SECURITY 2040

人工智能对于核战争风险意味几何?

EDWARD GEIST | ANDREW J. LOHN

视角
关于实时政策问题的专家见解

RAND CORPORATION

This is a Chinese translation (simplified characters) of *How Might Artificial Intelligence Affect the Risk of Nuclear War?* PE-296/1-RC

当今世界的核平衡依赖于几个不可持久的条件。计算能力和数据可用性的进步，让机器能够完成一度需要人类参与或被视为不可能的任务。这种人工智能（AI）也许会带来新的军事力量，从而引发军备竞赛，抑或增加国家在危机中有意或无意间动用核武的可能性。[1]兰德公司召开了一系列研讨会，集众多人工智能与核安全领域的专家于一堂，共同探讨在2040年之前人工智能会如何演变成一股稳定或不稳定的力量。

人工智能对核战略的影响，既取决于敌方对其军事应用能力的认知，也同样取决于其实际能力。例如，一个国家要发展出定位和锁定敌方所有核武发射器的能力，在技术上极具挑战性，但这项能力却能产生出巨大的战略优势。因此，各国都垂涎三尺，不顾一切技术困难追求这项能力，即使这样做有可能惊动对手，增加冲突的可能性。从技术上说，高级人工智能仍难以克服源自数据局限和信息论论证的障碍，但跟踪和瞄准系统只需要被视为具备这项能力，便可产生破坏稳定的作用。一种近乎实现的能力，也许比已经实现的能力更加危险。

人工智能的发展轨迹，连同辅助信息技术和其他方面的进步，将对未来25年的核安全问题产生重大影响。人工智能技术或延续近年来快速进化的趋势，抑或在现有技术成熟后趋于稳定。一些理论家认为，机器在未来某个时候可能会发展出提高自身智力的能力，形成“超级智能”，具备人类无法理解或控制的能力，但对于人工智能的进化之路（包括形成超级智能的可能性）仍然众说纷纭。有人设想取得初步突破之后会遭遇挫折；有人则猜测会走循序渐进的道路。

上述两种极端情况与核战争的前途关系不大。发展停滞（又称为人工智能的寒冬）只会导致当前核安全环境的微末变化。若超级智能成为现实，人工智能将令整个世界面目全非，在此过程中人类或将得到挽救或将走向毁灭。另外两种情况下，人工智能的发展取得巨大进步，实现许多新的功能，但至少在某些方面仍逊于人类、

容易出错，这似乎得到专家群体的更多认可，但对于这些功能给国家安全带来的影响，专家们莫衷一是。有些“乐观派”倾向认为，要达到能够执行某些足以打破核平衡的任务的人工智能是非常困难的，不大可能实现。相反，“谨慎派”则认为，人工智能能够执行某些任务，但不应该用于核战争的任何领域。第三种属于“颠覆派”，他们强调敌方篡改、误导、牵制或以其他方式欺骗人工智能的能力，这本身也可以起到稳定或打破稳定的作用。

研讨会上讨论了一个例子：作为决策支持系统的人工智能。即使不直接连接到核武发射器上，但人工智能仍可以就冲突升级问题为人类提供意见。鉴于人工智能在日益复杂和不甚明确的任务中取得持续进步，我们有理由相信，这项能力——至少在决策过程中的某些环节能够在2040年前实现。谨慎派担心，这项能力可能在还不可靠或尚不清楚其局限性的情况下就派上用场。然而，如果人工智能顾问经证实是有效的，则可以减小人为错误的可能性，做到彻底透明化，从而降低误判风险，提升稳定性。但许多专家担心，敌方会通过黑客攻击、破坏训练数据或操纵输入因素等手段，颠覆本领高强的人工智能。

在未来几十年维持战略稳定，需要在多极世界背景下重新审视威慑理论的理论基础。要做到有效的威慑，

即使不直接连接到核武发射器，人工智能仍可以就冲突升级问题为人类提供意见。

我们必须抗衡受人工智能进步所推动而快速迭代的各种能力。关键性的考量包括：实际能力的影响，相关能力的可预见潜能（不论是否存在），以及这些能力过早应用或出错性（特别是因敌对行动所致）。只要谨慎行事和一定的前瞻性思考，这些风险是有可能识别和减轻的。

未来核平衡重大变化的线索

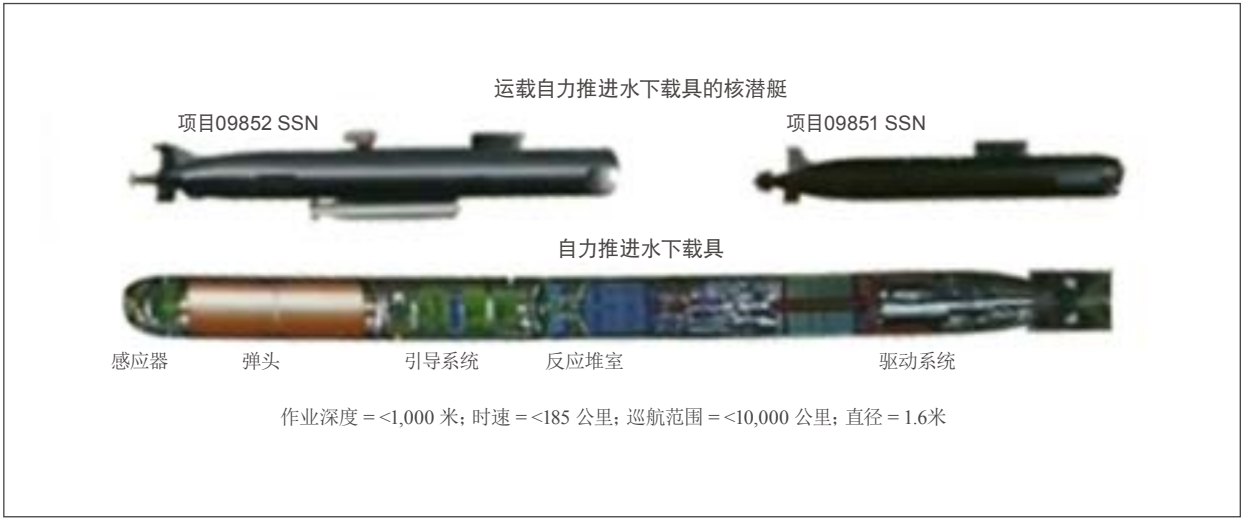
2015年11月，俄罗斯透露正在研发终极“杀手机器人”：核动力海底无人机，用于运载大型热核弹头。俄罗斯电视台在一次“意外”泄密中透露了这种可怕武器的存在，而大多数西方观察家认定这是故意为之。电视镜头有一瞬间停留在一张给总统普京的看似保密的简报幻灯片上，内容描述“海洋多用途系统斯塔图斯-6”。斯塔图斯-6外观像巨型鱼雷，由小型核反应堆提供动力（见第3页的图片），其巡航速度和范围在海洋里几无敌手，能够突破敌人的防御（Sutyagin，2016年）。假设美国先发袭击克里姆林宫，这部无人机将从俄罗斯北极地区的潜艇发射，以大约100公里的时速穿越海洋，同时自动避开反

潜防御系统，向美国海岸线投下致命的弹头。由于在水下通信困难，无人机需要具备一定程度的自主能力，而正是由于人工智能的进步，这直到最近才有可能实现。[2]

斯塔图斯-6不仅是人工智能的实际应用，还反映了人工智能对核威慑的潜在影响——利用报复威胁来阻拦对方攻击一个国家或其盟国。[3]在美国的核打击瞄准能力和导弹防御系统面前，俄罗斯报复力量的切实有效性堪忧，因此，核武无人机是俄罗斯领导人心底忧虑的最新体现。现今的俄罗斯无法抗衡这些军事力量，因而希望利用人工智能确保其威慑力量的切实有效性。克里姆林宫持续探索将人工智能诉诸于军事用途的创新方式，该计划有可能在2040年之前取得成功。这符合其数十年来推行的战略，即针对美国的优越军事力量发展“非对称性应对措施”。俄罗斯的海底“末日无人机”只是迄今为止这种现象最极端的例子。[4]

核威慑在2040年是否会被认可？斯塔图斯-6是一个严厉的警告，如果技术进步影响到核大国的安全感，它们可能会试图利用前所未见的新武器系统和军事姿态

斯塔图斯-6的组件



“海洋多用途系统斯塔图斯-6”外观像巨型鱼雷，由小型核反应堆提供动力，其巡航速度和范围在海洋里几无敌手，能够突破敌人的防御。

来补救其核威慑力量。与促使美苏暂时和平共处的安排相比，这些新兴战略安排可能不太稳定，而不稳定性会提高爆发核战争的概率。风险上升的幅度主要取决于人工智能进步的速度和程度，而人工智能的进步将带来投放核武器和防御核攻击的新方式。2017年5月至6月，兰德公司举办了三场研讨会，核安全专家和人工智能研究员在会上讨论了人工智能对核安全的影响。与会者似乎都认同，高级人工智能可能会严重破坏核战略的稳定性，加大核战争风险。然而，大家对人工智能为何及如何产生影响则莫衷一是，在不同意见派别内亦然。

研讨会的方法和情况简介

为了研究高级人工智能在未来25年对核安全的潜在影响，兰德公司在2017年5月和6月举办了系列研讨会。这些研讨会汇集了不同的专家群体，包括核安全专家和人工智能研究员，以及来自政府和行业的参与者，带来了各种各样的观点。

第一场研讨会

第一场研讨会于2017年5月1日在兰德公司圣莫尼卡总部举行，十六位参与者中有不少是兰德在核或人工智能领域的研究员。研讨会的目的是设想与人工智能互动的战略环境，并以未来地缘战略秩序的可预测性高于人工智

与促使美苏暂时和平共处的安排相比，这些新兴战略安排可能不太稳定，而不稳定性会提高爆发核战争的概率。

能技术发展作为前提。讨论围绕核大国冲突加剧的几个具体情景展开。这些情景包括：

1. “俄罗斯复兴”情景，即新削减战略武器条约失效，俄罗斯在本世纪三十年代初期前于战略核武器方面对美国取得重大优势
2. “中国崛起”情景，即中国逐步扩大战略核武器库，与美俄分庭抗礼
3. “成功有限度使用”情景，即巴基斯坦成功使用战术核武器迫使印度撤回入侵军队，从而打破“核禁忌”
4. “区域核战争”情景，即朝鲜政权崩溃，攻击中日韩，导致整个地区受到毁坏。

研讨会参与者被要求用有关大国核力量的技术细节充实这些情景，包括运载系统和C4ISR（指挥、控制、通信、计算机、情报及监视与侦察）系统的数量及能力，目的是识别拥核国可能感兴趣的人工智能应用。未来的作战系统被假定为与目前研发的系统相似，因为军队购置的速度很慢，但人工智能的进展要远远快于国防系统的

购置。参与者似乎认同，将高级人工智能应用于这些系统，在未来的对峙局面中可能会成为破坏稳定的影响因素。然而，参与者也假定，每种会破坏稳定的技术都存在一种具稳定作用的抗衡技术。这个主题在第二场研讨会上进一步展开。

第二场研讨会

第二场研讨会于2017年5月25日在旧金山举行，共有十九名参与者出席。其中七人从事人工智能研究，五人从事国家安全研究，三人同时从事这两个领域的研究，还有四人为其他学科研究人员。侧重于人工智能的参与者包括来自商业、学术和非营利人工智能研究机构的知名人士，以及人工智能政策圈的成员，而国家安全领域的参与者包括来自国家实验室的核武器专家。参与者分成多个小组，讨论了三个问题。

第一个问题是人工智能是否能让国家跟踪和瞄准敌方报复力量，从而削弱构成核战略理论主要根据（在后续环节进行了深入讨论）的保证报复的前提。以核安全专家为主的小组认为人工智能可以做到这点，但第二小组却不同意，小组成员包括一位生成对抗网络（通过生成中性网络与判别中性网络的互动创建趋于真实的虚假样本的技术）的著名专家。第二小组认为，多数机器的大部分学习技术都存在容易遭受敌方操纵攻击的弱点，因此国家将能够利用这些手段防止敌方跟踪其发射器。

这场研讨会讨论的第二个问题是，在决策支持系统中使用人工智能，在危机或冲突中就战略核问题为决策者提供建议。针对人工智能在这些任务中的应用，各小组存在较大的分歧，有些人表示人工智能应当严格由人类控制，而有些人则宣称这不切实际。这个话题将在稍后详细讨论。

最后，研讨会针对未来人工智能的应用概括了核武器控制的教训。参与者似乎认同，不可能复制用于制止核扩散的法律结构和准则来阻止人工智能的军事应用，因为核技术与人工智能相差甚远。在人工智能用于核战争相关任务的具体案例中，参与者指出控制人工智能也许很困难，但这些应用所必需的其他部件（即传感器平台）是可以监控的。参与者讨论了是否有可能通过控制数据、人才或处理资源来控制人工智能。几位研究人工智能的参与者主张，当前的人才短缺是暂时性的，而随着模拟的完善，训练数据的重要性将会下降，但届时硬件可能会成为制约因素。他们认为，生产图形处理器等部件的工厂数量有限，使得构建某种控制机制成为可能，但许多其他参与者则不以为然。

一个小组断言，未来的人工智能系统基本上可以充当军备控制机制，监察合规情况和判定违规，而无需人类参与。

第三场研讨会

第三场也是最后一场研讨会于2017年6月9日在兰德公司弗吉尼亚州阿灵顿分公司举行。这场研讨会共有十五名参与者，其中八人主要研究核问题，五人主要研究人工智能，虽然后者大多从事政策方面的研究，而不是人工智能研究人员。其余两名参与者提供了有关采购政策的宝贵专业知识。讨论小组包括兰德公司的研究员，以及来自美国军方、国防部长办公室和国务院军备控制、核查和合规局的代表。本场研讨会以前两次的成果为基础，询问出席的政策研究人员会如何应对前两场研讨会所确认的挑战。

第一场讨论聚焦于跟踪和目标锁定问题，要求参与者思考他们会尝试通过什么方法，来挫败敌方寻求利用人工智能削弱战略力量的意图。参与者建议尝试通过攻击相关传感器和通信网络来使这种能力瘫痪，而非攻击人工智能本身。在后续的讨论中，参与者考虑了人工智能研究员在第二场研讨会上强调的生成对抗技术所带来的挑战，不过没有形成共识（可能是由于大部分参与者不熟悉这些方法的技术细节）。

第二场讨论围绕美国是否需要重新考虑其核力量现代化计划的路径，因为人工智能有可能显著重塑战略格局。参与者指出，现有计划存在许多弱点，但显而易见的替代方案并非明显优于现有计划，即使它们大概至少值得分析一下。国内和国际的制度性压力使美国难以大幅背离由洲际弹道导弹、潜艇和有人驾驶轰炸机组成的现有“三元体系”。

第三场讨论围绕人工智能如何协助核军备控制。通过提高透明度和信任，人工智能可用于条约核查等任务。一个小组断言，未来的人工智能系统基本上可以充当军备控制机制，监察合规情况和判定违规，而无需人类参与。最后，参与者思考了是否可能或适宜对人工智能本身实施军备控制。大多数参与者对该目标的可行性或可取性保持怀疑；很多人认为这实际上不可能或需要极端和令人难以接受的干预，例如软禁人工智能研究员。

评估人工智能潜在影响力的理论和历史背景

冷战期间，美苏勉强接受了相互保证毁灭(MAD)的条件，即假定任何全面进攻均会遭到末日般的报复性打击，确保双方社会都会毁灭。MAD是一个条件，而非一项战略—而这两个超级大国都希望尽可能避免这种结

局（Buchan等人，2003年）。即使双方的弱点使得全面核交火的可能性下降，但意外或超级大国领导人的误判仍有可能导致战争的爆发。例如，罗纳德·里根要求科学家设计一种令核武器“失效和过时”的导弹防御系统，而苏联则制定了周密的民防系统计划（Garthoff，1987年；Geist，2012年)。MAD也不足以成为美国或苏联核战略的基础。虽然MAD能可靠地阻止苏联先发制人攻击美国，但也削弱了美国所做承诺的可信度。美国曾承诺，即使冒着爆发核战争的风险，也会保卫北约的欧洲盟友。如果华盛顿仅仅依赖MAD，苏联可以利用其传统的优势入侵西欧，而美国则面临要么投降要么全面爆发核战争的严酷选择。因此，美国战略学家和政府官员制定了更全面的保证报复原则，即针对任何敌方挑衅予以适当有效的回应（Long，2008年）。通过威胁对敌方可能发动的挑衅予以相应报复，“保证报复”原则寻求可靠地阻止小规模和全面的攻击。在冷战后期，此原则出现了一个称为抵消战略的变种，通过保证任何此类攻击将由于美国的报复而无法达成其目标的方式，寻求阻止一切形式的攻击，包括针对军事力量先发制人的攻击（Slocombe，1981年)。

核战略不止于威慑（见右表）。威慑是使用报复威胁阻止敌方攻击本国或盟友。威慑可划分为中心威慑（针对攻击本土的威慑）和延伸威慑（针对攻击战略伙伴的威慑）（Cimbala，2002年）。核武器也可以用于胁迫——迫使敌方做其不想做的事情（Long，2008年，第9页）。除了高压威慑和胁迫，核武器还可以用于作战，就像第二次世界大战结束时那样。核战略在实践中的复杂性源于保证—扩大威慑可信度的挑战。冷战期间，美国积攒了大量的战略和战术核武器，以便其盟友相信，美国愿意用核打击报复苏联对欧洲的常规攻击。正如英国国防部长Denis Healey指出的那样，“要威慑俄罗斯人，只需美国报复的可信度达到5%，但要让欧洲人放心，却需要达到95%”。然而，美国核武器库的规模令苏联领导人感到焦虑，他们认为美国人也许在试图针对苏

核战略目标类型

环节	定义
威压	
威慑	使敌方不敢做其想做的事情
胁迫	迫使敌方做其不想做的事情
保证	使盟友相信安全保证是可靠的
再保证	使敌方相信只要其不做出挑衅行为，就不会遭到攻击

联发展第一次核打击能力。这种猜疑彰显出有需要进行再保证——使敌方相信只要其不做出威慑所针对的行为，就不会遭到攻击（Schelling，1966年）。

当敌方缺乏发动挑衅行为的重大动机时，就存在战略上的稳定。[5] 战略稳定性有几种类型，可按不同的时间尺度区分。当没有国家能够不畏惧毁灭性报复而突袭对手时，便会存在第一次核打击稳定性。通过威胁以第二次打击力量进行全面自动报复，可以最有力地遏制这种可能性（Cimbala，2002年，第66页）。相反，危机稳定性旨在防止或管控危机期间的冲突升级，就像上世纪六十年代初在柏林和古巴发生的那样（Cimbala，2002年，第98页）。在这些情况下，国家领导人承受着不能退缩示弱的巨大压力，但由于国家试图调遣核力量来传递信号，意外升级的可能性会大大增加。在此背景下，适合用于最大化第一次核打击稳定性的大规模自动报复，可能会引发灾难。

最后，当敌方军事力量没有可利用的劣势时，便会存在军备竞赛稳定性。（Cimbala，2002年，第110页）。国家会避开这些劣势，以管理长期竞赛的风险和成本，同时避免影响未来的第一次打击稳定性和危机稳定性。核战略难以制定的原因，就是这些目标的相互关系很紧张。

在极端的案例中，人工智能可能会破坏MAD的条件，使打赢核战争变为可能，但破坏战略稳定性所需要的条件则少得多。在一定程度的冲突中，人工智能的进步只需要令人怀疑报复的可信度便足够。中美俄等核大国在维持中心威慑可信度上有着共同利益，但他们追求地区优势，以争取眼中的核心战略利益。可信度已然吃紧的领域，例如某些延伸威慑保证，尤其容易变得不稳定。日趋多极化的战略环境，也在推动各种形式的竞赛，从而对稳定构成威胁。例如，美国热衷于发展跟踪和锁定小型拥核国移动导弹发射器的能力，但中俄担忧这种技术发展成熟后，可能会威胁到其较先进的报复力量。在爆发危机的情况下，动用或提供人工智能情报、监测及侦察 (ISR) 或武器系统，有可能加剧局势紧张，增加冲突意外升级的可能性。最后，追求先进的军事力量会导致军备竞赛的不稳定，即使这些技术并不可行，导弹防御技术就是前车之鉴。

人工智能为战略稳定性带来的挑战并非这种技术所独有，但却更为严峻，因为人工智能技术进展迅速，而且与核战略多有相交。人工智能可能出现的大多数具体应用，例如ISR数据分析、控制自主传感器平台及自动目标

当敌方缺乏发动挑衅行为的重大动机时，就存在战略上的稳定。

识别(ATR)，数十年来一直备受追捧，但却超出了现有技术的能力范围。即使未取得进一步的突破，使用现有人工智能技术的增量进展，也能让这些追求已久的目标在可预见的将来成为现实。

中俄似乎都认为，美国试图利用人工智能威胁其战略核力量的生存能力，引发互相猜疑，而这在危机中可能会造成灾害性后果。如Paul Bracken所指出，人工智能等技术的持续完善有可能“削弱最低核威慑战略”和“模糊常规战争与核战争的界限”（Bracken，2017年）。

冷战中的人工智能

人工智能先驱马文·明斯基将人工智能定义为“使机器做到假如由人类来做则需要运用智慧的事情的科学”（Minsky，1968年，第v页）。自上世纪五十年代科学家开始研究人工智能以来，计算机重塑了人类对“智能”的理解，而这个领域的界限也随之改变。人工智能也随着各种理论范式的兴衰而持续演进。从上世纪五十年代到八十年代，旨在复制高级人类推理能力的“符号”范式占据了主导地位，但随后被“联结主义”范式所取代，后者寻求用人工神经网络模仿人类认知的生物学基础。在20世纪，这两种范式在实验室演示以外所取得的效果都不是很好。这导致在部分时期（有时称为人工智能寒冬），人工智能研究陷入经费短缺的窘况。归功于计算机科学近几十年来的发展、计算和通信硬件和软件的进步，以及云计算和大数据的崛起，人工智能在近年来突飞猛进，尤其以“深度神经网络”（即具有多个层级的神经网络）领域的进展最大（Goodfellow、Bengio及Courville，2016年）。深度神经网络的性能提升幅度之巨，使其几乎成为了人工智能的同义词，但事实上旧范式也持续取得进展，并广泛应用于商业和军事用途。近期一些令人印象深刻的人工智能系统，例如击败世界围棋冠军的Alphabet DeepMind的AlphaGo程序，就结合了深度神经网络与穷举搜索等旧技术。在人工智能的60年历史里，其支持者所寄予的厚望从未改变。有了足够的智能，是否有可能克服贫穷和疾病等看似无解的问题——甚至是打赢核战争？

人工智能与核战争的交集在50多年前就成为了科幻小说的陈词滥调，但两者在现实世界中的联系比这还更早。最早的人工智能研究员深度参与了国家安全工作，并声称其理论研究很快会转化为实际的军事应用，从而获得政府支持。Claude Shannon在1950年的论文“编程让计算机下象棋”中断言，让计算机玩这种历史悠久的游戏，可以提供理论见解，让“能在简单军事行动中做出战略决策的机器”可以“在短期内”问世（Shannon，1950年，第256页）。上世纪五十年代中期，研究人员在美国空军的支持下设计出最早生效的人工智能程序（Simon和Newell，1958年；Newell、Shaw和Simon，1959年）。这类机器的潜在应用不久便出现在战略理论家的著作里。五

十年代后期，Herman Kahn提出了“末日机器”的概念，利用经过编程的计算机来识别不可接受的敌方挑衅并还击（Kahn，1960年，第145–154页）。尽管Kahn原本打算把这作为思想实验，说明实施核战略应避免的做法，但科幻小说作者对智能计算机控制核武器的想法产生了浓厚兴趣，创作出无数小说和电影，例如《巨人》（1970年）、《战争游戏》（1983年）和《终结者》（1984年）。

虽然虚构的惊悚片编造了计算机支配核武器胡作非为的故事，但现实世界中运用人工智能处理核战略问题的尝试往往平平无奇。美苏官员都不愿把发射决策交给计算机，一方面原因是他们不情愿这么做，更希望这个特权能掌握在自己手中，另一方面则是用自动报复应对艰难的战略问题（例如胁迫或危机稳定性）并不合理。唯一一次值得注意的例外来自冷战末期的苏联。苏联领导人察觉美国期待掌握第一次打击能力，担心自己可能成为斩首行动的目标，所以想方设法确保资本主义侵略者无法全身而退。[6]据报道，苏联曾考虑研发一个系统，在遭到第一次打击后，万一联系不到苏联政治领导层，该系统会自动向美国发射尚存的洲际弹道导弹。看起来，昵称“死亡之手”的全自动化版本被否决了，改成另一个名为“Perimetr”的版本，会自动把发射权限授予战地指挥官，但过程中始终需要人类参与（Hoffman，2009年）。据俄罗斯媒体报道，Perimetr系统仍然健在并利用了某种人工智能。[7]而美国则在探索利用人工智能加强针对军事力量的作战能力。上世纪八十年代末期的一个研究项目可生存适应规划试验 (SAPE)寻求利用当时的人工智能技术，让美国能够瞄准苏联的移动洲际弹道导弹发射器。SAPE不会直接控制核武器，而是利用专家系统将侦察数据转化为核武瞄准计划，然后由有人驾驶的B2轰炸机执行。SAPE只是一套预想系统和军备的一部分，如果实现，将严重挑战苏联核武器的生存能力（Roland和Shiman，2002年，第305页；Long和Green，2012年）。

人工智能与新兴的地缘政治秩序

尽管20世纪的人工智能尚难以实现这些应用，计算能力近期的进步可望释放它们的潜力。深度学习等当代技术正推动机器视觉及其他信号处理应用显著发展，从而加强自主性和传感器融合。自主性和传感器融合也许具有极大的战略意义，因为这能大大提升人工智能情报、监测及侦察(ISR)、自动目标识别(ATR)和末端制导能力。这一切可能会严重削弱核大国保证其核力量生存能力的手段。由于武器精度的提升早已削弱了发射井式洲际弹道导弹的生存能力，中美俄都将核武器装备在被视为更有可能抵御第一次打击的潜艇和移动洲际弹道导弹上。如果技术使可生存力量（例如潜艇和移动导弹）能够被锁定和摧毁，则一个国家更有可能会发动第一次打击。这会破坏战略稳定性，因为即使掌握这些能力的国家无意实际

使用它们，敌方也无法肯定这点。因此，这些军事力量仍可能被用于在危机期间压迫潜在对手，从而令其让步。在危机期间，这种能力无需实际动用也能发挥政治作用。如Alfred T. Mahan所指出的，“力量在其存在为人所知，但尚未动用的时候，是最有效的”（Mahan，1912年，第105页）。只要敌方惧怕这种军事力量可能存在，就可以不战而屈人之兵——更强大的国家能够在危机中先发制人取得胜利。因此，核打击瞄准能力对许多国家很有吸引力，尽管它有可能破坏战略稳定性。

人工智能技术有助于在跟踪和瞄准及反潜战中实现新的突破，或者让高精度的常规弹药更容易摧毁加固的洲际弹道导弹发射井（Holmes，2016年）。这种能力破坏稳定的作用特别大，因为决策者威胁使用常规武器的可能性远高于任何类型的核攻击。在危机中，常规武器的威胁会令对手承受巨大压力，这可能会迫使其屈服，但也有可能卷入核战争。冲突可能升级的原因是，对手认为需要在被解除武装前使用核武，以反击未能成功解除武装的攻击，或者只是危机导致意外动用核武。

美国的潜在对手（例如俄罗斯）很严肃地看待美国利用人工智能等技术优势大幅提升打击能力的可能性。过去几年，俄罗斯军事分析家一直在军方媒体上争论其国家战略弱点的范围。[8]他们倾向于假定美国当前和未来的军事力量将严重威胁俄罗斯的安全，使这种焦虑火上浇油。

核战略的一大挑战是对手或将一个国家的安全报复力量视为第一次打击威胁或末日机器，并据此作出回应。例如，俄罗斯人大概会将斯塔图斯-6视为最后的第二次打击方案，利用人工智能自主绕过美国的防线，但西方观察家则认为它是奇爱博士的“钴弹”。人工智能的发展也促使俄罗斯加倍押注战略性能不理想的旧型系统。例如，俄罗斯研发出RS-28“萨尔马特”导弹后，重新投资于搭载分导式多弹头 (MIRV) 的发射井式大型洲际弹道导弹。根据现已废止的第二阶段削减战略武器条约，俄罗斯曾经计划弃用这类武器。西方战略理论普遍认为分导式多弹头大型洲际弹道导弹会破坏稳定，因为它们是先发制人的理想武器，也容易受到先发打击。

在新的千禧年之初，莫斯科认为能够通过主攻移动洲际弹道导弹，弃用苏联遗留的大型发射井式导弹，确保其核力量的生存能力。然而，俄罗斯领导人对美国可能威胁移动洲际弹道导弹生存能力的忧虑，似乎促使他

核战略的一大挑战是对手或将一个国家的安全报复力量视为第一次打击威胁或末日机器，并据此作出回应。

日趋多极化的核环境也扩大了人工智能的潜在战略影响。

们改变计划，尝试通过在美国攻击期间发射核弹来保证报复行动，而不是坐以待毙。这相当于采取攻击下即发射反击的核态势，这可能令俄罗斯领导人承受在危机中首先发射的巨大压力，从而增加冲突意外升级的可能性。俄罗斯人认识到萨尔马特发射井不太可能独自抵御先发制人的攻击，因此其生存能力依赖相关的主动防御系统。此系统代号为“莫济里”，其目的是在发射井一段距离外引爆敌方弹头，让其能够避免被核爆摧毁。

日趋多极化的核环境也扩大了人工智能的潜在战略影响。虽然冷战期间有六个国家拥有核弹，但其中五国将苏联视为主要敌人，因此战略秩序基本上是两极化。这种两极化为危机稳定性和军备竞赛稳定性提供了支持。现在有九个核武国家和多个间接互相影响的战略对手。美国担心俄罗斯和中国；俄罗斯计划对抗美国和中国；中国视美国、俄罗斯和印度为潜在对手；印度卷入与中国和巴基斯坦的战略竞争；而朝鲜几乎对每个国家来说都是一个麻烦。

发展适用于这些复杂多极冲突的战略稳定性理论，仍需要做大量工作。随着分析家研究这一难题的解决办法，他们需要考虑人工智能增加或减少蓄意或意外引发热核战争风险的不同方式。即使基于我们当前浅薄的理解，仍可以着手考虑某些新兴军事力量的影响及其相互作用。

专家对人工智能前途的普遍意见

如先前所讨论，有几种观点主导了研讨会的讨论。

预测人工智能的演化

关于人工智能的演化，主要有四种看法，很难提供严谨的证据来证明任何一种看法优于其他看法。尽管如此，其支持者常常为此争论不休。在大多数情况下，出席研讨会的专家都熟悉不同的看法，能够从所有角度提出论点。

对于人工智能可能出现的每种未来状态，第13页的表格概括了每类专家意见有关核安全的共同观点。这些观点和支持或反驳的论据在后续章节有进一步讨论。

超级智能

部分人预计超级智能是必然出现的状态，届时机器的智力将远胜人类。牛津哲学家Nick Bostrom（2014年）等理论家认为，一旦超级智能出现，可能会有两种结果：超级智能是善意的，解决了人类的所有问题，或者超级智能会毁灭人类，不论是恶意的还是出于意外。Bostrom相

信，反复自我完善的人工智能将极快地进化成超人类智能，同时几乎不会犯错。这种“智能爆炸”也许只需要几个小时或几分钟。

在此情况下，核安全的作用变得微不足道：超级智能如果是善意的，将从核战争中挽救人类；如果是恶意的，核打击不过是千万种灭绝人类的方法之一。

大多数人工智能专家似乎不认为超级智能即将或必然出现，但许多支持者认为这值得关注，原因是其代价和利益极其巨大，即使出现的可能性较低。

有限的突破

没有创造出真正的超级智能，人工智能也可能出现断续的跃进，机器的智力大幅提升，但至少在某些方面仍不及人类。举个例子，如果一个可反复改编程序的软件系统快速提升智能，达到其硬件所能支持的峰值，但无法再进一步提升，就会发生这种情况。

在此情况下，根据相对于人类形成的确切能力，人工智能最有可能被用于发掘其比较优势，而人类则会最大化自身的比较优势。与人类一样，人工智能仍然会出错，而可能出现结果的范围和影响主要取决于这种出错的性质。

循序渐进的发展

第三种可能性是与上述相似的状态，但并非通过断续的人工智能跃进达到，而是循序渐进地发展达到。这种演进取决于计算速度加快、硬件架构、算法发展、数据可

依专家意见判断的人工智能的不同前途

	人工智能的潜在前途		
专家意见派别	人工智能寒冬	有限的突破或持续递进的进展	超级智能
乐观派	有可能	数据不足，问题对高级人工智能来说仍然太复杂	不大可能出现，但大概比人类更安全
谨慎派	不大可能	几乎无效的算法会引起对手警戒，如果使用，有可能失败	最终必然出现，并有可能毁灭人类，不论有意或无意
颠覆派	中立	有可能造成人工智能彻底失败，或者造成人工智能失败的能力可为稳定性提供保证	超级智能可抵抗颠覆和人类控制

用性和成本的下降。这可以说是对人工智能近期趋势最靠谱的解读，当从任意一个时间点来看，人工智能能力的提升似乎平平无奇，但从几年的时间段来看，提升则相当巨大——当把预测的时间范围扩大超过20年直到2040年，有关进展将会是惊人的，好比今天的互联网于上世纪九十年代中期在普通人群中几乎无人知晓。

尽管从哲学或预防政策的角度来看也许区别明显，但就结果和对核安全领域的影响而言，20多年持续递进的进展与有限的突破基本别无二致。与有限的突破一样，人类和机器都可能存在有优势的方面，也仍然会出错，而这种出错性会带来风险。

人工智能的稳定期

几位研讨会参与者表达了最后一种观点，即人工智能在现有技术成熟后，可能趋于稳定。这种结果有别于过去的人工智能寒冬，当时人工智能研究继续取得进展，即使经费和大众的兴趣显著减少。例如，计算机硬件的发展可能会停滞不前，导致人工智能得不到实现理论潜力所需的计算资源。

当前活跃的人工智能研究员通常对这种想法持怀疑态度，不过他们承认这是可以想象的。目前全球对人工智能发展的直接投资水平空前高涨，未必可以持续，但私人公司和中国等国家政府明确的投入水平表明，经费在可见将来仍将充足。有了如此高水平的投资，对进展的预测可能会自我应验，就像摩尔定律数十年来推动半导体发展那样（Mack，2011年）。此外，“人工智能的稳定期”有可能在现今能力取得重大进展后出现，对核安全造成许多与前两种情景相同的挑战。

对核安全的预期影响

本期《视角》主要聚焦于有限的突破和持续递进的进展两种情况，因为另外两种从核安全的角度来看不太相关。有限的突破和持续递进的进展预测，可以概括为人工智能在更复杂和数据有限的任务中远超人类智能。对于这种能力会给核安全带来什么影响，专家们莫衷一是。

乐观派

一种普遍的观点是，人工智能除了提升效率和透明度之外，不会深远改变现状。这种“乐观”看法的支持者往往更注重技术问题，而非战略或政策。应当注意的是，虽然参与者包括全球最有才华的部分人工智能工程师，但与自身专业相比，他们很可能对人工智能安全性更感兴趣，因此这种看法的代表性也许不强。乐观派更有可能认为核战争太过复杂，导致人工智能难有大作为，因此人工智能对现有平衡的影响是可以忽略的。例如，乐观派断言即使到了2040年，人工智能也无法克服收集数

此外，“人工智能的稳定期”有可能在现今能力取得重大进展后出现，对核安全造成许多与前两种情景相同的挑战。

据和区分真实系统或行动与诱饵的挑战。他们也更有可能认为，识别和解读核升级相关决策输入数据的问题十分宽泛，足以令其成为人工智能完全问题。换句话说，凡是能够在此任务中做得比人类更好的计算机，将必须飞跃至能够胜过全部人类。

谨慎派

处于另一个极端的是谨慎派，他们倾向于认为人工智能将导致现有系统变得脆弱，或扰乱现有战略平衡以致产生严重问题。这一阵营包括绝不会将任何方面的核决策交托给一个计算程序。部分参与者过去曾经亲身尝试设计实现相关目标的计算程序。在某些情况下，这些程序不合理及其无法考虑决策的情感和道德因素，使参与者对人工智能与核问题的交集感到不安。谨慎派还认为，人工智能只需要被视为十分有效，便会产生破坏稳定的作用——例如在跟踪和瞄准对手的发射器方面。对手面对丧失第二次打击力量的可能性，将被迫先发制人发动第一次打击，或者扩充武器库，这两种结果都是不可取的。

颠覆派

第三种观点的立场介于乐观派与谨慎派阵营之间，立足点是对人工智能易受对抗行动影响的担忧。这源自有关对抗行动可能十分有效的理论考证。此观点所带来的结论未必与人工智能进展有限或被视为有效但实则不然的情况相同，但两者存在重叠之处。

在许多令人信服的求证中，旨在破坏机器学习算法的少量对抗工作极具成效。部分研究人员认为这是机器学习的普遍特性，并预期这个问题会持续多年。当有效的跟踪和瞄准人工智能破坏稳定，导致武器扩散甚至更糟糕的结果时，如果对手相信自己能够利用这些对抗方法防止被探测，则会恢复对第二次打击力量生存能力的信任，从而重建战略稳定性。另一方面，行动方可能认为自己能够破坏人工智能识别先发制人的第一次打击的能力，使这种打击成为可行的选项，从而破坏稳定。

示例：跟踪移动导弹发射器

人工智能可能破坏战略稳定性的原因，并非其效果太好，而是刚好足够增加不确定性。为说明这一点，我们会在本节介绍兰德公司早前研究瞄准移动导弹发射器问题得出的结果。

对可靠的第二次打击的影响

大多数核大国青睐移动导弹发射器，原因是它们难以跟踪和瞄准，所以被视为具备生存能力。这些导弹经常沿着公路或铁路移动，除非敌方一直追踪其位置，否则攻击它们的唯一方法（在武器部署到战场前将之摧毁的第一次打击除外）是尝试用核武器瞄准其较大的巡航区域。即使是这种轰炸战略，也只有在导弹的可能位置至少能够缩窄到一定范围，才真的切实可行。冷战时代针对苏联移动洲际弹道导弹发射器的计划，是把轰炸战略与苏联移动导弹规律的情报结合起来。

人工智能可为人工智能情报、监测及侦察(ISR)和分析系统做出关键贡献，推翻这些假设并使移动导弹发射器容易受到先发制人的打击。这种可能性令中俄的国防规划者忧心忡忡，因为这两个国家主要依赖移动洲际弹道导弹进行威慑。即使人工智能只是略微提升了整合敌方导弹部署数据的能力，也可能严重动摇一个国家的安全感，从而破坏危机稳定性。目前，发展自动目标识别(ATR)、传感器整合和信号处理所需的能力仍然极其困难。然而，这些挑战可能处于足以使这些武器完全过时与毫无可信度之间的尴尬处境。

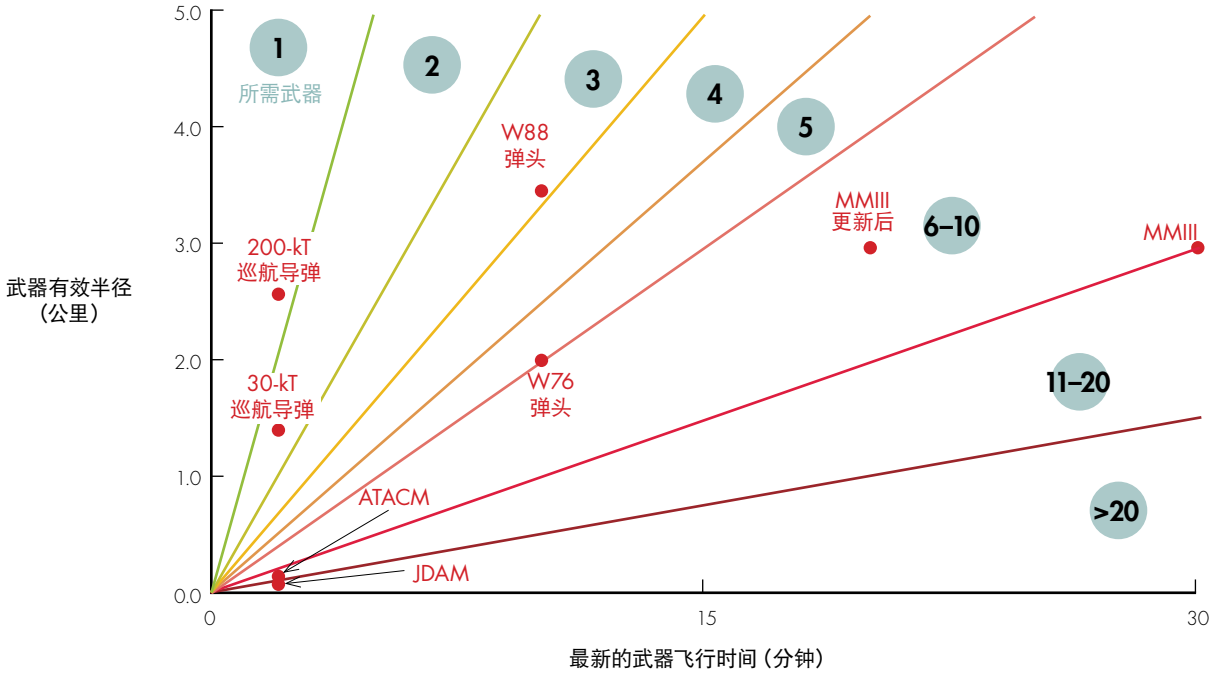
艰难的技术挑战

兰德公司构建了一个跟踪和瞄准对方力量的模型，当中包括传感、图像处理以及武器速度和杀伤半径。[9]上述

从极近距离发射，就连常规弹药都会变成可行的选项，从而显著增加先发制人打击军事力量的可靠性。

有些限制可以在较短时间内通过人工智能的进展克服，但有些限制则不太可能在2040年前得到解决。例如，即使对目标位置了如指掌，机动目标在武器发射与抵达期间能够移动。瞄准移动系统的武器也许能够飞得更快，更好地调整航向，但武器仍需要极其精密的末端制导能力，以大幅减少所需的武器数量。因此，即使图像处理和目标辨识取得进展，仍需要许多大型武器，或者小型武器需要从近距离发射。第17页的图显示摧毁移动目标所需的各类弹头数量，武器有效半径介于0至5公里。尽管其“杀伤半径”巨大，直径长达数公里，但要有把握地保证摧毁一个导弹发射器，仍需要由弹道导弹投放多枚热核弹头。例如，覆盖一个目标需要由飞行时间十分钟的三叉戟II型导弹投放三枚475-kT W88弹头，或者五枚100-kT W76弹头。然而，分析发现，从邻近目标的位置发射的精确巡航导弹（图中的30-kT巡航导弹和200-kT巡航导弹）只需一两枚弹头就能覆盖移动导弹发射器。从极近距离（即飞行时间只有几分钟）发射，就连常规弹

覆盖目标所需的最小武器数量



此图显示摧毁移动目标所需的各类弹头数量，武器有效半径介于0至5公里。尽管其“杀伤半径”巨大，直径长达数公里，但要有把握地保证摧毁一个导弹发射器，仍需要由弹道导弹投放多枚热核弹头。ATACM = 陆军战术导弹系统；JDAM = 联合直接攻击弹药；kT = 千吨；MMIII = 民兵III型。

药都会变成可行的选项，从而显著增加先发制人打击军事力量的可靠性。

这些发现表明，人工智能加上移动、甚至自主的传感器平台，可以威胁移动洲际弹道导弹发射器的生存能力，从而破坏战略稳定性，但也留下通过军备控制防止威胁的一线希望。即使跟踪和瞄准发射器的人工智能系统较为先进，要对移动洲际弹道导弹发射器造成可信的威胁，攻击部队的基地仍需要非常接近目标。即便在这种条件下，可实施攻击的“脆弱窗口”只会持续几分钟，机会一旦出现，攻击方必须抓住。担忧遭受解除武

装攻击的国家，将对这些部队出现在外围极其警惕，并可能让他们相信自己处于“不用则废”的境地。因此，这类举动可能会导致互不信任的恶性循环，反而激发并非任何一方本意的冲突。但是，这种不利的结果是可以避免的，方法是订立可核验的协定，禁止在移动导弹发射器的一定范围内建立基地或部署可用于解除武装攻击的武器。

示例——人工智能作为可信赖的顾问

除了可能削弱对第二次打击力量的信心外，人工智能也可能无意中影响国家的战争路径、升级和发射决策能力。自主控制不大可能在任何本土发射器或指挥控制平台上直接实施，但不需要这样做，人工智能也能发挥影响力。这种情况已经发生：计算机程序、模拟或数据分析程序被用于辅助人类决策。预计人工智能将更广泛用于协助决策（通常称为决策支持系统）。

人工智能在决策中可能扮演的角色

人工智能的发展一日千里，在日益复杂的任务中有着超越人类的表现。Alphabet DeepMind的AlphaGo击败世界围棋冠军，就连人工智能专家和战略专家也感到惊讶（Etherington，2017年）。当然，围棋中的决策比核战争中的决策要简单得多；走棋是有序可循的，并且有明确

人工智能的发展一日千里，在日益复杂的任务中有着超越人类的表现。

的规则。但DeepMind的研发人员一直努力设计一种能玩电脑游戏星际争霸的人工智能（Woyke和Kim，2017年），这款游戏模拟军事冲突，包括物流、基础设施，以及各种难以规定的行动和战略。星际争霸也比核战争简单得多，但到了2040年，预期人工智能系统也许能够以超人类的水平参与军事演习或演练的某些方面或阶段，似乎也合情合理。一旦这种能力得到证实，做出指挥决策的人类有可能将人工智能系统的建议视为等于或优于人类顾问的建议。这种可能未经验证的信任带来了必须考虑的新风险。

有些研讨会参与者深信，人类不愿意让计算机左右核战争的决策，而有些人则轻易就能预想到人类对这种想法会越来越放心。有趣的是，不同世代的参与者有着不同的观点，意味着到2040年上位的人可能会对让出一定的控制权感到更放心，特别是随着人工智能在未来数十年持续证明自己能够完成日趋复杂的日常任务。美国人已普遍依赖人工智能作出驾驶路线决策、协助安排任务和回复简单电子邮件。危险的是，这些成功会让人类形成无根据的信心，因为日常决策与核战争有着巨大差别。

对抗行动对效能的限制

发展人工智能用于核战争相关任务面临着两大挑战。第一，从1945年美国轰炸日本令其无条件投降以来，核武器未曾使用过，而且从未发生过核交火。因此，完全缺乏真实的训练数据。但是，模拟、演练和演习可能有助于减轻这个问题，而且不应忘记，真实数据的缺乏也同样限制了人类的学习和决策。

第二个显著的特点是，导航或行程安排等共同追求的参与各方有着成功完成任务的相同动机，而核战争天然具有对抗性。有各种方法能破坏人工智能系统，而在未来很长一段时间里，破坏手段似乎是一个有效的选项。我们会简要讨论黑客攻击、训练数据攻击和操纵输入因素，举例说明现存的几种问题。

黑客攻击

黑客攻击并非人工智能所独有，但只要人工智能涉及到计算机，则必须被视为容易被入侵。人工智能本身可以被黑客攻击（如下文所论述），但举例来说，数据也可能在输入、输出或从输出到显示的过程中被篡改。当然，在核战争中有一席之地的人工智能会得到谨慎保护，但也是一个高价值的目标。

训练数据攻击

破坏人工智能的另一种方法是篡改训练数据。有几种办法能实现这一目标：由内鬼替换数据、黑客入侵替换数据、在公开数据中加入错误样本，或对手精心选择行为方式充当套路。

各种研究已开始罗列破坏不同机器学习算法的训练数据的战略和效果（Anderson等人，2017年；Biggio、Nelson和Laskov，2012年；Kearns和Li，1992年），但还有很多工作有待完成，而且可以预期会有更多发现。当中许多研究工作是由杀毒专家群体牵头的，杀毒软件是其他具对抗性的少数应用领域之一——近年来，这个群体已从基于特征的传统方法转向机器学习。有些人想方设法确保机器学习在遭受数据操纵攻击后仍然有效，但这些研究工作仍处于初期阶段（Kegelmeyer等人，2015）。预计数据篡改在未来很长一段时间里都会构成威胁。

操纵输入因素

第三个破坏人工智能的机会来自其接受全面训练之后。巧妙地操纵输入因素，可促使高性能的人工智能系统生成攻击者期待的结论。在图像辨识领域，这已得到证实：通过对图像做出人类无法察觉的细微修改，促使人工智能将修改后的图像划分为攻击者所选择的类别（Karpathy，2015年）。这在核问题上也许更困难，人类

未必确切知道所有输入数据或可能的分类。在图像辨识的案例中，对手的输入数据范围很简单，局限于像素。对于其他任务，对手可能需要按特定规律调动部队或按特定顺序发表包含特定信息的声明，但仍然有可能——至少在原则上——“欺骗”经过全面训练的人工智能系统。重要的是，数据操纵攻击无须对手进入经训练的系统，因此即使是防护严密的人工智能仍然会受到攻击（Papernot、McDaniel和Goodfellow，2016年）。要了解弱点范围以及哪些输入值、输出值和数据需要保护，将需要进行更深入的研究。

有限效果对核安全的影响

在此前的跟踪和瞄准案例中，人工智能的威胁与破坏战略稳定性相关，因为对手过分相信其效果，但相反的情况也是成立的。在决策支持系统的案例中，利用人工智能的军队相信其有效的害处，比不相信其有效的害处更大。对手也有可能相信自己能够破坏人工智能和避免报复，导致其走上会令冲突升级的路线，包括先发制人发动第一次打击。例如，对手可能相信自己发现了一种发射和轨迹的模式，会让人工智能误判这是安全的，即使导弹正在飞向目标。

人工智能带来了一系列难以实时侦测的新弱点。但在战争路径、升级甚至发射决策中，人工智能几乎肯定—

人工智能的发展似乎不可阻挡，企业和政府纷纷扩大人工智能的应用，包括攻击性和防御性的用途。

最终或逐渐——获得更多重视。凡是承担这些职责的系统，都必须通过严格测试，包括对抗测试。只有测试者能够预想对手可能制造的所有攻击时，才能在测试中充分有效模拟对手。尽管如此，所有部署的军事系统都面临这项不可能完成的任务。

人工智能可能有一定的强化稳定作用

鉴于过去几十年来都没有发生核攻击，很容易把战略稳定性视为理所当然。虽然前述章节列出了人工智能的进步可能损害战略稳定性的方式，但事实未必如此。人工智能的发展似乎不可阻挡，企业和政府纷纷扩大人工智能的应用，包括攻击性和防御性的用途。人工智能对这些战略应用的影响只会逐步显现。人工智能有可能加剧核战略不同方面之间的紧张关系，但在有利的情况下也会缓和这些紧张关系，强化战略稳定性。有核国家虽然互不信任，但在自身利益的驱使下有可能为此展开协调。

高出错性过后的时期

研讨会参与者认同，在人工智能形成新的能力（例如跟踪和瞄准或有关冲突升级的决策支持）后是风险最大的时期。在这个试用期，较有可能出现错误和误会。随着时间推移和技术进步，这些风险有望减小。如果主要的启动能力是在和平时期开发出来的，则可以合理预期其会在原可首次投入战场的时间后持续取得进展，从而有时间提高其可靠性或充分了解其局限性。最终，人工智能系统所形成的能力，虽然可能会出错，但概率要低于人类，因此在长期具有稳定作用。

确保战略稳定性的潜在合作

使核战争风险持续存在的其中一项因素，是第一次打击稳定性所要求的保证报复（这促使政府采取“攻击下即发射反击”的核态势）与可能发生意外或失灵之间的矛盾。例如，1983年苏联的预警系统失灵，“探测到”不存在的美国攻击（Hoffman，2009年，第6至11页）。尤其是在危机情况下，这样的事故可能会促使官员下令对子虚乌有的袭击进行报复还击。人工智能可通过协助创建更可靠的预警系统缓和这个矛盾。而更强的第一次打击稳定性有助于减小危机意外升级的危险。即便如此，这种信任可能是好坏参半。自认为能够预测事态会否升级的侵略国，也许敢于采取囿于不确定性而不敢发动的挑衅行动。

情报收集和分析的准确性提高，也可能使威慑、保证和再保证更可信，从而巩固战略稳定性。如果潜在对手秘密谋划攻击的机会减少，威胁对本国或盟友动武的可能性将会下降。如果战略伙伴能获取更全面的情报和分析，则能更容易让他们放心。随着保证所需要的力量减少，美国这类核大国可以裁减核武器库规模，这将加强对敌方的再保证。这个过程可能会形成良性循环，最终极大降低战争风险。遗憾的是，这个结果的实现需要极幸运的条件，而不论人工智能技术处于什么状态。首先，所有行动方需要获得均等的情报和分析能力。在新兴情报不对称的局面中，较弱的国家很可能认为自身很容易受攻击，并对对手的疑心加重。此外，对手国家的意图必需是真正的善意。最后，官员对情报收集和分析系统（包括非人工智能的部分）的信心需要有充分的理由。为实现人工智能促进战略稳定性的潜力，随着这项技术的成熟，各国需要开始互相协调，以避免跌入陷阱。有关的讨论应包括外交和军方官员，以及技术专家。

彻底透明化

在一个非常乐观的可能性中，用于支持事态升级相关决策的人工智能算法有可能和对手分享。这种彻底透明

化伴随着多种风险。随后，对手或许会执行不可取的行动，直至升级到阈值的边缘。它也可能会破坏人工智能。同时，任何用于辅助上述决策的人工智能，都需要通过广泛的测试，包括对抗性测试。在任何情况下，设计人工智能时，努力保证它在敌方获得算法时仍然安全，是一个良好的做法；假定对手得不到它是很危险的想法（Kerckhoff，1883年）。如果人工智能计算机系统在投入实战前必须达到如此高的健全性标准，广泛散布也许能减轻顾虑，以及使错误判断的情况几乎不可能出现。

结论

整体而言，研讨会参与者认同，到2040年，人工智能有很大的潜力推翻核稳定的根基并削弱威慑力，在日趋多极化的战略环境中尤其如此。排除了好莱坞常见的邪恶人工智能试图用核武器毁灭人类的情况后，专家们转而讨论能力提升所产生的较为平凡的问题。所讨论的人工智能应用包括跟踪和瞄准对方发射器以打击军事力量的能力，以及将人工智能整合进决策支持系统，为决定是否使用核武器提供参考。

部分专家担忧，对人工智能的依赖加大可能会导致新型的毁灭性错误。有可能被迫在技术成熟前使用人工智能；人工智能也许很容易遭到对手破坏；或对手可能认为人工智能的能力比实际更强，导致他们犯下灾害性的错误。

另一方面，如果核大国成功建立与人工智能所带来的新能力相匹配的战略稳定性，机器可以减少猜疑和减缓国家之间的紧张关系，从而降低核战争风险。

目前，我们无法预计这些情景中哪一个（如有）会实现，但我们需要在这些挑战变得严峻之前，开始考虑人工智能对核安全的潜在影响。在未来数十年维持战略稳定性也许极其困难，所有核大国都必须参与制度建设，以帮助控制核风险。实现这个目标将需要极幸运的技术、军事和外交措施组合，以及对手国家的配合。我们希望这期观点能引发有关的讨论，并针对这些富有争议、经常两极分化的课题，开创通向实用主义和现实主义的道路。

附注

[1] 在本期《视角》中，我们使用非正式意义上的人工智能一词，它包括与人工智能广泛相关的研究项目所取得的许多计算机科学成果，即使这些成果最终与模拟人类智慧一事本身关系不大。这些项目产生了模式识别算法、新的编程语言、自然语言处理，以及各种其他在过去几十年被称为人工智能但早已进入主流计算技术的功能。

[2] 俄罗斯媒体账号称斯塔图斯-6采用人工智能实现自主能力。例子可见Tuchkov（2016年）和“Ros- siiskii proekt ‘Status-6’ meniaet sootnosheniia iadernykh sil v mire”（2016年）。后一篇文章称斯塔图斯-6“配备了人工智能”，能够沿着“原本不可达的路线”避开反潜作战措施，攻敌不备。

[3] 按照一般的说法，我们用通俗的语言包括符合人工智能定义的未来和近期发展，但不包括有很长应用历史的发展，即使这些任务在过去需要人类。

[4] 在冷战最后几年，苏联选择发展专门的导弹技术来反制美国未来的导弹防御设施。有关俄罗斯就苏联对罗纳德·里根战略防御计划的“非对称应对”的解释，见Oznobishev、Potapov和Skokov（2008年）。普京持续重复这种说法——例如，在他2012年的声明中，“俄罗斯对美国全球的反导弹防御设施及在欧洲的组成部分的军事技术应对将是有效和非对称的”（普京，2012年）。

[5] 2010年《核态势评论》指美国核战略的目标是“加强对朝鲜等地区对手的威慑力”，同时“巩固与中俄的战略稳定性”。然而报告没有提供“战略稳定性”的简要定义（美国国防部，2010年）。

[6] 美苏各自追求发展先发制人的打击力量，如果看来即将受到攻击，可用于解除对方的武装，但这应当与用于突袭的第一次打击力量区分开来。实际上，很难区分先发制人攻击与第一次打击所用的战略力量，导致两个超级大国的官员担心对方可能在谋划启动核战争。

[7] 关于Perimetr利用某种人工智能的说法反复在俄罗斯国家媒体上出现。例子见Timoshenko（2015年）和Valagin（2014年）。

[8] 例子参阅Akhmerov、Akhmerov 和Valeev（2016年）。

[9] 兰德公司未发表的研究报告，作者为Brien Alkire和Jim Powers。

参考文献

Akhmerov, D. E., E. N. Akhmerov, and M. G. Valeev, "'Uiazvimost' kontseptsii neiadernogo razoruzheniia strategicheskikh iaderniyi sil Rossii" ["The Dubiousness of the Concept of a Non-Nuclear Disarming Strike Against Russia's Strategic Nuclear Forces"], *Vestnik akademii voennykh nauk*, Vol. 54, No. 1, 2016, pp. 37–41.

Anderson, H. S., A. Kharkar, B. Filar, and P. Roth, *Evading Machine Learning Malware Detection*, blackhat.com, July 2017. As of August 15, 2017: https://www.blackhat.com/docs/us-17/thursday/us-17-Anderson-Bot-Vs-Bot-Evading-Machine-Learning-Malware-Detection-wp.pdf

Biggio, B., B. Nelson, and P. Laskov, "Poisoning Attacks Against Support Vector Machines," *Proceedings of the 29th International Conference on Machine Learning*, July 2012, pp. 1467–1474. As of August 15, 2017: https://arxiv.org/pdf/1206.6389.pdf

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014.

Bracken, P., "The Intersection of Cyber and Nuclear War," *The Strategy Bridge*, blog post, January 17, 2017. As of August 15, 2017: https://thestrategybridge.org/the-bridge/2017/1/17/the-intersection-of-cyber-and-nuclear-war

Buchan, G., D. Matonick, C. Shipbaugh, and R. Mesic, *Future Roles of U.S. Nuclear Forces: Implications for U.S. Strategy*, Santa Monica, Calif.: RAND Corporation, MR-1231-AF, 2003. As of March 8, 2018: https://www.rand.org/pubs/monograph_reports/MR1231.html

Cimbala, S. J., *The Dead Volcano: The Background and Effects of Nuclear War Complacency*, Westport, Conn.: Praeger, 2002.

Etherington, D., "Google's AlphaGo AI Beats the World's Best Human Go Player," TechCrunch, May 23, 2017. As of August 15, 2017: https://techcrunch.com/2017/05/23/googles-alphago-ai-beats-the-worlds-best-human-go-player/

Garthoff, R. L., "Refocusing the SDI Debate," *Bulletin of the Atomic Scientists*, Vol. 43, No. 7, September 1987.

Geist, E., "Was There a Real 'Mineshaft Gap'? Bomb Shelters in the USSR, 1945–62," *Journal of Cold War Studies*, Vol. 14, No. 2, Spring 2012, pp. 3–28.

Goodfellow, I., Y. Bengio, and A. Courville, *Deep Learning*, Cambridge, Mass.: MIT Press, 2016.

Healey, D., *The Time of My Life*, London: Michael Joseph, 1989, p. 243.

Hoffman, D. E., *The Dead Hand*, New York: Doubleday, 2009.

Holmes, J. R., "Sea Changes: The Future of Nuclear Deterrence," *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, 2016, pp. 228–233.

Kahn, H., *On Thermonuclear War*, Princeton, N.J.: Princeton University Press, 1960.

Karpathy, A., "Breaking Linear Classifiers on ImageNet," *Andrej Karpathy blog*, March 30, 2015. As of August 15, 2017: http://karpathy.github.io/2015/03/30/breaking-convnets/

Kearns, M., and M. Li, "Learning in the Presence of Malicious Errors," *SIAM Journal on Computing*, Vol. 22, No. 4, March 1992, pp. 807–837. As of August 15, 2017: https://doi.org/10.1137/0222052

Kegelmeyer, P., T. M. Shead, J. Crussell, K. Rodhouse, D. Robinson, C. Johnson, D. Zage, W. Davis, J. Wendt, J. Doak, T. Cayton, R. Colbaugh, K. Glass, B. Jones, and J. Shelburg, *Counter Adversarial Data Analytics*, Albuquerque, N.M.: Sandia National Laboratories, SAND2015-3711, May 2015. As of August 15, 2017: http://www.sandia.gov/~wpk/pubs/publications/cada-full-uur.pdf

Kerckhoff, A., "La Cryptographie Militaire," *Journal des Sciences Militaires*, January 1883.

Long, A., *Deterrence from Cold War to Long War: Lessons from Six Decades of RAND Research*, Santa Monica, Calif.: RAND Corporation, MG-636-OSD/AF, 2008. As of March 8, 2018: https://www.rand.org/pubs/monographs/MG636.html

Long, A., and B. R. Green, "Stalking the Secure Second Strike: Intelligence, Counterforce, and Nuclear Strategy," *Journal of Strategic Studies*, Vol. 38, Nos. 1–2, August 2012, pp. 38–76.

Mack, C. A., "Fifty Years of Moore's Law," *IEEE Transaction on Semiconductor Manufacturing*, Vol. 24, No. 2, May 2011, pp. 202–207. As of August 15, 2017: https://doi.org/10.1109/TSM.2010.2096437

Mahan, A. T., *Armaments and Arbitration: Or, The Place of Force in the International Relations of States*, New York: Harper & Brothers, 1912.

Minsky, M., *Semantic Information Processing*, Cambridge, Mass.: MIT Press, 1968.

Newell, A., J. C. Shaw, and H. A. Simon, *Report on a General Problem-Solving Program*, Santa Monica, Calif., RAND Corporation, Report P-1584, revised February 9, 1959.

Oznobishev, S. K., V. Ia. Potapov, and V. V. Skokov, *Kak gotovilisia "asymmetrichnyi otvet" na "Strategicheskuiu oboronnyiu initsiativu" R.Reigana. Velikhov, Kokoshin i drugie* [*How the "Asymmetric Response" to R. Reagan's "Strategic Defense Initiative" Was Prepared*], Moscow: Legand, 2008.

Papernot, N., P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples," arXiv, May 24, 2016. As of August 15, 2017: https://arxiv.org/abs/1605.07277

Putin, V., "Byt' sil'nymi: garantii natsional'noi bezopasnosti dlia Rossii" ["Being Strong Is the Guarantee of National Security for Russia"], *Rossiiskaia gazeta*, February 20, 2012. As of December 5, 2017: https://rg.ru/2012/02/20/putin-armiya.html/

Roland, A., and P. Shiman, *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*, Cambridge, Mass.: MIT Press, 2002.

"Rossiiskii Proekt 'Status-6' Meniaet Sootnosheniia Iadernykh Sil v Mire" ["The Russian Status-6 Project Is Changing the World's Nuclear Balance of Forces"], *Russkaia Politika*, November 14, 2016. As of December 4, 2016: http://ruspolitica.ru/post/rossiyskiy-proekt-status-6-menyaet-sootnoshenie-yadernyih-sil-v-mire/

Schelling, T., *Arms and Influence*, New Haven, Conn.: Yale University Press, 1966.

Shannon, C. E. "Programming a Computer for Playing Chess," *Philosophical Magazine*, Vol. 41, No. 7, 1950, pp. 256–275.

Simon, H. A., and A. Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research*, Vol. 6, No. 1, January–February 1958, pp. 1–10.

Slocombe, W., "The Countervailing Strategy," *International Security*, Vol. 5, No. 4, Spring 1981, pp. 18–27.

Sutyagin, I., "Russia's Underwater "Doomsday Drone": Science Fiction, but Real Danger," *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, June 2016, pp. 243–246.

Timoshenko, M., 'Mertvaia ruka' na strazhe perimetra Rossii" ["The 'Dead Hand' Guarding Russia's Periphery"], *Telekanal "Zvezda,"* February 18, 2015. As of August 15, 2017: http://tvzvezda.ru/news/krasnaya_zvezda/content/201502181414-gskc.htm

Tuchkov, V., *Status-6: Oruzhie Vosmezdia, Vognavshee Pentagon v Stupor* [*Status-6: The Retaliatory Weapon That Drove the Pentagon into a Stupor*], Svobodnaia Pressa, December 11, 2016. As of December 4, 2016: http://svpressa.ru/war21/article/162378/

U.S. Department of Defense, *Nuclear Posture Review Report*, Washington, D.C., April 2010.

Valagin, A., "Garantirovannoe vozmezdie: Kak rabotaet rossiiskaia sistema 'Perimetr'" ["Assured Retaliation: How the Russian 'Perimetr' System Works"], *Rossiiskaia gazeta*, January 22, 2014. As of August 15, 2017: https://rg.ru/2014/01/22/perimetr-site.html

Woyke, E., and Y. Kim, "Starcraft Pros Are Ready to Battle AI," *MIT Technology Review*, May 19, 2017. As of August 15, 2017: https://www.technologyreview.com/s/607888/starcraft-pros-are-ready-to-battle-ai/

关于本《视角》报告

我们谨此感谢兰德公司全球风险与安全中心以及中心主任Andrew Parasiliti发起这项研究工作，并感谢RAND Ventures的资助。对于Angela O'Mahoney和 Bill Welser全程提供的指导，Sonni Efron、Doug Irving和Greg Baumann协助搜集素材，Hosay Yaqub为成功举办有关活动所做出的努力，在此一并表示谢意。最后，我们还要感谢研讨会参与者，根据查塔姆宫守则略去与会者姓名。

2040年安全局势

本《视角》报告是RAND Ventures一项倡议的一部分，以预想2040年全球的关键安全挑战，考虑政治、技术、社会和人口趋势的影响，它们将在未来几十年左右这些安全挑战。本项研究在兰德公司全球风险与安全中心内进行。

兰德公司全球风险与安全中心

全球风险与安全中心在兰德公司开展多学科研究和政策分析，应对全球安全的系统性风险。中心利用兰德公司的专业知识补充和扩展兰德公司在许多领域的研究，包括安全、经济、卫生和技术。众多杰出的商业领袖、慈善家和前政策制定者提供建议并支持中心的工作，其研究重心日益趋向全球安全趋势以及颠覆性技术对风险和安全的影响。有关兰德公司全球风险和安全中心的详情，请访问www.rand.org/international/cgrs。

RAND Ventures

兰德公司是一家解决公共政策挑战的研究机构，旨在协助推进全球社区的安全、卫生与繁荣事业。兰德公司致力于公共利益，属于非营利性、无党派组织。

RANDVentures是投资于制定政策解决方案的机构。慈善捐款支持我们高瞻远瞩、处理棘手、经常富有争议的课题，以及通过创新和有吸引力的方式分享我们的发现。兰德公司的研究结果和建议以数据和证据为基础，因此不一定反映其委托人、赞助商或支持者的政策偏好或利益。

这项事业的资金来自兰德公司支持者的捐赠和经营收入。

关于作者

Edward Geist是兰德公司的政策研究副研究员。他此前是斯坦福大学国际安全与合作中心 (CISAC) 麦克阿瑟基金核安全研究员，以及兰德公司华盛顿分公司的斯坦顿核安全研究员，2013年5月从北卡罗来纳大学取得俄罗斯历史学博士学位。

Andrew J. Lohn是兰德公司的工程师。他运用各种数学和机器学习技术来提供对高技术性政策问题的新见解，例如，网络战、人工智能或无人机投递。Lohn持有加利福尼亚大学圣克鲁兹分校的电气工程学博士学位。