翻译

以下文件是“人工智能原则：美国国防部关于人工智能道德使用的建议”的中文译文，最初由美国国防部(DoD)国防创新委员会(DIB)于2019年10月31日以英文出版。为引起全球受众对美国国防创新委员会人工智能原则的关注，美国安全和新兴技术中心(CSET)特委托进行中文翻译。美国安全和新兴技术中心隶属华盛顿特区乔治城大学，是一个无党派智囊团，主要研究新兴技术的安全影响。美国安全和新兴技术中心不隶属于美国国防创新委员会或国防部。美国安全和新兴技术中心的翻译不是美国政府对国防创新委员会人工智能原则的官方翻译。美国安全和新兴技术中心对国防创新委员会人工智能原则的翻译并不构成对这些原则的认可。

查看本文档英文原件，请访问以下网址：
https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

人工智能原则：
美国国防部关于人工智能道德使用的建议

美国国防创新委员会

I. 目的

美国国防部（DoD）高层要求，国防创新委员会（DIB）应为国防部（DoD）提出人工智能（AI）道德原则，用于战斗和非战斗目的人工智能的设计、开发和部署。美国国防创新委员会在国防部现有道德、法律和政策框架的基础上，针对迅速发展的人工智能领域的复杂性，力图制定与国防部使命相一致的原则，遏制战争，确保国家安全。本文档是对国防创新委员会项目的总结，内容包括简要背景、国防部持久人工智能道德原则概要、人工智能道德原则提议以及为促进国防部采纳原则、推动人工智能安全性和稳健性目标实现而给出的建议。国防创新委员会的完整报告中列有详细说明，并针对这些建议提供了更广泛的历史、政策和理论背景。可登陆网站innovation.defense.gov/ai查找相关信息。

国防创新委员会是一个独立的联邦咨询委员会，为国防部高层提供建议；该委员会并不代表国防部。本报告试图在国防部内部和整个外部社会引发一次发人深思的对话。国防部全权负责确定如何以最佳方式执行本报告中提出的各项建议。

II. 背景

国防部为何要优先制定人工智能道德标准？人工智能正在改变人类社会，并影响着人类务业、社交和战争的方式。[1]从许多方面来看，人工智能领域正处于发展期。近年来，计算领域的快速发展推动了数十年来一直停留在理论层面的人工智能应用向前发展。然而，人工智能的实际应用往往是脆弱的，人工智能发展学科也在不断发展，其使用规范尚未形成。在全球范围内，公共部门、私营企业、学术界和民间团体就人工智能的前景、危机和恰当使用展开了持续的辩论。国家安全就是这些辩论的一个关键方面。目前，人工智能处于早期阶段，现在是时候就人工智能的发展规范及其在军事环境中的使用展开严肃讨论，*尽早*未雨绸缪。

我们的敌对者和竞争对手已经认识到人工智能的变革潜力，他们在积极参与全球挑衅活动的同时，还在大力投资于人工智能，使其力量现代化。中国公开、坚定承诺，将在2030年之前成为人工智能领域的全球领导者，并正花费数十亿美元获取优势地位。[2]俄罗斯同样在人工智能应用上投入巨资，并在实战场景中测试这些 系统。[3]联合人工智能中心(JAIC)主任Jack Shanahan中将清楚地认识到，美国国防部未来将具有压倒性优势:“我不希望看到未来我们的潜在对手拥有一支人工智能力量的部队，而我们没有……我没有那么多时间做决定。人工智能派上用场的时间可能只有几秒或几微秒。”[4]

《2018年美国国防战略》(NDS)呼吁，应加大对人工智能和自主系统的投资，为美国提供具有竞争力的军事优势。[5]与国家国防战略一致，国防部的人工智能战略声称国防部正在努力利用人工智能的潜力“积极变革所有部门职能,支持和保护服役人员,保卫美国公民,捍卫盟友和合作伙伴,提高负担能力、效率和速度。”[6]人工智能战略进一步指出，它“将以合法

---

[1]请参阅 2018年美国国防部人工智能战略报告：利用人工智能推动安定繁荣发展； 和2017年12月的美国国家安全战略报告。

[2]Savage, Luiza Ch., 和 Nancy Scola.“‘我们超支了。 我们正被赶超’:美国是否正在把 人工智能的未来拱手让给中国?”政客。2019年7月18日

[3]Konaev, Margarita, 和 Samuel Bendett.“俄罗斯人工智能战争:近在咫尺?”War on the Rocks 网站。2019年7月31日

[4] 请参阅“Lt. Gen Jack Shanahan关于美国国防部内部人工智能举措的媒体发布会。”美国国防部,2019年8月30日。

[5] 《2018年美国国防战略》。

[6]美国国防部， 2018年美国国防部人工智能战略报告：利用人工智能推动安定繁荣发展。第4页。（以下简称为“人工智能战略”）。

和合乎道德的方式阐明使用人工智能促进人类价值观的愿景和指导原则。”[7]国防部强调，有必要加强与学术界、私营企业和国际社会的合作，“促进军事领域的人工智能道德标准和安全”，并强调了其对开发和部署人工智能的道德和负责任的承诺。诚然，“*引领军事领域道德标准和人工智能安全*”是该战略的五大支柱之一。

美国国防部并不是第一个认识到为人工智能的开发和使用制定道德原则重要性的组织，但国防创新委员会注意到，许多现有的此类原则产生了很多问题，而不是关于人工智能使用限制的答案。在事关国家安全的高风险领域中，需要注意的是，美国发现自己正与威权主义国家进行技术竞争，而后者正以有悖于民主国家所期望的法律、伦理和道德规范的方式追求人工智能应用。我们的目标是依据国防部的长期道德框架确立这些原则——这一框架经受住了几十年来出现和部署的新型军事专用或两用技术，反映了人们的民主规范和价值观。

然而，我们承认，人工智能的独特性和脆弱性需要采用新方法，才能解决其潜在的意外负面后果。[8]在国家安全领域，考虑是否使用新兴技术时，关键是分析未预料到的行为。意外后果的不确定性并非人工智能独有；而是一直并始终与所有技术工程领域相关。例如，在土木工程和化学工程的相关领域成为正式学科之前，人类就开始建造桥梁和建筑物，对能源和物理材料进行操作，导致了许多无法预见的事故。[9]如今，尽管在如何利用人工智能最大化社会效益、减少意外后果方面缺乏共识，“人类仍在推进建设涉及机器、人类和环境的社会规模、推理和决策系统。”[10]因此，人工智能道德原则可以让如何以安全负责任的方式推进仍处于新生阶段的人工智能领域的讨论更加充实。土木工程和化学工程等工程学科发展其道德行为文化的方式与此类似。通过确定并维护其从业人员的技术卓越性义务做到了这一点。

全世界正在展开关于人工智能应用于国家安全的合适时机和条件的大讨论,需要注意的是,与竞争对手和敌对者将人工智能用于不符合美国国防部价值观的目的相比，国防部对美国人民及其盟友有责任维护其战略和技术优势。我们希望随后的原则可以为这项工作提供指导，同时国防部继续对合法和负责任的行为进行永久承诺；通过变革并调整人工智能领域的道德标准建立现有的道德基础；帮助塑造有关人工智能使用的国际新规范；并确保在利用好此项技术优势的同时减轻其潜在危害。

多元化视角。为了帮助国防部应对这一挑战，国防部高层要求国防创新委员会集思广益，就国防部可能采用的人工智能道德原则以及如何将这些原则整合到现有道德框架中执行任务提出建议。

国防创新委员会进行了一项为期15个月的研究，旨在使研究更加稳健、包容和透明。该过程包括收集线上线下的公众评论；在主要大学举行两场公开聆听会；并与来自学术界、工业界、民间团体和国防部的数十位课题专家进行三次专家圆桌讨论。圆桌会议的参与者包括获得图灵奖的人工智能研究人员、退休的四星上将、人权律师、政治理论家、军备控制活动家、科技企业家等。此外，国防部还成立了一个非正式的国防部原则和道德工作组，其中包括来自亲密伙伴国家的政府官员，协助国防创新委员会收集信息并促进合作。国防创新委员会还举行了一次机密的“红队”会议和一次桌面演示，在现实政策场景中对照当前有关人工智能在战争中的潜在应用情报，对这些原则进行压力测试。经过对100多名内外

[7]美国国防部（n6)8。
[8]有关人工智能较其他技术表现出何种不同的道德挑战的详细描述，以及对这些原则的进一步深入讨论，请参阅支持文档。
[9]有关人工智能作为潜在新工程学科的内容，请查看Michael Jordan博士的评论。
[10] Ibid。

部专家建议的仔细考量，并考虑到出广泛的观点和近200页的公众评论，[11]国防创新委员会制定了这一套人工智能道德原则及相关建议，供国防部长审议。这些专门针对人工智能的原则与国防部现有的道德、法律和政策框架相结合，将用于指导国防部的活动。

人工智能定义。人工智能是一门极为广泛的学科，不同的用途有不同的定义方法。为清楚起见并为了指导该项目，我们使用这一术语表示*各种执行目标明确任务的信息处理技术，以及在执行该任务时进行推理的方法*。在涉及到更广泛的情况时，我们使用术语人工智能(AI)；但在专门提及机器学习(ML)系统时，我们称之为ML。此外，我们会使用术语"AI系统"表示在整个系统内或某个子系统中具有人工智能组件的系统。[12]

我们使用人工智能的这一定义，是因为它符合国防部在过去40年观察、开发和部署人工智能系统的方式。该定义允许我们在遗留系统和使用机器学习的新系统之间进行更加细化的区分。使用该术语可以强调，国防部在以下现有道德框架下较早地开展了重要的人工智能工作。

我们还区分并明确了*人工智能不同于自主*。尽管某些自主系统可能在其软件架构中使用人工智能，但并非总是如此。例如，美国国防部第3000.09号指令强调了武器系统的自主性，但既没有强调人工智能本身，也没有涉及与武器系统无关的人工智能能力。[13]

最后，人工智能本身既不积极也不消极。[14]它是一种类似于电力、内燃机或计算机的赋能能力。因此，人工智能是推动还是破坏我们为世界安定繁荣而作出的努力，是由人类决定的。

III. 美国国防部现有道德框架和价值观

在非战斗和战斗应用方面，人工智能现在和将来都是美国国防部一项基本能力。事实上，人工智能只是国防部使用的众多技术之一，与国防部已安全、成功部署的其他大型技术复杂系统类似，同样面临着测试和现场挑战。在所有情况下，国防部及其服务运作[15]所依据的基于价值的框架以及国防部和美国民间团体运作所依据的法律构架，包括《美国宪法》、《美国法典》第10章和其他适用法律，都为各种人工智能道德原则提供了发挥作用的基础。下文概述的人工智能专用原则源自*现有并被广泛接受的道德和法律承诺*。

这一完善的道德框架及其附带的价值观为国防部如何制定并执行决策提供了指导。各种声明、政策文件和现有法律义务都反映了这一点。正式协议包括《战争法》和现有的国际条约，而国防部长的许多国防部备忘录均强调了道德行为对武装部队的重要性。从孤立和整体角度来看，这些证据均表明，国防部的道德框架反映了美国人民和美国宪法的价值观和原则。[16][17]

---

[11]参见国防创新委员会网站公共评论的链接和视频。

[12]一些文档偏好于使用短语"支持人工智能的系统"，但在我们的应用中，人工智能系统或支持人工智能的系统都是相同的。

[13]请参阅美国国防部指令3000.09，其中将自主武器系统定义为"激活后即可以选择和攻击目标的武器系统，无需操作员进一步操作。这包括由人类监督的自主武器系统，其设计旨在让人类操作员控制武器系统的操作，在系统激活后，无需进一步人工输入即可选择和攻击目标。"

[14]这并不是说技术和人工智能是价值中立的。像人工智能系统这样的技术产物反映了人类设计者、开发人员和用户的价值，以及他们所居住和制定决策的社会的价值。

[15]参见美国国防部核心价值观、美国空军核心价值观、美国陆军核心价值观、美国海军和 海军陆战队核心价值观、和美国海岸警卫队核心价值观。

[16]《战争法》是指适用于战争并为解决武装冲突行为的合法性提供完善框架的国际法律体系。

[17]参见国务卿Mark Esper的备忘录和前国务卿James Mattis的备忘录。

国防部尤其要遵守《战争法》，因为这是国际公认的所有武装部队行为的法律指南。[18]对美国而言，这一法律体系包括美国已接受的条约，比如1949年的《日内瓦公约》；各国出于法律义务意识而采取的普遍和一贯做法而产生的习惯国际法；以及《国防部战争法手册》。[19]

在武装冲突中使用新技术时，可以适用现有的《战争法》规则。[20]例如，2015年的《国防部战争法手册》反映了2012年与国防部指令3000.09相关的工作，详细说明了《战争法》如何适用于武器系统中的自主功能。[21]《战争法》的基本原则为战争期间的行为提供了一般性指导，不再适用更具体的规则，从而提供了诸如人工智能等新兴技术带来的法律和道德新问题的审议框架。例如，如果将人工智能应用到武器中，则对此类武器必须进行审查，确保与现有法律要求（例如，对武器不得造成不必要的痛苦或本质上无差别的痛苦的要求）相符。此外，根据《战争法》，指挥官和其他决策者必须本着诚意并根据他们所掌握的信息和当时的情况作出决策。使用人工智能支持指挥决策符合《战争法》的义务，包括采取可行的预防措施，减少对平民和其他受保护人员和物体造成伤害风险的义务。[22]

国防部拥有执行《战争法》的健全流程，包括培训、法规和程序、举报涉嫌违规事件、调查和审查事件、以及适当的纠正措施。[23]为补充和促进这些措施,国防部在过去半个世纪中已投入数千亿美元，确保武器系统和平台的安全性和可靠性,从而制造出更精确、更准确的武器,减少平民伤亡,保护民用基础设施,同时实现军事目标。此外，国防部不断鼓励转变培训方式，遵守这些标准并以负责任的方式使用这些工具。

还有一个示例值得注意: 美国核动力战舰自下水以来已安全航行了50多年，没有发生任何核反应堆事故，也没有释放出任何损害人类或海洋生物健康的放射性物质。核反应堆依靠核能安全运行超过1.62亿英里的里程，积累了超过6900多个反应堆安全运行年限。[24]

我们强调这个例子不是为了说明国防部应该将人工智能应用于其核能企业。相反，我们强调的是努力创造一种安全和精确的文化，完全遵守国防部为复杂系统工程建立的标准。这是国防部的一个重要基础，可以增强新型技术复杂工作（例如人工智能的开发和部署）的道德文化。[25]

IV. 美国国防部人工智能道德原则

我们重申，人工智能必须在现有的国防部道德框架范围内使用。在此基础上，我们提出专用于人工智能的原则，并注意这些原则适用于战斗和非战斗系统。人工智能是一个快速发展的领域，目前任何开发或研究人工智能系统或拥护人工智能道德原则的组织都无法声称已经解决了以下原则中包含的所有挑战。但是，国防部应设定人工智能系统的使用目标，具体如下:

1. **负责。**人类应进行适当的判断，并对国防部人工智能系统的开发、部署、使用和结果负责。

[18]虽然《战争法》是国防部的重要指南，但并不适用于国防部应用人工智能可能出现的所有情况。我们将在支持文档中更深入地描述这些适用情况。
[19]请参阅国防部战争法手册。
[20]这些规则依据五项基本原则（也是《战争法》的基础）:军事必要性、人道性、相称性、区别性和荣誉性。
[21]美国国防部（第18号）385.
[22]美国国防部（第18号）§ 5.2.3.2和5.3。
[23] 参见美国国防部指令2311.01E。
[24]美国国家核安全局和海军部情况说明书，“美国海军核 推进计划”。2017年9月。
[25] 有关美国国防部现行道德框架各方面的详细说明，请参阅有关支持文档。

2. **公平。**国防部应采取成熟措施，避免在开发和部署战斗或非战斗人工智能系统时出现会对人类造成伤害的意外偏差。
3. **可追溯。**国防部的人工智能工程学科应足够先进，确保技术专家能够对其人工智能系统的技术、开发过程和操作方法有适当的理解，包括透明和可审计的方法、数据源、设计程序和文档。
4. **可靠。**国防部人工智能系统应具有明确的使用范围，并且应在人工智能使用范围内测试并确保这些系统在整个生命周期中的安全性和耐用性。
5. **可控。**国防部人工智能系统的设计和工程应能实现其预期功能，同时具备探测和避免意外伤害或破坏的能力，并能对意外升级或其他行为的已部署系统进行人为或自动脱离或停用。

V. 建议

在制定人工智能道德原则的过程中，国防创新委员会已经确定了有助于阐明和实施这些原则的有用行动。国防部将最终确定希望采用的确切原则，但无论所批准原则的确切性质如何，以下十二项项建议将为此提供支持:

1. **通过国防部官方渠道将这些原则正式化。**联合人工智能中心应向国防部长提出适当的沟通和政策发布建议，确保这些人工智能道德原则的持久性。
2. **建立国防部范围的人工智能指导委员会。**国防部副部长应建立一个向其汇报的高级委员会，确保国防部人工智能战略的监督和执行，保证国防部的人工智能项目符合国防部人工智能道德原则。维护人工智能道德原则要求国防部将其整合到决策的许多基本方面，从概念层面(如DOTMLPF[26])到更具体的人工智能相关领域(如数据共享、云计算、人力资本和IT政策)。
3. **探索和发展人工智能工程领域。**研究与工程部副部长办公室(OUSD(R＆E)和服务实验室应通过借鉴国防部长期形成的良好工程实践、更广泛地与人工智能研究社区接触，为早期研究人员提供具体机会，并将国防部的安全和责任传统应用于人工智能领域，支持人工智能工程学科的发展和成熟，将人工智能技术整合到更复杂的工程系统中。
4. **增强国防部培训和劳动力计划。**每个部门、作战司令部、国防部长办公室、国防机构和国防野外活动都应制定与各自国防部人员相关的人工智能技能和知识的培训和教育计划。[27]从初级人员到人工智能工程师再到高层领导，均应广泛提供各种人工智能培训计划，并利用现有的数字内容，结合高层和专家的定制指导。[28]至关重要的是，初级军官、入伍服役人员和平民需要在其职业生涯的早期接受人工智能的培训和教育，国防部应通过正式的专业军事教育和实际应用为其整个职业生涯提供持续学习机会。
5. **投资研究新型人工智能安全。**政策副部长办公室和网络评估办公室应在了解人工智能时代竞争和威慑新方法方面进行投资，特别是与网络安全、量子计算、信息对抗或生物技术等其他领域相结合时。需要重点关注的领域包括人工智能的竞争和升级动态（避免危险扩散）、对战略稳定性的影响、威慑的选择、以及国家之间达成正

[26]这一术语指的是国防部的条令、组织、培训、材料、领导和教育、人员和设施。
[27]请参阅美国国防部人工智能战略第14页("提供全面的人工智能培训和培养劳动力人才")。
[28] Ibid。

和承诺的机会。

6. **投资开展研究，增强再现性。**研究与工程部副部长办公室应向提高人工智能系统再现性的研究进行投资。人工智能社区在这一领域所面临的挑战为国防部提供了一个机会，可帮助了解复杂人工智能模型的运行方式。此项工作也将有助于解决所谓的人工智能"黑匣子"问题。[29]
7. **确定可靠性基准。**研究与工程部副部长办公室应该探索如何以最佳方式制定适当的基准，以此衡量人工智能系统相对于人类的性能。
8. **加强人工智能测评技术。**国防部应在发展测试与评估办公室(ODT&E)的领导下，使用或改进现有的国防部测试、评估、验证和确认程序，并在必要时为人工智能系统创建新的基础架构。这些程序应遵循国防创新委员会软件获取与实践(SWAP)研究中详细介绍的T&E软件驱动指南。[30]
9. **制定风险管理方法。**联合人工智能中心应基于人工智能在道德、安全和法律风险方面的因素，创建国防部人工智能使用分类法。[31]这种分类法应鼓励在低风险的应用程序中迅速采用成熟技术，并在不尽成熟和/或可能导致更严重的不良后果的应用程序中强调并优先考虑更多的预防措施和审查。
10. **确保人工智能道德原则的正确执行。**联合人工智能中心应评估这些原则和任何相关指令的适当实施情况，作为《2019年国防授权法》第238条或未来其他指令所要求的治理和监督审查的一部分。
11. **扩展研究，了解人工智能道德原则的实施方法。**研究与工程部副部长办公室应联合服务研究办公室，组建一个关于人工智能安全性和稳健性的多学科大学研究计划(MURI)项目。该计划应作为在这些领域中进行持续性基础和学术研究的起点。[32]
12. **召开关于人工智能安全性和稳健性的年度会议。**鉴于人工智能领域的迅速发展，联合人工智能中心应当召开一次年度会议，审查内外部不同群体对人工智能安全性和稳健性中道德规范的看法。

VI. 结论

这些原则既不是为了掩盖有争议的问题，也不是为了限制国防部的能力。相反，我们希望这些原则能够通过人工智能系统和与国防部使命相一致的行动制止战争，保护我们的国家。此外，这些原则与现行的政策框架、《战争法》、国内法(如《美国法典》第10篇)以及反映民主价值观的持久道德规范相一致。国防创新委员会提出这些人工智能道德原则供国防部考虑，其中包括实施建议。毕竟，道德不仅是思想的集合，而且是一系列有目的的活动和持续进行的探究。

在研究技术和国防问题的三年中,我们发现美国国防部是一个具有深厚道德标准的组织,这并不是因为国防部可能会发布任何单一文件,而是因为他们凭借根深蒂固的信仰对生活和

[29]"黑匣子"问题是指由于算法评估各种输入时采用了许多隐藏的或无法解释的方式，人类无法理解人工智能系统如何得出特定结论，往往会导致人类对人工智能系统缺乏信任。
[30]关于国防部现有的人工智能测试和评估能力以及改进建议的更多细节，请参阅本报告附录四。
[31] 美国国防部高级研究计划局（DARPA）资助了2014年美国国家科学院的一项研究，该研究得出了一份报告，名为《*新兴和现成技术与国家安全: 道德、法律和社会问题解决框架*》，建议建立一个风险评估和缓解框架，解决针对国家安全目的的新兴技术研究所带来的道德、法律和社会问题。
[32]请参阅美国国防部人工智能战略第15页("投资弹性、稳健、可靠和安全人工智能的研发")。美国国家人工智能研发战略规划：2019年更新，尤其是战略1("对人工智能研究的长期投资")和战略4("确保人工智能系统的安全性")。

工作做出了持续承诺，有时甚至是战斗和死亡。这些价值观必须成为公开讨论和批判性思考的主题，保持相关性和真实性。在人工智能领域不断发展的同时，美国国防部仍继续对美国法律、《战争法》和民主价值观作出承诺。我们提出这些建议，希望能帮助国防部在人工智能等新兴技术背景下对现有承诺的解释进行重要的讨论。