

## 技术轮盘赌

# 各国军方追求技术优势背景下的失控管理

理查德·丹泽



## 关于作者

**理查德·丹泽**是约翰霍普金斯大学应用物理实验室的高级研究员，曾在克林顿政府时期担任海军部长。

## 致谢

我曾接受情报高级研究计划署（IARPA）的委托，研究美国应该如何应对挑战，在吸收和开发国家安全相关技术方面保持优势。本文由该研究的一节扩充而成。情报高级研究计划署署长Jason Matheny为前项研究提供了大力支持，并在研究过程中多次提出真知灼见。对于他对这项研究所倾注的智慧、心血和关注，我在此表示感谢。

Open Philanthropy Project为本文的扩展和深化提供了支持，并准备将其作为普通出版物出版。美国新安全中心负责其出版工作。对于来自Open Philanthropy Project的Helen Toner和Claire Zabel以及美国新安全中心的Paul Scharre和Loren DeJonge Schulman所做的贡献，我特别表示感谢。我还要感谢Kara Frederick、Maura McCarthy和Melody Cook，他们参与了本报告的审稿、制作和设计。

Open Philanthropy Project在今年早些时候曾主办研讨会，会上许多与会者讨论了本文的草稿，并提出了独特的见解。对于Jeff Alstott、Jim Baker、Jack Clark、Allan Dafoe、Craig Fields、Dan Geer、Larry Gershwin、Chris Inglis、Jason Matheny、Tim Maurer、Tyson Meadors、Jim Miller、Ernest Moniz、Anne Neuberger、David Relman、Peter Singer、Helen Toner、Renee Wegrzyn、Bob Work和Philip Zelikow的积极参与，我表示感谢。Craig Fields、Dan Geer和Phil Zelikow还在会后提供了详细的后续意见。

此外，Chris Beinecke、Steve Burton、Miles Brundage、Paul Gewirtz、Michael Hopmeier、Lou Pringle、Al Martin、Alex Montgomery和Jonathan Reiber都对原稿发表了意见，拓展了我的思路。Peter Levin不仅提出了尖锐批评和编辑建议，而且还给予我很大的鼓励。Jack Goldsmith教授还将本文分发给他的哈佛大学法学院的学生，并收集到很多有用的意见。我向他们表示感谢，其中也包括Chris Kirchhoff——当时我因波士顿发生暴风雪而无法到达讨论现场，是他帮我主持了那场讨论。Alex Downes针对关于规范的政治学文献提供了宝贵见解。Terry Leighton帮助探究了1957年H1N1全球流感大流行的挑战性问题。我还要感谢Gigi Kwik Gronvall和Tom Inglesby就此问题与我进行通信。

当然，以上所提人士均不对此处表达的观点或仍然存在的错误负责。再次感谢所有人的鼎力协助。

## 封面照片

Getty Images，美国新安全中心改编

# 技 术 轮 盘 赌

## 各国军方追求技术优势背景下的失控管理

2	执行摘要
4	引言
5	区分外行恐惧和专家恐惧
6	为什么关注军事发展和技术应用？
8	失控的原因
10	无法可靠和令人信服地衡量这些风险
11	人为介入
13	应该怎么办？
16	总结
18	附录
19	尾注

## 执行摘要

本报告认可激励美国军方追求技术优势并超过所有潜在对手的必要性，但同时又强调这种优势并不等同于安全。核武器、航空和数字信息系统方面的经验，也可用于探讨当前控制人工智能、合成生物学和自主系统的努力。从这个角度，最可能出现的结果是，如果引入了复杂、难懂、新颖和交互式技术，将会引发事故、产生涌现效应和造成破坏。总之，美国国家安全机构将会在许多场合和以多种方式，失去其对所创之物的控制。

我们不断追求技术优势存在一个强大的理由，即这种优势有助于增强我们的威慑力。但是，威慑这种战略旨在减少攻击，而非减少事故；旨在劝阻恶意，而非劝阻疏忽大意。事实上，技术创新之后，接踵而至的几乎必是技术扩散。由于技术扩散会将巨大的破坏力置于他人之手，而他们的安全优先顺序和标准可能不如我们完备，而且缺少资金，因此相互作用数量与复杂程度的增加将会给我们带来更多风险。

因此，当我们朝着自己的主要目标——即追求优势——奋力迈进，它势必会增加而不是减少失控的附带风险。很遗憾，本报告认为我们无法可靠判断产生的风险。更糟糕的是，我们还没有哪种途径可有效将其化解，甚至无法阻止风险的增加。以往人们常常提到一种辅助性的方法，就是在涉及新技术的运作中引入“人为介入环节”，其好处现在看来是微乎其微，而且在不断减少。

但是，我们对此并非束手无策。随着问题得到越来越多的关注，美国军方至少应该抑制事故增多的可能性，并减少事故发生时的不利后果。设定美国将会成为事故、涌现效应和破坏的受害者，它则应该改进其计划，以应对此类后果。这将涉及到重新分配我们军队的巨大能量，将原本用来应对恶意行为而投入的一部分转为做好计划以应对疏忽大意的行动以及相关活动。

美国国防部和情报机构设计的技术和系统不仅要能提高效率 and 效力，而且还要更多关注如何减轻失效的后果，并建设弹性的恢复能力。数字信息系统普遍缺乏安全性，应能客观说明当我们对一项新技术产生依赖后（而不是在进行设计的时候），进行失控管理不仅代价极高，而且有时无法做到。

最不容易的是美国还必须与对手合作，帮助他们做好控制并尽量减少发生事故的风险。二十一世纪的技术具有全球性，这不仅是指分布状况，而且也体现在后果方面。他人意外释放的病原体、计算机病毒、辐射以及推出的人工智能系统，同样也会危及我们。要化解共同面临的多种风险，我们必须执行一致的报告系统、共享的控制、共同的应急计划、规范和条约。鉴于采取这些重要措施存在困难，我们应认识到其中的最大挑战不是构建我们与技术的关系，而是构建我们彼此之间的关系。

这些观点虽针对的是国家安全领域，但相关反思和建议应超越特定领域，对新技术危险性控制方面的所有讨论产生影响。

**“与俄式轮盘这一界定明确、只要会计算6 的乘除的人便能预知风险的游戏不同，没有人能够看清现实的枪管。”**

**— 纳西姆·尼可拉斯·塔雷伯**

# 引言

技术创新与战争创新长期交织在一起，或许始终是相伴相随。20世纪的世界大战是对这种关系的实际见证，二次大战对我们当前观念的影响尤其大。每一位军事战略家当时都明白——而且现在还记得——这场战争开始于德国的“闪电战”，这一军事战略利用到了内燃机、信息加密和无线电；最后终止于英国布莱切利园的密码破译，以及美国利用科学和技术打造出的原子弹。

这场冲突结束时美国已拥有突出优势，其中包括毫发无损、不断复兴的工业企业，世界一半的国内生产总值，以及全球的大部分科技机构。四分之三个世纪以来，我们的国家安全战略凭借众多的战后机构、充满活力的商业部门、卓越的学术机构和资金充足的政府计划，充分扩展和利用了这些优势。<sup>1</sup>

## 技术创新与战争创新长期交织在一起，或许始终是相伴相随。

随着世界步入正常化，这种技术、经济和军事三重优势已经开始承受压力。依仗拥有全球四分之一的国内生产总值，我们仍占据着优势地位，但已不如以前显赫。其实到本世纪中叶，中国的国内生产总值预计将会高出我们50%。就近代而言，美国从未遇到过可与之抗衡的财富对手，更谈不上将其击败。<sup>2</sup>此外，科学家、技术知识和商业生产现已分布到了世界各地。其他国家、外国公司、甚至恐怖犯罪集团都拥有了必要的资金，来购买、创造、反向设计制造和窃取技术产品、见解和方法。可以说，他们在这方面正设法成为美国。面对这种局面，我们选择了通过加倍努力来保持技术优势。

至于美国为什么希望、如何以及能否保持优势，这方面的论述已经太多，本报告不再赘述。本报告不具体建议我们该如何维持优势地位，只是假定国家安全机构将会加倍努力，以维持优势地位，并能获得成功。

本报告关注的是，随着美国以及（在其带领下）其他许多国家致力于通过新技术增强其军力，一系列共生风险将随之而来。这些风险包括复杂系统的意外后果、分析或运行错误、单独开发系统之间的交互影响以及人为破坏所造成的系统漂移。

各国对先进技术能力的依赖，就像将越来越多的子弹装入越来越多的手枪，而枪口指向的正是人类。即使没有人希望直接开枪，却也可能意外走火。一个国家因疏忽大意而

释放的病原体、计算机病毒或人工智能自主本体，可能会对全球产生影响。对自身造成破坏——可能是灾难性破坏，一种无意的行为也可能引发其他反应。比如，俄罗斯、中国、美国、以色列或伊朗为了收集情报对计算机漏洞的利用，可能会意外破坏、或被认为能够破坏电网或金融系统的关键部分，从而招致灾难性报复。

这些例子并不是一种预测，而只是一种例证。本文重点探讨的是将会导致不可预知的结果及可能性的环境条件，尤其是那些可能产生大规模后果的环境条件。

正如引语中塔雷伯所言，我们无法计算枪筒中的弹药数量，但枪支和子弹的增多一定会增加我们的危险性。这种场景相当于俄式轮盘赌博游戏。由于这种赌博通常只用一颗子弹，本报告为了加以区分，特将我们的这种新游戏称为“技术轮盘”。这种游戏每天都在全球上演，玩家数量越来越多，武器种类越来越多，<sup>3</sup>拥挤的技术空间中的交互越来越多，一颗子弹走火便会引发众多（或全部）玩家的致命还击。

当然，名称并不重要，重要的是问题的动力学本质。对于这些危险人们早有认知，<sup>4</sup>尤其是通过分析美国的核指挥与控制系统。<sup>5</sup>而数字信息技术的脆弱性正是这些危险的另一种体现。第一个挑战是从这些经验中获得对风险的深入了解，并将其应用于更新的技术计划，其中包括人工智能、合成生物和自主系统。接下来，美国国家安全机构必须要能认识到，随着这些技术扩散至其他国家，我们面临的风险也将上升；这些国家的保障措施可能更少；我们双方的系统在互动时会带来技术性的风险。在此背景下，美国机构需要采取措施，降低我们自己的研发计划和运行中的风险，并思考我们该如何引导其他国家采取类似行动。本报告便是这个目的，肯定不够完美，但应该有望发挥一定的作用。

最重要的是，它传达了一种虽难以接受但在作者看来是非常重要的信息。国家安全专家渴望能争取到优势地位，强调军事优势能够最大程度地保障安全，其所赋予的威慑力是最可靠的工具。这些观点非常正确，也非常得当。它们能使这些专家专注于自己最擅长和最舒适的领域——赢得竞赛，而不是超越它或认识到它的危险。但是，这么做还不足够。事故和疏忽大意的“紧急”影响会在竞争框架之外产生风险。<sup>6</sup>尽管威慑必不可少，但它最多只能边际性地降低事故或意外效果的风险。简而言之，优势并不等同于安全。

创造出了技术机遇的科学家和工程师拥有自己的倾向性，这

种倾向性常会掩盖我们技术风险所带来的影响。与军事计划相比，他们时常选择讨论的各种环境下的危险并不紧张、复杂或具体，而且也缺少爱国主义的敏感性。那些关注当前最受争议的人工智能技术的人士，能够洞察到自主军事武器的风险，但是，他们认为未来的危险主要源于去除了人为“介入”，由此在无意间掩盖了当前迫在眉睫的问题。这在三个方面存在误导：它低估了机器决策对军事决策的影响程度；它高估了人类而不是自动化系统充当决策者可能带来的益处；它转移了人们对以下事实的注意力：其他技术也会带来人工智能所孕育的独特挑战——比如，生物学家创造出了具有自主性的新病原体。

本报告指出，我们的国家安全机构可能失去对技术的控制——由此产生他们不希望看到的结果——不只是未来将会面对风险，而是目前已经产生危险。不幸的是，新军事技术的利用和互动所产生的不确定性，并不取决于充满自信的估算或控制。还没有哪种途径能够有效化解这种风险，或阻止它们增加。通常情况下，加大行动决策中的人为控制并非可行之策，也不能令人满意。不过，我们能够采取一些技术和政策措施，来减轻失控的风险。对于这些措施，美国应给予更多的重视。采取这些措施最大挑战不是构建我们与技术之间的关系，而是构建我们彼此之间的关系。

## 优势并不等同于安全。

本报告将分六个部分来陈述观点和建议，最后做一总结。第一部分指出，外行常会担心新技术产生的后果，因为他们对现有的过程和权力关系感到不熟悉、紧张和不平衡。不过本报告也指出，正是那些通常不会产生这些顾虑的人们——也就是本报告专门指出的专家和新技术倡导者，现在出乎意料地在强调这些技术所固有的灾难性风险。第二部分讨论了为什么将这些技术用于军事——尽管拥有有价值的训练和常规预防程序——会产生特别大的风险，导致疏忽大意和不希望出现的结果。第三部分概述了产生——并可能重复产生——这些结果的五个主要途径。第四部分描述了为什么这些失控风险无法估算，即使是最小的信心指数也不行。第五部分剖析了“人为介入”的优势，并认为虽然这种安全保障具有优势，但这些益处目前非常有限，并可能在未来受到更多限制。第六部分认为，对于这种不确定性，我们能够而且必须进一步加强军事技术的控制。这一部分详细说明了五项具体的行动。总结部分阐述了可从研究分析中得出的一些基本经验。

## 区分外行恐惧和专家恐惧

在动人的演讲中警醒大家新技术会带来技术的末日，这是

很容易的一件事。其中大多数只是表达因未知而感到不安的人们的情绪。恐惧总与技术进步形影不离。新技术通常都是不请自来，就像被带入既定行动领域的特洛伊木马。<sup>7</sup>竞争压力让这些进步变得无法拒绝。一旦采用，它们就会发挥变革性作用，然后变得不可或缺。如果继续走下去，其效率会降低，然后变得衰弱无力，再然后变得不可信（甚至是不体面）。

新系统常常都是难懂、复杂的，而且需要持续修改和更新，这样它们才能与其他不断发展的技术系统进行正确的互动。这些属性增强了用户的失控感。随着机器的能力变得更为强大和重要，由此产生可以理解的焦虑：每个人都会担心自己的价值和个性正在遭到侵蚀——更糟糕的是我们人类正在失去权力和地位。普罗米修斯到来了，伊卡洛斯飞得太远了。弗兰肯斯坦博士创造了太多。

如果只有这些也就罢了。如果我们的恐惧只限于我们祖先在20世纪面对的自动化，或19世纪面对的工业化，那该多好。但是，在21世纪第二个十年结束时，或者说是（如果你愿意的话）自车轮发明以来技术进步迈入的第550个十年，一些不同寻常的东西总在伴随着这些根深蒂固、<sup>8</sup>经常闪现和不断膨胀的恐惧。现在不仅仅是那些受到这些技术影响的外行将会感到不安，就连它们的创造者和支持者也开始惶恐。引发我们关注的警报不是由头脑发热的卢德分子从工厂发出，而是由创造者从实验室拉响。<sup>9</sup>

## 现在不仅仅是那些受到这些技术影响的外行将会感到不安，就连它们的创造者和支持者也开始惶恐。

对于美国国家安全政策的制定者来说，最先引发专家担忧的是爱因斯坦、奥本海默及其同事制造出的原子武器；<sup>10</sup>而现在最明显的是人工智能和合成生物学。斯蒂芬·霍金及其同事表达了他们对人工智能的看法：“虽然它的短期影响取决于由谁控制，但长期影响却在于能否控制得住。”<sup>11</sup>一年后，他们被（根据最新统计）3,722位“人工智能/机器人研究人员”所取代，他们敦促“禁止使用超出了人类控制范围的攻击性自主武器。”<sup>12</sup>生物学家担心基因编辑方面的突破会引发类似担忧。<sup>13</sup>与几位前任相比，现在的国家情报总监对核武器方面的警告更关切。在参议院听证会上，他将基因编辑称为潜在的“大规模杀伤性武器”：

法规或道德标准不同于西方社会的国家进行基因组编辑研究，可能会增加产生潜在有害生物制剂或产品的风险。鉴于这种军民两用技术的分布广、成本低、发展快速，一旦出现故意或非故意滥用，便会对国家经

济和安全产生深远影响。基因组编辑研究在2015年的进展，迫使美国和欧洲众多知名的生物学家开始对不受约束的人类种系基因编辑提出疑问。<sup>14</sup>

数字技术专家现在也开始关注灾难性风险，只不过是出于其他原因和通过其他方式。引发他们担忧的与其说是未来不确定性，不如说是我们目前对极易遭到破坏的技术的普遍依赖性，而且这种依赖还在不断增强。专家懂得如何实现小型系统的信息完整性、可用性和机密性，但他们保护这些特性的能力却无法拓展。但是，从“小东西”到“云计算”，数字系统的普及和扩散已远超出了我们的控制范围。正如一位专家所言，“尽管安全水平正在提高，但事情变糟的速度却更快，其结果是我们不断改进，最终却节节败退。”<sup>15</sup>因此，另一位著名的思想家最近在撰写一篇重要论文时说：“毫不夸张地说，网络安全已同人类未来紧密联系在一起。”<sup>16</sup>

## 为什么关注技术的军事发展与应用？

民间研发活动往往与军事发展交织在一起。即使是独立进行，商业和学术工作——比如与病原体有关的商业和学术工作——也可能产生自己的风险。这些方面肯定值得关注。但是，考虑到军事武器的威力和后果，需要特别关注与这些系统有关的技术。技术发展不仅能决定这些武器的特性，而且也会影响传感器、通信和分析系统的特性，这些系统在很大程度上决定了武器的使用与否，以及如何使用。

除了这些考虑因素之外，还有一个不太明显但对本报告至关重要的因素。尽管军事系统拥有统一的指挥、严格的程序、紧张的训练以及对人员和计划的精心筛选，<sup>17</sup>但也存在特别容易出现人为错误、涌现效应、误用和误解的特性。它们包括先进武器开发及应用的保密性；运行互动和环境的不可预测性；专家技能与军事任务之间的不匹配性；军事系统之间存在的相互依赖性和脆弱性，尤其是美军方面；部署新技术，满足战场行动需求的紧迫性；以及军事竞争的无约束性。

本节将依次对这些问题展开讨论。

### 保密性

民间使用复杂技术通常需要进行审查，接受监督和监管。这些措施并不完美，而且需要付出一定成本。但是，当美国证券交易委员会对高速计算机交易、美国食品药品监督管理局对药品开发、联邦航空局对飞机创新……进行监管时，通

常都能降低风险，有时还可戳破过度自信的气泡。<sup>18</sup>可见性也能促进信息交流，以便开发商和运营商从事故和错误中吸取教训，改进自己的工作。<sup>19</sup>公开数据有助于第三方展开风险评估，以及通过诉讼、立法建议、可对品牌产生影响的宣传等方式发表反对意见。相比之下，军事系统在美国的可见性就非常有限，更不用提在威权国家。

即使在军事组织内部，机密技术经常也不会告知可能受到影响的人们，甚至只有到了战时或在其他紧急情况下才通知他们使用。这样导致的一个结果是主流技术无法与机密系统——尤其是高度机密系统——协调一致。具有复杂交互关系的民用组织都会努力优化相互操作和相互依赖的可见性，<sup>20</sup>而机密系统在过度保密倾向的推动下，<sup>21</sup>却会彼此隐藏自己的部分组织。

### 运行互动与环境的不可预测性

斯科特·萨根曾在25年前的一次经典分析中，将“高可靠性组织”描述为“相对封闭的系统，它们会极力减少行为者和组织外部环境对安全成果的影响。”<sup>22</sup>预先测试虽不完美，但却是解释和隔离外部变量的主要方式。因此，在对民用技术产生信赖并大规模部署之前，我们通常要求对它们进行广泛的应用环境测试。自动驾驶汽车投入常规使用前，需在现实世界条件下积累数千万英里的行驶经验。进行多年仔细标定的临床试验，是引入药物和医疗器械的先决条件。部署主要软件系统之前，需在各种使用环境中进行贝塔测试。军方也会广泛测试自己的产品，<sup>23</sup>但战争却是一种最不可预测的环境。使用的时间、地点和环境，甚至是对手和使用者的特征都是瞬息万变，根本无法完全预料。<sup>24</sup>大规模行动的隐藏性和复杂性<sup>25</sup>会因对手的出其不意而进一步加剧。因此，军事技术的测试和应用推断不如民用系统那么可靠。事故率通常在新系统中最高，然后会随着经验的积累和逐步改进而减少。<sup>26</sup>这种改进也会使军事创新受益——比如战斗机的更新换代，但军事技术被储备起来待用，或者很少运用，或者仓促引入，<sup>27</sup>并没有获益。

### 分配政策经常与主管和技术不匹配

军事训练的投入非常特殊，一些军事部门——特别是海军核武器部队——要求将技术优秀作为晋升的先决条件。但一般而言，根据不同的职位以及行动和地理条件，校级军官和高级士兵每两到三年会轮流调整一次，其稳定性远低于民用环境。分配和重新分配给军队的技术环境宽大而多变。晋升政策优先考虑的是管理宽度，而不是技术深度。训练常常不能很好地应对在不同环境下运行传统系统和现代系统的挑战。这样便会导致事故发生。<sup>28</sup>

## 行动相互依赖性

过去二十五年来，我们军队已将联合行动和建立“系统的系统”放在了首要位置。<sup>29</sup>其结果是，我们的许多防御系统的相互关联性和相互依存性比我们这个时代网络已经广泛普及的民用系统标准要超出许多。<sup>30</sup>美国的战略是朝着这个方向继续迈进，这也会被其他各国军队所效仿。正如一位记者最近所指出：“后勤主任都将目光瞄向了一种军事优势战略：连接一切。”<sup>31</sup>

这些系统非常复杂难懂，具有相互依存性<sup>32</sup>，而且遵守“CACE（Changing Anything Changes Everything）”原则，即改变任何一处，都将改变一切。<sup>33</sup>军事行动对单独系统组合的依赖，会增加涌现效应的脆弱性。它能“导致一种深度纠缠：如果残差与其他组件有更为强的关联，那么改进单个组件模型实际上可能削弱整个系统的精确度。”<sup>34</sup>商业企业的网络系统受到严格监管，非常讨厌分区运作。<sup>35</sup>相比之下，超越军事服务的网络通过连接孤立的官僚机构而建立，其历史可追溯至18世纪和19世纪。

## 不间断行动的紧迫性和重要性

军队履行使命的天职以及在紧急情况下“保证完成任务”的态度虽是老生常谈，但平民不一定能完全领悟。和平时期设计和推出军事武器和平台，目的就是为了规避风险。但在战争时期，甚至在“和平时期”的战斗中，军队通常会接受平民中无法容忍的后果。比如，第二次世界大战期间，美国希望从双翼飞机转向喷气式飞机，为了加快新兴技术的应用步伐，它推出了50多种战斗机和大量的轰炸机和支援飞机。<sup>36</sup>由于必须尽快吸收这些新的能力，安全问题遭到了忽视：战争期间，美国大陆的飞行事故“造成15,000多人死亡，相当于第二次世界大战的一个陆军师。”<sup>37</sup>

## 军事竞争的无约束性

企业之间的商业竞争非常激烈，但军事竞争所导致的利害关系更为严重：竞争对手若是拥有丰富资源的单一民族国家，行动起来必会更为偏执、无所顾忌。军事竞赛中，安全机构即使认定会造成破坏，但仍会自豪地执行，甚至破坏企图失败，还增加事故和意外影响的风险。此外，对军事对手的恐惧也会加剧冒险意愿：如果他们可能制造某种武器，我们必须抢先一步，或至少能了解和抵御他们可能使用的武器。



1954年3月1日，美国测试了有史以来威力最大的热核武器，当量（百万吨）比预期高出三倍，并产生了超出预料的放射性沉降物。本图为布拉沃城堡核试验。（美国能源部/国家海洋和大气管理局）

## 失控的原因

军事技术失控存在许多原因，本节将剖析五个主要问题：技术创造者的分析错误，技术使用者的操作失误，技术演化和互动所产生的意外涌现效应，对手对这些技术的破坏可能导致决策者渎职和态势感知失误，由此误导对相关技术的部署。

### 美国或对手的分析和操作失误

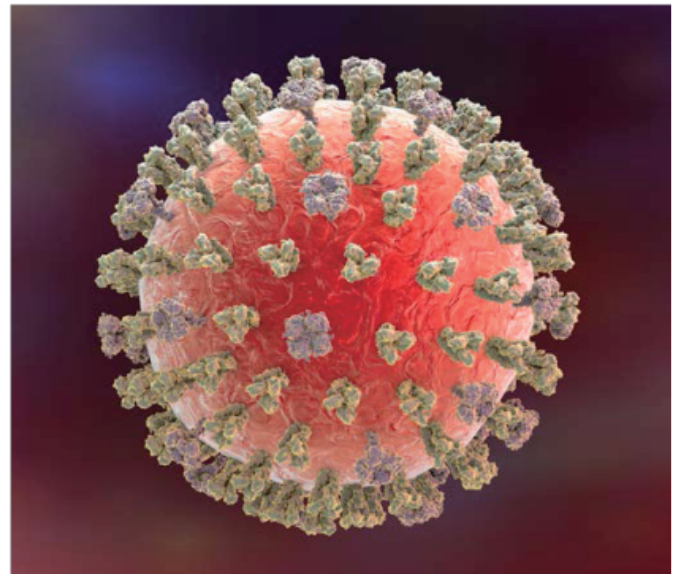
军事方面的分析错误事例包括1954年原子弹试验之前发生计算错误，导致当量比预期高出三倍，665人由此受到严重辐射。<sup>38</sup>我们军方的操作失误包括1961年，两颗原子弹从飞行中的B-52轰炸机机翼脱落；<sup>39</sup>1968年，一架携带4枚1.1百万吨威力炸弹的B-52轰炸机坠毁；<sup>40</sup>一次，由于没有移除代表核攻击的训练标识，致使警戒员误认为发生攻击，随即派遣飞机准备反攻，直至最终召回；<sup>41</sup>2015年，活炭疽菌由于贴错标签而被运至86个实验室，国防部副部长称之为“可导致生物毒素危机的大规模机构失误”。<sup>42</sup>

为了评估和化解事故风险，我们的国家安全机构必须剖析自己以及潜在对手的系统。即使其他人也同我们一样，是为了安全而投资，但技术扩散同样会增加技术风险。随着行为者数量以及他们互动次数的增加，这些风险也会增加。此外，由于其他人不可能像我们一样安全，这些问题将会变得更严重。<sup>43</sup>许多其他军队预算越来越紧张，设备越来越陈旧，控制越来越少，对事故造成的人员和环境损失也更冷漠。<sup>44</sup>

有些事故只会产生局部影响，比如1979年，苏联斯维尔德洛夫斯克军事设施中的炭疽被意外释放，造成大约80人死亡。<sup>45</sup>还有些事故会在局部产生主要后

果，并在全球形成次要后果，比如1986年，工作人员由于缺乏训练，面对切尔诺贝利反应堆——其拙劣设计出自苏联之手——发生爆炸而束手无策。<sup>46</sup>但是，还有一些事故会对他人及自己产生严重影响。比如，专家普遍认为

为，导致1977年全球H1N1流感大流行的病原体，应是源自中国的露天疫苗试验或俄罗斯的实验室事故。<sup>47</sup>



专家认为，1977年的H1N1流感大流行，可能是由露天疫苗试验或实验室事故释放的病原体所引起，如图所示。（Kateryna Kon/Science Photo Library/ Getty Images）

## 涌现效应

涌现效应是无法从系统的任何单个组件识别出来，但可在整个系统观察到的一种属性。意识是人类突发性行为的一个例子——我们无法从身体的任何一部分将它识别出来，但它却能从这些组合中“现身”。作为群体行为的一个突发性例子，早期的一个著名分析表明，如果一个社区所有家庭都希望60%的近邻具有种族一致性，那么这个社区将会形成完全隔离，即使他们不希望出现这种结果。<sup>48</sup>科幻作家亚瑟·C·克拉克的一篇短篇小说生动描绘了一种虚构军事技术（“The Field（战场）”）如何因突发性后果而变成了灾难：

第一次试验演习非常令人满意，设备似乎非常可靠。船员进行了许多次的模拟攻击，逐渐习惯了新技术。...[然而，战斗中]存在一种滞后效应，而且初始条件从来就无法完全复制，这是因为当“战场”开启时，船上已发生了数千次的电气变化和质量运动。这些不对称和失真会不断累积，虽然它们很少超过百分之一，但已足够。这意味着通信设备中的精密测距设备和调谐电路已完全失调。若有时间，我们有可能已克服了这些困难，但敌舰已开始攻击。...我们的舰队吃尽了自己科学的苦头，虽使出了浑身解数，但最终无力还击，被迫投降。<sup>49</sup>

本报告中，“涌现效应”这个术语不仅是指单个系统中的突发性影响，而且还指两个或多个系统交互时出现的影响。军事中的涌现效应例子并没有编辑在网站或列表中，因为它们都是事故的涌现效应。<sup>50</sup>但是，第一次世界大战可被视为技术适应的紧急后果。<sup>51</sup>在20世纪的第一个十年，欧洲国家安全规划者认识到必须对动员体系进行现代化改造，才能利用好铁路和电报。每个国家由此进行的“改进”本身都具有意义。但是，欧洲所有国家争相改进的结果是出现了这样一种体系，使本可得到控制的破坏性事件（暗杀大公）演变成了多个欧洲国家的先发制人。技术与战争决策出现了本末倒置。<sup>52</sup>这场浩劫最终导致2000万人丧生，2100万人受伤。<sup>53</sup>

虽然他们2017年的分析并未提及这一先例，但这位前五角大楼三级文职官员和一位同事指出，美国和俄罗斯最近和未来的技术发展也存在着相同趋势，即都增加了对脆弱的空间和网络资产的依赖。<sup>54</sup>这种情况催化了“非用即失”的第一次打击机会。

非核及核战略打击能力的进步及其相互作用的方式，将会对危机迅速升级为冲突灾难的“滑坡”论产生重大影响。这些因素各自都会出现这种现象，但尤其值得关注的是它们在危机和冲突早期的情形下的互动。这种可能性非常大，因为人们可能无法明确预期或了解使用这种新技术的本质、范围和后果，从而加剧本已严重的“战争迷雾”。<sup>55</sup>

## 破坏和间谍活动

与以前的模拟技术相比，网络化的现代技术使得现有武器和支持系统比过去更容易进入。这些技术更为复杂，遮掩破坏活动由此变得更容易，即使发现也更容易将其归咎为系统缺陷或事故。商业和军事系统之间的关系密不可分，再加上科学和商业技术的全球传播，进一步促进了逆向工程的发展。

由此产生的诸多后果中，有三个对于我们的目的至关重要。第一，负责开发新技术的军事官员定会认为这些技术也会落入他人口袋，<sup>56</sup>没有哪种战争技术能够永久单独持有。可以说，先发优势能够铸就持续的优势地位，或至少会在其他各方尚未获得这个技术的时期产生临时利益。

## 与以前的模拟技术相比，网络化的现代技术使得武器和支持系统更容易进入。

但是，先行者需要评估的不仅仅是竞争优势，他们除了计算这些利益，还需权衡相关举措渗入全球体系时将会产生的风险。<sup>57</sup>

第二，军事创新者需要认识到，任何重大的国家冲突中都会出现破坏活动，因此高科技武器的使用会产生很大的不确定性。毒杀、<sup>58</sup>误导<sup>59</sup>和破坏对手的系统是情报和军事工作的惯用手段。正如信息系统脆弱性方面的经验表明，与传统资产相比阻遏这些行动更加困难。<sup>60</sup>

在某些情况下，某些对手之间提前认识到这种不确定性，可能有助于稳定。比如，它可能会阻止中国和美国之间爆发全面冲突——它们都不能确定将会发生什么，以及谁会占上风。然而，那些对现状影响较小或脆弱感更强的国家，可能会发现不确定性可以接受，甚至乐于接受。<sup>61</sup>第三，正如破坏活动可被误解为事故一样，事故在紧张局势下也许会被误解为破坏活动，由此引发报复。<sup>62</sup>



密歇根大学博士学位候选人Kevin Eyholt和一组研究人员演示了“停止”标志上的贴纸位置如何歪曲机器感知。这种策略被称为对手物理扰动，能够导致对象分类不准确或不完整，由此产生现实伤害（比如，没有识别“停止”标志的自动驾驶车辆算法）。

（Kevin Eyholt等人/密歇根大学）

## 负责官员的误解

所知有限，会阻碍政策制定者讨论和决定是否开发和如何使用复杂能力，以及相关的依赖程度。越来越多的高级官员都需要做出技术决策，而这些技术在他们接受教育、积累经验时并不存在。入职流程可以弥补这些缺陷。大多数人都没有时间、能力和意愿来更新自己的知识。美国国家安全局和中央情报局前局长描述了他和白宫战情室讨论使用网络武器的经历：

那些参加会议的白宫官员并不了解这些武器。..我记得自己在任时，有一次网络行动出现了问题。..能从事后调查明显看出，在最后的审批阶段，两位高官离开战情室时并不清楚自己已批准了相关行动。<sup>63</sup>

我们对此不应感到惊讶。六十年前：

当杜鲁门询问奥本海默，俄罗斯何时会开发自己的原子弹时，奥本海默回答说...他不知道。杜鲁门随后说他知道。他自信地说，答案是“永远不会”。显然，杜鲁门并不明白奥本海默所说的话...以及洛斯阿拉莫斯的科学家试图表达的意思。<sup>64</sup>

一位同事回忆道：“杜鲁门所说的话以及表现出的不了解只会让[奥本海默]心碎。”<sup>65</sup>

## 无法可靠和令人信服地衡量这些风险

若能衡量或至少合理估计上述问题的风险，那就再好不过。不幸的是，塔雷伯的评论一针见血：“没有人能够看清现实的枪管。”虽然极具争议的前提和推论可以提供孤立的观点，但对这些技术风险的概率和后果做出合理而可靠的评估，目前还不具备这种能力。<sup>66</sup>

我们在数字信息系统方面的经验应能阐明这一点。美国对数字系统的依赖逐渐增强，而国家安全机构对数字漏洞的了解却非常有限。专家对一些基本问题的看法存在分歧，比如，数字系统的漏洞是多还是少。<sup>67</sup>他们经常会对没有预料到的漏洞目瞪口呆，有些甚至属于根本性漏洞。<sup>68</sup>独立行动者开发出了大量可以忍受的漏洞，维护者显然仍未发现它们，否则早就打了补丁。<sup>69</sup> 21世纪积累的数字信息技术经验只是一种极其有限的了解以及反复出现的惊愕，这种说法实不为过。<sup>70</sup>

建立危险病原体研究实验室之前，生物学家在评估事故发生可能性方面的努力，可从数学层面广泛印证我们预测能力的极限。在美国，生物安全防护三级实验室可以研究“可能通过呼吸传播，导致严重或致命疾病”的病原体。生物安全防护四级实验室能够提供：

最高级别的生物安全。...此类实验室中的微生物非常危险，且是外来之物，具有极高的气溶胶传播感染风险。这些微生物引起的感染通常都很致命，并且没有治疗途径或疫苗。埃博拉病毒和马尔堡病毒便是生物安全防护四级实验室研究的两种微生物。”<sup>71</sup>

2001年发生炭疽攻击事件之后，美国的此类实验室已大幅增多。现在，美国约有1,500个生物安全防护三级实验室和15个生物安全防护四级实验室。<sup>72</sup>

对于美国拟建的一处农业病原体生物安全防护四级实验室，相关风险评估明确指出：

根据国土安全部第一次风险评估....按50年寿命计算，实验室发生泄露的可能性超过70%，每年的基本概率....=2.4%。监督了这次风险评估的国家研究委员会评论说“...据估计，若按50年计算，[该实验室]释放[口蹄疫]而导致的感染概率接近70%...经济损失为90亿至500亿美元。该委员会发现，风险和成本可能远高于此...”国土安全部随后将50年间的逃逸风险降低至0.11%，但国家研

究委员会对新的评估提出了尖锐批评：“委员会认为极低的释放概率是基于对人为错误率的过度乐观和无据估算，低估了可能发生泄露的传染性物质，以及计算释放概率时没能正确判定依赖性、不确定性和敏感性。”<sup>73</sup>

做出该总结的生物学家也添加了自己的判断：“我们更相信国家研究委员会的结论，因为其中没有涉及他们的自身利益。”根据他们的计算，十年间，因实验室员工受感染却未被发现而造成至少一种病毒逃逸，十个运行良好的实验室发生概率为18%。如果发生这种情况，他们指出其他人随后感染能否引起大流行存在极大变数——这取决于病原体的再感染率，但他们认为这种情况下发生大流行的概率在1%至30%之间。<sup>74</sup>

### 对这些技术风险的概率和后果做出合理而可靠的评估，目前还不具备这种能力。

鉴于最近的一项调查指出国际生物安全控制状况不容乐观，其中包括“严重后果研究方面的生物安全规范存在缺陷，由此引发的事故可能导致病菌大流行，要提高全球实验室的安全水平必须解决这些问题。”那么究竟该如何根据这些评估来推断全球风险？<sup>75</sup>即使拥有可观和充分证据的生物技术评估仍存在这么大的不确定性，那么国家安全机构对开发、测试或部署新系统（比如人工智能、自主系统和太空武器）及其他技术（其中许多尚待开发）的风险能有多大信心？如果能做到这一切，他们又将如何评估技术盗窃和扩散、潜在破坏活动、平时和战时交互作用所带来的风险？

简而言之，美国国家安全领导人不应自欺欺人，称一切尽在掌控之中。<sup>76</sup>这会阻碍我们在降低新技术风险方面取得进展，以及对重要进展达成共识的步伐。为控制气候所作

的努力已激起大量的争论和抵制，但它们基于的是科学理论和证据，有助于区分无端恐惧和充分了解的事实。它们具有这种优势，因为其中涉及的（燃烧化石燃料）的技术已存在数十年，甚至数百年，而且所产生的后果可在某种程度上测量。即使如此，它们也受到了激烈争论。由于新技术的性质和后果更难以捉摸，解决这个问题变得更加困难。

## 人为介入

对重要的机器决策进行人工审查，通常被认为是一种可取之举。如果机器采取的行动造成了生命损失，通常会认为人类在其中扮演了必要角色。这已被我们的军事机构视为一种公理。“国防部观察家总是热衷于提醒人们，称官方政策是将人类置于指挥与控制的最高层面，监督致命性决策或至少保留否决权。”<sup>77</sup>

这种立场具有心理赋权性：人类能从人为控制的概念中获得安慰。将我们的分析与感受分开，我们能发现三个因素为人类参与机器决策系统提供了支撑。

第一个是人类能够为依赖机器的系统添加一种不同的推理。如果我们想让系统消除易用性导致的错误，就需要增加一个障碍。<sup>78</sup>比如，我们需要两位操作人员（而不是一位）才能发射带有核弹头的导弹。如果第二个障碍能与第一个很好地区分开来，安全水平则会进一步提高。机器系统可能遭到黑客入侵和受到误导，而破坏人类则需要完全不同的策略。机器与人类相结合，会减少对单独一方造成误导的脆弱性。<sup>79</sup>

应该注意的是，这种观点重视人类的角色是因为它具有辅助性和差异性，而不是它能提高正确决策的可能性。



2016年，谷歌DeepMind的人工智能程序AlphaGo击败了韩国专业棋手李世石。AlphaGo最初接受了人类比赛数据的训练，然后通过自我改进实现了超人性能。更新版本AlphaGo Zero在没有任何初始人类训练数据的情况下实现了超人性能。（Kim Min-Hee-Pool/Getty Images）

第二个论点认为，人类能为机器决策添加一种出色的定性因素：他们能考虑到程序员无法自信和全面地通过编程输入机器的东西，比如背景或道德。<sup>80</sup>因此，与单独的机器系统相比，引入人为监督可更好地防止不良行为的发生。

第三个论点源于我们对系统的不安，担心它会遵循自己设定的优先事项，而我们没有能力重新定向或将其关闭。这种担心与其说是源自机器的内在本质，不如说是来自我们完美编程的能力有限。如上所述，复杂系统的设计和操作会出现错误和意外影响。我们希望从后端进行审查，以降低这些系统失控的可能性。

所有这三个论点都有道理。然而，将人类决策作为一种控制机制的价值已被高估。其中的部分原因是人类自己也会犯错，很多情况下的行动偏见并不可取，<sup>81</sup>但主要是因为人类决策依靠的并不是独立或占主导地位的第二判断来源，其对机器的依赖程度超过了估计。因此，虽然人为干预能产生一些好处，但力度太过脆弱，且常会对速度非常珍贵的自动化系统持续控制产生反作用力。此外，随着时间的推移，这种安全保障效果将会逐渐下降。可从本报告得出的重要收获是对于这些系统，设计期间的控制要比运行期间的决策更有效。安全行动尽早展开，等到后期必然太晚。

本节将会从人工智能的角度阐明这些观点。但在此之前，值得注意的是人工智能方面的一些问题同样会波及其他技术。比如，生物系统会作为传感器、产品制造方式和武器，更广泛地渗入军事行动。<sup>82</sup>与人工智能一样，这些系统中的控制机会虽不完美，却也在前端占据着主导地位。一旦释放出来，生命系统便会对自身需求做出反应——它们会根据环境刺激自主行动。无论人类采取何种控制措施，它们主要来自我们在开始设计生物实体时的初始想法。

30年前一起备受争议的事故，为战时的人类/机器决策提

供了有用案例。1988年，文森斯号巡洋舰上的军官断定自己受到攻击，随即向一架伊朗民用客机发射两枚导弹，造成290名乘客和机组人员死亡。经过广泛调查，美国国防部对船员所犯的错误表示同情，甚至美化了船长和其他人的行为。尽管如此，这里存在的一个核心事实是目标飞机并没有像攻击机那样向下俯冲，而是在向上爬升（商用飞机都是如此）。对于这个问题，官方报告指出：

攻击之后...从文森斯号巡洋舰获得的数据磁带与它提交的情况报告之间存在直接矛盾...很显然，[相关官员]无法根据CRO [读出字符]显示的数据进行报告。[其他官员]的报告给出的解释最合理，即他的行为是因生理疲劳、战斗、压力和紧张所致，这些因素会对工作状态和任务执行能力产生不利影响。正如[先前报告]所述：“‘场景实现’的概念似乎适用于这种情况。”由于战术信息协调员毫不怀疑朝着舰艇飞来的是伊朗F-14战斗机，况且它对反复警告置之不理，这时“大脑会排斥数据的不一致性，并偏袒可促进内部一致性的错觉。”他发表的证词也反映出了这种精神焦虑——他强迫自己抓住第15巡回法庭的“每一次机会”，说明“每一位指挥决策者都能了解情况，明白正在发生的事情，以免他们会因其他事情而走神。”据报道，当发言接近尾声时，他已变成了竭力喊叫。<sup>83</sup>

文森斯号事故“仍是一个存在争议的问题”，<sup>84</sup>自动化系统无法明确提供数据、船舶积累的运行经验以及参与者的个性等另外一些因素也值得考虑。<sup>85</sup>我们的目的不是划分责任，而是希望评估人类面对压力时，在控制危险技术操作方面的作用。

要在五分钟之内<sup>86</sup>作出生死抉择，文森斯号上的决策者（1）受到了情绪和干扰的影响；（2）太过依赖于根据机器传感器推断威胁状态的程序；（3）报告关键的机器输出时发生了根本性错误。



1988年7月，美国海军文森斯号巡洋舰（CG-49）将一架伊朗民用飞机误认为攻击型战斗机，向其发射两枚导弹，造成机上290人全部死亡。本图拍摄于2003年。鉴于文森斯号巡洋舰的数据提供系统以及人类在导弹发射决策方面的作用，这一事故仍是人机整合与决策的重要研究对象。  
(Getty Images)

商业航空领域已对这种人机关系展开了深入探讨。这个行业正在逐步提高机器控制能力，究竟还有多少空间留给了人类控制——即使是训练有素且经验丰富的飞行员——引发了激烈争论。“空中客车的哲学似乎认为自动化通常比飞行员了解得更多，大多数情况下自动化都应拥有最终决策权。”<sup>87</sup>比如，“空中客车公司制定了严格限制，无论何种情况飞行员都不得超出此限制；

**目前，人类处理速度大约是机器的百万分之一。机器的处理速度变得越来越快，人类却不是这样。**

而相比之下波音公司设定的限制比较灵活，飞行员如果认为必要便可无视这些限制。”<sup>88</sup>尽管目前这两种系统哪个更可取属于一种封闭式问题，<sup>89</sup>但长期的变化趋势已很明显。如果机器决策的改进比人类决策更显著（这种可能性极大），那么顺从人类的空间就会进一步收缩。

无论是一般情境还是军事情境，处理速度和通信能力也越来越对人类决策者不利。目前，人类处理速度大约是机器的百万分之一。机器的处理速度变得越来越快，人类却不是这样。当即时响应迫在眉睫，即使国防部持人类应该介入观点的人也承认无法实现想要的人类控制效果。可以预见，随着机器能够完成的任务不断增多，处理速度不断加快（计算速度和行动速度）<sup>90</sup>以及自主操作能力不断增强，这个反例将会吞噬掉原有规则。正如两位观察家所断定：“下个世纪的军事超级大国必将拥有卓越的自主能力，否则便不会成为超级大国。”<sup>91</sup>

一位软件专家坦诚接受了这一现象，并强调称人类对远程系统的控制会对通信能力产生不可抗拒的依赖：

这些都是令人欣慰的术语，比如“半自主”和“人为介入”。但是....机器在各种情况下有效发挥功能需要完全的自主权，若需征求人类意见其实有害无益。..比如，当自主水下航行器发现关键目标...但意识到通信不畅时，我们如何期待它们采取行动。...当导弹进入舰艇2.5英里半径时，人类没有足够时间来做出反应；Phalanx系统必须以完全自主的方式运行，完全依靠自己来跟踪导弹、瞄准和射击。<sup>92</sup>

考虑到军事和平民的经验，我们应将最终决策权留给人类：我们总统拥有下令展开报复的专属权力。前国防部长阿什·卡特强调：“我们不要忘记，当需要使用武力来保护文明时，我们的一个原则是应让人类参与关键决策。我认为这个原则极其重要。”<sup>93</sup>但是，当卡特部长响当地作出承诺的时候，人们对自己的资格却含糊了。卡特部长

没有说，“为了保护文明”，应由总统（或者是总统周围的一群谋士）做出关键决策；而是说，应有“人类参与”。事实上，要求总统在收到警告后决定是否反击，他所处的状态与文森斯号舰长一样，决策时间大致相同。<sup>94</sup>只不过与海军上校相比，在决定是否反击时总统受到的训练和拥有的经验都更少。若被告知一枚导弹正向美国飞来，总统应基于什么来决定是否相信战局评估，并做出回应呢？

我们的总统并不是透过椭圆形办公室的窗户，看到了导弹发射。他依赖的只是几位顾问，而这些顾问又依赖于一系列的传感器、算法和模型，推断出导弹已经发射并在朝着我们飞来（更确切地说是为这种推断提供了支持）。在这几分钟内，我们的总统就像文森斯号舰长一样，依赖的是传感器、算法、数据输入和模型所提供的信息质量。对于这些，我们的总统及其周围的参谋可能只有一点点了解。<sup>95</sup>最多可以说是“人类参与其中”。机器和人这两个看似独立的系统其实是相互依赖，而机器对决策的影响更大。这种情况在当代非常普遍。人类决策者是穿越荒野的骑手，几乎或根本没有能力评估胯下的野兽将会奔向何方。<sup>96</sup>

人类在机器系统中的角色问题值得讨论。对于我们的目的而言，六个结论似乎非常合理：人类决策者是有价值的，因为他们能使攻击者破坏系统的能力变得更复杂；目前，人类可以添加使机器难以捉摸的环境；随着机器能力的提高，人类的角色空间将会收缩——特别是在评估环境时<sup>97</sup>或需要以机器的速度做出更多决策；即使在目前，决策者对机器输入的依赖非常大，作为一种控制机制机器通常不会增加显著的独立权力；在机器得到赋权范围内，人类决策者既可能导致意想不到的后果，也会控制影响；紧迫性会给人类决策带来压力和曲解，但即使拥有足够时间进行商议，机器的复杂性也会让人类决策者感到困惑。

## 应该怎么办？

分析和预测的不确定性以及人类参与决策的局限性，不能成为忽略问题的借口。认识到局限性后，更加迫切的是更好地了解（虽然不能完全了解）我们的军事技术风险，并制定预防措施。引入军事系统的每项新技术都具有独特的属性和风险。然而，这里所谈的技术拥有两个令人不安的特征：它们的发展速度比我们的控制机制更快，并有可能导致特别严重甚至可以说是灾难性的后果。

开发控制理论是为了保护我们，以免系统超出可接受的参数而对我们产生影响。<sup>98</sup>比如，恒温器会对许多熟悉的系统进行测量，一旦发现过冷、过热等异常状态便会启动恢

复平衡的步骤或将系统关闭。熟悉的闭合回路系统都是这种控制。我们关注的技术拥有陌生的属性，并有可能突破闭合回路的界限。<sup>99</sup>

它们也属于纳西姆·塔勒布（Nassim Taleb）所称的第四象限。

某些时候，你即使犯下大错仍会安然无恙。但在其他时候，错误即使微不足道也可能致命。如果你被利用，错误便会将你击倒；如果没有，你便可享受生活。...有些决定需要更加谨慎。...比如，你并不“需要证据”...来证明将枪上膛是为了避免玩俄式轮盘，或证明小偷正在寻机撬门。您需要的是安全证据，而不是缺乏安全的证据。<sup>100</sup>

在理解和记录新技术在军事环境中可能出现的问题方面，我们的国家安全机构显然投入不足。让人为介入的主张所提供的舒适度非常有限。人类能够提高机器的决策能力，但军官的判断现在通常会受到机器影响，而机器影响程度只会随着机器能力与速度要求的提高而增加。所需的解决办法不是让人类“介入”，而是在设计和部署新技术时要深思熟虑，最大限度地发挥创造力。

就此而论，下面将推荐五个举措（程度由轻到重）。

**1.增加国防和情报机构对事故和涌现效应风险的关注。比如，将这些关注作为四年审查、国家情报定期估计、净评估、作战演习和红军演习的重要部分。**任何事情都需付出代价。实施这项建议需要花费一些本会用于冲突评估和计划的边际投资。然而，本报告认为这种资源分配很有道理：虽然无法精确计算疏忽大意行动的风险，但一个更平衡的系统至少会让高级决策者了解这些问题、更多地考虑这些因素——这也是后续建议的思路。为了揭示那些通常会被计划支持者和日常操作者所忽视的脆弱性，红军攻击能够发挥极大作用。

## 我们关注的技术拥有陌生的属性，并有可能突破闭合回路的界限。

与此相反，有些人可能会回应说，故意发起的全球战争的后果远大于事故带来的后果，无论事故有多么严重。因此，根据这种观点，将任何预防性努力从前者转向后者的建议都不妥当。然而，这个论点设想了一种二元状态：一方面是造成可容忍创伤的事故，另一方面是可能导致不可容忍灾难的恶意行为。这种区别值得怀疑。历史和分析说明事故和涌现效应是灾难性恶意行为的重要原因。<sup>101</sup>举行演习和评估的一个价值是澄清这一点，并提出防火措施和

其他减轻这种风险的方法。

还有第二个考虑因素。如果发生灾难性事故，机构的新技术的开发和部署能力将会受损，尤其是当这些机构准备不足时。鉴于此，提出的举措远不是破坏用于战争和威慑的技术投资，而是加强对这种投资的保护。

除此之外，国家安全机构还需增加对这些技术投资所致风险的了解，能够公布这些风险和引起人们注意，以及事故发生时的响应流程。<sup>102</sup>

一些高层政策指导是平衡我们利益的先决条件，而这最有可能来自定期政策审查中的高度关注。此外，必须至少公开一些可反映这种风险认识的文件，并促使国会、白宫和其他相关的行政部门机构和办事处参与其中。国家安全机构内部的自然趋势是尽量减少这些问题的可见性，并避开具有潜在破坏性的外部行动者。但这样一来，技术举措所能得到的支持就会有限，当事故或暴露发生时它们会很容易受到反应过激的影响。情报机构应已吸取了这一教训——当网络文件和工具遭到黑客攻击时，他们面对强烈抵制所能获得的公众支持就非常微弱。

## 国家安全机构内部的自然趋势是尽量减少这些问题的可见性，并避开具有潜在破坏性的外部行为者。

仅靠领导关注和公众支持并不足够，领导者还必须加深对这些风险的了解。2008年发生金融危机，促使我们认识到一家银行的行为将会引起整个银行体系动荡。<sup>103</sup> 2013年发生埃博拉危机，促使我们加强了应对类似流行病风险的工作。<sup>104</sup> 上一节描述了我们对于新技术风险的认识不足，并指出这些风险很难准确评估，因为这些技术现在和将来都是新的。作战演习和模拟有助于提高我们的认识。

**2.开发、部署和使用战争技术时，经常考虑并努力降低扩散、敌对行为、事故和突发性行为所带来的风险。要将降低这些风险作为优先事项。**军事行动期间，人类决策者的作用将会非常有限。随着技术的进步，这个角色空间将会进一步收缩。最重要的控制机会应是在将其用于冲突之前，即在设计和实施计划与安全保障时。正如一位著名的历史学家所言：“决定性选择很少出现在最后阶段，通常情况下它们都是过去做出的选择。”<sup>105</sup>

对于这个观点人们非常认可。在新的武器系统部署之前，国防部门完善的测试和评估系统需要有一个“指定的审批机构”，对这些武器出现事故的风险进行评估。比如，在引

入一个新的系统之前，网络安全创新的支持者必须识别敌对攻击的机会，并同时考虑敌对攻击的风险和事故的风险。<sup>106</sup>但这种方法只能用于那些服从可信的概率计算的突发风险。它应能适用于整个新技术（生物学、机器人学等，而不只是信息技术），正在开发和将要部署的计划，以及评估对手采用我们开发的技术、利用我们的漏洞（因为我们增加了系统应用范围及其复杂性）时可能产生的成本。

在一定程度上受到本报告的启发，美国国防高级研究计划局和情报高级研究计划署已开始探索如何系统地将这些考虑因素纳入自己的计划。本报告还在附录提供了情报高级研究计划署署长Jason Matheny提出的问题，它们都是引入新计划之前必须回答的先决性问题。

虽然美国军事和情报机构正在努力推进我们国家安全所依赖的各种技术（尤其是破坏性技术），但重要的是领导人必须了解我们理解相关内容的局限性，机构应设计好应对失败的恢复力。<sup>107</sup>这需要了解各种条件下系统性能的模拟，这些条件包括模拟的攻击；从类似系统进行借鉴；<sup>108</sup>通过新系统的受控经验开发数据；<sup>109</sup>审查<sup>110</sup>这些系统在实践中维护、<sup>111</sup>更新和运行；并协助和培训操作人员 and 主管，以便他们进行干预并在必要时终止机器运行。

美国国家安全机构从我们对核技术的控制中获得了这种方法模型，我们将核技术当作了潜艇部队的推进手段、我们战略核力量的武器以及我们核电站的核心能源。就训练方案、程序和安全保障而言，我们未来对新技术的使用至少要像过去这些计划一样强大。但与过去相比，许多新技术已变得更不透明、发展更快、受到的人类控制更少。<sup>112</sup>此外，过去的核潜艇和核武器等计划相对孤立，因此更容易保护。对于拥有更多自主权、应用于人口更密集、交互更强的技术和物理环境的系统而言，互动造成的破坏和涌现效应风险将更大。因此，与以前的系统相比，我们需要更好地了解和控制这些系统。联合国一个关注致命自主武器的小组提出了四个相关问题：

- 现有系统应该如何核查（构建过程是否正确）和验证（构建的系统是否正确）？现有和计划中的自主系统是否可以了解（您了解什么，是如何了解的）？机器能描述它们的学习情况吗？
- 如何解决黑客和隐私等人类和社会安全问题？自主机器能否预防黑客攻击？
- 机器行为能否通过软件/硬件锁定，能否防止学习型机器绕过/更改它们？
- 人工智能的可变化的特征及其可能的普遍存在能否以某种方式限制有关致命自主武器系统的讨论，人工智能是否会像其他军民两用技术一样？<sup>113</sup>

美国国防高级研究计划局的两个计划含蓄指出了技术上的可能性。第一个计划指出“如果未来的作战人员要...适当地信任、有效地管理新一代人工智能机器合作伙伴，操作人员则必须了解不透明算法。”<sup>114</sup>“可解释人工智能计划（XAI）旨在创造...一种新的机器学习系统，该系统能够解释它们的基本原理、描述自己的优缺点并清楚它们未来将会如何表现。”<sup>115</sup>第二个计划与生物控制有关。“安全基因项目”旨在开发基因断开开关、<sup>116</sup>应对措施以及其他“可在新生物技术诞生之初打造生物安全性”的工具。<sup>117</sup>

**3.独立和定期评估我们的军事技术在事故和涌现效应方面的风险，并鼓励对手以及可靠的第三方做出同样努力。**上一节指出了无论一个人多么了解技术，识别和评估新技术风险必是十分困难。对组织内部此类技术的开发人员来说，持续展开潜在破坏性分析，并强制采取可最大程度地降低风险但却又会降低效力或提高成本的应对措施，这种困难程度将会以指数级增加。<sup>118</sup>两位最杰出的“常态性事故”<sup>119</sup>思想家特别说明了这一变量。斯科特·萨根（Scott Sagan）将他所调查事故中的军事倾向戏称为“抱团取暖”。<sup>120</sup>相比之下，查尔斯·佩罗（Charles Perrow）将商业航空安全水平的大幅提高归功于“感兴趣的正式组织相互监督”。<sup>121</sup>

就此而言，由支持者的组织之外的专家提供补充性的组织评估具有一定价值——也许是极高的价值。<sup>122</sup>为了检验这个命题，应该对比如五、六个新兴技术开发计划的样本进行外部评审。如果这种评审证明有用，那么它可能会在现有外部机构的支持下变得正规，比如白宫科技政策办公室、国家标准与技术研究院、总统情报咨询委员会、国防科学委员会或者监察长。<sup>123</sup>

判断外审的使用频率和力度之前，应充分吸收这些评审经验。如果外审的价值有限，应保持相当大的采用间隔或根本不予采用。如果极具价值，便可定期或以一定的连续性展开外部评审。

如果美国机构认为这种方法可取，就应鼓励其他国家建立类似的系统。也可鼓励第三方（比如联合国）<sup>124</sup>进行此类评估，最好是与国家展开合作，以便所有国家都能意识到至少在非机密背景下可以理解的风险。应与合规者分享一些民用技术应用，促进国家风险分析和降低风险方面的工作。<sup>125</sup>

**4.与我们的盟友和对手加强多边规划，以便我们更好地识别和应对事故、灾难性恐怖事件和意外的国家冲突。**美国政策制定者早就明白，如果只有我们保护自己的核武器，而

对手在这方面毫不在乎，那么我们的保护也毫无作用。他们需要从中吸取教训。比如，如果其他国家的基因改造泄漏到了全球环境，那么美国的转基因实验安全保障就毫无意义。这一点对于所有关键技术来说都是真理。如果天真地认为外国也会像我们一样，按照我们的参照标准仔细评估风险，这种想法未免太过危险。

所有采用先进破坏性技术的国家都需分享自己的风险和安全保障认知，<sup>126</sup>之前提供的建议会对此有所帮助。在此基础上，我们需要共同商议和应对意外事故。<sup>127</sup>莫斯科发生传染性病原体恐怖袭击或中国实验室发生事故，都将涉及到我们的安全利益。它们与我们的活动存在利害关系。我们需要制定计划，加强联盟内、外的合作。<sup>128</sup>

为实现这一目标，众多国家的安全机构应合作使用上文建议的分析、模拟和演习工具，这些工具原本用来加深美国的了解。这些努力将为合作奠定基础。风险不能只从我们的角度来理解，也不能仅仅依靠对手从他们的角度加以理解。<sup>129</sup>

**5.将新技术用作推进和验证条约与规范的手段。威慑是防范各国突破某些关键界限的大棒政策。**不过，美国国家安全官员还使用了一系列条约和规范来作为威慑手段的补充，这些条约和规范刻意某些技术的开发、部署和使用设置了障碍。<sup>130</sup>最明显的是，我们使用这些工具的成功和失败都体现在了这样一个事实：九个国家的核武库共有大约15,000件核武器。<sup>131</sup>但是，如果没有防扩散条约，这一数字将会变得更高。<sup>132</sup>虽然战略家将威慑当作防止核国家发生冲突的主要手段，但不首先使用核武器的规范<sup>133</sup>无疑有助于阻止核武国家使用这些武器打击非核武国家。<sup>134</sup>

这种依靠条约和规范加以约束的模式虽不完美，但可在控制其他新的军事技术方面发挥极大的作用。在减缓导弹<sup>135</sup>、激光致盲武器<sup>136</sup>、地雷<sup>137</sup>、太空武器<sup>138</sup>、化学武器<sup>139</sup>等技术的开发、扩散和使用方面<sup>140</sup>，各国取得了不同程度的成功。认识到这一点之后，生物学家<sup>141</sup>、人工智能领导者<sup>142</sup>、网络专家<sup>143</sup>以及其他与新技术相关的人员<sup>144</sup>正在鼓励制定新兴的规范<sup>145</sup>和条约，限制这些技术的开发、扩散和破坏性使用。

不幸的是，这些工具的能力和吸引力在很大程度上取决于参与国对我们观察违规行为能力的信心。<sup>146</sup>只有当各国能够找到监控试验、材料、设备、武器和使用的方法时，才会在规范和条约方面取得实质性进展。<sup>147</sup>负责的领导者不能盲目乐观，因为公开宣称已停止技术开发的对手可能仍会暗度陈仓。<sup>148</sup>里根总统主张的“信任，但应予以核查”似

乎是在极力暗示：对于无法核查的东西，我们不能信任。

我们的核心问题是无论按照目前哪种标准，本报告所描述技术的可见性都很低。以前需要工业设施和控制性材料的生物实验，现在在普通建筑物内就能展开，而且所需的微型设备和材料都能购买得到。<sup>149</sup>随着计算硬件的不断微型化和持续扩散，以及最基本的发展媒介——软件——极难接入和审查，人工智能方面的研究已变得更难追踪。

然而，在所关注的技术中可能存在固有机会。人工智能系统可在检测日常正常活动中的异常方面发挥极大价值。精心设计的生物系统可以充当环境扰动中的精密传感器。信息技术为探测对手的系统创造了机会。这些机制虽不完美，但都非常强大。随着这些技术的发展，美国国家安全官员应将扩大和加强对军备控制的作为优先事项，而不是只顾发展武器装备。

## 总结

在未来，美国有很好的理由仍像过去一样追求技术优势。但是，我们的国家安全领导人在朝这个方向尽力时，还必须认识到即使在这种竞赛中永葆第一，同样也会滋生风险。本文的鲜明观点是优势并不等同于安全：这种竞赛存在重大风险。

如果这种竞赛的可预见性结局良好，或者风险容易识别、可控而且轻微，那么还比较容易接受。但是，除非人类不再多疑，这种竞赛似乎没有止境。这个问题的根源不在于技术本身，而在于我们自己。正如莫里森（E. E. Morison）在半个世纪前所言，“你可从历史中明白的一点是，人类在与人相处的同时，也学会了如何与——也可能是更好地与——机器相处。”<sup>150</sup>

但技术让我们的情况变得更糟。许多人都已看到技术之力能够放大恶意行为之果，从而增加国家、团体和个人有意识地开启彼此毁灭进程的风险。本报告强调指出，我们避免这种结果是可能的，它之所以可能会出现，是因为国家、团体和个人无法很好控制自己创造的东西。再次回到本报告开始时所使用的比喻：我们正在将更多的子弹塞入枪支——这些枪支不仅广泛分布在我们系统之内以及盟友之间，而且而且分布在那些追随我们的领先技术的对手手里。由此产生的技术-人类系统非常危险，而且此类风险仍在日积月累，并且未来仍可能变得更大。美国并不是这种技术轮盘的唯一负责者，但鉴于它所具有的技术优势和主要机会，降低这种风险也就成了它的主要责任。这么做对于保护我们的安全至关重要。

本报告从一开始便指出，人类对新技术的很多担忧都是源于情感，而非理性。这样，人们甚至会轻易抛弃存在充分理由的担忧，认为它们缺少理性或充满了勒德分子的味道。希望这份报告能够有助于驳斥这种轻率的拒斥。在加强关注的同时，我们不应忽视情感化反应的相关性。我们至少应该认识到灾难性事故可能非常严重，因为它们可引起的强烈心理和政治反应，导致更直接的物质后果。（比如，海啸导致日本福岛反应堆发生泄漏之后，一些国家终止了所有的核项目。）

更乐观的一种态度是，恐惧可以成为我们的朋友。诚然，恐惧可以麻痹和误导人们，助长人们想要逆转时间之剑的非理性愿望，但如果适当考虑，它能提醒所有使用者——包括我们以及我们的对手——注意我们正在和将来会出现的错误。技术发展能让国家变得强大，使它被冲昏头脑，经常无法抗拒。但它们也会带来真正的风险，这些风险通常会被复杂性、官僚做法、误解以及个人和组织对增强能力的渴望所掩盖。正如一些人所言，这些风险并没有因“人为介入”而得到有效控制。本报告通过强调我们的控制能力是有限的，试图剥离虚假的安慰感。恐惧会让我们对通常忽视和不恰当地低估的风险能够更为敏感。

适度承认恐惧能够成为一种强力粘合剂。它可将关爱与自身利益结合起来，成为激励人们的三个重要因素之一。它在国际关系中更为重要，其原因温和地说，就是各国彼此缺少关爱。

国家安全官员经常会将盟友和对手视为不可改变的种类，它们对我们的关系产生确定的后果。但历史却给了我们一个不同的教训，美国在二战前夕的状况就能说明这一点。当希特勒在1940年和1941年发动袭击时，我们与西欧民主国家有的亲密关系，与英国更深层的特殊关系，但是这些并没有激励我们出手干预。但当美国开始为自己担心时，它甚至与苏联斯大林建立了联盟。我们现在面对的对手没有谁像斯大林那样令人讨厌和反感。我们愿与他携手，他也愿与我们合作，都是因为我们看到希特勒和纳粹国家能够带来更大风险。

科幻小说中，曾经不共戴天的国家面对外星人入侵也会化敌为友，这种想象其实并不夸张。我们不必等待外星人。如果人类能够认识到我们现在共同面临的巨大威胁正是出自我们之手，我们同样可以打开合作之门。在该领域展开合作，可成为更广泛合作的开始。换句话说，将军队明确视为疏忽大意的行动者，可能为驯服我们的毁灭能力开辟一条道路。也许这种期望有些过高，但也许不一定。

## 附录

### 可供计划管理者提出新建议的参考问题

情报高级研究计划署署长Jason Matheny博士

1. 按照您的估计，一个主要的竞争对手国家了解一项技术之后，需要多长时间便可将其转化为武器？拥有类似于本世纪初基地组织那么多资源的非国家恐怖组织需要多长时间？
2. 如果技术发生泄漏、失盗或遭复制，我们是否后悔将它开发了出来？如果竞争对手跟进之后，先发优势仍会持续下去将会如何？
3. 计划怎么能被外国情报误解？对于降低这种风险您有何建议？
4. 能否在进攻能力发展之前或发展过程中，我们就发展防御能力？
5. 能否降低技术失盗、复制和大量生产的可能性？哪些设计特征能够成为进入的壁垒？
6. 哪些红队活动有助于回答这些问题？谁的红军意见你最重视？

## 尾注

1. 军事技术优势来自于研究计划、吸收其他研究成果（包括民众和军事竞争对手开展的研究）、将发明转化为战争工具、培育可利用这些工具的战略和行动概念、为军队使用这些创新提供激励和培训。Malcolm Gladwell站在一般民众的角度，对这一过程的各个方面进行了简短描述：“Creation Myth: Xerox PARC, Apple, and the Truth About Innovation”，《纽约客》，2011年5月16日，<https://www.newyorker.com/magazine/2011/05/16/creation-myth>。本报告涉及到了所有这些方面，并未尝试分清各自的作用。
2. 中国人口为美国的三倍也是一个相关因素。这既是一种优势（比如，要求将技术共享作为进入市场的条件），也是一种劣势（比如，为人均社会支出增加而不是军事投资创造需求）。
3. 以及指挥与控制、传感器、分析算法等的支持系统。
4. 牛津大学人类未来研究所曾关注过这些问题，2008年的一次会议认为，下个世纪出现“技术引发全球灾难”的可能性为10%以上。Nick Beckstead等人的“Unprecedented Technological Risks”以几种技术为例，很好总结了人类未来研究所关注的风险（人类未来研究所，2014年），<https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Techno-logical-Risks.pdf>。
5. 两篇经典著作是Scott D. Sagan的“The Limits of Safety: Organizations, Accidents and Nuclear Weapons”（1993年），以及Charles Perrow的“Normal Accidents: Living with High-Risk Technologies”（1984年首次出版，1999年补充资料后再次出版；本报告引用的所有参考均出自后者）。Eric Schlosser, “Command and Control: Nuclear Weapons, the Damascus Accident and The Illusion of Safety”（2013年）针对特定事故提供了最新情况。高可靠性组织的分析仍层出不穷，尤其是加州大学伯克利分校Todd LaPorte领导的高可靠性组织项目。Mathilde Bourrier的“The Legacy of the Theory of High Reliability Organizations: An Ethnographic Endeavor”（Sociograph - Working Paper n° 6/2011年）记载了伯克利分校曾作的一些努力以及这些问题的反对观点，[https://www.unige.ch/sciences-societe/socio/files/4814/0533/5881/sociograph\\_working\\_paper\\_6.pdf](https://www.unige.ch/sciences-societe/socio/files/4814/0533/5881/sociograph_working_paper_6.pdf)。
6. 竞争中的破坏活动将会增加这些风险。
7. Calestous Juma, “Innovation and Its Enemies: Why People Resist New Technologies”（2016年），294页，提出了一系列有趣的案例研究。他强调说：“表面上的保守主义或对新思想的非理性拒绝可能代表了一种深层的社会稳定逻辑，这种逻辑是建立在道德价值观、合法性和经济利益来源的基础之上。”
8. 也许这些恐惧为人类所固有，我们的技术焦虑是对自然灾害和众神之怒的恐惧所致。根据这种观点，技术让人们更好地控制了自然环境，也许是它承担了以前对自然风险的恐惧，好像存在一些恐惧守恒的心理法则。这种观点认为我们应警惕情绪的作用，但并没有表明这里没有风险。
9. 正如Michael Hopmeier所指出，那些实验室便是今天的工厂。
10. “物理领域与核武器政策有着特殊的关系。物理学家发明和改进了核武器，并对限制其造成的危险作出了重大贡献。”Steve Fetter等人, “Nuclear Weapons”，《今日物理》，2018年4月，39页。
11. Stephen Hawking等人, “Stephen Hawking: ‘Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?’”《独立报》，2014年5月1日。
12. “Autonomous Weapons: An Open Letter from AI & Robotics Researchers”，2015年7月28日，<https://future-of-life.org/open-letter-autonomous-weapons/>。2017年的后续讨论和研究重点论述可参见“Asilomar AI Principles”，<https://futureoflife.org/ai-principles/>。
13. 比如参见Marc Lipsitch和Thomas V. Inglesby, “Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens”，（美国微生物学会，2014年），doi: 10.1128/mBio.02366-1412 December 2014 mBio vol. 5no.6 e02366-14: “旨在开发新的潜在大流行病原体的研究...属于传染病研究实验工作的一小部分，[却给公众带来了极大的潜在风险。”著名的遗传学家Kevin Esvelt说：“我所处的位置非常奇怪。...我可能是基因工程的首批批评者，但也行走在自己所批判对象的最前沿。”Kristen V. Brown, “This Scientist is Trying to Stop a Lab-created Global Disaster”，Splinter, 2016年6月27日，<https://splinternews.com/this-scientist-is-trying-to-stop-a-lab-created-global-d-1793857858>，引用了麻省理工学院Sculpting Evolution实验室负责人Kevin Esvelt的话。

14. James R. Clapper, "Statement for the Record Worldwide Threat Assessment of the US Intelligence Community" (参议院军事委员会, 2016年2月9日), 第9页 (着重强调), [https://www.dni.gov/files/documents/SASC\\_Unclassified\\_2016\\_ATA\\_SFR\\_FINAL.pdf](https://www.dni.gov/files/documents/SASC_Unclassified_2016_ATA_SFR_FINAL.pdf)。
15. Julie Bort引用Bruce Schneier的话说, "Now We Must 'Pledge Allegiance' To Apple Or Google To Stay Safe" 《商业内幕》, 2012年11月7日, [https://www.schneier.com/news/archives/2012/11/now\\_we\\_must\\_pledge\\_a.html](https://www.schneier.com/news/archives/2012/11/now_we_must_pledge_a.html)。
16. Daniel E. Geer Jr., "A Rubicon", (胡佛研究所, Aegis Series Paper, no. 1801, 2018年2月)。Geer接着说道: "有两类且只有两类网络风险会直接升级到国家安全水平: ...一类是根据其任务定义, 必会导致一个单点故障的关键服务; ...另一个是将会导致级联故障的关键服务或普适服务。"
17. 当然, 军方关心的是如何控制军队及其战争工具。这个等式的人类问题与技术问题同样重要, 但这不是本报告的主题。对于希望分析人类控制必要性的读者来说, 可先参阅Jim Frederick的"Black Heart: A Platoon's Descent into Madness in Iraq's Triangle of Death" (2012年)。
18. 比如参见"In the Matter of Knight Capital Americas LLC: Order Instituting Administrative and Cease-And-Desist Proceedings", 2013年10月16日, 第2-3页, 美国证券交易委员会对自动交易的描述是: "最近的事件和委员会的采取行动表明, 这项投资还必须优先考虑技术治理, 以便尽可能地防止软件故障、系统错误和故障、停机或其他突发事件, 并在此类问题出现时快速、有效地作出响应和降低风险。企业没有或不愿这样做, 可能会给公司、客户、对手、投资者和市场带来潜在的灾难性后果。委员会2010年11月通过的《交易法》第15c3-52条规定, 作为金融市场的守门员, 经纪人或交易商应"适当控制市场准入相关风险, 以免损害自身和其他市场参与者的财务状况、证券市场交易的完整性以及金融体系的稳定性。"
19. Perrow, "Normal Accidents", 第371-72页和382页, 区分错误诱导组织和错误避免组织。他将信息共享视为一个重要特征, 认为它会在确定特定系统的类型时发挥重要作用。关于这点, 他指出各种记录系统和"感兴趣的正式组织相互监查"为民用航空业提供了极大帮助。
20. 比如参见一个开发人员基础架构团队所做的努力, 该团队负责提高谷歌不同运作部门相互依赖关系的可见性。J. D. Morgenthauer等人, "Searching for Build Debt: Experiences Managing Technical Debt at Google" (第三届国际技术债务管理研讨会论文集, 2012年), <https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37755.pdf>。
21. 关于这个问题, 可参见美国科学家联盟, "What is Over-Classification?" 2013年10月21日, <https://fas.org/blogs/secrecy/2013/10/overclass/>。
22. Sagan, "The Limits of Safety", 第17页。
23. 称赞"五角大楼被众人嘲笑为创新障碍的谨慎和保守程序时, [国防部副部长强调了]运行测试和评估的重要性, 必须确保机器将能完全按照期望, 多次可靠地完成任务"。Sydney J. Freedberg Jr., "War Without Fear: DepSecDef Work On How AI Changes Conflict", "Defense One", 2017年5月31日, <https://breakingdefense.com/2017/05/killer-robots-arent-the-problem-its-unpredictable-ai/>。
24. 克劳塞维茨的著名观点仍然适用。他指出, 战争"会涉及到巨大的摩擦, 这些摩擦不能像机械一样减至几个点, 与机会存在着密切联系, 所带来到后果无法衡量, 这是因为它们主要取决于机会。"Karl von Clausewitz, 《战争论》, Michael Howard编辑, Peter Paret翻译 (1989年), 第120页。
25. Steven Johnson, "Emergence: The Connected Lives of Ants, Brains, Cities" (2001年), 第19页, 提供了深刻见解: "复杂性是一种系统或模型的行为特征, 它们的组件会以多种方式相互作用和遵循特定规则, 这意味着没有合理的高级指令来定义各种可能的互动...一个复杂系统...的特征在于其相互依赖性, 而复杂化系统的特征在于其层级。"
26. "第一代[空中客车]喷气式飞机的最低持续致命事故率约为每百万次战斗3.0起, 而第二代为0.7左右, 这意味着两代之间的致命事故减少了近80%。相比之下, 第三代喷气式飞机目前每百万次战斗大约发生0.2起事故, 减少了约70%。第四代喷气式飞机的事故发生率最低, 每百万次战斗平均保持在0.1左右, 与第三代相比进一步减少了50%。"参见空中客车公司, 《1958-2016商业航空事故的统计分析》。国际航空运输协会的《2016安全报告》(2017年4月)称"过去十年间, 全球商用"

- 航空业的整体安全水平提高了54%，”第6页，  
<https://www.skybrary.aero/bookshelf/books/3875.pdf>。
27. 将前一脚注中的民事记录与本报告第二部分的军事记录进行对比。
28. 最近一个例子：“许多情况下，推迟现代化已延误了安装现代综合桥接系统。一艘船舶的船员不能简单跨到另一艘同类船舶，期望能发现熟悉的设备或布局。低成本模拟器不能充分还原现实，完全复制高流量区域复杂操作的压力。”海军部，“Comprehensive Review of Recent Surface Force Incidents”，2017年10月26日，第14页。
29. William A. Owens，“The Emerging U.S. System-of-Systems”（1996年2月），是美国参谋长联席会议副主席早期发表的报告，很有影响力。
30. 十年前，一位睿智的观察家认为这种发展在商业和军事系统中均“不可避免”，并补充说：“随着‘系统的系统’潜在利益以及人们期望的增长，人们对它的需求也会增多。它们的数量将会继续增多，而且变得越来越重要。这种变化在美国国防部表现得更为明显。国防部在计算和通信技术进步的推动下越来越重视这种变革，并推出了被称为网络中心战（NCW）的‘系统的系统’。”参见David A. Fisher，“An Emergent Perspective on Interoperation in Systems of Systems”，卡耐基梅隆大学技术报告，2006年3月，第9页，<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.955.5917&rep=rep1&type=pdf>。
31. Patrick Tucker，“The Future the US Military is Constructing: a Giant, Armed Nervous System”，“Defense One”，2017年9月26日，<http://www.defenseone.com/technology/2017/09/future-us-military-constructing-giant-armed-nervous-system/141303/>。Tucker引用众多军队官员的话，为其观点提供佐证。比如，“正如海军作战部长John Richardson上将在七月份说道...“我想把一切都关联起来。”更早提出这种方法的是William A. Owens上将，参见“The Emerging U.S. System-of-Systems”（国防大学战略论坛，1996年2月）。
32. Geer认为“复杂性遮蔽住了相互依赖性，因此它是安全的天敌。”“A Rubicon”胡佛研究所，Aegis Series Paper，no. 1801（2018年），第2页。Geer引用Steven Johnson的“Emergence: The Connected Lives of Ants, Brains, Cities, and Software”（2001年），第19页，解释道“复杂性是一种系统或模型的行为特征，它们的组件会以多种方式相互作用和遵循特定规则，这意味着没有合理的高级指令来定义各种可能的交互。”
33. David Sculley等人：“Hidden Technical Debt in Machine Learning Systems”，<https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>。
34. 同上。
35. 有关组织做出的努力，可参见J. David Morgenthauer等人，“Searching for Build Debt: Experiences Managing Technical Debt at Google”（2012年），<https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/37755.pdf>。
36. 内容列表可参见Stephen Sherman，“World War Two Aircraft Specs of Fighter Planes by Model and Type”（2002年5月，2012年1月更新），<http://acepilots.com/planes/specs.html>。
37. Marilyn R. Pierce，“Earning their Wings: Accidents and Fatalities in the United States Army Air Forces During Flight Training in World War II”，待发表博士论文，堪萨斯州立大学，2013年，摘要，第208页。这种“保证完成任务”的态度几乎与Pierce描述的历史无关。美国海军调查2017年太平洋舰队事故时，专门讨论了这一话题。对于事故率高得无法接受的原因，调查报告还给出了这样一个事实：“行动所占据的首要地位掩盖住了风险评估的严谨性”。海军部，“Comprehensive Review of Recent Surface Force Incidents”，2017年10月26日，第102页。
38. Thomas Kunkle和Byron Ristvet，“CASTLE BRAVO: Fifty Years of Legend and Lore”，国防威胁降低局（D-TRIAC SR-12-001，2013年1月），第127页，<https://web.archive.org/web/20140310004623/http://blog.nuclearsecrecy.com/wp-content/uploads/2013/06/SR-12-001-CASTLE-BRAVO.pdf>。一艘“出于某种奇怪的巧合...在发射前的空中搜索中未被发现”的日本渔船（载有23名船员）也遭到了辐射。
39. “Broken Arrows: Nuclear Weapons Accidents”，[http://www.atomicarchive.com/Almanac/Brokenarrows\\_](http://www.atomicarchive.com/Almanac/Brokenarrows_)

- [static.shtml](#)，记述了这次事件以及其他31起事件：  
“1961年1月24日：执行空中警戒任务时，B-52轰炸机的右翼出现结构损坏，造成两件核武器意外释放。一件安全着陆，几乎没有造成伤害；另一件自由落体后在北卡罗来纳州戈尔兹伯勒镇附近爆裂，武器中的一些铀无法回收，不过这个地区未发现放射性污染。”这个网站完整记录了外国和美国飞机、船舶和潜艇之间发生的碰撞事件。另见Patricia Lewis等人，“Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy”（2014年），第1页：“1945年以来发生了许多令人不安的侥幸事件，差点出现疏忽大意发射核武器的情况。许多解密文件、证词和访谈提供的证据表明世界是幸运的产物。”<http://www.europeanleadershipnetwork.org/medialibrary/2016/02/04/d2106a19/Is%20Trident%20safe%20from%20cyber%20attack.pdf>。另见Schlosser，“Command and Control”。
40. Sagan, “The Limits of Safety”, 第156页及之后，第238页及之后。
41. 前美国国防部长威廉·佩里在回忆录《我在核战争边缘的历程》（2015年），第52-3页描述了1979年的这一事件。
42. “国防部副部长罗伯特·普雷斯周四在媒体吹风会上说：‘我们对这些失败感到震惊。国防部对失败负全部责任。’”参见Heath Druzin, “DOD: Live anthrax sent to labs was ‘a massive institutional failure,’”，《星条旗报》，2015年7月23日，<http://www.stripes.com/news/us/dod-live-anthrax-sent-to-labs-was-a-massive-institutional-failure-1.359510>。《今日美国》对此进行了广泛调查，具体参见<http://www.usatoday.com/topic/9ee9e5de-b702-4ft>c-9e5d-tb595adcf938/biolabs/>。更概括地说，可参见2015年联邦政府对291个受监管生物实验室的审查，这些实验室报告了“12起潜在损失和233起潜在释放事故”- 根据评估没有一起对人类造成了伤害。参见美国卫生及公共服务部等，“2015 Annual Report of the Federal Select Agent Program”（2016年6月），[http://www.selectagents.gov/resources/FSAP\\_Annual\\_Report\\_2015.pdf](http://www.selectagents.gov/resources/FSAP_Annual_Report_2015.pdf)；总审计局，“HIGH-CONTAINMENT LABORATORIES: Improved Oversight of Dangerous Pathogens Needed to Mitigate Risk”，<https://www.gao.gov/assets/680/679392.pdf>；以及“CDC 90 Day Internal Review of the Division of Select Agents and Toxins”，<http://www.cdc.gov/phpr/dsat/documents/full-report.pdf>。“内部审查工作组承认，在理解如何更好地加强生物安全、适当平衡生物选择剂和毒素救生研究与确保公众及这些机构员工安全方面存在不确定性和差距。”
43. 在其回忆录“State Secrets: An Insider’s Chronicle of the Russian Chemical Weapons Program”中（2009年，第196-7页，第202页），Vil.Z. Mirzayanov指出（比如）“只有魔鬼才可能在意化学武器除气部门负责人对于安全的态度...员工的无知或粗心...[以及]过时的安全程序。”一般来说，“苏联军用化学综合体也没有恰当地处理废水和废气。”根据各种情况、各种原因引起的全球民用航空事故变化，我们应能看出美国 and 外国安全记录之间可能存在的差异。比如，非洲每商业里程的事故次数要比美国多数倍。比如参见国际民用航空组织，“State of Global Aviation Safety”（2013年，引用2012年数据），第16页，[https://www.icao.int/safety/State%20of%20Global%20Aviation%20Safety/ICAO\\_SGAS\\_book\\_EN\\_SEPT2013\\_final\\_web.pdf](https://www.icao.int/safety/State%20of%20Global%20Aviation%20Safety/ICAO_SGAS_book_EN_SEPT2013_final_web.pdf)。2014年的报告（引用2013年数据）对各国进行了“区域航空安全分组”，发现撒哈拉以南非洲地区的故事率为美洲的四倍半。国际民用航空组织，“State of Global Aviation Safety”（2014年，引用2013年数据），第8页，附录II对区域进行了界定：[https://www.icao.int/safety/Documents/ICAO\\_2014\\_Safety\\_Report\\_final\\_02042014\\_web.pdf](https://www.icao.int/safety/Documents/ICAO_2014_Safety_Report_final_02042014_web.pdf)；参见2016年报告<https://www.skybrary.aero/bookshelf/books/3681.pdf>。
44. 应该指出的是，美国在自我克制上的敏感性并非最高。比如，“从民意调查和政府政策来判断，可能除法国之外，大多数西方民主国家的反核意识都比美国更强。”参见Nina Tannenwald, “The Nuclear Taboo: The United States and the Normative Basis of Nuclear NonUse”, “International Organization”, 53, no. 3 (1999年夏)，第464页。
45. Jeanne Guilleman, “Anthrax: The Investigation of a Deadly Outbreak”（1999年）。Milton Leitenberg和Raymond A. Zilinskas, “The Soviet Biological Weapons Program: A History”（2012年，第129-131页）描述了一场露天释放天花病毒导致10人死亡的事故。
46. “一人受伤后当场死亡，另一人在医院死亡。根据报道，还有一人当时死于冠状动脉血栓症。最初，237名现场人员和参与清理的人员被诊断为急性放射综合征，后来134例得到确诊。在这之中，28人在事故发生

- 几周后死于该病；1987年至2004年间又有19人死亡，但原因不一定是由于辐射。虽然没有场外人员受到急性辐射，但事故发生后很多儿童被确诊为甲状腺癌，其中大部分可能是因为摄入了放射性碘沉降。此外，白俄罗斯、乌克兰、俄罗斯等大部分地区都受到了不同程度的污染。”世界核协会，“Chernobyl Accident, 1986”（2016年11月更新），<http://www.world-nuclear.org/information-library/safety-and-security/safety-of-plants/chernobyl-accident.aspx>。
47. 1977年的病毒DNA几乎与1950年相同——这种情况极不可能发生在自然界，因为病毒经常会因自然选择而发生变异。有关这种意外情况发生原因的讨论，可成为生物归因困难方面的一个案例研究。这方面的主流观点是“1950年以来，这种病毒一直被冷冻在实验室，1977年被故意或意外泄漏。中国和俄罗斯科学家否认了这种可能性，但到目前这仍是唯一合理的科学解释。”“Origin of Current Influenza H1N1 Virus”，“Virology Blog”，2009年3月2日，<http://www.virology.ws/2009/03/02/origin-of-current-influenza-h1n1-virus/>。另见R.G.Webster等人，“Evolution and Ecology of Influenza A Viruses”，“Microbiology Review”，56 nos.152-179（1992年），170；以及Marc Lipsitch和Alison P. Galvani，“Ethical Alternatives to Experiments with Novel Potential Pandemic Pathogens”，“PLOS Medicine”，2014年5月，<http://journals.plos.org/plos-medicine/article/asset?id=10.1371/journal.pmed.1001646.PDF>。最近公布的证据审查认为：“...中国东北产生的类似分离菌存在三个来源，无法为这是单一实验室事故的结论提供支持。更有可能的是，这些实验室生产的疫苗或用于疫苗开发试验的病毒本身拥有剧毒，从而引发了1977年的流行病。大部分证据都是依赖这些可能性：非自然起源，病毒表现温和，病例在短时间内广泛传播，样本对温度敏感，当代的观察，以及当时会进行活病毒疫苗试验。”参见Michelle Rozo和Gigi Kwik Gronvall，“The Reemergent 1977 H1N1 Strain and the Gain-of-Function Debate”（2015年7月/8月）mBio 6(4):e01013-15. doi:10.1128/mBio.01013-15。第三种比较可靠的观点认为，由于数据非常稀少而且容易出错，可能无法确定最初案例的时间和来源。参见Joel O. Wertheim，“The Re-Emergence of H1N1 Influenza Virus in 1977: A Cautionary Tale for Estimating Divergence Times Using Biologically Unrealistic Sampling Dates”，“PLOS One”，2010年6月，e11184。通过研究一些参考文献，奥克兰儿童医院的Terry Leighton教授评论道：“很明显，需要更详细的系统发育和系统地理数据集来限制相互抵触的爆发源假设。”2018年4月2日通过邮件与作者交流。有关归因于生物袭击的问题可参见Anne L.Clunan等人，“Terrorism, War or Disease”（2008年）。
  48. Tom Schelling最初对这种现象进行了论证，Frank McCown对其进行了详细解释，“Schelling’s Model of Segregation”，<http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/>。Joshua Epstein和Robert Axtel，“Growing Artificial Societies”（1996年），通过模拟含糖细胞的相互作用证明了这种现象。因此，它通常被称为糖域主体模型。
  49. Arthur C. Clarke，“Superiority”（1951年），[http://www.mayo-family.com/RLM/txt\\_Clarke\\_Superiority.html](http://www.mayo-family.com/RLM/txt_Clarke_Superiority.html)。
  50. 商业环境下智能机器之间的交互可为我们提供其他一些例子。马克威尔逊，“AI Is Inventing Languages Humans Can’t Understand.Should We Stop It?”（2017年7月14日），<https://www.fastcodesign.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it>，讨论了由Facebook计算机发明的、可提高沟通效率的语言。谷歌的David Sculley等人指出：“完全独立的系统之间可能存在着隐式回圈。比如，两家不同投资公司的两种股票市场预测模型就属这种情况；一方的改进（或更可怕的是出现错误）可能会影响另一方的竞价购买行为。”“Hidden Technical Debt in Machine Learning Systems,” <https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>。
  51. 这并不是说技术是唯一或者主要的原因，而是说涌现效应导致了冲突的爆发，而这种冲突并不是交战国决策者所希望的结果。
  52. 经典的说法是Barbara Tuchman的“The Guns of August”（1962年）。最近的参考是Christopher Clark，“The Sleepwalkers: How Europe Went to War in 1914”（2012年），它在第305、308、322、336、352和420页阐述了铁路方面的见解。
  53. Nadege Mougel，“World War I Casualties”（REPERES），<http://www.centre-robert-schuman.org/userfiles/files/REPERES%20%E2%80%93%20module%201-1%20-%20explanatory%20notes%20%E2%80%93%20World%20War%20I%20casualties%20%E2%80%93%20>

54. James N. Miller, Jr.和Richard Fontaine, “A New Era in US-Russian Strategic Stability: How Changing Geopolitics and Emerging Technologies are Reshaping Pathways to Crisis and Conflict” (哈佛大学, 贝尔弗中心, 2017年)。
55. 同上, 第16页。
56. “军事革命的支持者和怀疑者都认为创新具有蔓延性, 但在创新的容易程度和速度方面的评估不同。”Leslie C. Eliason和Emily O. Goldman, “Theoretical and Comparative Perspectives on Innovation and Diffusion”, 出自Leslie C. Eliason和Emily O. Goldman编辑的“The Diffusion of Military Technology and Ideas” (2003年), 第6页。
57. 当然, 即使我们限制自己的投资, 也要考虑其他人能否同样快速地 (或甚至更快地) 走上发展道路。当我们宣布不会研发某种武器时, 俄罗斯似乎会加快推进自己的进攻性生物武器计划, 部分原因是他们会认为我们在瞒天过海。比如参见, Milton Leitenberg和Raymond A. Zilinskas, “The Soviet Biological Weapons Program: A History” (2012年)。
58. Miles Brundage等人, “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation” (人类未来研究所, 2018年), 在第17-18页写到: “今天的人工智能系统存在许多新的、尚未解决的脆弱性。这包括数据中毒攻击 (引入可导致学习系统发生错误的训练数据)、对抗样本 (旨在诱导机器学习系统错误分类的程序) 以及利用自主系统目标设计中的缺陷。这些脆弱性不同于传统的软件漏洞 (比如缓冲区溢出), 人工智能系统虽能在许多方面超越人类, 但也可能出现人类永远不会犯的错误。”
59. 美国国家安全局/中央安全局研究理事会负责人Deborah Frincke博士说: “这方面的担忧是在大数据的环境中, 人们可能发现对手可控制足够多的数据, 从而对你产生误导。”我们可以看到, Frincke所称的对抗性机器学习‘正在一点点出现, 我们有理由相信这种发展道路将会持续下去’, 她说。参见Stilgherian, “Machine learning can also aid the cyber enemy: NSA research head” (Full Tilt, 2017年3月15日), <http://www.zdnet.com/article/machine-learning-can-also-aid-the-cyber-enemy-nsa-research-head/>。通过“Attacking Machine Learning with Adversarial Examples” (2017年2月24日), <https://blog.openai.com/adversarial-example-research/>), Ian Goodfellow等人很好阐述了中毒机器学习。另见Nicolas Papernot等人, “Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples”, arXiv preprint arXiv: 1602.02697 (2016年): “[假设]对手可依据所选的既定输入观察到目标[深度神经网络]的输出, 攻击者无需获得模型、参数或训练数据, [便可]控制远程托管的[深度神经网络]。我们引入了以这种方式将替代模型拟合到输入-输出对的攻击策略, 然后根据这种辅助模型制作对抗样本。”
60. 窃取原子弹技术具有挑战性, 而窃取空军最先进的战斗机技术或相当数量的NSA网络攻击工具显然更容易。美国司法部新闻稿, “Chinese National Pleads Guilty to Conspiring to Hack into U.S. Defense Contractors’ Systems to Steal Sensitive Military Information”, 2016年3月23日, <https://www.justice.gov/opa/pr/chinese-national-pleads-guilty-conspiring-hack-us-defense-contractors-systems-steal-sensitive>。
61. 我在Richard Danzig, “Surviving on a Diet of Poisoned Fruit: Reducing The National Security Risks of America’s Cyber Dependencies”中讨论了这一点 (美国新安全中心, 2014年)。
62. 指出俄罗斯和美国“都有强烈的动机, 在冲突初期利用网络和反太空能力获得优势”之后, Miller和Fontaine (“A New Era”) 接着说: “这种新兴局面可能极大增加因意外或疏忽大意而陷入冲突的风险。”对于围绕南海发生重大网络冲突的假想场景, Jonathan Reiber和Arun Mohan Sukumar没想到: “解放军秘密侵入胡志明市Cho Ray医院的网络, 意图监视关键领导人和收集相关信息。在此期间, 一名年轻的解放军少校无意间测试了一款恶意软件, 它可破坏两个病房的用药和给药时间数据。数据破坏性网络攻击造成的第一起已知死亡事故中, Cho Ray医院有四名患者因医疗设备无法提供足量药物而死亡。”当美国通过法医分析确定这些攻击的来源时, Reiber和Sukumar预计会出现升级效应, 但美国情报界无法归因于他们的动机。“Asian Cybersecurity Futures: Opportunity and Risk in a Rising Digital World” (长期网络安全中心, 2017年12月), 第37页。

63. Michael V. Hayden, “Playing to the Edge: American Intelligence in the Age of Terror” (2016年), 第147页。讨论这份报告时, Paul Scharre提出由于缺乏了解, 决策者可能只会看到新技术将能带来的好处, 而不是可能引发的不利后果。
64. Ray Monk, “Robert Oppenheimer: A Life Inside the Center” (2012年), 第494页。
65. Monk, “Robert Oppenheimer”, 第494页。
66. 正如其标题所示, Lee Clarke, “Mission Improbable: Using Fantasy Documents to Tame Disaster” (2003年) 肯定了这一点。另见他在“Perrow”第374页及之后对自己著作的讨论。计算极限方面的问题在许多领域都很重要。Peter Lewin, “What Do We Know for Certain About Uncertainty,” (主题演讲, 查尔斯街列格坦研究所研讨会, 2012年6月) 详细讨论了这个问题在经济学中的中心地位。
67. 参见Bruce Schneier, “Should U.S. Hackers Fix Cybersecurity Holes or Exploit Them?” 《大西洋》, 2014年5月19日, <https://www.theatlantic.com/technology/archive/2014/05/should-hackers-fix-cybersecurity-holes-or-exploit-them/371197/>。
68. 最近的例子可参见Peter Bright, “As predicted, more branch prediction processor attacks are discovered”, Ars TechnicaU, 2017年3月26日: “[这些攻击]证明, 人们一直认为存在的隔离边界在推测执行硬件(它们对于高性能处理器至关重要)面前已变得不堪一击。此外, BranchScope显示, Spectre并不是这种推测执行能力的唯一用途。...研究人员可能需要多年时间才能确定推测执行硬件泄漏信息的各种方式, 而开发出强大和通用的防御技术来阻止这种攻击还需更长时间。”
69. “对于几乎任何目标, 所有较真的攻击者都可找到一个“零日漏洞”...这种漏洞的平均预期寿命为6.9年。”参见Lily Ablon和Andy Bogart, “Zero Days, Thousands of Nights: The Life and Times of Zero Day Vulnerabilities and Their Exploits” (RAND, 2017年), xiii。
70. “在一些重要方面, 美国政府领导人对网络技术问题的回应, 就像两百年前美国人及其领导者回应我们的西部边疆问题一样。我们对经历过的事情感到兴奋, 并相信未来的发现和开发将会产生巨大的变革效应。与此同时, 我们也充满了迟疑和矛盾——新疆域将会是什么样? 该如何治理? 未来几十年它定会动荡不安吗? 我们该如何协调旧的价值、利益、权力关系和实践, 与边疆的陌生风险、要求和做事方式?” Richard Danzig, “Forward”, Richard M. Harrison和Trey Herr, “Cyber Insecurity: Navigating the Perils of the Next Information Age” (2016年), xi。
71. 美国疾病控制与预防中心, “Quick Learn Lesson: Recognizing the Biosafety Levels”, <https://www.cdc.gov/training/quicklearns/biosafety/>。
72. Keith Rhodes (美国审计总署技术与工程中心首席技术专家) 在“High-Containment Biosafety Laboratories: Preliminary Observations on the Oversight of the Proliferation of BSL-3 and BSL-4 Laboratories in the United States” (2007年10月4日) 中提到了生物安全防护四级实验室的数量。我们无法精确统计美国生物安全防护三级实验室的数量, 最好的依据是2010年, 生物安全防护三级实验室的注册数量为1,495个。Jocelyn Kaiser, “Taking Stock of the Biodefense Boom”, 《科学》, 2011年9月2日, 第1214页。值得注意的是, 即使美国审计总署似乎也采信了这位记者根据八年前的注册数值估计的总量。美国审计总署评论道: “美国的高级别防护实验室缺少可靠的统计数字。”美国审计总署, “High-Containment Laboratories: Assessment of the Nation’s Need Is Missing”, 2013年2月25日, 第6页, <https://www.gao.gov/assets/660/652308.pdf>。
73. Lynn C. Klotz和Edward J. Sylvester, “The Consequences of a Lab Escape of a Potential Pandemic Pathogen”, Front Public Health, no. 2 (2014年), 第116页, 10.3389/fpubh.2014.00116, PMID: PMC4128296。
74. 同上。
75. Gigi Kwik Gronvall和Michelle Roza, “Synopsis of Biological Safety and Security Arrangements”, [http://www.upmchealthsecurity.org/our-work/pubs\\_archive/pubs-pdfs/2015/Synopsis%20of%20Biological%20Safety%20and%20Security%20Arrangements%20UPMC%20072115.pdf](http://www.upmchealthsecurity.org/our-work/pubs_archive/pubs-pdfs/2015/Synopsis%20of%20Biological%20Safety%20and%20Security%20Arrangements%20UPMC%20072115.pdf)。

76. 仍拿我们对边疆的探索作为例子，“就像我们的先辈一样，我们只能部分了解和绘制未知的[网络安全领域]。我们依赖的是从各个定居点过滤回来的报告，以及探险者就连自己也摸不清的野径。根据这些杂乱信息，我们会尝试着治理至少已稳定了下来的地区，偶尔会通过突袭来维护一些关键地区的秩序，并会或多或少地默认其他地方的无政府状态。我们几乎无法了解将要发生的事情，因此只能根据内心对未来的描绘来孕育希望和恐惧。”Richard Danzig, “Cyber Insecurity”, xi。
77. Patrick Tucker, “The Future the US Military is Constructing: a Giant, Armed Nervous System,” “Defense One”, 2017年9月26日, <http://www.defenseone.com/technology/2017/09/future-us-military-constructing-giant-armed-nervous-system/141303/>。二十年前，一位学术观察家提出了同样的观点：“最持久的军事传统观念是相信人类是最终的负责者，并能最终控制一切，直到/除非他们的装备完全失效。”参见Gene I. Rochlin, “Trapped in the Net: The Unanticipated Consequences of Computerization” (1997年)，第167页。
78. 当然，我们也可通过编程将这种偏见输入机器。IBM在展示Watson的强大时（曾在智力竞赛节目“Jeopardy”中击败对手），“这个人工智能系统的答题机能可被调整为更积极（错误率高）或更保守（准确率高）。”Erik Brynjolfsson和Andrew McAfee, “The Second Machine Age” (2014年)，第26页。
79. 出于这些原因，我认为应提出一个强大假定，即关键系统会在自己的访问和运行中整合非网络安全措施。非网络组件也被称为“带外”措施，可以包括（比如）让人类参与决策，使用模拟设备检测数字设备，以及可在网络系统遭到破坏时提供非网络替代方案。推动信息技术发展的数字信息系统只会招来数字漏洞。我们可强制攻击者应对计算机代码无法覆盖的系统属性，从而保护自己。Richard Danzig, “Surviving on a Diet of Poisoned Fruit: Reducing the National Security Risks of America’s Cyber Dependencies”, (美国新安全中心, 2014年7月)，第21页。Dan Geer极力主张将模拟系统保留下来，作为关键数字系统的一种替补。Daniel E. Geer, Jr., “A Rubicon” (胡佛研究所, Aegis Series Paper, no. 1801, 2018年2月)。Peter Levin在与作者的交流中另外提到，虽然一个普通漏洞有可能破坏一个复杂系统——比如数字机器，但单一策略在与人类决策者打交道时并没有成功的把握，因为每个人都有自己的特质。
80. 训练标识引发的美国警戒员误判，以及苏联军官置机器判定于不顾，拒绝报告即将发生的核攻击，似乎基于的便是这种情境理解。
81. 比如在战斗中，我们既害怕漏报（将攻击者误当作无辜者），也担心误报（将无辜者误当作攻击者）。
82. 包括美国在内的许多国家都使用过生物武器。发表这一声明后，苏联仍在推进一项庞大的生物武器计划。Milton Leitenberg和Raymond A. Zilinskas将苏联的这个计划写入了“The Soviet Biological Weapons Program: A History” (2012年)。
83. 美国海军, “Formal Investigation into the Circumstances Surrounding the Downing of Iran Airflight 655 on 3 July 1988”, 调查报告，第13页。
84. Nancy C. Roberts和Kristen Ann Dotterway, “The Vincennes Incident: Another Player on the Stage?” “Defense Analysis”, 11, no. 1 (1995年)，第31页。
85. “文森斯号事件还有许多未解之谜，各种听证会、报告和文章都证明了这一点。”Roberts和Dotterway，第42页。Roberts和Dotterway指出，调查人员可能夸大了人为故障：“调查人员发现系统数据和机组人员的回忆之间存在差异，而且无法解释之后，便推测这个问题一定是由任务定位、场景实现和战斗压力所致。考虑到时间压力，相比于留下一个无解和令人尴尬的关键问题，或质疑NTDS等战术支持系统的有效性，也许将失误归咎于操作人员会更简单 (Link 11)。”Roberts和Dotterway，第43页。另见David Evans, “Vincennes: A Case Study”, 美国海军研究所论文集, 1993年8月, <https://www.usni.org/magazines/proceedings/1993-08/vincennes-case-study>。Scott Sagan也将这个事例当作一个案例研究，写入了建立交战规则的挑战报告。Sagan强调指出，斯塔克号护卫舰遭袭事件发生之后，海军的交战规则已变得更宽松。Scott D. Sagan, “Rules of Engagement”, 《安全研究》，1991年，第78-108页, <https://doi.org/10.1080/09636419109347458>。
86. 美国海军, “Airflight 655 Investigation Report”, 第4页。
87. Clint R. Balog, “Airbus v. Boeing: Whose Automation

Philosophy is Best?”Aerospace America, 2015年7月至8月, 第33页及之后。

88. 同上, 第34页。

89. Balog更喜欢波音的立场。“虽然人脑可能被计算机超越, 但它仍是世界上最强大、最灵活、适应性最强的复杂处理器。”Balog, “Airbus v. Boeing”, 第35页。对比分析2009年的两起空中客车事故时, William Langewiesche也将焦点对准了人类飞行员。第一个是“Fly by Wire”(2009年), 描述了美国航空公司的一位飞行员在发动机被飞鸟撞击失灵之后, 设法将飞机安全降落在纽约哈德逊河的故事。第二个描述了一架法航班机从里约热内卢飞往巴黎, 因飞行员人为错误而导致228人丧生的事故。“The Human Factor,”《名利场》, 2014年10月, <https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash>。“Fly by Wire”的第98-115页介绍了“fly-by wire”数字技术的演变, 其中指出“这项新技术具有革命性。...飞机有史以来第一次拥有了有别于莱特兄弟飞机的根本性特征”Balog, “Airbus v. Boeing”, 第9页。

90. 比如, 未来十年可能出现的高超音速导弹巡航变体(飞行速度超过声速五倍)能够“危及半径1000公里之内的目标...并可在几分钟内发起攻击。”一般来说, “防御者如果拥有强大的地面和空间传感器, 只需几分钟便可知道这些导弹已经入境。”Richard H. Speier等人, “Hypersonic Missile Nonproliferation: Hindering the Speed of a New Class of Weapons”(RAND, 2017年), 第12页和14页。

91. Cara LaPointe和Peter L. Levin, “Automated War: How to Think About Intelligent Autonomous Systems in the Military”, 《外交事务》, 2017年9月5日。值得注意的是, 作者利用“在下个世纪”这句话插入了自己的观点。与许多预测一样, 趋势非常明显, 但时间无法确定。在其作者看来, 验证这个看法的时间不会超过一、二十年。

92. Amir Husain, The Sentient Machine: The Coming Age of Artificial Intelligence (2017年) 第100-101页。

93. Nicholas Thompson, “The Former Secretary of Defense Outlines the Future of Warfare”, 《连线》, 2017年2月19日, <https://www.wired.com/2017/02/>

[former-secretary-defense-outlines-future-warfare/?](https://www.wired.com/2017/02/former-secretary-defense-outlines-future-warfare/?mbid=social_gplus)

[mbid=social\\_gplus](https://www.wired.com/2017/02/former-secretary-defense-outlines-future-warfare/?mbid=social_gplus)。另见David Emery, “Robots with Guns: The Rise of Autonomous Weapons Systems”, Snopes, 2107年4月21日, [https://](https://www.snopes.com/2017/04/21/robots-with-guns)

[www.snopes.com/2017/04/21/robots-with-guns](https://www.snopes.com/2017/04/21/robots-with-guns): “现任国防部副部长罗伯特·沃克(也曾在奥巴马政府期间担任该职)坚信人工智能可在战场发挥巨大作用, 同时也强烈反对将人类的介入排除在外。他...没有理会‘机器人杀手’的言论, 以及与‘终结者’进行比较。‘美国观念中, 人类永远都是致命武器的决策者’, 他说。‘故事结束。’”

94. Jeffrey Lewis, “Is Launch Under Attack Feasible?”《核威胁倡议》, 2017年8月24日, <http://nti.org/6687A>, 提供了一种宝贵的决策过程, 这也包括他的判断“美国总统最多只有2-3分钟来权衡和考虑备择方案。”随着武器速度的增加, 文森斯号指挥官曾面对的、总统也将面对的个位数决策时间(分钟)将会变得越来越普遍。Richard H. Speier, “Hypersonic Missile Nonproliferation”, 第12页和14页。

95. 即使接受过更多技术培训、日常经验更丰富的操作人员也观察到了这一点。Nadine Sarter被William Langewiesche称为“密歇根大学的工业工程师、这个领域的杰出研究人员之一”, 后者在报道中写道: “Sarter撰写了大量自动化意外方面的文章, 事故原因通常都是飞行员没有完全了解控制模式, 或者飞机可能被切换到了自动驾驶状态, 飞行员却未留意到飞机的消息提醒。这些因素无疑加重了法航班机447的混乱状况。...Sarter说, 这个系统性问题现已变得非常复杂, 而不仅限于一家制造商。我能轻松列出10个或更多制造商的事件, 这些问题都与自动化和混乱有关。错综复杂意味着你拥有大量的子部件, 它们有时会以意想不到的方式进行交互。飞行员并不知道这些, 因为他们没有经历过系统内部的边缘状态。我曾看到五位工程师聚在一起研制飞机, 便问道: ‘这个/那个部件有何作用?’对于这个问题, 他们无法提供一致答案。所以我在想, 就连五位工程师都不能统一看法, 那么可怜的飞行员如遇到这个问题可该怎么做...只能祝他好运了。”参见William Langewiesche“‘The Human Factor’”, 《名利场》, 2014年10月, <https://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash>。

96. Cathy O'Neil, “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens

- Democracy”（2016年）深入探讨了算法的力量，以及日常生活如何会不加批判地接受它们。Christian Sandvig等人，“Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms”，2014年5月22日国际交流协会第64届年会预备会议“Data and Discrimination: Converting Critical Concerns into Productive Inquiry”论文，该论文深入调查了用户可能无法察觉的算法偏差。
97. 当人类在逻辑思维和场景线索分析方面明显优于机器时，便可避开人类如何增加效益的问题。然而，随着机器在国际象棋、围棋、“Jeopardy”和扑克等游戏中屡屡战胜人类，这一基础已开始动摇。这些都是具体的任务，但“...甚至在最近的突飞猛进之前，这一部分已在随着时间的推移稳步扩展。此外，经常出现的情况是，一旦人工智能系统在特定任务方面（比如国际象棋）达到了人类的表现水平，他们便会继续进化，超越最具有才华的人类。根据一项调查，几乎所有人工智能研究人员都期望人工智能系统能在所有调查任务中，最终达到并超过人类的表现水平。大多数人认为，这种超越很可能在未来五十年实现。”参见Miles Brundage等人，“The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation”（人类未来研究所，2018年），第16页。
98. 更多历史背景可参见James R. Beniger，“The Control Revolution: Technological and Economic Origins of the Information Society”（1986年）。
99. 安全基因项目旨在将控制理论引入基因工程。这方面的一般性讨论可参见Domitilla Del Vecchio等人，“Control Theory Meets Synthetic Biology”（皇家学会出版社，2016年6月），<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971224/pdf/rsif20160380.pdf>。
100. Nassim Nicholas Taleb，“The Fourth Quadrant: A Map of the Limits of Statistics”，[https://www.edge.org/conversation/nassim\\_nicholas\\_taleb-the-fourth-quadrant-a-map-of-the-limits-of-statistics](https://www.edge.org/conversation/nassim_nicholas_taleb-the-fourth-quadrant-a-map-of-the-limits-of-statistics)。
101. 参见本报告第二部分引用的有关第一次世界大战的讨论，以及Miller和Fontaine对意外导致空间和网络冲突风险的评论。
102. 2011年，日本爆发的超级海啸几乎摧毁了福岛第二核电站，美国在提供支持方面发挥了核心作用。但是，我们虽能根据美日战时周密行动计划采取应急措施，但面对意外我们几乎没有应急计划，不得不临时提供指挥和通信关系、运输流量等。参见Richard Danzig和Andrew Saidel，“Beyond Fukushima: A Joint Agenda for U.S. and Japan Disaster Management”（CNAS，2012年）<https://www.cnas.org/publications/reports/beyond-fukushima-a-joint-agenda-for-us-japanese-disaster-management>。另一个例子是2013年和2014年，美国领导人认识到美国在对抗西非埃博拉病毒方面存在极大的利益关系。美国国家安全机构在利比里亚发挥了核心作用，为英国和法国在几内亚和塞拉利昂的工作提供了支持。但是，疫情爆发的后果将会非常严重，如果受灾国家不喜欢西方干预，美国安全机构便会面临更高的要求。
103. Stefano Battiston等人在“Complexity theory and financial regulation: Economic policy needs interdisciplinary network analysis and behavioral modeling”（《科学》，2016年2月19日）中提供了简短而精彩的总结，<http://polymer.bu.edu/hes/rp-battiston16.pdf>。
104. Hans Heesterbeek等人，“Modeling Infectious Disease Dynamics in the Complex Landscape of Global Health”，《科学》，2015年3月13日，<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445966/>。
105. Arnold J. Toynbee引用自AZQuotes.com，<http://www.azquotes.com/quote/945879>。
106. 国防部指令8500.01，主题：网络安全，2014年3月14日，[https://rmf.org/images/stories/rmf\\_documents/850001\\_2014.pdf](https://rmf.org/images/stories/rmf_documents/850001_2014.pdf)；国防部指令8510.01，主题：国防部信息技术风险管理框架，2014年3月12日，[https://rmf.org/images/stories/rmf\\_documents/851001\\_2014.pdf](https://rmf.org/images/stories/rmf_documents/851001_2014.pdf)。
107. 这并不是说所有用户都必须成为专家。与此相反，它强调的是我们需要确定并不断寻找一些非常了解我们正在使用的重要系统的专家，以及使用这些系统的人应对相关技术的限制和缺陷能有一定了解。
108. 比如，情报高级研究计划署启动了“威胁的功能基因

- 组和计算评估”(Fun GCAT) 计划...为核酸序列的筛选、特定基因的功能注释和特性描述开发新的方法和工具, 目的在于防止意外或故意制造生物威胁。”它的目的是开发“可从计算和功能层面增强核酸序列分析能力的工具, 通过与已知威胁功能的比较, 将威胁潜力归因于已知和未知基因。”IARPA-BAA-16-08, 2016年9月, <https://www.iarpa.gov/index.php/research-programs/fun-gcat>。
109. “我们该如何期望自主系统自己决定——尤其是能够支持或实施致命行动的系统? 我们如何向它们灌输符合人类价值观的判断标准? 基本上这也是我们培训员工的方式: 利用我们对过去事件和先前状况的了解来预测未来结果。”Cara LaPointe和Peter Levin, “Automated War: How to Think About Intelligent Autonomous Systems in the Military”, 《外交事务》, 2016年9月5日, <https://www.foreignaffairs.com/articles/2016-09-05/automated-war>。另见O’Neil, “Weapons of Math Destruction”, 第209页: 我们“只有当获得了拥有正反馈回路的生态系统时”才能改进算法。
110. O’Neil在“Weapons of Math Destruction”第207-8页指出“数学模型应是我们的工具, 而不是主人”, 呼吁进行“算法审核”。Sam Arbesman在“Overcomplicated Technology And The Need For Biological Thinking: Complex Systems Like Stock Trading Software Need To Be Studied Like An Ecosystem” (Ars Technica, 2016年11月6日, <http://arstechnica.com/business/2016/11/overcomplicated-technology-and-the-need-for-biological-thinking/>) 中建议道: “野外生物学家非常清楚他们正在做的假设, 知道任何时刻他们看到的只是复杂性的一小部分...同样, 当遇到复杂的技术系统时...如果我们试图将优雅或简洁的感觉强加给它的整体, 那么物理思维将只能带领我们到达目前的位置。我们若想了解自己的技术系统、预测它们的行为, 就需成为技术方面的野外生物学家。”
111. 软件系统复杂性方面的经典描述指出“修正一处缺陷, 便有20-50%的机会引发另一个缺陷。因此, 整个过程就是走两步退一步...所有修正都会破坏结构, 增加系统的熵和无序性...“事务的最佳状态都是最开始时。”Frederick OP.Brooks, “The Mythical Man-month: Essays on Software Engineering” (1975年, 周年版, 1995年), 第122-3页。
112. 比如增材制造所面临的攻击风险。“没有哪项商用技术能在视觉上确保打印层完全达到设计者的意图。尽管利用机器视觉来监控构建方面出现了一些新兴研究, 但目前仅会根据STL文件来核查打印部分。如果攻击者修改了打印部分的STL文件或NC代码, 即使打印层与设计者的意图不同, 打印机也会报告无误, 因为它是在按照修正文件进行核查。Hamilton Turner等人, “Bad Parts: Are Our Manufacturing Systems at Risk of Silent Cyberattacks?”[www.computer.org/security](http://www.computer.org/security)。
113. “Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects: Food-for-thought paper Submitted by the Chairperson” 2017年9月4日, [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/2117A10B-536751D2C1258192004FD7EA/\\$file/FoodforthoughtPa-per\\_GGELAWS\\_Final.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/2117A10B-536751D2C1258192004FD7EA/$file/FoodforthoughtPa-per_GGELAWS_Final.pdf)。
114. David Gunning, “Explainable Artificial Intelligence (XAI)”, (2016年) <http://www.darpa.mil/program/explainable-artificial-intelligence>。情报高级研究计划署署长Jason Matheny敦促人工智能系统开发人员“从一开始就讨论可解释性”, 因为这样做有助于提高这些系统结果对政策制定者的说服力。<https://www.youtube.com/watch?v=z-vpdFWPYBs> (Jason Matheny在23:20发表的演讲。) 他认为“可解释性”即使导致了效率降低也非常值得, 因为它能增强可用性。
115. Gunning, “Explainable Artificial Intelligence (XAI)”, Will Knight, “The Dark Secret at the Heart of AI”, 《麻省理工科技评论》, 2017年4月11日, 进一步评论道: “从2018年夏开始, 欧盟可能要求公司必须为用户解释自动化系统所做的决定。这可能无法做到, 即使表面看来相对简单的系统也是如此, 比如使用深度学习来提供广告或推荐歌曲的应用和网站。运行这些服务的计算机会为自己编程, 而其中的方式我们无法理解。甚至是这些应用的编程工程师也无法解释清楚它们的行为。”最近一篇具有开创性的论文探讨了可解释性人工智能的一种工作方式。证明机器可根据完整的面部图像来判断同性恋者之后, 研究人员通过向

- 其提供部分信息（比如面部的象限）来深入了解机器决策的基础。系统变化揭示了机器所依赖的关键变量。Yilun Wang和Michal Kosinski, “Deep Neural Networks Can Detect Sexual Orientation from Faces”, 《人格与社会心理学》, 2017年, <https://osf.io/fk3xr/>。另见Andrew Lohn等人, “How We Can Overcome the Risks of AI”, 《时代》, 2015年10月22日, <http://time.com/4080577/artificial-intelligence-risks/>。
116. 威斯研究所描述了它在这方面的努力, 参见Benjamin Boettner, “Kill Switches for Engineered Microbes Gone Rogue”, 2017年11月16日, <https://wyss.harvard.edu/kill-switches-for-engineered-microbes-gone-rogue/>。Stuxnet拥有这样的一个断开开关, 可在规定日期终止程序。但要注意, 任何时候都会出现意外。对于Stuxnet来说, 将其终止取决于计算机内时间和日期的正确设置。Andrew Leedom, “Stuxnet: Risk & Uncertainty in the First Salvo of Global Cyber Warfare”, “SAIS Europe Journal”, 2016年4月1日, <http://www.saisjournal.org/posts/stuxnet>。
  117. “基因编辑领域一直在狂速发展, 为以前不可实现的遗传解决方案打开了大门, 但没有高度重视如何减轻潜在的负面影响。”安全基因项目经理Renee Wegrzyn说。美国国防高级研究计划局的新闻稿继续写道: “美国国防高级研究计划局推出安全基因项目是为了完善这些能力: 所有潜在应用都要确保安全第一, 通过提供防止和减少滥用工具来推动科学可靠发展。七个团队...都将努力实现以下三个技术目标中的一个或多个: 开发基因构建体(即生物分子“指令”), 为生命系统中的基因组编辑提供空间、时间和可逆控制; 制定新的药物对策, 提供预防和治疗方案, 以限制生物体中的基因组编辑、保护生物群体中的基因组完整性; 创造一种能力, 剔除系统中不需要的工程基因并将其恢复到遗传基线状态。”参见美国国防高级研究计划局, “Building the Safe Genes Toolkit”, <https://www.darpa.mil/news-events/2017-07-19>。另见Joseph Garthwaite, “U.S. Military Preps For Gene Drives Run Amok: DARPA Researchers Are Developing Responses For Accidental Or Malicious ‘Genetic Spills’” 《西雅图时报》, 2016年11月18日。
  118. Lynn Eden, “Whole World on Fire: Organizations, Knowledge and Nuclear Weapons Devastation” (2004年) 从历史角度分析了美国空军如何将相当多的注意力放在核武器爆炸效应方面, 但极大忽略了可能比爆炸损害严重二至五倍的火灾后果。她总结了一个教训, 即组织总在努力“解决组织已决定要解决的问题...灾难发生的可能性显而易见...人们对物理过程可能知之甚少或不曾预料到。”(第3页和5页)。组织惯例和选择是一种自我强化: “学习效果、高研究成本、相互依赖和期望的自我强化会巩固已经做出的选择。对于爆炸损害, 专家知识被编入了惯例, 不断凝聚更多的组织能力来预测爆炸损害。对于火灾损害, 专家知识没有转化为组织惯例, 没有建立起预测能力。...也许最重要的是, 参与者是通过了解组织能力的差异正日益加剧(能预测爆炸损害, 却无法预测火灾损害), 来证明事实上无法预测火灾损害的。”(第285-286页)。
  119. 参见尾注5。
  120. Sagan, “The Limits of Safety”, 第237页。
  121. Perrow, “Normal Accidents”, 第382页。
  122. 在讨论潜在的大流行病病原体的发展时, 两位专家写道: “鉴于这一过程中的风险, 风险评估过程应由那些在结果中没有明确个人利害关系的人指导, 正如科学的同行评议是由没有直接的人进行的那样。对结果感兴趣。风险评估的可信度将取决于定量过程的严格程度和过程的客观性。”Marc Lipsitch和Thomas V. Inglesby, “暂停研究旨在创造新的潜在流行病病原体”(美国微生物学会, 2014年), doi: 10.1128/mBio.02366-1412 December 2014 mBio vol. 5no.6 e02366-14。
  123. 国防核设施安全委员会是一个依法于1988年成立的独立组织, 为定期进行外部审查提供了先例。
  124. 人工智能领域的这样一个开端可参见“Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects: Food-for-thought paper Submitted by the Chairperson”, 2017年9月4日, [https://www.unog.ch/80256EDD006B8954/\(httpAssets\)/2117A10B536751D2C1258192004FD7EA/\\$file/Food-forthoughtPaper\\_GGELAWS\\_Final.pdf](https://www.unog.ch/80256EDD006B8954/(httpAssets)/2117A10B536751D2C1258192004FD7EA/$file/Food-forthoughtPaper_GGELAWS_Final.pdf)。

125. Nick Beckstead等人, “Unprecedented Technological Risks” (人类未来研究所, 2014年), 第10页, <https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf>, 说明了如何使用这种技术来促进核不扩散制度的遵守, 并解释了联合国助理秘书长Stephen Stedman教授提出的倡议, 从而“帮助发展中国家获得安全技术, 继而建立安全和监测系统来防止[生物技术]事故和恐怖活动的发生。”
126. “纳恩-卢格计划”有助于推动核武器的这种努力。“负责任的治理体系需将明确的国际对话机制引入管理局, 或许还有使用潜在基因驱动技术和生物安全可比标准方面的正式或非正式协议。”美国国家科学院, “Gene Drives on the Horizon: Advancing Science, Navigating Uncertainty, and Aligning Research with Public Values”, (2016年), 第163页。
127. 我们对自然事件的回应说明了需求的存在。参见上文有关福岛和埃博拉危机的讨论, 尾注104。
128. 国防部全球新发传染病监测与反应系统是对健康风险共同特征的一种认可, 也是一种国际合作模式。联合国政府专家组和2013年的一份美俄协议提出了打击某些犯罪或恐怖主义行为的网络合作设想。Thomas Remington等人, “Working Group on the Future of U.S.-Russia Relations: Toward U.S.-Russian Bilateral Cooperation in the Sphere of Cybersecurity”, Working Group Paper #7, 2016年5月, 总结了联合国的观念: “报告提出的实用建议包括...各国之间展开合作, 适当减轻从其领土发生恶意网络活动的可能性。比如, 如果其他国家向美国或俄罗斯系统发起攻击, 恰如其分的回应是受害者(通过国家的计算机应急响应小组)与攻击国的对应机构取得联系, 要求提供详细解释, 比如: ‘我们认为系统X受到的网络攻击来自系统Y (IP地址: aa.bb.cc.dd), 相关IP属贵国所有。能否告知一下贵国是否掌握有本次攻击的相关证据, 以便我们共同分析? 根据该协议, 双方将能行使权利, 通过适度击破假定攻击者的ICT空间来化解攻击。两个行动者能否共同针对网络威胁展开调查(分析)和化解攻击, 是对协议稳健性的一种测试。”
129. 比如, 中国虽承认基因编辑可导致不可预测的“脱靶”后果, 但与美国的现行政策相比它似乎更愿意尝试。参见David Nield, Science Alert, 2016年4月11日,

<http://www.sciencealert.com/scientists-genetically-modify-an-em-bryo-for-only-the-second-time-ever>; “中国科学家再次对人类胚胎进行基因改造: 他们试图通过胚胎基因改造来抗击艾滋病病毒。人类胚胎基因改造方面的伦理问题虽已引发争议, 但中国科学家已第二次成功进行了这种改造。该团队这次使用了CRISPR/Cas9基因编辑器, 尝试创建抗HIV胚胎。”这个问题应放在中国优生观念的大背景下来看待。参见Geoffrey Miller, “What Should we Be Worried About? Chinese Eugenics”, Edge, <https://www.edge.org/responses/q2013>: “三十多年来, 中国实施了全球规模最大、最成功的优生计划, 推动中国迅速成为超级大国。...今年夏天当我了解到中国的优生计划时, 我很惊讶其人口政策受到的关注竟如此之少。无论是文化历史还是政府政策, 中国对自己的优生野心都毫无掩饰。”与美国相比, 俄罗斯和中国似乎更愿意转向不存在人为控制的快速自动化军事系统。这些重要主题需要更多的关注和评估。威斯康星大学的Alta Charo对生物技术风险国际应对方法的多样性进行了深入讨论, “Comparative Approaches to Biotechnology Regulation”, 2015年12月1日发表于国际人类基因编辑峰会, <https://vimeo.com/album/3703972/video/14918256>。作为文化如何塑造技术使用方面一个相对温和但具启发性的例子, 人类学家Jennifer Robertson建议道: “...关键文化因素影响着日本人对机器人的看法。首先是神道教, 即生死方面的一种本地化泛灵信仰。一神教从未在日本生根, 与三大一神教不同, 神道教缺乏复杂的形而上学和神学理论, 主要关注的是纯洁和污染的概念。神道教认为, 世界和宇宙的各个方面都存在“kami”(神道教的神)的重要能量或力量。有些“kami”存在于宇宙, 有些则注入了树木、溪流、岩石、昆虫、动物和人类, 以及玩偶、汽车和机器人等人类创造物。”出自Jennifer Robertson, “ROBO SAPIENS JAPANICUS: Humanoid Robots and the Post-human Family”, 《批判性亚洲研究》, 39, no. 3 (2007年), 第377页。另见Aubrey Belford, “That’s Not A Droid, That’s My Girlfriend”, Countdown, 2013年2月21日, <http://www.countdown.org/en/entries/news/s-not-droid-s-my-girlfriend/>。

130. 这些反过来能得到一些工具的支持, 比如信任建立措施、出口管制、分类和制裁等。条约和规范的力量很容易被低估或高估。Robert Keohane进行了深入探讨, 并指出这些认知形成了“一种被社会视为常识的

实践标准，行动者能够遵从这些实践不是因为它们唯一或最好，而是因为他人也会遵从...这些安排...形成了对他人行为模式的共同期望...建立起了合作关系。..违背承诺的成本将会增多，遵照这些框架的成本将会降低。”“After Hegemony: Cooperation and Discord in the World Economy”（1984年），第89页。

131. 国际废除核武器运动，“Nuclear Arsenals”，<http://www.icanw.org/the-facts/nuclear-arsenals/>。最近的一次详细评估指出：“《不扩散核武器条约》对于限制核武器的扩散发挥了作用”，不过仍不完美。
132. Matthew Fuhrmann和Yonatan Lupu, “Do Arms Control Treaties Work? Assessing the Effectiveness of the Nuclear Nonproliferation Treaty”, 《国际研究季刊》，2016年9月28日，<http://yonatanlupu.com/Fuhrmann%20Lupu%20NPT.pdf>。
133. 只要遵守，美国拒绝宣布其对这一准则的承诺并不会消除它的约束力。“不首先使用核武器”方面的规范可参见Nina Tannenwald, “The Nuclear Taboo: The United States and the Normative Basis of Nuclear Non-Use”, 《国际组织》，53, no. 3 (1999年夏)，第433页及之后；Nina Tannenwald, “The Nuclear Taboo: The United States and the NonUse of Nuclear Weapons since 1945” (2007年)，以及T. V. Paul, “The Tradition of Non-Use of Nuclear Weapons” (2009年)。
134. 在国际关系中胁迫性使用核武器方面，最有分量的定量研究指出：“胁迫性核威胁在核时代的前七十年间缺乏可信度，部分原因是任何出于胁迫目的而使用核武器，都会引发强烈抵制。..领导人如有一天认为不会引发强烈抵制，那么使用核武器进行胁迫的阻碍就会减少。”Todd S. Sechser和Matthew Fuhrmann, Nuclear Weapons and Coercive Diplomacy (2017年)，第257页。
135. 与核不扩散方面的努力一样，40年前制定的《导弹技术控制制度》也是有得有失的一个例子，只是使用了一种截然不同的机制。《导弹技术控制制度》是“各国之间的一种非正式政治谅解”，现有35个“志同道合的国家”“单方面遵守”一般出口限制，但不承担法律责任。“Frequently Asked Questions (FAQs)”, [mtrc.info/frequently-asked-questions-faqs/](http://mtrc.info/frequently-asked-questions-faqs/)。军备控

制协会，“The Missile Technology Control Regime at a Glance”，2017年7月，<https://www.armscontrol.org/factsheets/mtrc>，对导弹控制制度进行了简要概述：“自制定以来，《导弹技术控制制度》已成功减缓或中止了几个导弹计划。它能加大潜在买家实现愿望的难度，或使某些活动和计划显得让人不齿。阿根廷、埃及和伊拉克放弃了它们的Condor II弹道导弹联合研发计划。巴西、南非、韩国和台湾地区也搁置或取消了自己的导弹或太空运载火箭计划。波兰和捷克等一些东欧国家为顺利加入《导弹技术控制制度》，也部分摧毁了弹道导弹。该协定进一步阻碍了利比亚和叙利亚的导弹研发工作。但是，这份协定也有局限性。伊朗、印度、朝鲜和巴基斯坦仍在推进它们的导弹计划。这四个国家都不同程度得到了外国援助，已成功部署了可飞行1000多公里的中程弹道导弹，并正在探索射程更远的导弹。印度正在测试自己的洲际导弹。除印度之外，其它三个国家均未加入《导弹技术控制制度》，它们不只是全球军火市场的买家，而且也是卖家。比如朝鲜已成为当今世界弹道导弹的主要扩散源，伊朗也向叙利亚提供了导弹制造物资。”值得注意的是，中国已宣布将会遵守《导弹技术控制制度》指导方针，其2004年的加入申请因向朝鲜出口导弹技术而被拒绝。

136. 参见美国国务院，“Blinding Laser Weapons (Protocol IV)”，<https://www.state.gov/documents/organization/190580.pdf>。
137. Richard Price, “Reversing the Gun Sights: Transnational Civil Society Targets Land Mines”, 《国际组织》，52, no. 3 (1998年夏)，第613-44页。
138. 参见美国国务院，“Blinding Laser Weapons (Protocol IV)”，<https://www.state.gov/documents/organization/190580.pdf>。我们也成功阻止了北极南极军事竞争。有关最近北极合作（及其面临的挑战）方面的讨论，可参见Stephanie Pezard等人，“Maintaining Arctic Cooperation with Russia: Planning for Regional Change in the Far North” (RAND, 2018年)，[https://www.rand.org/pubs/research\\_reports/RR1731.html](https://www.rand.org/pubs/research_reports/RR1731.html)。
139. Richard Price, “The Chemical Weapons Taboo” (1997年)。

140. 参见Peter Katzenstein (编辑), “The Culture of National Security: Norms and Identity in World Politics” (1996年)。
141. 2015年12月1日至3日, 美国国家科学院、中国科学院和英国皇家学会共同举办了“人类基因编辑国际峰会”。<http://www.nationalacademies.org/gene-editing/Gene-Edit-Summit/index.htm>。美国生物安全协会正在尝试制定可适用于全球的规范。参见<https://www.absa.org/>。同样值得注意的是“Presidential Commission for the Study of Bioethical Issues, New Directions: The Ethics of Synthetic Biology and Emerging Technologies” (2010年)。
142. “Autonomous Weapons: An Open Letter from AI & Robotics Researchers”, 2015年7月28日, <https://future-of-life.org/open-letter-autonomous-weapons/>。Alina Selyukh, “Tech Giants Team Up To Tackle The Ethics Of Artificial Intelligence”, NPR, 2016年9月28日, <http://www.npr.org/sections/alltechconsidered/2016/09/28/495812849/tech-giants-team-up-to-tackle-the-ethics-of-artificial-intelligence>, 报道称亚马逊、Facebook、谷歌、微软和IBM正在展开合作, 努力推进“合乎伦理、安全和可靠技术的研究, 提供帮助而不是造成伤害 - 同时消除这方面的恐惧和误解。”Nicholas Bostrom在“Superintelligence: Paths, Dangers, Strategies” (2014年) 中对这些问题进行了深入讨论。
143. Tim Maurer等人, “Toward A Global Norm Against Manipulating the Integrity of Financial Data”, <http://carnegieendowment.org/2017/03/27/toward-global-norm-against-manipulating-integrity-of-financial-data-pub-68403>。Martha Finnemore和Duncan B. Hollis, “Constructing Norms for Global Cybersecurity” 《美国国际法杂志》 (即将出版), 总结了这方面的一些努力, 并指出“联合国政府专家组和更具包容性的‘London Process’已在倡导适用于各国的通用网络标准。其他网络安全工作针对的是特定范围的行动者 (比如志同道合的国家、主要大国) 或特定利益领域 (比如出口管制、数据保护) 的规范制定。它们的脚注提供了详细参考。此外, 北约还制定了《塔林网络战国际法手册》, 北约网络合作防御卓越中心, <https://ccdcoe.org/tallinn-manual.html>。
144. 参见Elisa Catalano Ewers等人, “Drone Proliferation: Policy Choices for the Trump Administration” (美国新安全中心, 2017年6月), <http://drones.cnas.org/wp-content/uploads/2017/06/CNASReport-Drone-Proliferation-Final.pdf>; Richard H. Speier等人, “Hypersonic Missile Nonproliferation: Hindering the Speed of a New Class of Weapons” (RAND, 2017年)。
145. Martha Finnemore和Duncan B. Hollis, “Constructing Norms for Global Cybersecurity”, 《美国国际法杂志》, 2016年, 对如何构建规范做出了系统概述。虽然Finnemore和Hollis将重点放在了网络安全方面, 但他们的观点覆盖到了所有技术。Rebecca Crootof和Frauke Renz, “An Opportunity to Change the Conversation on Autonomous Weapon Systems”, <https://www.lawfareblog.com/opportunity-change-conversation-autonomous-weapon-systems>, 认为“对话法”能对行为产生影响, 有时效果等同于条约。Dimitri Kusnezov和Wendell B. Jones, “Are Some Technologies Beyond Regulatory Regimes?” (未发表文稿, 2017年) 指出廉价技术的迅速普及造成了混乱局面, 其中众多的竞争参与者无法找到稳定的位置 (纳什均衡)。他们认为“无论其他领域是否成功”, 我们都需要“新方法”来实现某种稳定——或许可从混沌理论发掘出来。
146. 在本节脚注所引用的著作中, 政治学家辩论了国际机构检查、国家之间的作弊行为监督、内部禁忌在多大程度上限制了国际竞争。禁忌虽能形成了一种独立于监督之外的国际约束, 但很难相信重要的技术活动仍属一种禁忌——除非能切实证明其他人也在遵守同样的禁忌, 或无论如何都无法从违规中受益。
147. 我们能从技术信息流通限制中获得一些短暂和次要的好处, 但当相关信息可被用于军民两用性商业项目时, 这些好处通常很小, 非常不足。
148. “苏联代表进行《禁止生物武器公约》最后阶段谈判的同时, ..该国也在推进建造当今世界最大的生物武器。”Milton Leitenberg和Raymond A. Zilinskas, “The Soviet Biological Weapons Program: A History” (2012年), 第9页。
149. 这种趋势也在侵蚀对化学武器的限制: “微型部件的组装需要专业知识, 而且自身就存在一些问题, 尤其

是当它们的微小通道被杂质堵塞时。然而，它们却能用于热控制、混合、层流、反应堆容器和催化，实现高温和高压条件下的反应，促使试剂和反应物即时混合，充当稳定的中间体，以及严格控制放热和吸热反应。通过笔记本电脑对流化学控制，现在的桌面微反应器每年能提供大量的连续合成输出。”参见Richard Danzig等人，“Aum Shinrikyo: Insights Into How

Terrorists Develop Biological and Chemical Weapons”（美国新安全中心，第2版，2012年12月），fn. 206, 64。本注解中的许多引用都提供了符合这种描述的商业设备例子。

150. Elting E. Morison, *Men, Machines and Modern Times* (1966年, 1989年), 第89页。

## 关于美国新安全中心

美国新安全中心（CNAS）致力于制定强大、务实和有原则的国家安全和防务政策。它以员工和顾问的专业知识和经验为依托，吸引政策制定者、专家和公众共同参与具有创新性的、基于事实的研究、创意和分析，以塑造和推动国家安全讨论。美国新安全中心的一个关键使命是为当前和未来的国家安全领导人提供信息，帮助他们做好准备。

美国新安全中心位于华盛顿，由联合创始人Kurt M. Campbell和Michele A. Flournoy设立于2007年2月。

美国新安全中心是一家501(c)3免税非营利组织，它独立开展研究，不隶属于任何党派。作为一家机构，美国新安全中心不对政策问题采取立场。因此，本出版物表述的所有观点、立场和结论仅代表作者个人意见。

© 2018 Center for a New American Security。

版权所有。

1152 15th Street, NW Suite 950 Washington, DC 20005  
t. 202.457.9400 | f. 202.457.9401 | [info@cnas.org](mailto:info@cnas.org) | [cnas.org](http://cnas.org)



**Bold. Innovative. Bipartisan.**