研讨会

人道行动中负责任使用技术

2025年12月4日至5日 | 中国北京

研讨会报告

清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

ICRC

报告

目录

研讨会主办单位

清华大学战略与安全研究中心

清华大学战略与安全研究中心成立于2018年11月7日，是聚焦国际战略与安全领域研究的清华大学校级智库。中心有两大目标：一是研究国际秩序、国际关系及战略与安全问题，跟踪国际形势变化，为决策提供参考和建议，向社会传递理性专业声音；二是通过国际交流与合作阐释和传播中国的理念和政策主张，增进国际社会对中国的了解，提升清华大学在战略与国际关系领域的影响力。

中心下设美欧研究、全球治理研究、欧亚研究、人工智能治理等研究方向，设有致力于国际传播的子品牌“中国论坛”以及覆盖国内外一流高校的学生学术项目“战略青年”，指导清华大学学生全球战略研究协会和中美新声人文交流社团。

红十字国际委员会

红十字国际委员会是一个公正、中立和独立的组织，其独特的人道使命在于保护武装冲突和其他暴力局势受难者的生命与尊严，并为其提供援助。

红十字国际委员会东亚地区代表处于2005年7月在北京设立。过去几十年来，随着中国的国际地位愈发显著，该代表处将鼓励并支持中国为国际人道事业作出贡献作为其使命。东亚地区代表处积极推进并维护与中国政府、军队、警方、中国红十字会、智库与学术机构、媒体以及企业界的战略对话，就人道议题和国际人道法的推广展开深入交流。同时，东亚地区代表处同时也是红十字国际委员会在全球最大的采购中心之一。

红十字国际委员会卢森堡全球网络事务代表处的设立，旨在应对数字化转型对武装冲突和人道行动日益加深的影响。红十字国际委员会依托该代表处，将网络和数字层面的考量纳入其中立、独立的人道行动中，并始终优先考虑受冲突影响民众的需求。该代表处的使命是确保红十字国际委员会在网络空间及整个数字时代，始终作为中立、公正且独立的人道组织开展行动。为践行这一使命，该代表处积极开展战略研究、推动法律与政策层面的接触沟通、实施外联举措并推进行动创新。

执行摘要

“人道行动中负责任使用技术研讨会”汇聚了来自近20个国家的80多名人道实践者、学术研究人员及产业界代表，共同讨论了新兴技术正在如何重塑人道行动，以及在危机环境中负责任使用这些技术所需要采取的措施。

第一场圆桌会议围绕“负责任的人工智能”主题展开，重点探讨治理、伦理与问责三个关键问题。与会者一致认为，人的控制与对人的问责仍是人工智能治理的核心，因为法律和伦理责任不能转嫁给机器。讨论指出，尽管不同地区监管方式存在差异，各方在安全性、透明度、人类监督与包容性等共同原则上正逐步形成共识。与会者同时强调，仅靠原则已不足以应对当前挑战。推动负责任的人工智能，需要建立具体的保障措施、测试与评估机制、明确的责任链条以及可执行的问责体系。

第二场圆桌会议围绕“科技向善”主题展开，探讨技术如何为高效协同的危机应对提供支持。与会者指出，健全的数字基础设施是“科技向善”最重要的体现形式，涵盖网络联通、电力供应和支持离线运行的系统。讨论随后聚焦于颇具争议的技术中立性问题。部分与会者主张，代码可以保持中立。而另一些与会者则强调，技术始终反映权力结构与激励机制，会形成人道行动者应予管理的依赖关系。与会者一致认为，与科技行业开展对话是使“科技”成为“向善力量”的关键所在——而这需要责任担当、保障措施和退出策略，以及敢于对造成伤害、依赖或破坏人道原则的技术说“不”的勇气。

第三场圆桌会议围绕“人工智能的安全与保障”议题，探究如何在危机环境中负责任地部署人工智能。与会者一致认为，人工智能风险程度与使用情境高度相关，比如在物流或后台职能中应用风险较低，而影响分诊或预警系统的人工智能系统则风险更高。由于人工智能系统将数据直接嵌入算法，基于“安全设计”理念，人们必须在数据、透明度和模型训练方面做出艰难抉择。讨论强调，必须让工作人员和社区具备识别偏见与风险的能力，使其在必要时能够对人工智能系统提出质疑并保持审慎；人道组织无法单独管理这些风险，因此与学术界、产业界及公共部门的合作至关重要。

第四场圆桌会议围绕“数字信任”展开，重点探讨在互联世界中如何构建和维护信任。与会者将信任界定为一种由实践驱动的关系型概念，并一致认为，信任建立在一致、透明且可预测的行为基础上。在冲突环境中，数字信任尤为脆弱。敏感数据的滥用或泄露不仅可能直接危及民众安全，也会阻碍人道援助的准入与接纳。展望未来，与会者强调，构建并维系信任需要对未来风险做好准备和预判，包括制定系统发生故障或被滥用情况下的应急预案，并推行如数字红十字、红新月及红水晶标志等务实举措。

引言

在数字化及其他新兴技术加速发展的当下，人道领域正处于关键的十字路口。从网络互联与数据管理、到人工智能与卫星系统，技术正在重塑人道组织评估需求、规划行动、沟通联络、提供援助以及开展保护的方式。这类工具有助于提升人道行动的效率、精准度与覆盖面，但同时也会引入新的风险、依赖关系和两难困境。例如，网络有害信息的快速传播可能助长分裂、煽动暴力，加重冲突给民众带来的伤害。数据泄露则会破坏平民及武装冲突各方对人道组织的信任，不仅妨碍人道援助有效触达受武装冲突及其他暴力影响的人群，也可能危及人道工作者的安全。此外，军事技术与自主武器系统的研发，也会在战场上引发大量人道、法律及伦理层面的挑战。

随着人道行动日益依赖主要由外部相关方开发的数字系统，治理、问责、中立性与保护等问题已经从人道实践的边缘走向核心。这意味着，人道领域难以独自解决这些技术挑战，而是需要与产业界、学术界、民间社会及人道领域的利益相关方一道，开展网络化、持续性的协作。

为此，红十字国际委员会（ICRC）在2022年通过其位于卢森堡的全球网络事务代表处发起了一系列研讨会，以应对网络空间与数字时代人道行动面临的挑战。该系列研讨会均遵循查塔姆规则，构建一个开放、可信赖的平台，围绕新技术、网络安全、数据保护与人道行动的交汇点，开展跨领域和跨学科的对话。研讨会此前已先后于2022年和2024年在卢森堡、2024年11月在肯尼亚内罗毕、2025年6月在瑞士日内瓦、2025年11月在奥地利维也纳举行。更多详情可访问红十字国际委员会官网www.icrc.org/zh。

以下报告总结了该系列研讨会于2025年12月在中国北京举办的一场会议的关键发现与成果。

人道行动中负责任使用技术研讨会

本次研讨会聚焦人道行动中负责任使用技术，在内容深度与合作广度上均成为该系列研讨会的重要里程碑。研讨会由红十字国际委员会卢森堡全球网络事务代表处、红十字国际委员会位于北京的东亚地区代表处及清华大学战略与安全研究中心（CISS）共同主办。研讨会将有关技术的人道辩论置于学术环境中，围绕安全、技术政策与全球治理的交汇领域展开深入探讨。本次活动汇聚了来自亚、欧、非及北美近20个国家的80余位人道实践者、政策专家、产业界代表及研究人员，有力推动人道一线关切与人工智能治理与安全、网络风险及数字化转型等相关议题深度融合。

作为红十字国际委员会迄今在中国举办的规模最大的技术领域活动，本次研讨会也为与中国的科研与技术界开展合作营造了良好氛围。研讨会汇聚众多学术精英，聚焦长期研究、治理模式及战略风险思维等议题。同时，讨论始终立足于现实人道实践，将数字信任、中立性与独立性等问题与数字技术在冲突环境中给平民带来的机遇与风险相结合。

开幕致辞及主旨演讲

开幕式围绕新兴技术展开探讨，指出其既为人道行动提供重要助力，也带来伦理、人道与治理层面的重大风险。开幕式由清华大学战略与安全研究中心副主任肖茜主持，清华大学校务委员会副主任杨斌教授和红十字国际委员会主席中国特使、东亚地区代表处主任史德林（Balthasar Staehelin）先生分别代表主办方发言。

> 在人类文明的漫长历程中，
> 技术始终是社会进步的驱动力。
> ……但正因为技术蕴含着巨大的潜力，
> 我们才必须确保其得到负责任地使用。
> 如果没有适当的规范和伦理框架，
> 技术在人道领域的应用可能会达不到预期效果，
> 甚至给弱势群体带来新的风险。
>
> — 杨斌

> 我们在战争中面临的这类[数字]困境，
> 就像放大镜一样，
> 将某些问题以极其鲜明的方式凸显出来。
> — 史德林

两位发言人均将研讨会置于快速的技术变革与人道责任的交汇点，强调虽然人工智能、大数据和云计算正在以前所未有的速度改变社会和人道行动，但在危机情境中采用这些技术仍需要审慎的治理。杨斌指出，中国在人工智能符合伦理的发展和全球治理倡议议程方面参与日益深入，这也是选择北京作为研讨会举办地的动因之一。史德林则将人道场景喻为“放大镜”，指出数据敏感性、算法偏见、网络虚假信息与数字包容等困境在此类场景中尤为凸显，并可能直接引致人身伤害。两位发言人一致强调，应加强跨部门、跨区域的对话与合作，充分纳入中国及“全球南方”的多元视角。

主旨演讲嘉宾从互补角度对这些议题作了进一步阐释。中国新一代人工智能发展战略研究院执行院长龚克强调，人工智能已应用于人道实践，助力更快速、更高效的危机应对。他呼吁进一步优先发展契合人道需求的人工智能应用，并强调以人道和负责任的方式发展人工智能的重要性。相关保障措施可包括将伦理原则贯穿于人工智能整个生命周期——从初始设计和部署阶段就嵌入伦理原则，而不是仅在事故发生后才处理伦理问题。他还提议建立人工智能事故学习平台、无过错人工智能赔偿基金以及构建负责任人工智能的问责框架。

> 人工智能能帮助我们更有效、更迅速、更准确地应对危机。
> 但只有将人工智能牢牢植根于人道主义的核心原则——
> 人道、中立和独立中，这一潜力才能真正得以实现。
> 弱势群体应当是首先受益于人工智能的人群，而不是最后一批。
> — 龚克

中国军控与裁军协会秘书长戴怀成警示了人工智能军备竞赛的持续升级的风险，强调人工智能的军事应用必须遵守国际人道法，呼吁加强全球治理、推动主要大国开展自我约束，并为不可接受的使用方式划定明确界限。北京前瞻人工智能安全与治理研究院院长曾毅教授指出，人工智能模型不会自动变得更安全，他呼吁在全球范围内为人工智能的高风险使用及军事应·用划定红线。红十字国际委员会卢森堡全球网络事务代表处主任埃尔斯·德布夫（Els Debuf）强调，数字技术已成为人道响应的生命线；她呼吁在设计初期就融入人道原则，并持续开展跨领域合作，深化与中国伙伴的对话。

开幕小组讨论：为研讨会奠定基调

开幕小组讨论由史德林主持，将主旨演讲的信息转化为关于技术治理和应用的更具实践性的讨论。陈琪（清华大学战略与安全研究中心副主任）指出，禁止致命性自主武器系统并不现实，重点在于必须将国际人道法嵌入此类系统，并始终确保人类在决策过程中的主导地位。

> 与其只着眼于借助人工智能实现优化，
> 不如关注一个更有趣的问题：
> 在（考虑到）这些系统已经可用的情况下，
> 重新思考机构和组织本身。
> 一个已有160年历史的组织，
> 如何向人工智能借力，
> 通过使用人工智能以另一种方式实现组织的既有宗旨。
>
> — 安德烈亚·卡瓦拉罗教授

> 人道部门今天面临的技术碎片化，
> 恰似互联网发展早期的环境。
> 各大技术公司研发了强大但封闭的数据系统，
> 形成了‘小院高墙’。
> 在危机情境下，
> 这种数据不可互操作性将构成严峻的人道挑战。
>
> — 李晓东

安德烈亚·卡瓦拉罗（伊迪亚研究所主任、洛桑联邦理工学院教授）强调，尽管技术取得了巨大进步，但复杂精密的人工智能系统依然存在脆弱性。例如，一个拥有70亿参数的大语言模型，仅需操控其中的4个参数，就可能引发整个系统的崩溃。他观察到，学术界和人道行动遵循着截然不同的时间框架——前者以72个月为周期，后者则以72小时为单位。然而，他主张在长期的“前瞻性思考”和短期的“高效执行”之间找到平衡。李晓东（伏羲智库创始人、主任）强调了技术碎片化“小院高墙”带来的相关风险，并提出可从互联网治理中借鉴两项可用于人工智能治理的指导原则：基本共识，由以实

践为基础的广泛共识推动进步，而非追求形式上的全体一致；可执行的代码，优先通过测试、学习和逐步改进来推动工作，而非等待形成完备的监管方案。布莱兹·罗贝尔（红十字国际委员会数字转型和数据部门全球人工智能顾问）总结了以上嘉宾发言，并再次强调数字技术本质上具有两用性，有必要在即时行动需求和长期治理之间取得平衡，同时将人道原则作为实践指南。

开幕致辞、主旨演讲与小组讨论共同为研讨会四个下设分论坛圆桌会议（遵守查塔姆规则）奠定了理论框架，突出了创新与克制之间的张力，并确立了负责任设计、治理及跨部门对话是人道行动中负责任使用技术的先决条件。

分论坛一

负责任的人工智能：伦理、安全与问责治理

背景

各国政府、私营企业和人道组织日益认识到，构建全面的人工智能治理框架，对于指导人工智能负责任开发与部署至关重要。人工智能系统——广义上指使用计算机系统来完成通常需要人类认知与推理能力完成的任务的任一工具或技术——为各领域带来了宝贵机遇，人道领域亦不例外。但是，由于数据保护关切、系统局限性以及人机交互相关问题的存在，人工智能系统也带来了严重风险。各地区在人工智能治理路径上的差异，给行动遍及全球的人道组织带来了难题。本圆桌会议旨在探讨各类治理模式，明确共识交汇点，同时探索通过跨部门协作最大化人工智能效益并降低其风险。

圆桌讨论概要

讨论围绕两条相互交织的主线展开：一是如何在系统自主性不断增强的情况下，维持人的责任；二是如何构建既具有全球一致性、又能够在地方层面落地的治理框架。开场发言中，与会者重点提及了若干问题：人工智能系统的局限性，此类系统带来的风险，治理理念上的显著分歧（部分地区趋于审慎预防，其他地区则更倾向创新驱动），人工智能能力差距的扩大，以及可能会加剧各地区碎片化的全球紧张局势。与此同时，与会者也强调，全球范围内在以下方面已达成广泛共识：坚持人工智能以人为本的方法，将问责制和透明度作为共同原则，以及责任和控制权必须始终属于人类而非机器。

决策的复杂性被视为是一大挑战，与会者也就此展开了讨论：这是否意味着无法实现问责，还是说人工智能会使逃避责任变得更容易。多名发言者反对“问责缺失”的说法，提出设计、授权及部署此类系统的人是出于主观意图行事的，必须始终承担相应责任。

与会者还围绕军民两用技术问题展开讨论。在人工智能用于人道行动和冲突场景的背景下，这一议题尤为重要。与会者指出，人工智能正日益被整合入武器系统，交战行为的不可预测性随之上升，这可能会削弱确保遵守相关法律义务的能力，这一情况令人担忧。与会者还注意到，关键数字基础设施对人工智能的依赖性与日俱增，针对数据中心或电力网络发动的攻击可能会模糊民用物体和军事目标之间的界限，这也引发了新的关切。卫星系统、网络安全和生物研究也存在类似的两用技术问题。因此，一名与会者提议，应明确考虑“人工智能催生的危机”，完善灾害管理工具以适应人工智能情境，包括事件信息共享、监测、演习以及危机后调查等。

讨论多次回到“包容性”这一议题。多名与会者指出，“人工智能鸿沟”在不断扩大，世界上相当一部分人口难以接触到人工智能技术，而资源匮乏的国家既没有能力研发人工智能系统，也无法在人工智能治理中发挥实质性影响。另有与会者指出，来自中国、非洲国家及更广泛的全球南方国家的行为体，他们的声音在全球讨论中仍未得到充分重视。上述观察进一步强化了相关呼吁，即应使人工智能服务全球公共利益，并构建更加包容的人工智能治理体系。

主要结论：

人类控制和问责制始终是负责任人工智能治理的基石。监管和规范约束的对象是人，而非机器；不能让人工智能系统削弱或替代人类判断，更不能模糊责任链条。法律和伦理责任必须始终由人承担。

尽管各地区治理模式不尽相同，本次圆桌会议发现，围绕多项核心价值的共识正不断加深，例如在安全、透明度、人类监督和包容性等方面。政府、业界、学术界及人道参与者等各方之间的协作，始终是避免治理碎片化、确保人工智能负责任开发与使用的关所在。

仅有原则已不足够，治理还需依托具体的工具、程序和安全保障措施。多名与会者强调，国际社会不能止步于宽泛的原则（如安全、透明度和责任），而必须转向测试和评估、可追溯机制、明确责任分配以及人工智能全生命周期问责制落地等实操举措。

分论坛二

技术向善：技术如何支持高效协同的危机应对？

背景

主要科技企业日益重视“技术向善”倡议，陆续推出了聚焦社会责任、数字包容与可持续性的项目。这为人道行动带来了机遇，使医疗影像支持工具、供应链溯源以及卫星或移动网络连接等创新技术得以用于危机应对行动。在此背景下，本圆桌会议旨在探讨以下问题：技术如何能够为人道应对行动提供有效支持；脆弱局势下哪些技术创新最为重要；以及人道组织、政府、学术界和私营部门如何通过合作伙伴关系，负责任且可持续地利用技术。

圆桌讨论概要

讨论反复回到一个根本性问题：人道组织在危机情境下开展工作时，“技术向善”首先依托的是基本基础设施，而非先进工具。多名与会者强调，网络连接和电力系统，以及可在断网断电条件下运行的系统，共同构成了人道行动的根基。有发言者指出，对身处危机之中的民众而言，网络连接的重要性堪比水电供应，一旦缺失，再先进的技术方案都将失去价值。

随后，圆桌会议围绕技术中立性展开讨论，气氛十分热烈。部分发言者表示，至少在代码层面，技术本身是中立的。另一些发言者则认为，技术从来不是中立的，其设计理念、企业激励、地缘政治环境、出口管制和制裁制度等，都会深刻影响工具的开发和部署方式。在将该议题与人道行动相关联时，与会者对如何处理所谓的技术中立性问题、以避免与人道组织所秉持的中立性原则产生冲突表示了关切。

由此引出了关于权力、问责制和信任问题的广泛讨论。多位发言者指出，数字工具或将重塑科技企业、人道组织和受影响社区之间的关系。因此，加强与技术行业的对话被视为关键，以确保科技企

业所设计的技术和产品能够成为向善之力，而非伤害之源。生物识别技术即为一个例证，表明技术会造成责任倒置，致使受影响民众不得不向相关组织证明自身身份，而非加强相关组织对其所服务民众负责的问责机制。与会者还援引人种学研究指出，数字工具可能会影响社区对人道组织的看法和信任，尤其是互动方式从面对面接触转向使用屏幕和系统之后，这一影响更为显著。

讨论通过具体案例，展示了技术的潜力与风险。有与会者提到，在宏福苑火灾发生后，专门开发了数字工具来连接受影响民众和志愿者。在这一案例中，开发者特意选择了非实名注册制，防止给受灾民众带来额外负担，并避免引发更多数据保护关切。与会者也强调了一些值得警惕的案例，例如某些生物识别数据库在组织退出使用时，缺乏妥善删除数据的机制，这可能会导致高度敏感数据遭到暴露。

最后，会议回顾了重点案例，并达成共识，认为仅凭良好意图不足以实现技术“向善”。与会者强调，必须要建立问责制，在技术设计之初就融入伦理考量，以及具备拒绝使用那些造成依赖性、暴露敏感数据或削弱信任的技术的能力。多位发言者呼吁加强人道组织、学术界和私营部门之间的协作。与会者总结到，评判技术不应以复杂精密程度为标准，而应以其是否能在危机情况下切实运行、是否真正服务于有需要的人群——往往最简单的方案才是最恰当的方案。

主要结论

强健的数字基础设施是“技术向善”中最关键的一环。在危机情境下，“技术向善”始于有韧性的基础条件，即网络连接、电力保障和可离线运行的系统。如果没有强健的基础设施，即便是最先进的技术也不能发挥其价值。

技术中立性存在争议。部分与会者认为技术可以保持中立；但也有与会者认为，技术绝不可能中立，始终受到政治、法律和权力等因素的影响。对人道组织而言，维护其自身的中立性，关键在于妥善管理外部依赖、应对供应商锁定问题，同时防范受影响民众造成长期依赖的风险。

仅凭良好意图仍不足够。善意和创新并不能保证积极的结果。技术行业、公共部门和人道工作者之间进行有意义的对话对于让“技术”成为“向善”之力至关重要——这需要明确责任担当、强化安全保障、制定退出策略，以及向可能带来伤害、造成依赖或有损于人道原则的技术说“不”的勇气。

分论坛三

人工智能安全保障：危机环境负责任部署的可行防范措施

背景

随着人工智能系统在世界范围内得到部署，其在危机环境中的应用也日益增多。从开源医疗助手、卫星影像分类器等模型，到运用算法监测冲突，都展现出人工智能在促进人道应对方面的潜力。然而，高风险环境也暴露出诸多严重隐患：算法偏见与歧视、数据保护不足、决策过程不透明，以及一旦出现故障，甚至可能直接危及弱势民众的安全与尊严。鉴于人工智能研究与技术大多源于商业领域，如需将这些工具安全地部署于人道场景，就必然会面临更多复杂挑战。因此，亟需建立可行的综合保障措施，加强数据治理，提升透明度和增进稳健性。本次圆桌会议旨在探讨以下议题：如何评估危机情形下人工智能应用的相关风险；需要设定哪些“红线”并采取哪些风险减轻策略；以及如何通过多利益相关方合作，确保人工智能切实服务于受影响社区。

圆桌讨论概要

会议伊始，与会者就一致认为人工智能可以有效支持人道行动，但其风险也因具体情境而差异显著。相较于后勤或后台管理工具，影响伤员分诊、早期预警或救济物资分配的系统会带来更为严重的风险，因其所用工具压缩了决策时间，可能固化偏见和现有歧视，因而必须实施严格的保障措施。虽然针对具体情境开展人工智能设计与部署的方法原则上大有裨益，但若需要构建涵盖所有可能情境及相关风险因素的完美分类体系，反而可能削弱其实际效果。因此，与会者认为，风险框架应根据不伤害、中立、允许人类干预等明确原则制定，并根据具体情形灵活调整。

与会者指出，网络安全、数据保护不足以及基础设施薄弱等诸多风险并非人工智能领域所独有。但由于人工智能系统直接将数据嵌入算法之中，就会产生额外的脆弱性。训练数据不仅影响输出结

果，还可能泄露敏感信息，甚至存在被篡改的风险。与会者呼吁采取基于风险评估的采购机制，对供应商的数据实践进行评估；制定防止人道数据二次使用的限制措施；并优先选择可予以审查和约束的模型。他们强调，数据最小化原则（即仅收集绝对需要的数据）应成为核心保障措施，同时也承认在现实紧急情况下落实这一原则极具挑战。例如，知情同意原则上至关重要，但在重大危机局势中却往往难以实现。此外，多位与会者还指出，必须由人类来决定何时可作出例外处理，而非人工智能自动决定。

会议重点提出，提升透明度与培养人工智能素养同等重要。首先，与会者呼吁引入人工智能用户交互界面，以揭示不确定性，解释输出结果的产出过程，并提高人工干预模型的便捷性。其次，员工培训不应仅止步于工具使用层面，还应帮助员工了解其中可能存在的偏见、脆弱性和操作风险。员工在觉察异常时，必须具备“敢于质疑人工智能”的能力。

最后，与会者们一致认为，人道组织无法仅凭自身力量有效管控人工智能安全。他们需要依托学术界、技术行业和公共部门提供专业支持，共同开展安全模型研发、评估工作，建立事件信息共享机制，携手推动能力建设。正在逐步启用人工智能工具的人道组织以及正在部署此类工具的社区，必须要及早参与相关工具的设计过程，确保工具始终安全、适当且符合人道原则。

主要结论

人工智能在安全和保障方面的风险有高度的情景依赖性。与会者一致认为，无论是在危机前、危机中还是危机后，人工智能均具有重要应用价值，但其相关风险特征却大不相同。后台管理和后勤工具带来的风险较为可控；而影响伤员分诊或早期预警的一线系统则需受到更为严格的审查。

“安全设计”理念意味着要在数据、透明度及模型训练等方面作出艰难的运营抉择。由于人工智能系统直接将数据嵌入算法之中，模型行为因而会映射训练该模型时所使用的数据。因此，除传统网络安全问题和数据保护风险外，人工智能还会带来新的薄弱环节和风险。数据与算法的深度耦合，导致系统容易受数据投毒的影响，并可能因数据质量低下而产生偏见。

相关培训不仅要培养员工和社区使用人工智能工具的能力，还应帮助他们了解其固有偏见和风险，学会质疑人工智能系统。与会者呼吁，不应仅依靠单个组织的努力，而应推动在更广泛的社会层面提升人工智能素养，使相关社区了解人工智能的应用场景与使用局限，并明晰何时应质疑其输出结果。

分论坛四

数字信任：构建和维护互联世界中的信心

背景

数字信任，指公众对数字技术、服务及其提供方的预期，即认为相关技术和服务能够安全、可靠地运行，并符合问责、监督等社会价值观和共同期待。随着人道行动日益依赖数字化及数字赋能服务，民众对这些系统的运作方式、数据处理方式以及相关组织行为模式的信任，对于开展行之有效、恪守原则的人道应对行动至关重要。在危机和冲突背景下，受影响民众遭受剥削、歧视与伤害的风险显著增加，若无法建立数字信任，其安全、尊严和人道准入均将受到严重影响。本次圆桌会议将数字信任视为在人道行动中负责任地使用数字技术的基础条件，重点探讨如何将技术保障措施、治理框架和跨部门合作相结合，在快速演变且充满争议的数字环境中建立、维护和保障数字信任。

圆桌讨论概要

本次圆桌会议将数字信任定位为，在日益互联且充满争议的数字环境中开展人道行动的重要基石。会议提议对数字信任这一概念加以界定，与会者提出了各不相同但又互为补充的定义。他们并未将其限定于某种固定含义，而是将其视为一种具有关系性、情境化、由实践驱动的概念。有与会者将数字信任定义为在冲突环境中安全开展工作，与病患、当局及武装行为体维系关系所需的信任。另有与会者提出了涵盖政策、技术、伦理及法律四个维度的信任分类法。同时，还有与会者将其定义为对某一数字工具、平台或服务的信心，即相信其会按照预期方式运作且相关数据会得到负责任的处理。整个讨论过程中，数字信任从未被视为纯粹的技术概念，而更多体现为人们对数字工具的运行方式、数据的处理方式，以及相关组织能否按预期负责任行事所抱有的信心。

基于这一共识，与会者探讨了数字信任何以日益难以建立与维系。多位发言者指出，若数字系统出现故障、遭到滥用或未得到充分了解，信任就会变得非常脆弱。他们进一步指出，认知与教育鸿沟、

网络安全威胁与数据泄露、错误信息，以及技术的快速迭代，都是不断侵蚀信任的因素。与会者还指出，上述挑战在冲突和危机环境中更为严峻，因为人道组织往往需要在充满不安全因素、存在跨境数据流动但监管保护薄弱的环境中，处理涉及弱势民众的高度敏感信息。他们对冲突期间人道数据遭到滥用、恶意利用或挪用尤为担忧，指出此类事件会导致受影响民众受到伤害，也会导致信任迅速瓦解，致使社区放弃使用数字服务，从而破坏人道准入与接受度。

关于建立信任的讨论指出，网络安全、加密以及通过设计保障隐私等技术措施至关重要，但仅靠技术本身尚有不足。信任建立在始终如一、公开透明且可预测的行为之上。与会者强调，治理与政策承诺，以及多因素认证等可被观察到的网络安全实践，都是构建信任的重要措施。有发言者表示，在解释组织的决策过程和风险权衡时保持公开透明本身就有助于建立信任。针对军民两用数据带来的挑战，以及人道组织在资源上无法与各交战方匹敌的现实，有与会者建议，最佳的应对方法是从源头上避免存储军民两用数据或采用非数字形式进行保存。

最后，会议提及红十字国际委员会的“数字标志”项目，将其作为一项与信任相关的具体倡议。这一新举措为的是构建一个有助于在网络空间实现人道保护目的的标识系统。尽管数字红十字、红新月及红水晶标志本身并不免受网络攻击的保护，但其背后的理念是使用这些标志来标示人道与医疗数字资产的受保护地位。与会者们还讨论了制定清晰明确且普遍适用的技术标准的必要性，以及数字标志应具备的功能，包括数字证书的可追溯性。

主要结论

对人道行动的数字信任，既来源于安全可靠的技术系统**始终如一、透明及可预测的表现**，也来源于跨境协作，尤其是在高风险环境下处理敏感数据时。

在**冲突背景下，数字信任尤为脆弱。**人道数据的滥用、挪用或泄露会直接危及受影响民众，破坏中立性、准入，以及相关社区对人道工作的接纳度。

构建并维系数字信任，需要前瞻未来风险并做好准备，包括为技术故障或滥用制定应急预案，并实施红十字国际委员会“数字标志”项目等实用机制。该项目旨在通过技术手段标示人道与医疗数字资产所享有的法律保护。

致谢

本报告汇总2025年12月4日至5日在中国北京举办的“人道行动中技术负责任使用研讨会”上探讨的主要内容与观点。本次研讨会由红十字国际委员会卢森堡全球网络事务代表处、红十字国际委员会东亚地区代表处（北京）与清华大学战略与安全研究中心联合策划主办。

主办方谨向来自近20个国家的80余位发言嘉宾及参会人员致以诚挚谢意，感谢他们齐聚北京，参加此次为期两天的深入研讨。同时，衷心感谢红十字国际委员会卢森堡全球网络事务代表处、东亚地区代表处及清华大学众多工作人员为此次研讨会成功举办所付出的努力。

本报告由红十字国际委员会研讨会组织团队编写。作者谨向所有参与报告筹备工作的人员致以诚挚谢意。本报告基于作者根据查塔姆规则整理的会议记录撰写而成。作者虽力求准确呈现讨论的核心要义，但疏漏或不当之处在所难免。报告内容并不代表活动主办方、参会人员或协调员的官方意见。主办方亦不对其中任何建议、观点或其他信息的准确性或可靠性作任何保证。报告中的任何错误或表述不当之处，概由作者自行负责。

红十字国际委员会通常携手其红十字和红新月的合作伙伴，帮助世界各地受武装冲突和其他暴力影响之人，竭尽所能保护他们的生命与尊严，减轻他们的苦难。该组织还通过推广并加强人道法，弘扬普遍人道原则，来尽力防止苦难的发生。

民众知道他们可以信赖红十字国际委员会在冲突地区开展一系列挽救生命的行动，并与当地社区紧密合作，以理解并满足他们的需求。该组织具有相关经验和技术专长，故而能够迅速、有效并公正地进行应对。

CISS 微信　CISS 头条　CISS 官网　ICRC 官网　ICRC 电子资源库　ICRC 微信　ICRC 微博　ICRC 哔哩哔哩官号

www.icrc.org
facebook.com/icrc
x.com/icrc
instagram.com/icrc

红十字国际委员会东亚地区代表处
中国北京市建国门外大街9号
齐家园外交公寓3-2
邮编：100600
电话：+86 10 8532 8500
传真：+86 10 6532 0633
邮箱：bej_beijing@icrc.org　www.icrc.org

ICRC

2529/006 2026