人工智能伦理问题建议书

序言

联合国教育、科学及文化组织（教科文组织）大会于 2021 年 11 月 9 日至 24 日在巴黎召开第四十一届会议，

认识到人工智能（AI）从正负两方面对社会、环境、生态系统和人类生活包括人类思想具有深刻而动态的影响，部分原因在于人工智能的使用以新的方式影响着人类的思维、互动和决策，并且波及到教育、人文科学、社会科学和自然科学、文化、传播和信息，

忆及教科文组织根据《组织法》，力求通过教育、科学、文化以及传播和信息促进各国间之合作，对和平与安全作出贡献，以增进对正义、法治及所确认之世界人民均享人权与基本自由之普遍尊重，

深信在此提出的建议书，作为以国际法为依据、采用全球方法制定且注重人的尊严和人权以及性别平等、社会和经济正义与发展、身心健康、多样性、互联性、包容性、环境和生态系统保护的准则性文书，可以引导人工智能技术向着负责任的方向发展，

遵循《联合国宪章》的宗旨和原则，

考虑到人工智能技术可以对人类大有助益并惠及所有国家，但也会引发根本性的伦理关切，例如：人工智能技术可能内嵌并加剧偏见，可能导致歧视、不平等、数字鸿沟和排斥，并对文化、社会和生物多样性构成威胁，造成社会或经济鸿沟；算法的工作方式和算法训练数据应具有透明度和可理解性；人工智能技术对于多方面的潜在影响，包括但不限于人的尊严、人权和基本自由、性别平等、民主、社会、经济、政治和文化进程、科学和工程实践、动物福利以及环境和生态系统，

又认识到人工智能技术会加深世界各地国家内部和国家之间现有的鸿沟和不平等，必须维护正义、信任和公平，以便在公平获取人工智能技术、享受这些技术带来的惠益和避免受其负面影响方面不让任何国家和任何人掉队，同时认识到各国国情不同，并尊重一部分人不参与所有技术发展的意愿，

意识到所有国家都正值信息和通信技术及人工智能技术使用的加速期，对于媒体与信息素养的需求日益增长，且数字经济带来了重大的社会、经济和环境挑战以及惠益共享的机会，对于中低收入国家（LMIC）——包括但不限于最不发达国家（LDC）、内陆发展中国家（LLDC）和小岛屿发展中国家（SIDS）而言尤为如此，需要承认、保护和促进本土文化、价值观和知识，以发展可持续的数字经济，

还认识到人工智能技术具备有益于环境和生态系统的潜能，要实现这些惠益，不应忽视而是要去应对其对环境和生态系统的潜在危害和负面影响，

注意到应对风险和伦理关切的努力不应妨碍创新和发展，而是应提供新的机会，激励合乎伦理的研究和创新，使人工智能技术立足于人权和基本自由、价值观和原则以及关于道义和伦理的思考，

又忆及教科文组织大会在 2019 年 11 月第四十届会议上通过了第 40 C/37 号决议，授权总干事“以建议书的形式编制一份关于人工智能伦理问题的国际准则性文书”，提交 2021 年大会第四十一届会议，

认识到人工智能技术的发展需要相应提高数据、媒体与信息素养，并增加获取独立、多元、可信信息来源的机会，包括努力减少错误信息、虚假信息和仇恨言论的风险以及滥用个人数据造成的伤害，

认为关于人工智能技术及其社会影响的规范框架应建立在共识和共同目标的基础上，以国际和国家法律框架、人权和基本自由、伦理、获取数据、信息和知识的需求、研究和创新自由、人类福祉、环境和生态系统福祉为依据，将伦理价值观和原则与同人工智能技术有关的挑战和机遇联系起来，

又认识到伦理价值观和原则可以通过发挥指引作用，帮助制定和实施基于权利的政策措施和法律规范，以期加快技术发展步伐，

又深信全球公认的、充分尊重国际法特别是人权法的人工智能技术伦理标准可以在世界各地制定人工智能相关规范方面发挥关键作用，

铭记《世界人权宣言》（1948 年）；国际人权框架文书，包括《关于难民地位的公约》（1951 年）、《就业和职业歧视公约》（1958 年）、《消除一切形式种族歧视国际公约》（1965 年）、《公民及政治权利国际公约》（1966 年）、《经济社会文化权利国际公约》（1966 年）、《消除对妇女一切形式歧视公约》（1979 年）、《儿童权利公约》（1989 年）和《残疾人权利公约》（2006 年）；《反对教育歧视公约》（1960 年）；《保护和促进文化表现形式多样性公约》（2005 年）；以及其他一切相关国际文书、建议书和宣言，

又注意到《联合国发展权利宣言》（1986 年）；《当代人对后代人的责任宣言》（1997 年）；《世界生物伦理与人权宣言》（2005 年）；《联合国土著人民权利宣言》（2007 年）；2014 年联合国大会关于信息社会世界峰会审查的决议（A/RES/70/125）（2015 年）；联合国大会关于“变革我们的世界：2030 年可持续发展议程”的决议（A/RES/70/1）（2015 年）；《关于保存和获取包括数字遗产在内的文献遗产的建议书》（2015 年）；《与气候变化有关的伦理原则宣言》（2017 年）；《关于科学和科学研究人员的建议书》（2017 年）；互联网普遍性指标（2018 年获得教科文组织国际传播发展计划认可），包括立足人权、开放、人人可及和多利益攸关方参与原则（2015 年获得教科文组织大会认可）；人权理事会关于“数字时代的隐私权”的决议（A/HRC/RES/42/15）（2019 年）；以及人权理事会关于“新兴数字技术与人权”的决议（A/HRC/RES/41/11）（2019 年），

强调必须特别关注中低收入国家，包括但不限于最不发达国家、内陆发展中国家和小岛屿发展中国家，这些国家具备能力，但在人工智能伦理问题辩论中的代表性不足，由此引发了对于地方知识、文化多元化、价值体系以及应对人工智能技术的正负两方面影响需要实现全球公平的要求受到忽视的关切，

又意识到在人工智能技术的伦理和监管方面，目前存在许多国家政策以及由联合国相关实体、政府间组织（包括地区组织）和由私营部门、专业组织、非政府组织和科学界制定的其他框架和倡议，

还深信人工智能技术可以带来重大惠益，但实现这些惠益也会加剧围绕创新产生的矛盾冲突、知识和技术获取不对称（包括使公众参与人工智能相关议题的能力受限的数字和公民素养赤字）以及信息获取障碍、能力亦即人员和机构能力差距、技术创新获取障碍、缺乏适当的实

体和数字基础设施以及监管框架（包括与数据有关的基础设施和监管框架）的问题，所有这些问题都需要解决，

强调需要加强全球合作与团结，包括通过多边主义，以促进公平获取人工智能技术，应对人工智能技术给文化和伦理体系的多样性和互联性带来的挑战，减少可能的滥用，充分发挥人工智能可能给各个领域特别是发展领域带来的潜能，确保各国人工智能战略以伦理原则为指导，

充分考虑到人工智能技术的快速发展对以合乎伦理的方式应用和治理人工智能技术以及对尊重和保护文化多样性提出了挑战，并有可能扰乱地方和地区的伦理标准和价值观，

1. 兹于 2021 年 11 月 23 日**通过**这份《人工智能伦理问题建议书》；

2. **建议**会员国在自愿基础上适用本建议书的各项规定，特别是根据各自国家的宪法实践和治理结构采取适当步骤，包括必要的立法或其他措施，依照包括国际人权法在内的国际法，使建议书的原则和规范在本国管辖范围内生效；

3. **又建议**会员国动员包括工商企业在内的所有利益攸关方，确保他们在实施本建议书方面发挥各自的作用；并提请涉及人工智能技术的管理部门、机构、研究和学术组织、公共、私营和民间社会机构和组织注意本建议书，使人工智能技术的开发和应用做到以健全的科学研究以及伦理分析和评估作为指导。

I. 适用范围

1. 本建议书述及与人工智能领域有关且属于教科文组织职责范围之内的伦理问题。建议书以能指导社会负责任地应对人工智能技术对人类、社会、环境和生态系统产生的已知和未知影响并相互依存的价值观、原则和行动构成的不断发展的整体、全面和多元文化框架为基础，将人工智能伦理作为一种系统性规范考量，并为社会接受或拒绝人工智能技术提供依据。建议书将伦理视为对人工智能技术进行规范性评估和指导的动态基础，以人的尊严、福祉和防止损害为导向，并立足于科技伦理。

2. 本建议书无意对人工智能作出单一的定义，这种定义需要随着技术的发展与时俱进。建议书旨在探讨人工智能系统中具有核心伦理意义的特征。因此，本建议书将人工智能系统

视为有能力以类似于智能行为的方式处理数据和信息的系统，通常包括推理、学习、感知、预测、规划或控制等方面。这种方式有三个重要因素：

(a) 人工智能系统是整合模型和算法的信息处理技术，这些模型和算法能够生成学习和执行认知任务的能力，从而在物质环境和虚拟环境中实现预测和决策等结果。在设计上，人工智能系统借助知识建模和知识表达，通过对数据的利用和对关联性的计算，可以在不同程度上实现自主运行。人工智能系统可以包含若干种方法，包括但不限于：

(i) 机器学习，包括深度学习和强化学习；

(ii) 机器推理，包括规划、调度、知识表达和推理、搜索和优化。

人工智能系统可用于信息物理系统，包括物联网、机器人系统、社交机器人和涉及控制、感知及处理传感器所收集数据的人机交互以及人工智能系统工作环境中执行器的操作。

(b) 与人工智能系统有关的伦理问题涉及人工智能系统生命周期的各个阶段，此处系指从研究、设计、开发到配置和使用等各阶段，包括维护、运行、交易、融资、监测和评估、验证、使用终止、拆卸和终结。此外，人工智能行为者可以定义为在人工智能系统生命周期内至少参与一个阶段的任何行为者，可指自然人和法人，例如研究人员、程序员、工程师、数据科学家、终端用户、工商企业、大学和公私实体等。

(c) 人工智能系统引发了新型伦理问题，包括但不限于其对决策、就业和劳动、社交、卫生保健、教育、媒体、信息获取、数字鸿沟、个人数据和消费者保护、环境、民主、法治、安全和治安、双重用途、人权和基本自由（包括表达自由、隐私和非歧视）的影响。此外，人工智能算法可能复制和加深现有的偏见，从而加剧已有的各种形式歧视、偏见和成见，由此产生新的伦理挑战。其中一些问题与人工智能系统有能力完成此前只有生物才能完成、甚至在有些情况下只有人类才能完成的任务有关。这些特点使得人工智能系统在人类实践和社会中以及在与环境和生态系统的关系中，可以发挥意义深远的新作用，为儿童和青年的成长、培养对于世界和自身的认识、批判性地认识媒体和信息以及学会作出决定创造了新的环境。从长远看，人工智能系统可能挑战人类特有的对于经验和能动作用的感知，

在人类的自我认知、社会、文化和环境的互动、自主性、能动性、价值和尊严等方面引发更多关切。

3. 秉承教科文组织世界科学知识与技术伦理委员会（COMEST）在 2019 年《人工智能伦理问题初步研究》中的分析，本建议书特别关注人工智能系统与教育、科学、文化、传播和信息等教科文组织核心领域有关的广泛伦理影响：

(a) 教育，这是因为鉴于对劳动力市场、就业能力和公民参与的影响，生活在数字化社会需要新的教育实践、伦理反思、批判性思维、负责任的设计实践和新的技能。

(b) 科学，系指最广泛意义上的科学，包括从自然科学、医学到社会科学和人文科学等所有学术领域，这是由于人工智能技术带来了新的研究能力和方法，影响到我们关于科学认识和解释的观念，为决策创建了新的基础。

(c) 文化特性和多样性，这是由于人工智能技术可以丰富文化和创意产业，但也会导致文化内容的供应、数据、市场和收入更多地集中在少数行为者手中，可能对语言、媒体、文化表现形式、参与和平等的多样性和多元化产生负面影响。

(d) 传播和信息，这是由于人工智能技术在处理、组织和提供信息方面起到日益重要的作用；很多现象引发了与信息获取、虚假信息、错误信息、仇恨言论、新型社会叙事兴起、歧视、表达自由、隐私、媒体与信息素养等有关的问题，而自动化新闻、通过算法提供新闻、对社交媒体和搜索引擎上的内容进行审核和策管只是其中几个实例。

4. 本建议书面向会员国，会员国既是人工智能行为者，又是负责制定人工智能系统整个生命周期的法律和监管框架并促进企业责任的管理部门。此外，建议书为贯穿人工智能系统生命周期的伦理影响评估奠定了基础，从而为包括公共和私营部门在内的所有人工智能行为者提供伦理指南。

II. 宗旨和目标

5. 本建议书旨在提供基础，让人工智能系统可以造福人类、个人、社会、环境和生态系统，同时防止危害。它还旨在促进和平利用人工智能系统。

6. 本建议书的目的是在全球现有人工智能伦理框架之外，再提供一部全球公认的准则性文书，不仅注重阐明价值观和原则，而且着力于通过具体的政策建议切实落实这些价值观和原则，同时着重强调包容、性别平等以及环境和生态系统保护等问题。

7. 由于与人工智能有关的伦理问题十分复杂，需要国际、地区和国家各个层面和各个部门的众多利益攸关方开展合作，故而本建议书的宗旨是让利益攸关方能够在全球和文化间对话的基础上共同承担责任。

8. 本建议书的目标如下：

(a) 依照国际法，提供一个由价值观、原则和行动构成的普遍框架，指导各国制定与人工智能有关的法律、政策或其他文书；

(b) 指导个人、团体、社群、机构和私营部门公司的行动，确保将伦理规范嵌入人工智能系统生命周期的各个阶段；

(c) 在人工智能系统生命周期的各个阶段保护、促进和尊重人权和基本自由、人的尊严和平等，包括性别平等；保障当代和后代的利益；保护环境、生物多样性和生态系统；尊重文化多样性；

(d) 针对与人工智能系统有关的伦理问题，推动多利益攸关方、多学科和多元化对话并建立共识；

(e) 促进对人工智能领域进步和知识的公平获取以及惠益共享，特别关注包括最不发达国家、内陆发展中国家和小岛屿发展中国家在内的中低收入国家的需求和贡献。

III. 价值观和原则

9. 首先，人工智能系统生命周期的所有行为者都应尊重下文所载的价值观和原则，并在必要和适当的情况下，通过修订现行的和制定新的法律、法规和业务准则来促进这些价值观和原则。这必须遵守国际法，包括《联合国宪章》和会员国的人权义务，并应符合国际商定的社会、政治、环境、教育、科学和经济可持续性目标，例如联合国可持续发展目标。

10. 价值观作为催人奋进的理想，在制定政策措施和法律规范方面发挥着强大作用。下文概述的一系列价值观可以激发理想的行为并为各项原则奠定基础，而各项原则则更为具体地阐明作为其根本的价值观，以便更易于在政策声明和行动中落实这些价值观。

11. 下文概述的所有价值观和原则本身都是可取的，但在任何实际情况下这些价值观和原则之间都可能会有矛盾。在特定情况下，需要根据具体情况进行评估以管控潜在的矛盾，同时考虑到相称性原则并尊重人权和基本自由。在所有情况下，可能对人权和基本自由施加的任何限制均必须具有合法基础，而且必须合理、必要和相称，符合各国依据国际法所承担的义务。要做到明智而审慎地处理这些情况，通常需要与广泛的相关利益攸关方合作，同时利用社会对话以及伦理审议、尽职调查和影响评估。

12. 人工智能系统生命周期的可信度和完整性，对于确保人工智能技术造福人类、个人、社会、环境和生态系统并且体现出本建议书提出的价值观和原则至关重要。在采取适当措施降低风险时，人们应有充分理由相信人工智能系统能够带来个人利益和共享利益。具有可信度的一个基本必要条件是，人工智能系统在整个生命周期内都受到相关利益攸关方适当的全面监测。由于可信度是本文件所载各项原则得到落实的结果，本建议书提出的政策行动建议均旨在提升人工智能系统生命周期各个阶段的可信度。

III.1. 价值观

尊重、保护和促进人权和基本自由以及人的尊严

13. 每个人与生俱来且不可侵犯的尊严，构成了人权和基本自由这一普遍、不可分割、不可剥夺、相互依存又彼此相关的体系的基础。因此，尊重、保护和促进包括国际人权法在内的国际法确立的人的尊严和权利，在人工智能系统的整个生命周期内都至关重要。人的尊严系指承认每个人固有和平等的价值，无论种族、肤色、血统、性别、年龄、语言、宗教、政治见解、民族、族裔、社会出身、与生俱来的经济或社会条件、残障情况或其他状况如何。

14. 在人工智能系统生命周期的任何阶段，任何人或人类社群在身体、经济、社会、政治、文化或精神等任何方面，都不应受到损害或被迫居于从属地位。在人工智能系统的整个生命周期内，人类生活质量都应得到改善，而“生活质量”的定义只要不侵犯或践踏人权和基本自由或人的尊严，应由个人或群体来决定。

15. 在人工智能系统的整个生命周期内，人会与人工智能系统展开互动，接受这些系统提供的帮助，例如照顾弱势者或处境脆弱群体，包括但不限于儿童、老年人、残障人士或病人。

在这一互动过程中，人绝不应被物化，其尊严不应以其他任何方式受到损害，人权和基本自由也不应受到侵犯或践踏。

16. 在人工智能系统的整个生命周期内，必须尊重、保护和促进人权和基本自由。各国政府、私营部门、民间社会、国际组织、技术界和学术界在介入与人工智能系统生命周期有关的进程时，必须尊重人权文书和框架。新技术应为倡导、捍卫和行使人权提供新手段，而不是侵犯人权。

环境和生态系统蓬勃发展

17. 应在人工智能系统的整个生命周期内确认、保护和促进环境和生态系统的蓬勃发展。此外，环境和生态系统也是关乎人类和其他生物能否享受人工智能进步所带来惠益的必要条件。

18. 参与人工智能系统生命周期的所有行为者都必须遵守适用的国际法以及国内立法、标准和惯例，例如旨在保护和恢复环境和生态系统以及促进可持续发展的预防措施。这些行为者应减少人工智能系统对环境的影响，包括但不限于碳足迹，以确保将气候变化和环境风险因素降到最低，防止会加剧环境恶化和生态系统退化的对自然资源的不可持续开采、使用和转化。

确保多样性和包容性

19. 在人工智能系统的整个生命周期内，应依照包括人权法在内的国际法，确保尊重、保护和促进多样性和包容性。为此，可以促进所有个人或群体的积极参与，无论种族、肤色、血统、性别、年龄、语言、宗教、政治见解、民族、族裔、社会出身、与生俱来的经济或社会条件、残障情况或其他状况如何。

20. 对于生活方式的选择范围、信仰、意见、表达形式或个人经验，包括对于人工智能系统的任选使用以及这些架构的共同设计，在人工智能系统生命周期的任何阶段都不应受到限制。

21. 此外，应作出努力，包括开展国际合作，以弥补而绝非利用某些社区所面临的必要的技术基础设施、教育和技能以及法律框架缺乏的情况，特别是在中低收入国家、最不发达国家、内陆发展中国家和小岛屿发展中国家。

生活在和平、公正与互联的社会中

22. 人工智能行为者应为确保建设和平与公正的社会发挥参与和促进作用，这种社会的根基是惠及全民、符合人权和基本自由的互联的未来。在和平与公正的社会中生活的价值观表明，人工智能系统在整个生命周期内都有可能为所有生物之间及其与自然环境之间的互联作出贡献。

23. 人与人之间互联的概念是基于这样一种认识，即每个人都属于一个更大的整体，当这个整体中的组成部分都能够繁荣兴旺时，整体才会蒸蒸日上。在和平、公正与互联的社会中生活，需要一种有机、直接、出自本能的团结纽带，其特点是不懈地寻求和平关系，倾向于在最广泛的意义上关爱他人和自然环境。

24. 这一价值观要求在人工智能系统的整个生命周期内促进和平、包容与正义、公平和互联，人工智能系统生命周期的各种进程不得隔离或物化人类和社区或者削弱其自由、自主决策和安全，不得分裂个人和群体或使之相互对立，也不得威胁人类、其他生物和自然环境之间的共存。

III.2. 原则

相称性和不损害

25. 应该认识到，人工智能技术本身并不一定能确保人类、环境和生态系统蓬勃发展。况且，与人工智能系统生命周期有关的任何进程都不得超出实现合法目的或目标所需的范围，并应切合具体情况。在有可能对人类、人权和基本自由、个别社区和整个社会，或者对环境和生态系统造成损害时，应确保落实风险评估程序并采取措施，以防止发生此类损害。

26. 应从以下方面证明选择使用人工智能系统和选用哪种人工智能方法的合理性：（a）所选择的人工智能方法对于实现特定合法目标应该是适当的和相称的；（b）所选择的人工智能方法不得违背本文件提出的基本价值观，特别是其使用不得侵犯或践踏人权；（c）人工智能方法应切合具体情况，并应建立在严谨的科学基础上。在所涉决定具有不可逆转或难以逆转的影响或者在涉及生死抉择的情况下，应由人类作出最终决定。人工智能系统尤其不得用于社会评分或大规模监控目的。

安全和安保

27. 在人工智能系统的整个生命周期内，应避免并解决、预防和消除意外伤害（安全风险）以及易受攻击的脆弱性（安保风险），确保人类、环境和生态系统的安全和安保。通过开发可持续和保护隐私的数据获取框架，促进利用优质数据更好地训练和验证人工智能模型，可以实现有安全和安保保障的人工智能。

公平和非歧视

28. 人工智能行为者应根据国际法，促进社会正义并保障一切形式的公平和非歧视。这意味着要采用包容性方法确保人工智能技术的惠益人人可得可及，同时又考虑到不同年龄组、文化体系、不同语言群体、残障人士、女童和妇女以及处境不利、边缘化和弱势群体或处境脆弱群体的具体需求。会员国应努力让包括地方社区在内的所有人都能够获取提供本地相关内容和服务且尊重多语言使用和文化多样性的人工智能系统。会员国应努力消除数字鸿沟，并确保对人工智能发展的包容性获取和参与。在国家层面，会员国应努力在人工智能系统生命周期的准入和参与问题上促进城乡之间的公平，以及所有人之间的公平，无论种族、肤色、血统、性别、年龄、语言、宗教、政治见解、民族、族裔、社会出身、与生俱来的经济或社会条件、残障情况或其他状况如何。在国际层面，技术最先进的国家有责任支持最落后的国家，确保共享人工智能技术的惠益，使得后者能够进入和参与人工智能系统生命周期，从而推动构建在信息、传播、文化、教育、研究、社会经济和政治稳定方面更加公平的世界秩序。

29. 人工智能行为者应尽一切合理努力，在人工智能系统的整个生命周期内尽量减少和避免强化或固化带有歧视性或偏见的应用程序和结果，确保人工智能系统的公平。对于带有歧视性和偏见的算法决定，应提供有效的补救办法。

30. 此外，在人工智能系统的整个生命周期内，需要解决国家内部和国家之间的数字和知识鸿沟，包括根据相关的国家、地区和国际法律框架解决技术和数据获取及获取质量方面的鸿沟，以及在连接性、知识和技能以及受影响社区的切实参与方面的鸿沟，以便让每个人都得到公平对待。

可持续性

31. 可持续社会的发展，有赖于在人类、社会、文化、经济和环境等方面实现一系列复杂的目标。人工智能技术的出现可能有利于可持续性目标，但也可能阻碍这些目标的实现，这取决于处在不同发展水平的国家如何应用人工智能技术。因此，在就人工智能技术对人类、社会、文化、经济和环境的影响开展持续评估时，应充分考虑到人工智能技术对于作为一套涉及多方面的动态目标（例如目前在联合国可持续发展目标中认定的目标）的可持续性的影响。

隐私权和数据保护

32. 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。重要的是，人工智能系统所用数据的收集、使用、共享、归档和删除方式，必须符合国际法，契合本建议书提出的价值观和原则，同时遵守相关的国家、地区和国际法律框架。

33. 应在国家或国际层面采用多利益攸关方办法，建立适当的数据保护框架和治理机制，将其置于司法系统保护之下，并在人工智能系统的整个生命周期内予以保障。数据保护框架和任何相关机制应参鉴有关收集、使用和披露个人数据以及数据主体行使其权利的国际数据保护原则和标准，同时确保对个人数据的处理具有合法的目的和有效的法律依据，包括取得知情同意。

34. 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。

人类的监督和决定

35. 会员国应确保始终有可能将人工智能系统生命周期的任何阶段以及与人工智能系统有关的补救措施的伦理和法律责任归属于自然人或现有法人实体。因此，人类监督不仅指个人监督，在适当情况下也指范围广泛的公共监督。

36. 在某些情况下，出于效率性的考虑，人类有时选择依赖人工智能系统，但是否在有限情形下出让控制权依然要由人类来决定，这是由于人类在决策和行动上可以借助人工智能系

统，但人工智能系统永远无法取代人类的最终责任和问责。一般而言，涉及生死抉择的情况，不应任由人工智能系统决定。

透明度和可解释性

37. 人工智能系统的透明度和可解释性往往是确保人权、基本自由和伦理原则得到尊重、保护和促进的必要先决条件。透明度是相关国家和国际责任制度有效运作的必要因素。缺乏透明度还可能削弱对根据人工智能系统产生的结果所作决定提出有效质疑的可能性，进而可能侵犯获得公平审判和有效补救的权利，并限制这些系统的合法使用领域。

38. 在人工智能系统的整个生命周期内都需要努力提高人工智能系统（包括那些具有域外影响的系统）的透明度和可解释性，以支持民主治理，但透明度和可解释性的程度应始终切合具体情况并与其影响相当，因为可能需要在透明度和可解释性与隐私、安全和安保等其他原则之间取得平衡。在所涉决定系参考或依据人工智能算法作出的情况下，包括在所涉决定关乎民众安全和人权的情况下，民众应充分知情，并且在此类情况下有机会请求相关人工智能行为者或公共部门机构提供解释性信息。此外，对于影响其权利和自由的决定，个人应能够了解据以作出该决定的理由，并可以选择向能够审查和纠正该决定的私营部门公司或公共部门机构指定工作人员提出意见。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。

39. 从社会—技术角度来看，提高透明度有助于建设更加和平、公正、民主和包容的社会。提高透明度有利于开展公众监督，这可以减少腐败和歧视，还有助于发现和防止对人权产生的负面影响。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。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）。在存在会对人权产生不利影响的严重威胁的情况下，透明度要求可能还包括共享代码或数据集。

40. 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。因此，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。

41. 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。

责任和问责

42. 人工智能行为者和会员国应根据国家法律和国际法，特别是会员国的人权义务，以及人工智能系统整个生命周期的伦理准则，包括在涉及其有效疆域和实际控制范围内的人工智能行为者方面，尊重、保护和促进人权和基本自由，并且还应促进对环境和生态系统的保护，同时承担各自的伦理和法律责任。以任何方式基于人工智能系统作出的决定和行动，其伦理责任和义务最终都应由人工智能行为者根据其在人工智能系统生命周期中的作用来承担。

43. 应建立适当的监督、影响评估、审计和尽职调查机制，包括保护举报者，确保在人工智能系统的整个生命周期内对人工智能系统及其影响实施问责。技术和体制方面的设计都应确保人工智能系统（的运行）可审计和可追溯，特别是要应对与人权规范和标准之间的冲突以及对环境和生态系统福祉的威胁。

认识和素养

44. 应通过由政府、政府间组织、民间社会、学术界、媒体、社区领袖和私营部门共同领导并虑及现有的语言、社会和文化多样性的开放且可获取的教育、公民参与、数字技能和人工智能伦理问题培训、媒体与信息素养及培训，促进公众对人工智能技术和数据价值的认识和理解，以确保公众的有效参与，让社会所有成员都能够就使用人工智能系统作出知情决定，避免受到不当影响。

45. 了解人工智能系统的影响，应包括了解、借助以及促进人权和基本自由。这意味着在接触和理解人工智能系统之前，应首先了解人工智能系统对人权和权利获取的影响，以及对环境和生态系统的影响。

多利益攸关方与适应性治理和协作

46. 对数据的使用必须尊重国际法和国家主权。这意味着各国可根据国际法，对在其境内生成或经过其国境的数据进行监管，并采取措施，力争在依照国际法尊重隐私权以及其他人权规范和标准的基础上对数据进行有效监管，包括数据保护。

47. 不同利益攸关方对人工智能系统整个生命周期的参与，是采取包容性方法开展人工智能治理、使惠益能够为所有人共享以及推动可持续发展的必要因素。利益攸关方包括但不限于政府、政府间组织、技术界、民间社会、研究人员和学术界、媒体、教育、政策制定者、私营部门公司、人权机构和平等机构、反歧视监测机构以及青年和儿童团体。应采用开放标准和互操作性原则，以促进协作。应采取措施，兼顾技术的变化和新利益攸关方群体的出现，并便于边缘化群体、社区和个人切实参与，同时酌情尊重土著人民对其数据的自我管理。

IV. 政策行动领域

48. 以下政策领域所述的政策行动，是对本建议书提出的价值观和原则的具体落实。主要行动是会员国出台有效措施，包括政策框架或机制等，并通过开展多种行动，例如鼓励所有利益攸关方根据包括联合国《工商企业与人权指导原则》在内的准则制定人权、法治、民主以及伦理影响评估和尽职调查工具，确保私营部门公司、学术和研究机构以及民间社会等其他利益攸关方遵守这些框架或机制。此类政策或机制的制定过程应包括所有利益攸关方并应考虑到各会员国的具体情况和优先事项。教科文组织可以作为合作伙伴，支持会员国制定、监测和评估政策机制。

49. 教科文组织认识到，各会员国在科学、技术、经济、教育、法律、规范、基础设施、社会、文化和其他方面，处于实施本建议书的不同准备阶段。需要指出的是，这里的“准备”是一种动态。因此，为切实落实本建议书，教科文组织将：（1）制定准备状态评估方法，以协助有关会员国确定其准备进程各个方面在特定时刻所处的状态；（2）确保支持有关会员国制定教科文组织人工智能技术伦理影响评估（EIA）方法，分享最佳做法、评估准则以及其他机制和分析工作。

政策领域 1：伦理影响评估

50. 会员国应出台影响评估（例如伦理影响评估）框架，以确定和评估人工智能系统的惠益、关切和风险，并酌情出台预防、减轻和监测风险的措施以及其他保障机制。此种影响评估应根据本建议书提出的价值观和原则，确定对人权和基本自由（特别是但不限于边缘化和弱势群体或处境脆弱群体的权利、劳工权利）、环境和生态系统产生的影响以及伦理和社会影响，并促进公民参与。

51. 会员国和私营部门公司应建立尽职调查和监督机制，以确定、防止和减轻人工智能系统对尊重人权、法治和包容性社会产生的影响，并说明如何处理这些影响。会员国还应能够评估人工智能系统对贫困问题产生的社会经济影响，确保人工智能技术在目前和未来的大规模应用不会加剧各国之间以及国内的贫富差距和数字鸿沟。为做到这一点，尤其应针对信息（包括私营实体掌握的涉及公共利益的信息）获取，实行可强制执行的透明度协议。会员国、私营部门公司和民间社会应调查基于人工智能的建议对人类决策自主权的社会学和心理学影响。对于经确认对人权构成潜在风险的人工智能系统，在投放市场之前，作为伦理影响评估的一部分，人工智能行为者应对其进行广泛测试，包括必要时在真实世界的条件下进行测试。

52. 会员国和工商企业应采取适当措施，监测人工智能系统生命周期的各个阶段，包括用于决策的算法的性能、数据以及参与这一过程的人工智能行为者，特别是在公共服务领域和需要与终端用户直接互动的领域，以配合开展伦理影响评估。人工智能系统评估的伦理方面应包含会员国的人权法义务。

53. 各国政府应采用监管框架，其中特别针对公共管理部门提出人工智能系统伦理影响评估程序，以预测后果、减少风险、避免有害后果、促进公民参与和应对社会挑战。评估还应确立能够对算法、数据和设计流程加以评估并包括对人工智能系统的外部审查的适当监督机制，包括确定可审计性、可追溯性和可解释性。伦理影响评估应透明，并酌情向公众开放。此类评估还应具备多学科、多利益攸关方、多文化、多元化和包容等特性。应要求公共管理部门通过引入适当的机制和工具，监测这些部门实施和/或部署的人工智能系统。

政策领域 2：伦理治理和管理

54. 会员国应确保人工智能治理机制具备包容、透明、多学科、多边（包括跨界减轻损害和作出补救的可能性）和多利益攸关方等特性。特别是，治理应包括预测、有效保护、监测影响、执行和补救等方面。

55. 会员国应通过实施有力的执行机制和补救行动，确保人工智能系统造成的损害得到调查和补救，从而确保人权和基本自由以及法治在数字世界与现实世界中同样得到尊重。此类机制和行动应包括私营部门公司和公共部门公司提供的补救机制。为此，应提升人工智能系统的可审计性和可追溯性。此外，会员国应加强履行这项承诺的机构能力，并应与研究人员和其他利益攸关方合作调查、防止并减少对于人工智能系统的潜在恶意使用。

56. 鼓励会员国根据应用领域的敏感程度、对人权、环境和生态系统的预期影响以及本建议书提出的其他伦理考量，制定国家和地区人工智能战略，并考虑多种形式的柔性治理，例如人工智能系统认证机制和此类认证的相互承认。此类机制可以包括针对系统、数据以及伦理准则和伦理方面的程序要求的遵守情况开展不同层面的审计。同时，此类机制不得因行政负担过重而妨碍创新，或者让中小企业或初创企业、民间社会以及研究和科学组织处于不利地位。这些机制还应包括定期监测，以便在人工智能系统的整个生命周期内确保系统的稳健性、持续完整性和遵守伦理准则，必要时可要求重新认证。

57. 会员国和公共管理部门应对现有和拟议的人工智能系统进行透明的自我评估，其中尤其应包括对采用人工智能是否适当进行评估，如果适当则应为确定适当的方法开展进一步评估，并评估采用这种方法是否会导致违反或滥用会员国的人权法义务，如果会导致则应禁止采用。

58. 会员国应鼓励公共实体、私营部门公司和民间社会组织让不同利益攸关方参与其人工智能治理工作，并考虑增设一个独立的人工智能伦理干事岗位或某种其他机制，负责监督伦理影响评估、审计和持续监测工作，确保对于人工智能系统的伦理指导。鼓励会员国、私营部门公司和民间社会组织在教科文组织的支持下，创设独立的人工智能伦理干事网络，以便在国家、地区和国际层面为这一进程给予支持。

59. 会员国应促进数字生态系统的发展和获取，以便在国家层面以合乎伦理和包容各方的方式发展人工智能系统，包括消除在人工智能系统生命周期准入方面的差距，同时推动国际合作。此类生态系统尤其包括数字技术和基础设施，在适当情况下还包括人工智能知识共享机制。

60. 会员国应与国际组织、跨国公司、学术机构和民间社会合作建立机制，以确保所有会员国积极参与关于人工智能治理的国际讨论，特别是中低收入国家，尤其是最不发达国家、内陆发展中国家和小岛屿发展中国家。可以通过提供资金、确保平等的地区参与或任何其他机制来实现这一目标。此外，为确保人工智能论坛的包容性，会员国应为人工智能行为者的出入境提供便利，特别是中低收入国家，尤其是最不发达国家、内陆发展中国家和小岛屿发展中国家的行为者，以便其参加这些论坛。

61.修正现行的或制定新的有关人工智能系统的国家立法，必须遵守会员国的人权法义务，并在人工智能系统的整个生命周期内促进人权和基本自由。随着人工智能技术的发展，还应采取以下形式促进人权和基本自由：治理举措；关于人工智能系统的合作实践的良好范例；国家及国际技术和方法准则。包括私营部门在内的各个部门在其关于人工智能系统的实践中必须利用现有的和新的文书以及本建议书，尊重、保护和促进人权和基本自由。

62. 为人权敏感用途（例如执法、福利、就业、媒体和信息提供者、卫生保健和独立司法系统等）配置人工智能系统的会员国应提供机制，由独立的数据保护机关、行业监督机构和负责监督的公共机构等适当监督部门监测人工智能系统的社会和经济影响。

63. 会员国应增强司法机构根据法治以及国际法和国际标准作出与人工智能系统有关决定（包括在其审议中使用人工智能系统的决定）的能力，同时确保坚持人类监督原则。司法机关如若使用人工智能系统，则需要有足够的保障措施，以便尤其确保对基本人权的保护、法治、司法独立以及人类监督原则，并确保司法机关对人工智能系统的开发和使用值得信赖、以公共利益为导向且以人为本。

64. 会员国应确保政府和多边组织在保障人工智能系统的安全和安保方面发挥主导作用，并吸收多利益攸关方参与其中。具体而言，会员国、国际组织和其他相关机构应制定国际标准，列出可衡量及可检测的安全和透明度等级，以便能够客观评估人工智能系统并确定合规水平。此外，会员国和工商企业应对人工智能技术潜在安全和安保风险的战略研究提供持续

支持，并应鼓励对透明度、可解释性、包容和素养问题开展研究，在不同方面和不同层面（例如技术语言和自然语言）为这些领域投入更多资金。

65. 会员国应实施政策，在人工智能系统的整个生命周期内确保人工智能行为者的行动符合国际人权法、标准和原则，同时充分考虑到当前的文化和社会多样性，包括地方习俗和宗教传统，并适当考虑到人权的优先性和普遍性。

66. 会员国应建立机制，要求人工智能行为者披露并打击人工智能系统的结果和数据中任何类型的陈规定型观念，无论是设计使然还是出于疏忽，确保人工智能系统的训练数据集不会助长文化、经济或社会不平等和偏见，不会散播虚假信息和错误信息，也不会干扰表达自由和信息获取。应特别关注数据匮乏地区。

67. 会员国应实施政策，促进并提高人工智能开发团队和训练数据集的多样性和包容性，以反映其人口状况，确保人工智能技术及其惠益的平等获取，特别是对农村和城市地区的边缘化群体而言。

68. 会员国应酌情制定、审查并调整监管框架，在人工智能系统生命周期的不同阶段对其内容和结果实施问责制和责任制。会员国应在必要时出台责任框架或澄清对现有框架的解释，确保为人工智能系统的结果和性能确定责任归属。此外，会员国在制定监管框架时，应特别考虑到最终责任和问责必须总是落实到自然人或法人身上，且人工智能系统本身不应被赋予法人资格。为确保这一点，此类监管框架应符合人类监督原则，并确立着眼于人工智能系统生命周期不同阶段的人工智能行为者和技术流程的综合性方法。

69. 为在空白领域确立规范或调整现有的法律框架，会员国应让所有人工智能行为者（包括但不限于研究人员、民间社会和执法部门的代表、保险公司、投资者、制造商、工程师、律师和用户）参与其中。这些规范可以发展成为最佳做法、法律和法规。进一步鼓励会员国采用政策原型和监管沙箱等机制，以便加快制定与新技术的飞速发展相适应的法律、法规和政策，包括对其进行定期审查，确保法律法规在正式通过之前能够在安全环境下进行测试。会员国应支持地方政府制定符合国家和国际法律框架的地方政策、法规和法律。

70. 会员国应对人工智能系统的透明度和可解释性提出明确要求，以协助确保人工智能系统整个生命周期的可信度。此类要求应包括影响机制的设计和实施，其中要考虑到每个特定人工智能系统的应用领域的性质、预期用途、目标受众和可行性。

政策领域 3：数据政策

71. 会员国应努力制定数据治理战略，确保对人工智能系统训练数据的质量进行持续评估，包括数据收集和选择过程的充分性、适当的数据安全和保护措施以及从错误中学习和在所有人工智能行为者之间分享最佳做法的反馈机制。

72. 会员国应采取适当的保障措施，根据国际法保护隐私权，包括应对人们对于监控等问题的关切。会员国尤其应通过或实施可以提供适当保护并符合国际法的法律框架。会员国应大力鼓励包括工商企业在内的所有人工智能行为者遵守现行国际标准，特别是在伦理影响评估中开展适当的隐私影响评估，其中要考虑到预期数据处理产生的更广泛的社会经济影响，并从其系统设计开始即实施保护隐私原则。在人工智能系统的整个生命周期内应尊重、保护和促进隐私。

73. 会员国应确保个人可以保留对于其个人数据的权利并得到相关框架的保护，此类框架尤其应预见到以下问题：透明度；对于处理敏感数据的适当保障；适当程度的数据保护；有效和实际的问责方案和机制；除符合国际法的某些情况外，数据主体对访问和删除其在人工智能系统中个人数据的权利和能力的充分享有；数据用于商业目的（例如精准定向广告）或跨境转移时完全符合数据保护立法的适度保护；切实有效的独立监督，作为推动个人掌控其个人数据并促进国际信息自由流通（包括数据获取）之惠益的数据治理机制的一部分。

74. 会员国应制定数据政策或等效框架，或者加强现有政策或框架，以确保个人数据和敏感数据的充分安全，这类数据一旦泄露，可能会给个人造成特殊损害、伤害或困难。相关实例包括：与犯罪、刑事诉讼、定罪以及相关安全措施有关的数据；生物识别、基因和健康数据；与种族、肤色、血统、性别、年龄、语言、宗教、政治见解、民族、族裔、社会出身、与生俱来的经济或社会条件、残障情况或任何其他特征有关的个人数据。

75. 会员国应促进开放数据。在这方面，会员国应考虑审查其政策和监管框架，包括关于信息获取和政务公开的政策和监管框架，以便反映出人工智能特有的要求，并促进相关机制，

例如为由公共资金资助或公共持有的数据和源代码以及数据信托建立开放式存储库，以支持安全、公平、合法与合乎伦理的数据分享等。

76. 会员国应推动和促进将优质和稳健的数据集用于训练、开发和使用人工智能系统，并在监督数据集的收集和使用方面保持警惕。这包括在可能和可行的情况下投资建立黄金标准数据集，包括开放、可信、多样化、建立在有效的法律基础上并且按法律要求征得数据主体同意的数据集。应鼓励制定数据集标注标准，包括按性别和其他标准分列数据，以便于确定数据集的收集方式及其特性。

77. 按照联合国秘书长数字合作高级别小组报告的建议，会员国应在联合国和教科文组织的支持下，酌情采用数字共享方式处理数据，提高工具、数据集和数据托管系统接口的互操作性，并鼓励私营部门公司酌情与所有利益攸关方共享其收集的数据，以促进研究、创新和公共利益。会员国还应促进公共和私营部门建立协作平台，在可信和安全的数据空间内共享优质数据。

政策领域 4：发展与国际合作

78. 会员国和跨国公司应优先考虑人工智能伦理，在相关国际、政府间和多利益攸关方论坛上讨论与人工智能有关的伦理问题。

79. 会员国应确保人工智能在教育、科学、文化、传播和信息、卫生保健、农业和食品供应、环境、自然资源和基础设施管理、经济规划和增长等发展领域的应用符合本建议书提出的价值观和原则。

80. 会员国应通过国际组织，努力为人工智能促进发展提供国际合作平台，包括提供专业知识、资金、数据、领域知识和基础设施，以及促进多利益攸关方之间的合作，以应对具有挑战性的发展问题，特别是针对中低收入国家，尤其是最不发达国家、内陆发展中国家和小岛屿发展中国家。

81. 会员国应努力促进人工智能研究和创新方面的国际合作，包括可以提升中低收入国家和其他国家（包括最不发达国家、内陆发展中国家和小岛屿发展中国家）研究人员的参与度和领导作用的研究和创新中心及网络。

82. 会员国应通过吸收国际组织、研究机构和跨国公司参与，促进人工智能伦理研究，可以将这些研究作为公共和私营实体以合乎伦理的方式使用人工智能系统的基础，包括研究具体伦理框架在特定文化和背景下的适用性，以及根据这些框架开发技术上可行的解决方案的可能性。

83. 会员国应鼓励在人工智能领域开展国际合作与协作，以弥合地缘技术差距。应在充分尊重国际法的前提下，在会员国与其民众之间、公共和私营部门之间以及技术上最先进和最落后的国家之间，开展技术交流和磋商。

政策领域 5：环境和生态系统

84. 在人工智能系统的整个生命周期内，会员国和工商企业应评估对环境产生的直接和间接影响，包括但不限于其碳足迹、能源消耗以及为支持人工智能技术制造而开采原材料对环境造成的影响，并应减少人工智能系统和数据基础设施造成的环境影响。会员国应确保所有人工智能行为者遵守有关环境的法律、政策和惯例。

85. 会员国应在必要和适当时引入激励措施，确保开发并采用基于权利、合乎伦理、由人工智能驱动的解决方案抵御灾害风险；监测和保护环境与生态系统，并促进其再生；保护地球。这些人工智能系统应促进地方和土著社区参与人工智能系统整个生命周期，并应支持循环经济做法以及可持续的消费和生产模式。例如，在必要和适当时可将人工智能系统用于以下方面：

(a) 支持对自然资源的保护、监测和管理。
(b) 支持与气候有关问题的预测、预防、控制和减缓。
(c) 支持更加高效和可持续的粮食生态系统。
(d) 支持可持续能源的加速获取和大规模采用。
(e) 促成并推动旨在促进可持续发展的可持续基础设施、可持续商业模式和可持续金融主流化。
(f) 检测污染物或预测污染程度，并协助相关利益攸关方确定、规划并实施有针对性的干预措施，以防止并减少污染及暴露风险。

86. 会员国在选择人工智能方法时，鉴于其中一些方法可能具有数据密集型或资源密集型特点以及对环境产生的不同影响，应确保人工智能行为者能够根据相称性原则，倾向于使用数据、能源和资源节约型人工智能方法。应制定要求，确保有适当证据表明一项人工智能应用程序将产生这种预期效果，或一项人工智能应用程序的附加保障措施可以为使用该应用程序的合理性提供支撑。假如做不到这一点，则必须遵循预防原则，而且在会对环境造成极其严重的负面影响的情况下，不得使用人工智能。

政策领域 6：性别

87. 会员国应确保数字技术和人工智能促进实现性别平等的潜能得到充分发挥，而且必须确保在人工智能系统生命周期的任何阶段，女童和妇女的人权和基本自由及其安全和人格不受侵犯。此外，伦理影响评估应包含横向性别平等视角。

88. 会员国应从公共预算中划拨专项资金，用于资助促进性别平等的计划，确保国家数字政策包含性别行动计划，并制定旨在支持女童和妇女的相关政策，例如劳动力教育政策，以确保她们不会被排除在人工智能驱动的数字经济之外。应考虑并落实专项投资，用于提供有针对性的计划和有性别针对性的语言，从而为女童和妇女参与科学、技术、工程和数学（STEM）领域，包括信息和通信技术（信通技术）学科，以及为她们的就业准备、就业能力、平等的职业发展和专业成长，提供更多机会。

89. 会员国应确保人工智能系统推动实现性别平等的潜能得到实现。会员国应确保这些技术不会加剧模拟世界多个领域中已经存在的巨大性别差距，而是消除这些差距。这些差距包括：性别工资差距；某些职业和活动中不平等的代表性；人工智能领域高级管理职位、董事会或研究团队中的代表性缺失；教育差距；数字和人工智能的获取、采用、使用和负担能力方面的差距；以及无偿工作和照料责任在社会中的不平等分配。

90. 会员国应确保性别陈规定型观念和歧视性偏见不会被移植入人工智能系统，而且还应对其加以鉴别和主动纠正。必须努力避免技术鸿沟对以下方面产生复合性负面影响：实现性别平等和避免暴力侵害，例如针对妇女和女童以及代表性不足群体的骚扰、欺凌和贩运，包括在线上领域。

91. 会员国应鼓励女性创业、参与并介入人工智能系统生命周期的各个阶段，具体做法是提供并促进经济和监管方面的激励措施以及其他激励措施和支持计划，以及制定目的是在学术界的人工智能研究方面实现性别均衡的参与、在数字和人工智能公司高级管理职位、董事会和研究团队中实现性别均衡的代表性的政策。会员国应确保（用于创新、研究和技术的）公共资金流向具有包容性和明确性别代表性的计划和公司，并利用平权行动原则鼓励私人资金朝着类似方向流动。应制定并执行关于无骚扰环境的政策，同时鼓励传播关于如何在人工智能系统的整个生命周期内促进多样性的最佳做法。

92. 会员国应促进学术界和产业界人工智能研究领域的性别多样性，具体做法包括为女童和妇女进入该领域提供激励措施，建立机制消除人工智能研究界的性别陈规定型观念和骚扰行为，以及鼓励学术界和私营实体分享关于如何提高性别多样性的最佳做法。

93. 教科文组织可以协助建立最佳做法资料库，以鼓励女童、妇女和代表性不足的群体参与人工智能系统生命周期的各个阶段。

政策领域 7：文化

94. 鼓励会员国酌情将人工智能系统纳入物质、文献和非物质文化遗产（包括濒危语言以及土著语言和知识）的保护、丰富、理解、推广、管理和获取工作，具体做法包括酌情出台或更新与在这些领域应用人工智能系统有关的教育计划，以及确保采用针对机构和公众的参与式方法。

95. 鼓励会员国审查并应对人工智能系统产生的文化影响，特别是自动翻译和语音助手等自然语言处理（NLP）应用程序给人类语言和表达的细微差别带来的影响。此类评估应为设计和实施相关战略提供参考，通过弥合文化差距、增进人类理解以及消除因减少使用自然语言等因素造成的负面影响，最大限度地发挥人工智能系统的惠益。减少使用自然语言可能导致濒危语言、地方方言以及与人类语言和表达有关的语音和文化差异的消失。

96. 随着人工智能技术被用于创造、生产、推广、传播和消费多种文化产品和服务，会员国应促进针对艺术家和创意专业人员的人工智能教育和数字培训，以评估人工智能技术在其专业领域的适用性，并推动设计和应用适当的人工智能技术，同时铭记保护文化遗产、多样性和艺术自由的重要性。

97. 会员国应促进当地文化产业和文化领域的中小企业对于人工智能工具的认识和评价，避免文化市场集中化的风险。

98. 会员国应吸收技术公司和其他利益攸关方参与进来，促进文化表现形式的多样化供应和多元化获取，特别要确保算法建议可以提高本地内容的知名度和可见性。

99. 会员国应促进在人工智能和知识产权的交叉领域开展新的研究，例如确定是否或如何对通过人工智能技术创作的作品给予知识产权保护。会员国还应评估人工智能技术如何影响作品被用于研究、开发、培训或实施人工智能应用程序的知识产权所有者的权利或利益。

100. 会员国应鼓励国家级博物馆、美术馆、图书馆和档案馆使用人工智能系统，以突出其藏品，强化其图书馆、数据库和知识库，并允许用户访问。

政策领域 8：教育和研究

101. 会员国应与国际组织、教育机构、私营实体和非政府实体合作，在各个层面向所有国家的公众提供充分的人工智能素养教育，以增强人们的权能，减少因广泛采用人工智能系统而造成的数字鸿沟和数字获取方面的不平等。

102. 会员国应促进人工智能教育“必备技能”的掌握，例如基本读写、计算、编码和数字技能、媒体与信息素养、批判性思维和创意思维、团队合作、沟通、社会情感技能和人工智能伦理技能，特别是在这些技能的教育存在明显差距的国家及国内地区或区域。

103. 会员国应促进关于人工智能发展的一般性宣传计划，其中包括数据、人工智能技术带来的机会和挑战、人工智能系统对人权（包括儿童权利）的影响及其意义。这些计划对于非技术群体和技术群体而言都应方便可及。

104. 会员国应鼓励开展关于以负责任和合乎伦理的方式将人工智能技术应用于教学、教师培训和电子学习等方面的研究活动，以增加机会，并减轻这一领域的挑战和风险。在开展这些研究活动的同时，应充分评估教育质量以及人工智能技术的应用对于学生和教师的影响。会员国还应确保人工智能技术可以增强师生的权能和体验，同时铭记关系和社交层面以及传统教育形式的价值对于师生关系以及学生之间的关系至关重要，在讨论将人工智能技术应用于教育时应考虑到这一点。当涉及到监测、评估能力或预测学习者的行为时，用于学习的人

工智能系统应符合严格的要求。人工智能应依照相关的个人数据保护标准支持学习过程，既不降低认知能力，也不提取敏感信息。在学习者与人工智能系统的互动过程中收集到的为获取知识而提交的数据，不得被滥用、挪用或用于犯罪，包括用于商业目的。

105. 会员国应提升女童和妇女、不同族裔和文化、残障人士、边缘化和弱势群体或处境脆弱群体、少数群体以及没能充分得益于数字包容的所有人在各级人工智能教育计划中的参与度和领导作用，监测并与其他国家分享这方面的最佳做法。

106. 会员国应根据本国教育计划和传统，为各级教育开发人工智能伦理课程，促进人工智能技术技能教育与人工智能教育的人文、伦理和社会方面的交叉协作。应以当地语言（包括土著语言）开发人工智能伦理教育的在线课程和数字资源，并考虑到环境多样性，特别要确保采用残障人士可以使用的格式。

107. 会员国应促进并支持人工智能研究，特别是人工智能伦理问题研究，具体做法包括投资于此类研究或制定激励措施推动公共和私营部门投资于这一领域等，同时承认此类研究可极大地推动人工智能技术的进一步发展和完善，以促进落实国际法和本建议书中提出的价值观和原则。会员国还应公开推广以合乎伦理的方式开发人工智能的研究人员和公司的最佳做法，并与之合作。

108. 会员国应确保人工智能研究人员接受过研究伦理培训，并要求他们将伦理考量纳入设计、产品和出版物中，特别是在分析其使用的数据集、数据集的标注方法以及可能投入应用的成果的质量和范围方面。

109. 会员国应鼓励私营部门公司为科学界获取其数据用于研究提供便利，特别是在中低收入国家，尤其是最不发达国家、内陆发展中国家和小岛屿发展中国家。这种获取应遵守相关隐私和数据保护标准。

110. 为确保对人工智能研究进行批判性评估并适当监测可能出现的滥用或负面影响，会员国应确保人工智能技术今后的任何发展都应建立在严谨和独立的科学研究基础上，并通过吸收除科学、技术、工程和数学（STEM）之外的其他学科，例如文化研究、教育、伦理学、国际关系、法律、语言学、哲学、政治学、社会学和心理学等，促进开展跨学科的人工智能研究。

111. 认识到人工智能技术为助力推进科学知识和实践提供了巨大机会，特别是在传统上采用模型驱动方法的学科中，会员国应鼓励科学界认识到使用人工智能的惠益、局限和风险；这包括努力确保通过数据驱动的方法、模型和处理方式得出的结论完善可靠。此外，会员国应欢迎并支持科学界在推动政策和促进人们认识到人工智能技术的优缺点方面发挥作用。

政策领域 9：传播和信息

112. 会员国应利用人工智能系统改善信息和知识的获取。这可包括向研究人员、学术界、记者、公众和开发人员提供支持，以加强表达自由、学术和科学自由、信息获取，加大主动披露官方数据和信息的力度。

113. 在自动内容生成、审核和策管方面，会员国应确保人工智能行为者尊重并促进表达自由和信息获取自由。适当的框架，包括监管，应让线上传播和信息运营商具有透明度，并确保用户能够获取多样化的观点，以及迅速告知用户为何对内容进行删除或其他处理的相关程序和让用户能够寻求补救的申诉机制。

114. 会员国应投资于并促进数字以及媒体与信息素养技能，以加强理解人工智能系统的使用和影响所需的批判性思维和能力，从而减少和打击虚假信息、错误信息和仇恨言论。此种努力应包括加强对推荐系统的积极和潜在有害影响的了解和评估。

115. 会员国应为媒体创造有利的环境，使媒体有权利和资源切实有效地报道人工智能系统的利弊，并鼓励媒体在其业务中以合乎伦理的方式使用人工智能系统。

政策领域 10：经济和劳动

116. 会员国应评估并处理人工智能系统对所有国家劳动力市场的冲击及其对教育要求的影响，同时特别关注经济属于劳动密集型的国家。这可以包括在各级教育中引入更广泛的跨学科“核心”技能，为当前的劳动者和年轻世代提供可以在飞速变化的市场中找到工作的公平机会，并确保他们对人工智能系统的伦理问题有所认识。除专业技术技能和低技能任务之外，还应教授“学会如何学习”、沟通、批判性思维、团队合作、同理心以及在不同领域之间运用知识的能力等技能。关键是在有关高需求技能方面保持透明度，并围绕这些技能更新学校课程。

117. 会员国应支持政府、学术机构、职业教育与培训机构、产业界、工人组织和民间社会之间的合作协议，以弥合技能要求方面的差距，让培训计划和战略与未来工作的影响和包括中小企业在内的产业界的需求保持一致。应促进以项目为基础的人工智能教学和学习方法，以便公共机构、私营部门公司、大学和研究中心之间能够建立伙伴关系。

118. 会员国应与私营部门公司、民间组织和其他利益攸关方（包括劳动者和工会）合作，确保有风险的员工可以实现公平转型。这包括实施技能提升计划和技能重塑计划，建立在转型期内留住员工的有效机制，以及为无法得到再培训的员工探索“安全网”计划。会员国应制定并实施计划，以研究和应对已确定的各项挑战，其中可能包括技能提升和技能重塑、加强社会保障、积极的行业政策和干预措施、税收优惠、新的税收形式等。会员国应确保有足够的公共资金来支持这些计划。应仔细审查并在必要时修改税制等相关法规，消解基于人工智能的自动化造成的失业后果。

119. 会员国应鼓励并支持研究人员分析人工智能系统对于当地劳动环境的影响，以预测未来的趋势和挑战。这些研究应采用跨学科方法，调查人工智能系统对经济、社会和地域因素、人机互动和人际关系产生的影响，以便就技能重塑和重新部署的最佳做法提出建议。

120. 会员国应采取适当措施，确保竞争性市场和消费者保护，同时考虑可在国家、地区和国际各级采取何种措施和机制来防止在人工智能系统的整个生命周期内滥用与人工智能系统有关的市场支配地位，包括垄断，无论是数据、研究、技术还是市场垄断。会员国应防止由此造成的不平等，评估相关市场，并促进竞争性市场。应适当考虑中低收入国家，尤其是最不发达国家、内陆发展中国家和小岛屿发展中国家，这些国家由于缺乏基础设施、人力资源能力和规章制度等要素，面临滥用市场支配地位行为的风险更大，也更容易因之受到影响。在已制定或通过人工智能伦理标准的国家从事人工智能系统开发的人工智能行为者，在出口这些产品以及在可能没有此类标准的国家开发或应用其人工智能系统时，应遵守这些标准，并遵守适用的国际法以及这些国家的国内立法、标准和惯例。

政策领域 11：健康和社会福祉

121. 会员国应努力利用有效的人工智能系统来改善人类健康并保护生命权，包括减少疾病的暴发，同时建立并维护国际团结，以应对全球健康风险和不确定性，并确保在卫生保健领域

采用人工智能系统的做法符合国际法及其人权法义务。会员国应确保参与卫生保健人工智能系统的行为者会考虑到患者与家属的关系以及患者与医护人员关系的重要性。

122. 会员国应确保与健康、特别是精神健康有关的人工智能系统的开发和部署适当关注儿童和青年，并受到监管，使其做到安全、有效、高效、经过科学和医学证明并能促进循证创新和医学进步。此外，在数字健康干预的相关领域，大力鼓励会员国主动让患者及其代表参与系统开发的所有相关步骤。

123. 会员国应特别注意通过以下方式规范人工智能应用程序中用于卫生保健的预测、检测和治疗方案：

(a) 确保监督，以尽可能减少和减轻偏见；
(b) 在开发算法时，确保在所有相关阶段将专业人员、患者、护理人员或服务用户作为“领域专家”纳入团队；
(c) 适当注意因可能需要医学监测而产生的隐私问题，并确保所有相关的国家和国际数据保护要求得到满足；
(d) 确保建立有效机制，让个人数据被分析的人了解对其个人数据的使用和分析并给予知情同意，同时又不妨碍其获取卫生保健服务；
(e) 确保人工护理以及最终的诊断和治疗决定一律由人类作出，同时认可人工智能系统也可以协助人类工作；
(f) 必要时确保人工智能系统在投入临床使用之前由伦理研究委员会进行审查。

124. 会员国应研究人工智能系统对心理健康的潜在危害所产生的影响及如何加以调控的问题，例如深度抑郁、焦虑、社会隔离、成瘾、贩运、激进化和误导等。

125. 会员国应在研究的基础上，针对机器人的未来发展，制定关于人机互动及其对人际关系所产生影响的准则，并特别关注人类身心健康。尤其应关注应用于卫生保健以及老年人和残障人士护理的机器人、应用于教育的机器人、儿童用机器人、玩具机器人、聊天机器人以及儿童和成人的陪伴机器人的使用问题。此外，应利用人工智能技术的协助来提高机器人的安全性，增进其符合人体工程学的使用，包括在人机工作环境中。应特别注意到利用人工智能操控和滥用人类认知偏差的可能性。

126. 会员国应确保人机互动遵守适用于任何其他人工智能系统的相同价值观和原则，包括人权和基本自由、促进多样性和保护弱势群体或处境脆弱群体。应考虑与人工智能驱动的神经技术系统和脑机接口有关的伦理问题，以维护人的尊严和自主权。

127. 会员国应确保用户能够轻松识别与自己互动的对象是生物，还是模仿人类或动物特征的人工智能系统，并且能够有效拒绝此类互动和要求人工介入。

128. 会员国应实施政策，提高人们对于人工智能技术以及能够识别和模仿人类情绪的技术拟人化的认识，包括在提及这些技术时所使用的语言，并评估这种拟人化的表现形式、伦理影响和可能存在的局限性，特别是在人机互动的情况下和涉及到儿童时。

129. 会员国应鼓励并促进关于人与人工智能系统长期互动所产生影响的合作研究，特别注意这些系统对儿童和青年的心理和认知可能产生的影响。在开展此类研究时，应采用多种规范、原则、协议、学科方法，评估行为和习惯的改变，并认真评估下游的文化和社会影响。此外，会员国应鼓励研究人工智能技术对卫生系统的业绩和卫生成果产生的影响。

130. 会员国和所有利益攸关方应建立机制，让儿童和青年切实参与到关于人工智能系统对其生活和未来所产生影响的对话、辩论和决策中。

V. 监测和评估

131. 会员国应根据本国具体国情、治理结构和宪法规定，采用定量和定性相结合的方法，以可信和透明的方式监测和评估与人工智能伦理问题有关的政策、计划和机制。为支持会员国，教科文组织可以从以下方面作出贡献：

(a) 制定以严谨的科学研究为基础且以国际人权法为根据的教科文组织人工智能技术伦理影响评估（EIA）方法，关于在人工智能系统生命周期各个阶段实施该方法的指南，以及用于支持会员国对政府官员、政策制定者和其他相关人工智能行为者进行伦理影响评估方法培训的能力建设材料；

(b) 制定教科文组织准备状态评估方法，协助会员国确定其准备进程各个方面在特定时刻所处的状态；

(c) 制定教科文组织关于在事先和事后对照既定目标评估人工智能伦理政策和激励政策效力和效率的方法；

(d) 加强关于人工智能伦理政策的基于研究和证据的分析和报告；

(e) 收集和传播关于人工智能伦理政策的进展、创新、研究报告、科学出版物、数据和统计资料，包括通过现有举措，以支持最佳做法分享和相互学习，推动实施本建议书。

132. 监测和评估进程应确保所有利益攸关方的广泛参与，包括但不限于弱势群体或处境脆弱群体。应确保社会、文化和性别多样性，以期改善学习过程，加强调查结果、决策、透明度和成果问责制之间的联系。

133. 为促进与人工智能伦理有关的最佳政策和做法，应制定适当的工具和指标，以便根据商定的标准、优先事项和具体目标，包括关于处境不利者、边缘化群体和弱势群体或处境脆弱群体的具体目标，评估此类政策和做法的效力和效率，以及人工智能系统在个人和社会层面产生的影响。人工智能系统及相关人工智能伦理政策和做法的影响监测和评估，应以与有关风险相称的系统方法持续开展。这项工作应以国际商定的框架为基础，涉及对于私营和公共机构、提供方和计划的评估，包括自我评估，以及开展跟踪研究和制定一系列指标。数据收集和处理工作应遵守国际法、关于数据保护和数据隐私的国家立法以及本建议书概述的价值观和原则。

134. 尤其是，会员国不妨考虑可行的监测和评估机制，例如伦理问题委员会、人工智能伦理问题观察站、记录符合人权且合乎伦理的人工智能系统发展情况或在教科文组织各职能领域通过恪守伦理原则为现有举措作出贡献的资料库、经验分享机制、人工智能监管沙箱和面向所有人工智能行为者的评估指南，以评估会员国对于本文件所述政策建议的遵守情况。

VI. 本建议书的使用和推广

135. 会员国和本建议书确定的所有其他利益攸关方，应尊重、促进和保护本建议书提出的人工智能伦理价值观、原则和标准，并应采取一切可行步骤，落实本建议书的政策建议。

136. 会员国应与在本建议书的范围和目标范畴内开展活动的所有相关国家和国际政府组织及非政府组织、跨国公司和科学组织合作，努力扩大并充实围绕本建议书采取的行动。制定教科文组织伦理影响评估方法和建立国家人工智能伦理委员会，可以作为这方面的重要手段。

VII. 本建议书的宣传

137. 教科文组织是负责宣传和传播本建议书的主要联合国机构，因此将与其他相关联合国实体合作开展工作，同时尊重它们的任务授权并避免工作重复。

138. 教科文组织，包括其世界科学知识与技术伦理委员会（COMEST）、国际生物伦理委员会（IBC）和政府间生物伦理委员会（IGBC）等机构，还将与其他国际、地区和分地区政府组织和非政府组织开展合作。

139. 尽管在教科文组织范围内，促进和保护任务属于各国政府和政府间机构的职权范围，但民间社会仍将是倡导公共部门利益的重要行为者，因此教科文组织需要确保和促进其合法性。

VIII. 最后条款

140. 应将本建议书作为一个整体来理解，各项基本价值观和原则应被视为相互补充、相互关联。

141. 本建议书中的任何内容既不得解释为取代、改变或以其他方式损害各国根据国际法所负义务或所享权利，也不得解释为允许任何国家、其他政治、经济或社会行为者、群体或个人参与或从事任何有悖人权、基本自由、人的尊严以及对生物和非生物的环境与生态系统所抱之关切的活动或行为。