化学和生物武器禁令背景下的人工智能焦虑

人工智能（AI）有可能彻底改变不同行业的科学、经济和社会活动。然而，人们对于人工智能双重用途的应用和结果可能带来的负面影响越来越有顾虑。这些“人工智能焦虑”之中既有真实也有想象，根源在于现在和未来的可能性。

在实施化学和生物武器（CBW）禁令的背景下，也出现了一些关于人工智能潜在双重用途的顾虑，人们开始考虑人工智能促成 CBW 设计、开发、部署和检测的能力。

一些新的说法正在吸引公众的注意力，这些说法认为人工智能可能会推动超级毒剂的开发，或者为国家和非国家行为者开发和使用 CBW 提供低成本的途径。

本篇简报介绍了四种新出现的关于 CBW 的人工智能焦虑，以现有负责任的人工智能伦理原则为指导，思考我们的规范和价值观如何帮助解决这些挑战。

我们可以通过确定这些原则试图减少人工智能的哪些特征，更合理地设想这些特征可能在实施 CBW 禁令的背景下带来哪些具体的挑战。这是一套早期方案，有助于从本质上理解这些新出现的挑战。

要点

- 人工智能并非独立实体，而是一个大型系统的组成部分，与其他技术结合后可以增强数据处理和决策。
- 虽然人工智能会带来新的挑战和焦虑，但其中许多与其他技术或实践的相关挑战和焦虑并无本质区别。人工智能可能会放大和改变现有的挑战，必须放在其所处系统的更广泛背景下看待。
- 减轻人工智能的潜在有害影响并不需要全新的治理架构，而是通过识别和解决潜在的人工智能焦虑，帮助修正和增强现有框架。

UNIVERSITY OF SUSSEX

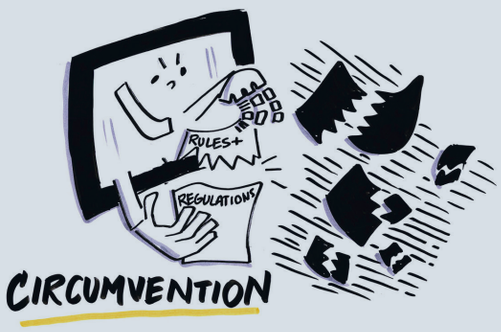
HSP | THE HARVARD SUSSEX PROGRAM

人工智能 +（技术）

人工智能与其他技术的结合对 CBW 禁令提出了重大挑战。符合这一焦虑的各种场景都围绕着一个核心，即人工智能如何增加或扩大现有双重用途技术的风险。例如，在自动驾驶车辆中使用人工智能瞄准来投放防暴剂 (RCA) 可能会导致《禁止化学武器公约》中 RCA 的安全合法使用更加模棱两可。另外，人工智能与先进机器人技术的结合可以帮助人力资源有限的行为者远程设计和合成化学品或生物制品。在这种情况下，伦理原则可能更符合“安全可靠”和“以人为本”的价值观，其中基础设施、治理和意图评估有助于厘清人工智能使用中潜在的治理措施。

形形色色的规避途径

人工智能可以帮助克服阻碍开发、储存和部署 CBW 的传统限制、障碍与控制。例如，人工智能可能会帮助恶意行为者通过识别或模拟未列入清单的化学合成前体来规避进出口管制。此外，人工智能的使用可以减少开发或部署这些武器所需的人力资源，从而降低成本并避免潜在的“道德困境”。更进一步，还可以利用人工智能设计更难检测的药剂，而且可能更难防御。在这些情况下，“可控制”和“可问责”的价值标准可能会反映出在人工智能开发中进行审查治理的必要。



人工智能

可解释且可问责

“能够解释如何以及为何达到了特定的结果；数据输入、设计结构和操作系统、整体运作和结果均应允许问责。”

可控制

“可以控制输入、操作和结果以便改变、修正、覆盖或关闭作业；输出也必须可控（如防止自动传播）。”

应当……

以人为本

“人类最终要对人工智能的整个设计和运行周期及其产出负责，了解人类在哪些环节以哪些方式牵涉其中，有助于对意图进行评估。”

安全保障

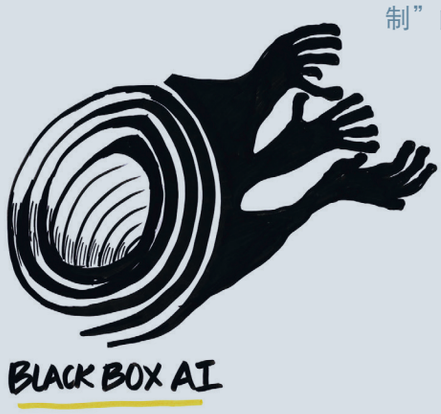
“人工智能需要网络和基础设施保障来防止恶意行为的影响；需要内部应变计划以便停止/覆盖；还需要采用审查机制以确保结果不会对人类造成危险。”

暗中合法化

通过特定的研究和开发路线暗中打破法律禁令和社会禁忌，并不是新近才有的问题。然而，人工智能有可能为处于合法和非法之间的研究推动、加速和创造新的路径，这可能会为不法行为者提供新的机会，开发可能挑战 CBW 条约的制剂和技术。例如，人工智能对现有假设的强化可能会产生无意或模糊的 RCA 部署，松动法律的界限。此外，为实现“无死亡战争”结局而对生物生命过程进行的军事研究，可能会因人工智能技术提供的机会而恢复活力。在这里，“可问责”和“以人为本”的原则可以让我们了解不同的行为者可以如何被治理（或进行治理），以根据新的能力维持法律和社会界限。

黑箱

人工智能的输入、方法和输出的不透明为开发者和用户带来了额外的澄清负担。人工智能系统这种无形、不透明和解释能力不足的性质，会渗透到许多领域的挑战中，并成为其特点。例如，在核实使用指控时，如果使用现成的人工智能工具，但对其确切出处、训练数据和推断方法了解有限，那么责任追踪可能会难以进行。另外，在有多名行为者能够获得所需工具和数据的情况下，使用人工智能工具开发和生产潜在的 CBW，可能会影响归因。在这些情况下，对于理解如何通过更广泛的公开透明和信息共享来监测人工智能算法的创建和使用，“可解释”、“可问责”和“可控制”的原则显得尤为突出。



影响

这四种人工智能焦虑表明，人工智能技术的整合不一定会产生全新的威胁或风险类别，但会嵌入到现有的威胁或风险中。重要的是，根据人工智能与现有技术和治理模式的互动方式，具体场景可符合多个焦虑类别。这一认识要求基于在 CBW 禁令范畴内运作的行为者已经熟悉的背景和挑战来考虑人工智能焦虑。在这些背景下，人工智能的影响可能有：

- 加快研发进程；
- 开辟新的未来研究路径；
- 降低透明度；
- 使信息的提供、关联和后果变得复杂。

下一个阶段是评估现有的治理形式是否能够通过现有能力或通过修正来减轻人工智能对现有和新出现挑战的影响。

负责任的人工智能的四项伦理原则引导我们关注，在设想、描述和最大限度减少人工智能技术负面影响的过程中应当提出什么样的问题。

更多信息：

本研究系由萨塞克斯大学商学院科学政策研究组的哈佛-萨塞克斯项目开展。本研究项目得到英国外交、联邦和发展事务部反扩散和军备控制中心的慷慨资助。感谢 Shaunna McIvor 和 Boaz Chan 为本项目提供的出色协助。

首席研究员： Joshua R Moon 博士
电子邮件： J.R.Moon@sussex.ac.uk

如何在 CBW 背景下更好地管理人工智能？

1. 将人工智能的发展和应用与背景结合，以便更好地理解特定于政策的挑战。
2. 通过确定参与开发和使用目的的行为者，对人工智能进行社会化。
3. 从人工智能焦虑可能会如何阻碍或削弱当前的 CBW 禁止工作出发，对其进行研究。
4. 评估现有的治理行为者和架构，以了解在解决这些人工智能焦虑方面存在的机会和差距，并在可能的情况下对其进行纠正。
5. 重点关注现有的工具和机制，以解决特定于挑战的人工智能焦虑。

改进人工智能政策的后续步骤：

- 建构更详细的人工智能对 CBW 禁令影响的潜在场景，特别是提供技术使用的细节以及不同行为者在这些场景中的角色，有助于更好地理解如何降低风险；
- 了解人工智能在背景和实践中的影响和后果，有助于更深入地了解如何治理人工智能以保持其对社会的积极影响；
- 具体参照详细场景，描述并评估 CBW 禁令的现有治理方式，有助于决策者更好地了解人工智能与不同防御网络的互动，以及如何发展这些网络以适应这一新的科技领域。