披露有关他们使用的系统的信息，即使在这些系统的工作方式不能得到完全解释的情况下。从这个意义上讲，透明度是一个比人工智能系统的可解释性（explainability）的狭义概念更广泛的概念。[96]

---

Framing Guiding Principles into Legal Obligations in the GDPR”, *European Data Protection Law Review*, Vol. 4, No. 2, 2018; Giovanni Buttarelli, *Opinion 5/2018: Preliminary Opinion on Privacy by Design*, 31 May 2018, available at: https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf.

[90] Lee Bygrave, “Data Protection by Design and by Default: Deciphering the EU’s Legislative Requirements”, *Oslo Law Review*, Vol. 4, No. 2, 2017.

[91] C. Kuner and M. Marelli, above note 58.

[92] *Ibid*., p. 60.

[93] *Ibid*., p. 84.

[94] Arvind Narayanan and Vitaly Shmatikov, “Robust De-anonymization of Large Sparse Datasets”, *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 18-22 May 2008.

[95] Luc Rocher, Julien M. Hendrickx and Yves-Alexandre de Montjoye, “Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models”, *Nature Communications*, Vol. 10, No. 1, 2019.

[96] Stefan Larsson and Fredrik Heintz, “Transparency in Artificial Intelligence”, *Internet Policy Review*, Vol. 9, No. 2, 2020.

例如，考虑这样一种情况：人工智能系统被用来对难民进行生物特征身份核验（biometric identity verification），作为在难民营派发援助的条件。[97]在这种情况下，使用这种人工智能系统的人道参与方应当告知难民他们正在这样做。同样重要的是，他们应向这些难民披露他们是如何使用人工智能系统的，以及其所带来的影响。例如，他们应该披露将收集什么类型的信息、出于什么样的目的、数据将储存多长时间以及谁将访问这些信息。同样，他们应就将采取哪些保障措施来避免网络安全漏洞进行沟通。

可归责性被理解为要求某人对自己的作为或不作为负责的行为。[98]这是一种以评估一个人或一个实体的作为或不作为是否必要或合理，以及该人或该实体是否可能对其作为或不作为的后果承担法律责任或义务为目的的过程。[99]可归责性也是一种涉及解释和合理化其行为的义务的机制。[100]

在人道背景下，可归责性应体现在人道参与方及其受益人之间的关系中——特别是当人工智能系统被用于支持人道行动时，因为这些技术可能给他们的人权带来风险。例如，人道参与方应告知其受益人任何可能暴露受益人个人信息的数据安全漏洞，并说明为补救这种情况所采取的措施。红十字国际委员会最近对一个数据泄露事件作出的迅速反应，树立了这一领域良好实践的榜样。该机构以直接和全面的努力解释其所采取的行动，并向世界各地的受影响社区通报这一网络安全事件的后果。[101]

最后，如果决策对个体的权利产生了不利影响，那么个体应该能够对自动化决策或人工智能系统支持下的人类决策提出质疑。[102]因此，司法或非司法的申诉机制都可以提供获得补救的法律途径，特别是在人道援助的受益人受到无意伤害的情况下。行政申诉或其他争议解决方案等非司法机制可能有助于那些或无力支付司法诉讼费用的个体。

## 结论

数据驱动的人工智能技术正在逐步改变人道领域。它们有潜力支持人道参与方，因其实现了从反应性到预测性的人道行动方式的范式转变。因此，人工智能可能有助于人道行动的三个主要方面：准备、应对和恢复。

人工智能技术可以支持人道准备工作。它们可以通过快速分析大量的多维数据，识别数据中的模式，进行推断，并在危机或人道灾难发生之前提供关于潜在风险的重要判断。人工智能技术还可以为支持有效的人道救济应对行动和促进恢复方案提供机会，特别是在旷日持久的冲突局势中。

人道组织目前正在使用和测试若干基于人工智能的举措。其中包括使用用于预测人口流动、绘制受人道危机影响地区的地图和查明失踪人员的人工智能系统，从而为一线的人道行动提供信息和便利。然而，使用这些系统并非没有风险。本文分析了三大关切：用于训练人工智能算法的数据质量、贯穿于人工智能系统设计和开发中的算法偏见以及对数据隐私的尊重和保护。

虽然这些问题不是人道领域所独有的，但它们可能会严重影响因冲突或危机而本已处于脆弱境况的群体。因此，如果不希望以牺牲人道为代价使用人工智能系统，那么至关重要的是，人道参与方在实施这些技术时必须符合“不伤害”的人道要求。风险分析和影响评估可能有助于落实“不伤

---

[97] 生物特征是指“现代统计方法在生物学中的应用”，涉及“为进行生物特征识别可以从中提取出可区分、可重复的生物点的个体的”生物特征或“生物和行为特征”，如指纹、虹膜特征和面部特征。International Organization for Standardization, “Information Technology - Biometrics - Overview and Application”, ISO/IEC TR 24741:2018, 2018, available at: www.iso.org/obp/ui/#iso:std:iso-iec:tr:24741:ed-2:v1:en.

[98] Richard Mulgan, “‘Accountability’: An Ever Expanding Concept?”, *Public Administration*, Vol. 78, No. 3, 2000.

[99] Ivo Giesen and François G. H. Kristen, “Liability, Responsibility and Accountability: Crossing Borders”, *Utrecht Law Review*, Vol. 10, No. 3, 2014, p. 6.

[100] Mark Bovens, “Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism”, *West European Politics*, Vol. 33, No. 5, 2010, p. 951.

[101] 红十字国际委员会，《网络安全事件：我会受到何种影响？》，2022 年 2 月 7 日，载：www.icrc.org/zh/document/cyber-security-how-it-affect-me; 红十字国际委员会，《红十字国际委员会遭到网络攻击：分享我们的分析结果》，2022 年 2 月 16 日，载：www.icrc.org/zh/document/icrc-cyber-attack-analysis。

[102] M. Pizzi, M. Romanoff and T. Engelhardt, above note 2, p. 179.

害”的要求。这两个程序对于减少风险和最大限度地减少或避免对受影响群体的负面影响可能都是有价值的。

在诸如目前蹂躏乌克兰、导致欧洲 400 多万人流离失所的武装冲突局势中，“不伤害”的要求尤为重要。[103]在这种情况下，人工智能技术可以在战场内外以有益和有害的方式被使用。例如，人工智能可用以支持分析社交媒体数据和评估信息的准确性，[104]但它也可用以支持使用深度伪造技术制作虚假视频，从而助长虚假信息运动。[105]

因为人工智能系统本身并不是中立的，所以由于其使用方式，其可能会给本已处于脆弱境况的群体带来新的、不必要的风险。例如，人工智能驱动的聊天机器人可以帮助简化签证申请程序，以应对逃离冲突的大规模人员流动，[106]但如果这些系统在没有适当监督的情况下使用，它们可能会使个体的个人信息面临不必要的网络安全风险和潜在的数据泄露。因此，为了让人工智能服务于人道行动，在充分利用其益处的同时抵消其风险，人道组织应该意识到，没有现成的、适用于所有情况的“一刀切”的人工智能解决方案。其还应当评估在某些情况下是否应当使用人工智能系统，因为这种系统对其受益人造成的伤害可能大于其益处。在某些情况下，一项技术的存在并不意味着也必须使用它。

最后，在使用这些技术时，人道组织必须建立适当的框架，以促进在人道背景下使用人工智能的可归责性和透明度。总之，这种机制将有助于实现在人道行动中负责任地利用人工智能潜力的目标。

[103] See International Organization for Migration, “IOM Ukraine Situation Reports”, available at: www.iom.int/resources/iom-ukraine-situation-reports.
[104] Craig Nazareth, “Technology Is Revolutionizing How Intelligence Is Gathered and Analyzed - and Opening a Window onto Russian Military Activity around Ukraine”, *The Conversation*, 14 February 2022, available at: https://theconversation.com/technology-is-revolutionizing-how-intelligence-is-gathered-and-analyzed-and-opening-a-window-onto-russian-military-activity-around-ukraine-176446.
[105] Hitoshi Nasu, “Deepfake Technology in the Age of Information Warfare”, *Articles of War*, 1 March 2022, available at: https://lieber.westpoint.edu/deepfake-technology-age-information-warfare/.
[106] A. Beduschi and M. McAuliffe, above note 51.