"hd Centre for Humanitarian Dialogue

*Mediation for peace*

# 军事系统人工智能行为准则

# 行为准则背景

该人工智能军事系统行为准则草案是人道主义对话中心（HD)亲自和在线召集的中国、美国和国际专家为期两年的磋商过程的产物。磋商过程的目标是确定是否可以就具有重要人工智能组件的武器和相关军事系统达成某些原则和限制,特别是在该领域技术和部署最先进的国际参与者之间。对话的参与者包括来自美国、中国和一个来自欧洲和拉丁美洲的国际代表团,具有军事、外交、情报、武器设计和法律背景的现任学者和前任官员。虽然一些专家参加了以前与先进武器限制有关的联合国会议 例如联合国 CCW 的致命自主武器系统 (LAWS) 政府专家组 但磋商的目的是摆脱公共立场,看看是否谨慎在启用人工智能的军事系统变得如此普遍以至于使未来的限制变得不切实际之前,流程可以找到限制的共同点。

本守则没有定义 LAWS 或限制对武器系统的考虑,而是考虑 AI 对武器以及相关情报和目标系统的影响,因为它们将用于现实世界的冲突。在咨询过程中,专家们参与了强调:

- 证明当事故发生时启用人工智能的系统在意外或意外情况下运行的挑战
  方法;
- 曲解人工智能军事系统行动的风险,以及有人与无人资产之间的信号差异;
- 对向对手提供与测试和评估相关的任何透明度的根深蒂固的偏见
  先进武器系统的进程,即使这种透明度将促进共同商定的目标,例如安全、安保或问责制;和
- 需要保持对武器的人工控制,以防止在瞄准或发射时出现错误或事故,
  特别是在敏感的冲突环境和核领域。

在与官员分享本准则时,HD 中心希望这些概念能够激发国家行为者的思考,即关于未来设计、部署和评估具有重要人工智能能力的军事系统可能达成哪些协议。在安全、安保、防扩散和其他领域确定了共同利益。我们希望在人工智能相关系统变得如此普遍以至于限制变得不切实际之前,促进官方考虑对人工智能相关系统的可能限制。同样,我们希望官方对限制的考虑可以更容易,因为本守则的要素已经被主要国家的国防和安全意识专家考虑和同意,同时考虑到竞争激烈的国际环境和当前的国家安全挑战在现今世界里。

# 前言

1. 本行为准则草案的目的是阐明一套原则,主要国家可以考虑采用这些原则来设计、部署、使用和评估包含一个或多个通过人工智能（AI)产生的重要组件的军事系统。 ),尤其是机器学习 (ML)。国际社会中的其他国家以及相关的非国家行为者也可以同意这些原则是可取的。

2. 这不是一个全面的或具有法律约束力的行为准则,但可以作为一个开始
未来制定此类行为准则的要点。

3. 本行为准则从基于风险的角度审视支持人工智能的军事系统和组件,特别考虑了设计/测试、部署/使用和使用后评估的各个阶段。在确定要纳入的要素时，本行为准则侧重于决策者感兴趣或具有重要价值的主题,参与设计和部署人工智能武器系统的主要国家可能能够在这些主题上

达成共识。

4. 本行为准则的起草考虑了以下目标：

一个。遵守国际法,特别是国际人道主义法（IHL）；

湾。尊重《联合国宪章》的原则；

C。防止生命损失并避免无意冲突或冲突升级；

d。保护国家安全、主权和国防；

e.鼓励基础科学的持续发展和人工智能的生产性民用
和用于人类发展的机器学习应用程序;和

F。确保人类对支持人工智能的军事系统及其行动的后果负责。

5. 本行为准则的重点是支持人工智能的军事系统和组件,因为它们会影响安全关系。本守则涉及和相关的军用人工智能系统和组件的类型包括但不限于武器系统;指挥和控制系统;预警系统和其他系统,以支持可能导致使用武力的人类决策;情报、监视和侦察（ISR)系统;以及其他可能影响国际和平与安全的人工智能系统。

本规范也适用于最初为商业目的设计但随后被用于军事用途的系统。

# 一般原则

6.人工智能的合法使用。各国拥有为民用、经济发展和国家安全开发包括人工智能在内的技术的合法权利。各国应以合理的方式开发和使用支持人工智能的军事系统
符合国际法，不损害国际安全与稳定。各国应考虑发展
以旨在减少平民伤亡和人类痛苦并避免武装冲突的方式操作和使用支持人工智能的系统。

7.合法、合乎道德和正当。各国应将人工智能军事系统的开发和使用置于
最高的法律和道德标准。人工智能军事系统的设计、部署和发射以及相关人工智能设备和技术的开发背后应该有正当理由。军事人工智能研究和开发应支持国际和平与安全并符合国际法，包括具有法律约束力的联合国

文书，以及适用的军备控制和裁军文书。

8.鲁棒性和可靠性。各国应确保支持人工智能的军事系统应可靠、稳健，并按照其设计规范以可预测的方式运行。在部署支持人工智能的军事系统之前，各国应确保对系统的可靠性以及国家根据国际法和指挥官意图开展行动的能力高度信任。

9.人的控制和责任。任何使用支持人工智能的军事系统都应遵守明确的军事指挥和控制线以及人的责任。由于可能导致生死攸关的后果，各国应设计支持人工智能的军事系统，以确保人类对人工智能系统的开发、部署和使用负责，并避免自动化偏见或将所有人类判断拱手让给人工智能系统。人类的判断也是问责制所必需的，因为人类而非计算机系统受法律规定的权利和责任的约束。

10.技术上可限制。人工智能军事系统应在时间和空间领域具有一定的合理限制，以避免误用和意外平民伤亡，并允许终止已通过预定目标的人工智能武器。

11. AI 无武器区。各国不妨考虑限制部署某些形式的 AI 启用或 au
某些地理区域的单调系统（例如，“AI 无武器区”）。示例可能包括同意不在关键民用基础设施（包括水坝、机场）附近部署支持人工智能的军事系统或对其造成损害
或高使用率的商业航道和水道，并且不在该地区部署致命或未经测试的人工智能军事系统
政治和军事紧张局势加剧。

12.禁止某些类型的人工智能武器系统。由于人工智能是一项新技术，各国应考虑
禁止某些类型或领域的人工智能军事应用。可能的系统类型示例
声明的人工智能禁区包括但不限于：

- 小武器和轻武器，因为它们具有重大的扩散风险并且在国与国冲突中的作用不那么突出；
- 核指挥和控制，由于其破坏力和人类对核控制的要求
  发射决定；和
- 致命武器系统，(a) 无视人类控制，或 (b) 具有自主性并能够在不可预见的情况下进化
  或部署后的危险方式。

13.能力。决策者和军方领导层必须对人工智能系统有扎实的了解，包括其能力、局限性和风险。各国应承诺确保其相关的政治和军事领导层理解人工智能系统使用的基础技术，尤其是它们的局限性。军事、技术和法律能力都与那些指挥人工智能武器系统的人相关。

# 设计和开发

## 设计原则

14. 各国应将支持人工智能的军事系统设计为可靠、稳健、可追溯、可治理、无意外偏见、明显符合目的，并以增加透明度和可解释性的方式设计。

系统功能。各国应尽可能在符合其国家安全的情况下披露
假设、限制和对人工智能军事系统性能的综合评估。

15. 各国应为其内部目的全面记录设计、开发、
与人工智能系统开发相关的测试和评估。这应该包括所有代码和数据的正确时间戳和版本控制副本以及对实验平台/架构的访问

在开发过程中使用。各国应考虑共享其系统稳健性的文件
在符合其国家安全的范围内。

16. 各国应承认决策速度和质量之间的紧张关系是人工智能军事系统设计的一个要素。各国应在不牺牲决策质量或人类判断作用的情况下开发提高决策效率
的系统。旨在提高人工智能军事系统决策效率的设计参数不得妨碍人类控制或其他法律或

技术要求。

17. 各国应建立并遵守人工智能武器系统的性能指标，包括准确度
racy、可靠性、稳健性、安全性和其他性能指标。在正常情况下，攻击性人工智能武器系统的安全性要求应该高于防御性人工智能武器系统。

## 故障保护和信令

18. 各国应设计其支持人工智能的武器系统，以包括嵌入式故障保护功能，以便在发生故障时
安全事件发生时，它们将可能导致平民伤害或意外军事伤害和潜在意外升级的事故的可能性和后果降至最低。这些功能应该转移对 AI-en 的控制

在可能的情况下，将有能力的武器系统提供给人类。这对于启用人工智能的致命武器系统尤其重要，各国应考虑设计嵌入式自毁或

其他停用机制。

19. 在设计人工智能军事系统时，各国应考虑与自主和人工智能系统如何可能无意中发出或表明敌对意图有关的独特问题。各国应考虑设计机制，使系统能够以
清晰、人类可识别的方式发出非敌对意图的信号。系统应该是钻机

经过严格测试，以确保仅在有意时才显示可能被理解为敌对意图的行为。
当此类系统投入使用时，各国应考虑人工智能系统特有的额外信号和通信机制何时将互惠互利，以减少意外升级或事故风险。

## 测试和评估

20. 各国应确保对支持人工智能的军事系统进行全面测试、评估、验证和验证
部署前的实际操作条件。

21. 各国应建立适当的制度框架来测试和评估人工智能系统,其中
包括通过咨询相关专家及早考虑法律和技术问题。

22. 为了遵守国际条约和习惯法,各国应对新武器进行法律审查,以确定它们在某些或所有情况下是否会违反其根据国际法承担的义务。这一义务同样适用于支持人工智能的武器系统。因此,各国应确保法律审查尽早成为人工智能系统测试和评估的一个组成部分。

23. 各国应善意地设计测试环境和测试参数,以提供系统性能的真实印象及其在符合适用法律和道德标准的情况下使用的能力。在没有关于人工智能武器系统的可靠性和可预测性的国际规则或标准的情况下,各国应考虑建立自己的最低标准,以此衡量正在测试和审查的系统的性能。

24. 各国应确保测试和评估,包括法律审查,是一个持续和反复的过程,应在开发和/或采购和采用阶段尽早启动,并在系统的整个生命周期中重复,以包括考虑持续部署,例如当现场系统收到定期 AI 算法或模型更新时。

25. 各国应建立监测系统功能变化的程序,以便能够决定是否以及何时需要额外或补充审查,以确保遵守国际人道法和其他相关规则
的国际法。

# 部署和使用

## 监督

26. 各国应确保负责使用人工智能武器系统的人应采取一切必要措施
采取预防措施,限制对军事目标和战斗人员的攻击,避免或尽量减少附带的平民生命损失和平民财产损失。根据个别攻击的情况,必要的预防措施可能包括但不限于:

一个。对经营地域范围的限制;

湾。对操作的时间范围/持续时间的限制;

C。对目标类别的限制;

d。对目标识别标准的限制;和

e.需要人工操作员进行目标批准或任务监督。

27. 各国应确保人类有责任对武力的使用作出判断,特别是对可能导致人命丧生的武力使用的决定。各国应确保决不将生死决定权委托给机器。

28. 在网络空间部署支持人工智能的军事系统可能会带来独特的挑战,尤其是在归属、问责和人为控制方面。尽管如此,在网络空间部署支持人工智能的军事系统应遵守与其他领域相同的原则和限制。

29. 由于核武器的破坏潜力,各国必须确保人类继续控制核发射决定。各国应确保核指挥和控制系统的设计如此

启动核发射需要积极的人类行动,并且技术事故不会导致
一次不经意的发射。例如,早期预警系统应该与核发射系统在物理上分开,在潜在的核武器发射序列的每一步,人类都应发挥关键的监督作用。

同样,无人车辆(例如无人机)绝不应配备核武器或用作核运载平台,因为存在事故风险和对核武器失去控制的风险。

## 能力和责任

30. 各国应确保具有军事专业知识的人员理解和批准设计要求。

31. 各国应确保关于使用人工智能军事系统的决定仍由军事指挥官负责,并仅由具有适当权力的人作出。各国应确保指挥官具有解释、解释和使用其指挥下的人工智能军事系统的技术能力,包括评估不确定性对人工智能军事系统性能的影响、错误的可能来源和规模部署,以及此类错误的成本。各国应提供能够支持指挥官主导决策的人类专家。

# 使用后评估和问责制

32. 各国应确保在部署支持人工智能的军事系统导致发生任何意外、非法、不道德或意外活动时进行调查。各国不妨考虑在符合其国家安全的范围内分享所有人工智能相关事件的事实和结果。

33. 负责部署支持人工智能的军事系统的国家应确保在意外、非法、不道德或意外使用后尽快采取任何必要的补救、纠正和再培训行动，并在所有情况下在系统之前采取问题被重新投入使用。

34. 各国应在符合其国家安全的最大可能范围内，与其他有关各方分享有关纠正和再培训步骤的信息，包括那些涉及意外、意外、不道德或非法使用事件的各方。

# 信息共享、信任建立和国际合作

## 透明度和识别

35. 各国应考虑对自主系统的功能采取透明度措施，以避免计算错误和事故的风险，并允许在事后调查时对安全措施进行核查。

36. 为减少起源和指挥方面的不确定性风险，各国应承认并遵守现有的国际法律义务，将独特的国徽贴在支持人工智能的军事系统及其相关组件上。各国应考虑如何确保此类标志可识别，包括人类对手和其他支持人工智能的系统。

37. 各国不妨考虑识别并向其他各方传达已部署人工智能军事系统的各种特征的方法，包括但不限于：车辆/船只/飞机有人驾驶（“有人驾驶”）或无人居住的程度（“无人”）；系统自主功能的级别和类型以及人工监督的程度；以及系统是否布防/未布防。这可以通过各种识别和标记措施来实现，包括但不限于物理标记，如油漆、旗帜、灯或其他标志和/或电磁装置，如射频信标

或留言。

38. 各国不妨考虑为人工智能军事系统开发通用信号
撤防或恢复到故障安全模式。

## 信息共享

39. 各国应在符合其国家安全的范围内，共享其程序和结果
(a) 他们的法律审查和 (b) 他们对人工智能军事系统的测试和评估。

40. 各国应考虑采用一套通用的性能指标和评估标准来评估人工智能系统的有效性、稳健性和安全性。各国应确保并证明其支持人工智能的军事系统达到商定的最低性能。

41. 各国应考虑是否有必要修订内部交战规则以适应交战
军事人工智能系统，并可能希望考虑共享可能减少事故或误判的此类修正的各个方面。各国不妨考虑在记录和共享与支持人工智能的军事系统相关的指挥结构方面是否存在互惠互利。

42. 各国应就机器学习可解释性和鲁棒性的基础科学交换研究，并应在符合其国家安全的范围内，考虑交换有关应用的信息，以促进实现确保人类安全的目标。监督。长期以来，公开发表 AI/ML 系统学术进展的传统可以成为透明度的重要来源，在国家安全阻止共享应用系统细节的情况下提供可以共享的细节水平。

## 国际合作

43. 各国应就人工智能的军事理论、能力和用途建立定期对话，以减少误解
并解决有关特定人工智能系统及其属性如何破坏所有国家的安全和保障的问题。对于在附近部署支持人工智能的军事系统的国家，各国应在符合其国家安全的情况下制定危机沟通的可靠方法。

44. 各国应在反扩散努力上进行合作，以提高关于向国家或非国家行为者扩散具有潜在危险的人工智能系统的透明度。具体措施可能包括标记国家所有权系统、共享有关国际武器转让、技术保障的信息和/或同意不向某些参与者扩散某些人工智能系统或应用程序。

45. 各国不妨考虑创建、支持和使用一个中立的、技术上称职的第三方组织化，可以：

一个。推荐可靠性、稳健性、保证和验证的国际标准和最佳实践
支持人工智能的军事系统；

湾。为支持人工智能的军事系统提出一套国际认可的建立信任措施 (CBM)。

C。为支持人工智能的军事系统提出一套通用的测试和评估指标和标准；

d。以支持保护敏感信息的方式参与测试、评估、验证和确认
信息和能力；

e.建立培训计划，为指挥官及其团队提供技术和法律知识
与对支持人工智能的军事系统进行负责任的指挥和控制有关；和

F。澄清对 AI 能力或学说的不明智表述（所谓的“神话”）。

Centre for Humanitarian Dialogue

*Mediation for peace*