弱势世界假设[1]

(2018) Nick Bostrom
Future of Humanity Institute
University of Oxford

[Working Paper, v. 3.15]
[www.nickbostrom.com]

Translated by Xiaohu Zhu 朱小虎
University AI
[xiaohuzhu.xyz]

推荐语

科技进步常常会破坏文明稳定发展进程。本文作者 Nick Bostrom 提出了“弱势世界假设”，讨论了如何去稳定脆弱世界。我们需要提升全体人类的预防性警务和全球治理的能力。弱势世界假设成为一种深入探讨人类未来发展的新视角，从中可以有效评估现有技术的风险收益，从而有助于我们在推进科学进程中时时审视并有效控制，能够保证技术的一般安全性。

Foreword by translator

Scientific and technological progress often undermines the steady development of civilization. The author, Nick Bostrom, proposes the “Vulnerable World Hypothesis” and discusses how to stabilize a vulnerable world. We need to improve the ability of preventive policing and global governance. The VWH becomes a new perspective to explore the future development of human beings further, from which we can effectively evaluate the risk benefits of existing technologies. It will help us to review and effectively control the scientific process from time to time, and ensure the general safety of technology.

摘要

科技进步可能会以破坏文明稳定的方式改变人们的能力或激励。例如，DIY生物黑客工具的进步可能使任何接受生物学基础培训的人都能轻松杀死数百万人；新型军事技术可以引发军备竞赛，其中任何先发制人的人都具有决定性的优势；或者可能发明一些经济上有利的过程，这些过程会产生难以调节的灾难性负面全球外部性。本文介绍了一个弱势世界的概念：粗略地说，

[1] For comments, discussion, and useful critique, I’m grateful to Nick Beckstead, Miles Brundage, Ben Buchanan, Owen Cotton-Barratt, Niel Bowerman, Paul Christiano, Allan Dafoe, Jeff Ding, Eric Drexler, Peter Eckersley, Owain Evans, Thomas Inglesby, Gregory Lewis, Matthijs Maas, Jason Matheny, Michael Montague, Luke Muehlhauser, Toby Ord, Ben Pace, Anders Sandberg, Julian Savulescu, Stefan Schubert, Carl Shulman, Tanya Singh, Andrew Snyder-Beattie, Helen Toner, and to the audiences of several workshops and lectures where earlier versions of this work were presented); and to Carrick Flynn, Christopher Galias, Ben Garfinkel, and Rose Hadshar for help with the manuscript and many useful suggestions.

其中有一定程度的技术发展，文明几乎肯定会在默认情况下被破坏，即除非它已经退出“半无政府状态的默认条件”。将分析几种反事实的历史和投机未来弱势性并将其排列成类型。稳定脆弱世界的一般能力需要大大放大预防性警务和全球治理的能力。因此，弱势世界假设提供了一种新的视角，从中可以评估对普遍监视或单极世界秩序的发展的风险效益平衡。

存放可能发明的罐子里是否存在一个黑球？ **3**

一个思想实验：简单的核武器 **4**

弱势世界假设 **7**

弱势性的类型 Typology of vulnerabilities **9**

1-型弱势性（简单的核武器）Type-1 (“easy nukes”) 9

2a-型弱势性（安全的首次攻击）Type-2a (“safe first strike”) 10

2b-型弱势性（更糟糕的全球变暖） Type-2b (“worse global warming”) 12

0-型弱势性 Type-0 (“surprising strangelets”) 13

达成稳定 Achieving stabilization **15**

技术放弃 Technological relinquishment 15

偏好修改 Preference modification 16

一些具体的对策及其局限性 Some specific countermeasures and their limitations 18

治理差距 Governance gaps 20

预防性警务 Preventive policing **20**

全球治理 Global governance **22**

讨论 Discussion **23**

结论 Conclusion **26**

参考文献 References **27**

存放可能发明的罐子里是否存在一个黑球？

一种看待人类创造性的方式是看做从一个巨大的罐子里面取出里面的球。[2] 这些球代表可能的想法、发现和技术发明。循着历史的进程，我们已经取出了相当多的球了——大多数是白色的（有益的）但也有不同的灰色的球（中度有害的和混合着祝福的）。在人类条件上的累积效应目前为止仍是非常积极的，也有可能在未来仍旧会变得更好。[3] 全球人类总数已经在过去上万年里面增长了三个数量级，在过去的两个世纪里，人均收入、生活水平和预期寿命也有所提高。[4]

到目前为止，我们还没有取出一个黑球——某一种始终或默认的技术会破坏那个发明它的文明。原因并不是我们对自身的技术策略特别小心或者明智。我们仅仅是幸运罢了。

并没有出现任何人类文明已经被自己发明摧毁了的情况——均是被转变的。[5]我们确实有被其他地方发明的文明毁坏的例子。 例如，支持越洋旅行和兵力投送的欧洲发明可被视为美洲、澳大利亚、塔斯马尼亚和其他一些地方的土着居民的黑球事件。 智人在技术上优势可能促进了古老的人类种群（例如尼安德特人和杰尼索瓦人）的灭绝。 但到目前为止，我们似乎没有看到任何足够自我破坏的发明可被视为人类的黑球。[6]

假如在罐子里有一个黑球会怎样？如果科学和技术研究持续，那么我们会最终达到黑球并取出来。我们的文明有能力将其取出，但去没有能力放回罐中去。我们能够发明但不会**反发明**（un-invent）。我们的策略是希望那里没有黑球。

本文发展出几个能够帮助我们思考技术黑球的可能性和这样的场景可能发生的不同形式的概念。我们还从全球角度讨论了对政策的一些影响，特别是关于如何看待大规模监控的发展以及如何实现更有效的全球治理或更单极的世界秩序。 这些影响决不会解决有关这些宏观战略变量变化的可取性的问题——因为确实还有其他这里没有涉及强相关的因素需要加入到平衡中。然而，在未来关于这些问题的政策辩论中，这些因素应该被考虑在内。

在进入论文的更多概念部分之前，构想一个技术黑球可能看起来像什么的更具体的图景将是有用的。 最明显的是一种可以很容易地发挥出巨大的破坏性力量的技术。核爆炸是我们掌握的最明显的破坏性力量。 因此，让我们考虑在发动这种力量非常容易时将会发生什么。

---

[2] 显然，罐子隐喻有重要的局限性。我们稍后会讨论其中一些。
[3] Bostrom (2008).
[4] 对非人类动物的条件的净影响更难以评估。 特别是，现代工厂化养殖涉及对大量动物的虐待。
[5] 然而，有一些“文化”或当地人口的例子可能是由他们自己的技术实践（至少部分地）带来的，例如复活节岛（Rapa Nui）人和梅萨维德（Anasazi）的祖先普韦布洛人 谁削减了自己的森林，然后遭受了环境的破坏（Diamond 2005）。
[6] 然而，如果我们将进化适应视为发明，则可以在其他物种中找到实例。 例如，有一篇文献探讨了进化死亡如何导致一个物种或一个群体的灭绝，可能来自有利的进化变化，例如专业化的变化（适应狭窄的生态位可能涉及不可逆转的丧失 在更广泛的环境中生存所需的特性）（Day et al.2016），近交系统的出现（例如社会蜘蛛）（Aviles和Purcell 2012），或转向自交（例如开花植物物种转变） 从异交到自体受精）（Igic和Busch 2013）

一个思想实验：简单的核武器

在1933年9月12日伦敦灰色的早晨，利奥·西拉德（Leo Szilard）正在阅读报纸，当时他看到了尊敬的现被公认为核物理学之父的卢瑟福勋爵刚发表的一份报告。[7]卢瑟福在演讲中驳回了从核反应中提取有用能量为“月光”的想法。 这个说法让西拉德很恼火，他就出去散步舒缓心情了。 在行走期间，他得到了核链式反应的想法 —— 核反应堆和核反应堆的基础。 后来的调查表明，制造一种原子武器需要几公斤的钚或高浓缩铀，这两种铀都非常难以生产。 然而，假设它的确已经出现了：那就有了一些非常简单的方法来释放原子的能量 —— 比如通过在两片玻璃之间放置金属物体发送电流。

因此，让我们考虑一个反事实的历史，其中西拉德发明核裂变，并意识到核弹可以用一块玻璃、一个金属物体和一个特定配置的电池制成。 接下来发生什么？ 西拉德变得非常担心。 他认为必须不惜一切代价保密他的发现。 但是怎么保密呢。 他的见解必然会发生在其他人身上。 他可以与他的一些物理学家朋友交谈，他们最有可能偶然发现这个想法，并试图说服他们不要发表关于核链式反应或导致危险发现的任何推理步骤的任何内容。 （这就是西拉德在实际历史中所做的。）

在这里，西拉德面临着一个两难的境地：要么他没有解释这个危险的发现，但就很难有效地说服他的许多同事停止发表研究成果; 要么他告诉他们自己担心的原因，但这样他就是将危险的知识进一步传播开来。 无论哪种方式，他都在打一场失败的战斗。 科学知识的普遍进步最终将使危险的洞察力更容易获得。 很快，弄清楚如何利用金属、玻璃和电力发起核链式反应将不再需要天赋，但凡具有创造性思维的STEM学生都将可以有能力实现。

让我们向后进一步看。 情况看起来毫无希望，但西拉德并没有放弃。 他决定让一位朋友信任他，这位朋友是世界上最著名的科学家——阿尔伯特 爱因斯坦。 他成功地说服了爱因斯坦认可这个危险（再次遵循实际的历史）。 现在西拉德得到了一个可以让他和任何政府进行听证的人的支持。 两人写信给富兰克林·罗斯福总统。 在一些委员会的争吵和报告撰写之后，美国政府的最高层最终确信已准备好采取认真的行动。

美国可以采取什么行动？让我们首先考虑一下究竟发生了什么。[8]美国政府在消化了爱因斯坦和西拉德提供的信息之后做了什么，并且在接受了一些正在调查此事的英国人的进一步推动之后，发起了曼哈顿计划。为了尽快武器化核裂变。炸弹准备好后，美国空军用它来摧毁日本的人口中心。许多曼哈顿科学家通过指出纳粹德国首先获得炸弹时可能出现的致命危险，证明了他们的参与是正当的;德国战败后，他们继续致力于该项目。[9]西拉德主张在一个无人居住的地区而不是在一个城市展示“小工具”，但没有成功。[10]战争结束后，许多科学家赞成国际控制原子能和积极参与核裁军运动;但是，由于核政策已被取消，他们的观点并不重要。四年后，苏联引爆了自己的原子弹。苏联的努力得到了曼哈顿计划中的间谍的帮助，但即使没有间谍活动

[7] Rhodes (1986).
[8] Rhodes (1986).
[9] 虽然参与该项目的大多数科学家赞成像巴鲁克计划这样将核能置于国际控制之下的提案，但他们在这一点上几乎没有保留决策权。
[10] Franck et al. (1945).

，它也会在一两年内成功。[11]随后的冷战，在其高峰时有 70,000 枚核弹头随时准备释放，将会造成全球毁灭，人类颤抖的手指旋在两边的“红色按钮”上方。[12]

幸运的是，人类文明在广岛和长崎遭到破坏之后，没有其他原子弹被愤怒引爆。七十三年后，部分由于国际条约和防扩散努力，只有九个国家拥有核武器。任何非国家行动者都不会相信会拥有核武器。[13]

但是，如果有一种简单的方法可以制造核武器，那会怎么样呢？也许西拉德和爱因斯坦可以说服美国政府禁止所有核物理研究（高安全性政府设施之外）？这种对基础科学的禁令将受到巨大的法律和政治挑战 —— 因为禁止的原因不能在没有造成不可接受的信息危害的情况下公开披露。[14]

但是，我们假设罗斯福总统可以某种方式动员足够的政治支持来推动禁令，并且美国最高法院可以某种方式找到一种将其视为宪法有效的方式。然后，我们面对一系列艰巨的实际困难。所有大学物理系都必须关闭，并启动安全检查。大量的教师和学生将被迫离开。大家都会激烈猜测所有这些严厉措施的原因。物理学博士生和教师被禁止进入他们的研究领域，他们会坐下来猜测秘密危险可能是什么。他们中的一些人会想出来。在那些想出来的人中，有些人会被迫使用这些知识来打动他们的同事；那些同事想要告诉其他人，以表明他们是知情人。或者，反对禁令的人会单方面决定公布这个秘密，也许是为了支持他们认为禁令无效或者出版物的好处超过风险的观点。[15] [16] 政府实验室的粗心或心怀不满的员工最终也会让信息泄露，间谍会把这个秘密带到外国的首都。即使由于某种奇迹，这个秘密永远不会泄漏到美国，其他国家的科学家也会独立发现它，从而使它可以传播的来源倍增。迟早 —— 可能更早 —— 秘密将不再是秘密。在当今时代，当人们可以在互联网上即时和匿名发布时，限制科学秘密的传播将更加困难。[17]另一种方法是消除所有玻璃，金属或电流源（节省也许在一些高度戒备的军事仓库中）。鉴于这些材料无处不在，这样的事业将是非常艰巨的。确保对这些措施的政治支持绝不比关闭物理教育更容易。然而，在蘑菇云在几个城市上升之后，可能会出现集中力量进行尝试的政治意愿。金属的使用几乎是文明的同义词，并不是消除现实的目标。玻璃生产可能被禁止

---

[11] Holloway (1994)

[12] Norris和Kristensen（2015年）。 当然是隐喻的。 但可以说，在这个比喻中，由于广泛授权的命令和控制，每侧应该有不止一个颤抖的手指（Schlosser 2013; Ellsberg 2017）。

[13] 但是，在一个特定的州内，有权发动核攻击的行动者人数可能非常多。 Ellsberg（2017）声称，至少在冷战的很大一部分中，发射核武器的权力被授予了美国指挥系统的多个级别。 具有发射核武器的物理能力的军官人数虽然不是这样做的权力，但也必然更大。 在苏联，在1991年8月对米哈伊尔·戈尔巴乔夫的政变期间，苏联的所有三个棋子（“核公文包”）都掌握在政变领导人手中（Sokoski and Tertrais 2013; Stevenson 2008）。

[14] “信息危害”是传播真实信息所产生的风险，例如，因为这些信息可能使某些代理人造成伤害。 Bostrom（2011）更广泛地讨论了信息危害。

[15] 他们可能会争辩说开放性会带来好处，因为它可以让更多的人参与对策（参见围绕流感病毒获得功能工作的争论；（Fauci等人2011; Duprex等人2015; Sharp 2005）。 他们也可能会争辩说，只要政府继续通过引用秘密信息来证明激烈的行为是合理的，那么对其公民来说将是危险的不负责任。类似的信念激发了美国杂志The Progressive在70年代后期的决定。 即使在美国能源部面临法律挑战的情况下，也会发布有关氢弹制造的秘密。该作者的作者霍华德莫兰德写道：“保密本身，特别是一些指定'专家'的力量 宣布某些主题不受限制，有助于建立核设施可以照常营业的政治气候，保护和延续这些恐怖武器的生产。“Morland（1979）。

[16] 通常，在多个参与者各自具有单独采取行动的独立概率的情况下，随着参与者数量的增加，采取行动的概率趋于一。 当这种现象出现在具有共同目标但不一致的判断的行动者身上时，由于他们所接触的证据的随机性或他们所进行的推理，就会产生“单边主义者的诅咒”（Bostrom等人，2016）。 诅咒意味着即使是一个非常不明智的决定，例如决定发布核武器设计，如果有足够的行动者能够单方面采取这种决定，也很可能做出决定。

[17] Cf. Swire (2015); Greenberg (2012).

，现有的玻璃窗被没收;但是玻璃碎片会长时间散落在整个景观中。可以查封电池和磁铁，但有些人会先将这些材料藏起来，然后才能被当局收集。许多城市将被虚无主义者，勒索者，复仇主义者甚至只想“看看会发生什么”的个人所摧毁。[18]

人们会逃离城市地区。最终，许多地方将被核沉降物摧毁，城市将被废弃，不会使用电力或玻璃。拥有可能用于制造它们的被禁材料或设备将受到严厉的处罚，可能是现场执行。为了执行这些规定，社区将受到严格的监视 —— 线人网络的鼓励，这些网络受到大奖励，频繁的警察袭击私人住所，持续数字监控等等的鼓励。这是乐观的情景。在一个更悲观的情况下，法律和秩序将彻底崩溃，社会可能分裂为发动内战与核武器，产生饥荒和瘟疫的派系。只有当社会变得如此渺茫以至于再也没有人能够从储存的材料或废弃的城市废墟中放置炸弹和延迟雷管时，解体才会结束。即使在那时，危险的洞察力 —— 一旦它的重要性得到如此惊人的展示 —— 将被铭记并传承下去。如果文明开始从灰烬中升起，知识将处于等待状态，一旦人们再次学会如何制造平板玻璃和电流发生器，就会立即扑向。如果文明开始从灰烬中升起，知识将处于等待状态，一旦人们再次学会如何制造平板玻璃和电流发生器，就会立即扑向。 即使知识被遗忘，一旦核物理研究恢复，它将被重新发现。

我们很幸运，制作核武器的确很难。

[18] 今天，许多同样的动机在进行恶意网络攻击的“黑帽”黑客中显而易见。 例如，作为敲诈勒索的方法，一些匿名黑客已证明自己愿意取消城市向居民提供重要服务的能力（Blinder和Perloth 2018）。 经济上破坏性网络攻击的动机似乎也包括政治意识形态和好奇心。 由于当代网络攻击的破坏性远远低于使用核武器的攻击，因此那些愿意使用核武器或威胁使用核武器的行动者肯定比愿意参与恶意黑客攻击的行动者小得多。 然而，与这两种情况相关的社会和心理因素可能相似。

弱势世界假设

我们现在知道，只用一块玻璃、一些金属和一块电池并不会引发核爆炸。制造原子弹需要几千克的裂变材料，这很难生产。那个时候我们拿出一个灰色的球。然而，对于每一项发明行为，我们都是重新将手伸进这个罐子。

让我们来介绍一个假设，即创造力的罐子至少包含一个黑球。我们可以将此称为***弱势世界假设***。直观地说，假设存在一定程度的技术，文明几乎肯定会被破坏，除非实施非常特殊且历史上前所未有的预防性警务和/或全球治理。更准确地说：

> VWH：如果技术发展继续下去，那么在某种程度上可以获得一系列能力，使文明的毁灭极有可能，除非文明充分退出半无政府状态的默认条件。

“半无政府状态默认条件”是指以三个特征为特征的世界秩序：[19]

> （a）预防性警务能力有限。各国没有足够可靠的实时监视和拦截手段，使其境内的任何个人或小团体几乎不可能采取非法行动 —— 特别是对99%以上的人口非常不喜欢的行动。
> （b）全球治理能力有限。没有可靠的机制来解决全球协调问题和保护全球公域 - 特别是在涉及重要国家安全利益的高风险情况下。
> （c）不同的动机。人口分布广泛且可识别地由大量行动者（在个人和州一级）代表 —— 特别是，有许多行动者在很大程度上通过感知到的自身利益来激励（例如金钱、权力、地位、舒适和便利）并且有一些行动者（“世界末日残余”）将以摧毁文明的方式行事，即使对他们自己也要付出高昂代价。

上述定义中的“文明毁灭”一词可以用各种方式解释，产生不同版本的 VWH。例如，人们可以定义一个存在风险的脆弱世界假设（x-VWH），它表明在某种技术水平上，默认情况下会发生存在性灾难，涉及地球智能生命或永久性枯萎的灭绝。我们未来实现价值的潜力。但是，在这里我们将设置较低的栏。在目前情况下，一个关键的问题是文明在目前的半无政府状态违约情况下持续的后果是否灾难性地超过了对退出这一条件所需的急剧发展的合理反对意见。如果这是标准，那么缺乏人类灭绝或存在性灾难的阈值就足够了。例如，即使那些对政府监督高度怀疑的人，如果真的有必要防止偶尔的区域性破坏，也可能会大大增加这种监视。同样，那些重视生活在一个主权国家的人可能会合理地宁愿生活在世界政府之下，因为他们认为这种选择会带来像核浩劫一样可怕的东西。因此，我们规定 VWH 中的“文明破坏”一词指（除非另有说明），任何破坏性事件至少与世界 15% 的人口死亡或全球GDP减少 > 50%持续一样糟糕十多年来。[20]

---

[19] 这一概念不同于国际关系领域的国际无政府状态。目前的概念强调无政府状态是一个程度问题，并且在国际关系中的不同思想流派之间是相对中立的（参见 Lechner 2017）。更重要的是，它包括缺乏治理，不仅是“在顶部”，而且在“底部”。也就是说，半无政府主义的违约条件指的是，在当前的世界秩序中，由于缺乏全球治理或其他完全有效的制约因素，国际层面不仅存在一定程度的无政府状态。国家行动和解决全球协调问题，但在个人（和其他次国家行动者）层面也存在一定程度的无政府状态，即使是功能强大的国家目前也缺乏完全规范这些小行动者行为的能力。例如，尽管许多州试图在其领土内防止强奸和谋杀，但强奸和谋杀仍然以非零频率发生。如果个人获得更大的破坏能力，那么这种无政府状态在底层的后果可能会大大放大。

[20] 相比之下，当前世界人口的 15% 的死亡人数是第一次世界大战，西班牙流感和第二次世界大战的综合影响的两倍多，占全球人口的百分比（绝对值差异更大）。 世界 GDP 下降 50%，大于有史以来

本文的主要目的不是争论VWH是真的。（我认为这是一个悬而未决的问题，尽管在现有证据的情况下，我认为 VWH 是错误的，但在我看来这是不合理的。）相反，这里声称的主要贡献是 VWH 以及相关概念和解释，有助于我们提出有关人类巨大战略态势的重要考虑和可能性。但是，这些考虑和可能性需要进一步分析，并结合本文范围之外的其他考虑因素，然后才能产生任何明确的政策含义。

在我们继续之前再做一些澄清。本文从最广泛的意义上使用“技术”一词。因此，原则上，我们不仅要计算机器和物理设备，还要考虑其他类型的仪器有效的模板和程序 —— 包括科学思想、制度设计、组织技术、意识形态、概念和模因 —— 构成潜在的技术黑球。[21]

我们可以谈论开放和关闭的弱势性。在“易核武器”情景中，脆弱时期始于发现核爆炸的简便方法。当达到某种程度的技术使其能够合理地承受能够阻止核爆炸造成不可接受的损害 - 或者再次使其不可能产生核爆炸（因为技术倒退）时，它就结束了。[22] 如果没有可能的保护技术（如例如，它可能不是核武器的情况，并且技术倒退不会发生，那么世界就会变得永久脆弱。

如果退出半无政府状态默认条件以防止弱势性导致实际灾难，我们也可以说世界正在稳定（关于某些弱势性）。为实现稳定而必须改变半无政府状态默认条件的方式取决于所涉及的弱势性的具体情况。在后面的部分中，我们将讨论可以稳定世界的可能方法。现在，我们只是注意到 VWH 并不意味着文明注定要失败。

的最大跌幅。 例如，在大萧条期间，世界国内生产总值下降了大约 15% 或更少，并且大部分在几年内恢复（尽管一些模型表明它对贸易产生了长期持久的抑制作用，而这种影响长期影响了贸易。 世界经济）（Bolt et al.2018; Crafts and Fearon 2010）。

[21] 本文重点关注技术弱势性。也可能存在与人类文明进步无关的自然弱势性，例如在未来某个日期影响我们星球的暴力流星弹幕。一旦我们的技术能力水平超过某个阈值（例如偏转流星的能力），一些自然弱势性就可以稳定下来。可能的是，技术弱势性的风险大于自然弱势性的风险，尽管对于这种情况而言，与文明破坏的严重性截止相比，如果将截止设置为存在性灾难，则不那么明显（Bostrom 2013; Bostrom and Ćirković2011）。

这种主张（技术弱势性占主导地位）的（大）附带条件是，它预先假定世界并非如此以显着的速度大幅降低价值潜力。相反，如果我们从一个可以称之为流血世界假设的观点来评估事物，那么很可能是由自然（即非人类）过程引起的默认破坏主导了等式。如果，例如，（a）评估者关心现有人（包括自己和家人）并且他们自然会以相当大的速度（例如从衰老）死亡，那么流血世界假设就可以成立，从而失去了能够继续享受生活，并有机会获得更高水平的福祉，例如在技术成熟时可能实现；（b）评估员非常关心避免使用更先进技术可以避免的痛苦和痛苦，但目前正在发生这种情况，从而加剧了负效用；（c）存在一些实质性的外部文明破坏率（例如自然灾害，与我们的活动无关的随机模拟终止），虽然我们允许时间流逝，但在最大限度地发挥我们的技术潜力之前，我们会累积被销毁的风险；（d）有些方法，使用我们目前无法理解的物理学，开始快速增长的价值创造过程（例如通过建立一个指数级联的婴儿宇宙，其居民将非常高兴），评估员关心对这种创造有一种比例敏感的方式；（e）其他超级智能选区，他们能够极大地影响评估人员关心的事情，对我们达到某种进步感到不耐烦，但他们对此的价值随着时间的推移迅速衰退。

[22] 在经历了深刻的技术倒退之后，世界可能仍然处于脆弱状态，例如，如果许多预制核武器甚至在文明退回到无法制造新核武器之后仍然存在。

弱势性的类型 Typology of vulnerabilities

我们可以辨别出四种文明弱势性的类型。

1-型弱势性（简单的核武器）Type-1 (“easy nukes”)

第一种类型是，在“简单的核武器”场景中，它对于造成大规模杀伤的个人或者小团体来说变得太容易了：

> 1-型弱势性：有一些技术具有破坏性且易于使用，鉴于半无政府状态的默认情况，世界末日残余中的参与者的行为极有可能造成文明破坏。

请注意，在确定方案是否存在1型弱势性时，可能导致事件的容易程度与事件的破坏性之间存在反比关系。单个事件的破坏性越大，引起此类事件就越不容易，以便我们诊断出1型弱势性的存在。

因此，考虑一个“非常容易的核武器”情景，其中任何一名半傻瓜可以在一个下午的过程中在厨房水槽上创建一个易于携带的热核武器：这绝对可以当作一个文明的弱势性。将此与“适度轻松的核武器”情景进行对比，其中需要一个由五人组成的半技术人员整整一年辛苦工作以生产单个庞大的千吨级装置：这可能不会达到这个水平文明的弱势性。在“适度轻松的核武器”情景中，似乎有可能绝大多数城市将逃脱破坏，尽管资源充足的恐怖组织，如奥姆真理教 Aum Shinrikyo anno 1995 或基地组织 Al-Qaeda anno 2001 所构成的威胁将基本增加。然而，考虑另一种情况，“适度容易生物毁灭”，其中再次需要一个半熟练的五人团队工作一年才能实施黑球技术，除了这次它是一个生物智能体，单点释放足以杀死数十亿。在“适度容易生物毁灭”中，将达到1-型弱势性的门槛。如果破坏文明只需要一个小组在中等水平的任务中取得成功，那么在半无政府状态的默认条件下，文明可能会在几年内被摧毁。事实上，奥姆真理教和基地组织都试图获得核武器和生物武器，并可能选择使用它们。[23]

因此，如果它非常容易造成中等程度的伤害或者适度容易造成极大的伤害，则存在1型弱势性。[24]如果技术足够易于使用，那么每次事件只能实现适度伤害的黑球技术可以算作一类1-型弱势性的原因 —— 几乎肯定会发生大量此类事件。采取这样的场景，一般人可以轻松制作一个破坏大都市的氢弹。这不一定是一个人可能破坏文明的情景。即使制造单个炸弹相当容易，制造数百枚炸弹并将它们运送到数百个城市而不被抓住仍然是一项艰巨的任务。然而，“简单的核武器”情景呈现出一种文明的弱势性，因为事实上有数百人会在这种情况下摧毁至少一个城市，这似乎是合理的。

这几乎是从大数定律和可能的假设中得出的结论，即对于任何随机选择的人来说，有一些小但可观的机会，他们会被激发触发这种破坏 —— 无论是出于意识形态的仇恨，虚无主义的破坏性，报复不公正，作为一些勒索情节的一部分，或因妄想或精神疾病，或者甚至只是为了看看

[23] See e.g. Danzig et al. (2012), Olson (1999), Mowatt-Larssen and Allison (2010).

[24] 对我们最初的“简单核武器”情景来说很重要的是，每次核使用只需要一个人或一小群人的努力。 虽然可能需要数百名行动者共同努力在这种情况下破坏文明 - 毕竟，破坏一个城市或一个大都市区与破坏文明不同 - 这些数百名行动者无需协调。 这允许世界末日的残余发挥作用。

会发生什么。鉴于人类性格和环境的多样性，对于任何如此不谨慎，不道德或自我毁灭的行为，有一些残余的人类会选择采取这种行动。如果所涉及的行动代表了一种文化上显着的可承担性，那么这一点尤为合理 - 因为在世界任何地方发生过一次这样的核攻击之后，它就会到处发生。换句话说，“简单的核武器”是一个脆弱世界的例证，因为它看起来像世界末日残余与一组赋权的行动者有足够大的交集，人们会期望造成毁灭性文明的破坏。

2a-型弱势性（安全的首次攻击）Type-2a (“safe first strike”)

一种“民主化”大规模毁灭的技术并不是唯一一种可以从罐子中悬挂出来的黑球。另一种技术是强烈激励强大的行动者利用其力量造成大规模杀伤的技术。我们再次可以转向核历史以供说明。

在原子弹的发明和美国短暂的核垄断之后，美国和苏联之间发生了军备竞赛。竞争对手的超级大国积累了惊人的武器库，在1986年超过7万枚核弹头，足以摧毁文明。[25] 虽然自1991年和平结束以来，公众对冷战风险的认识似乎已经消退，但学术界受益匪浅从以前分类的档案的开放和退休政策制定者，官员和分析家的证词 —— 发现了令人不安的一系列做法和事件，这些做法和事件似乎一再使世界陷入困境。[26] 我们离我们的距离有多近争议话题。一些学者认为，只有运气好，才能避免核浩劫。[27]

无论在冷战中幸存还是需要一点运气，我们都可以很容易地想象出一种反事实，即避免核爆燃的可能性会大大降低。即使我们假设核武器只能由技术先进的大型国家生产（从而将这种情况与“易核武器”的1-型弱势性区分开来），这种情况仍然存在。反事实可能涉及技术可能性边界的变化，这将使军备竞赛不那么稳定。

例如，核战略家普遍认为，20世纪60年代中期两个超级大国都发展了相当安全的二次打击能力，为“战略稳定”创造了条件。[28] 在此之前，美国的战争计划反映了很多在任何危机情况下，更加倾向于对苏联的核武库发动先发制人的核打击。基于核潜艇的洲际弹道导弹的引入被认为特别有助于确保二次打击能力（从而“相互确保破坏”），因为人们普遍认为，侵略者在初期就无法消除对手的临时工队。[29] 也可以采用其他确保二次打击能力的策略，但它们有缺点。例如，美国短暂使用的一种选择是在持续的空中警报中安装一队远程核轰炸机。[30] 该计划成本非常高，增加了意外或未经授权攻击的风险。另一种选择是建造硬化的陆基导弹发射井：足够数量，原则上可以保证一方的二次打击能力；然而，如此庞大的军火库将威胁到提供对另一方安全第一次打击的能力，从而再次破坏任何危机的稳定。道路移动的洲际弹道导弹发射器比基于筒

---

[25] Norris and Kristensen (2015).

[26] 鲍姆（2018年）提供了最新的核事故清单以及考虑使用核武器的场合。 Sagan（1995）对整个冷战期间的危险实践进行了更全面的描述。 Schlosser（2009）研究了近乎意外事故，特别关注一起导致Titan-II洲际弹道导弹无核爆炸的事件。

[27] 虽然很少有学术上的尝试来评估避免这种结果所涉及的运气程度，但最近的一项估计，从近似未命中数据集中得出，将美国和苏联避免核战争的可能性降至50%以下（Lundgren 2013）） 。 这符合一些具有核危机内幕知识的官员的观点，例如约翰·肯尼迪总统，他表示相信，事后看来，古巴导弹危机介于 ½ 到 ⅓ 几率导致核战争的机会。 尽管如此，一些著名的国际安全学者，如肯尼思华尔兹和约翰穆勒认为，核战争的可能性一直很低（Sagan and Waltz 2012; Mueller 2009）

[28] Gerson (2013)

[29] 或许错误地相信。 根据美国太平洋舰队前指挥官的说法，在冷战期间，反潜监视变得非常有效：“[美国]可以通过船体编号识别苏联潜艇的身份，因此我们可以进行人体计数 并确切地知道他们在哪里。 在港口或海上。 .....所以我觉得我们有能力在几乎任何情况下都能在非常短的时间内对苏联的 SSBN 部队采取相当严肃的措施。“（引自Ford 和 Rosenberg 2015）。

[30] Sagan (1995)

仓的导弹更难攻击，最终在苏联于 1985 年（冷战结束前几年）部署的时候提供了一些稳定性。[31]

因此，考虑一个反事实，其中先发制人的反击力更可行。想象一下可以轻松跟踪弹道导弹潜艇的一些技术。我们还可以想象，核武器更加脆弱，因此核武器被另一枚核武器引爆所摧毁的半径远大于实际上的大小。[32] 在这种情况下，它可能是不可能的。确保第二次打击能力。此外，假设该技术使得很难发现导弹发射，使得发射警告策略完全不可行。

冷战的危机不稳定将会大大放大。无论哪一方最初都会相对毫发无损地生存（或者至少可以相信它会这样，因为当时战争规划者很大程度上忽视了核冬天的可能性。）[33] [34] 不那么激进的一方将被完全摧毁。在这种情况下，相互的恐惧很容易引发全力以赴的战争。[35]

其他技术参数变化同样可以增加攻击的可能性。在现实世界中，核第一次打击的主要“吸引力”在于它可以减轻人们可能会成为这种罢工受害者的恐惧；但我们可以想象一个反事实，其中除了消除负面之外，还有利于核侵略。假设以某种方式可能从发起大规模核攻击中获得巨大的经济收益。[36] 可能很难看出情况如何，但人们可以想象一些自动化制造技术或能源技术使物质资源更有价值；或者技术支持的人口增长可以再次使农业土地成为一个更重要的资源。[37] 一些国际关系学者认为征服后的净经济效益在后工业时代大幅度下降，而这种下降是和平的主要推动因素。[38] 如果强有力的国家经济动机再次被加入战争的其他原因（例如关心自己的安全，非经济价值的争端，维护国家声誉，特别是好战的特殊利益集团的影响，除其他外），那么武装冲突可能会变得更加普遍，更有可能发生大规模的核战争。在这些例子中，弱势性不是来自破坏变得容易，而是来自导致破坏的行动得到更强有力的激励措施的支持。我们将称之为这些 2-型弱势性。具体来说，像“安全第一次打击”这样的场景，其中一些非常具有破坏性的行为被激励，我们将其称为 2a-型：

> 2a-型弱势性：有一定程度的技术使强大的行动者能够产生文明毁灭性的伤害，并且在半无政府状态的默认情况下，面临使用该能力的激励。我们将在下面看到更多关于 2a-型弱势性的例子，其中“文明毁灭性的危害”采取风险外部性的形式。

---

[31] Brower (1989).
[32] 事实上，遥感，数据处理，人工智能，无人机和核输送系统的进步现在有可能破坏核威慑，特别是对于拥有相对较小和不成熟的核武库的国家（Lieber and Press 2017）。
[33] Ellsberg (2017); Badash (2001).
[34] 当然不是没有受到伤害：放射性沉降会影响盟友，并在某种程度上影响祖国; 经济影响将对市场造成严重破坏，并引发全球经济萧条。 尽管如此，作为攻击的目标（特别是如果我们搁置核冬天）将更为可取。
[35] Schelling (1960).
[36] 另一种可能性是会产生政治上的好处，例如在表现出使用核武器的意愿后，对第三方进行核胁迫的能力增强。
[37] Drexler (1986).
[38] Brooks (1999); Gartzke (2007); Gartzke and Rohner (2010). For a dissenting view, see Liberman (1993)

2b-型弱势性（更糟糕的全球变暖） Type-2b (“worse global warming”)

还有另一种方式可以使世界变得脆弱；我们可以通过与气候变化相关的反事实来说明这一点。

在现实世界中，我们观察到全球平均温度的长期上升，人们普遍认为这主要是由人为排放的温室气体如二氧化碳，甲烷和氧化亚氮引起的。[39] 根据排放情景和建模，预测会有所不同。假设，但预测这意味着在没有任何减少排放的重大行动的情况下，2100 年（与 2000 年相比）平均气温上升 3 摄氏度至 4.5 摄氏度是非常典型的。[40] 通常预计海平面，天气模式，生态系统和农业的变暖将对人类福利产生净负面影响。[41] 温室气体通过广泛的活动排放，包括工业、运输、农业和电力生产，以及来自世界各地，特别是来自工业化国家或工业化国家。到目前为止，遏制排放的努力未能在全球范围内产生太大影响。[42]

现在我们可以想象一种情况，即全球变暖问题实际上似乎更加可怕。例如，瞬态气候敏感性（衡量由某种强迫导致的地球平均全球表面温度的中期变化，例如大气中 CO_2 加倍）可能已经是比它本身大很多了。[43]如果比它的实际值大几倍，那么我们应该会遇见温度上升，比如 15 ℃ 或 20 ℃ 而不是 3 ℃ ——这会是一个远远超过预期的具有更大的文明毁灭潜力的后果。[44]

我们还可以想象其他可能导致全球变暖问题的现实偏离。化石燃料可能比它们更加丰富，并且可以以更便宜的可开采矿床获得，这将鼓励更多的消费。与此同时，清洁能源替代品可能更昂贵，技术上更具挑战性。如果有更强的正反馈回路和非线性，全球变暖也可能是一个更糟糕的问题，例如大气层逐渐加载温室气体而没有太多可观察或有害影响的初始阶段，其次是温度的第二阶段突然猛烈地射击。为了从全球变暖中获得真正的文明威胁，可能还需要反事实地规定通过地球工程进行减缓是不可行的。

这种“更糟糕的全球变暖”情景所说明的弱势性不同于像“安全第一次打击”这样的 2a-型情景。在 2a-型弱势性中，某些行动者有能力采取某些行动 - 例如发动核首次攻击 - 这种破坏足以破坏文明。在“更糟糕的全球变暖”情景中，不存在这样的行动者。相反，在我们称之为 2b-型弱势性的情况下，有大量个别微不足道的参与者，每个人都被激励（在半无政府状态的默认条件下）采取某些行动，这些行动对累积成为毁灭性文明的行为略有贡献问题：

> 2b-型弱势性：有一定程度的技术，在半无政府状态的默认情况下，许多参与者都会采取激励措施采取一些有些破坏性的行动，这些行动的综合效果是文明的破坏。

---

[39] Stocker et al. (2014)
[40] See table 12.2, Stocker et al. (2014).
[41] See figure 10-1, Field et al. (2014).
[42] Friedlingstein et al. (2014).
[43] Shindell (2014).
[44] 如果地球的气候对二氧化碳敏感，那么人类文明可能永远不会出现，因为过去的二氧化碳水平（寒武纪时期的 4,000 ppm，而今天约为 410 pmm）可能会严重破坏复杂的进化。 生活。 较不遥远的反事实可能会涉及一些化合物，这种化合物在自然界中不会大量出现，而是由人类文明产生的，例如氯氟烃。氟氯化碳已经通过“蒙特利尔议定书”逐步淘汰，因为它们对臭氧层具有破坏性影响，但它们在每公斤基础上也是非常有效的温室气体。 因此，我们可以考虑一种反事实，其中氟氯化碳在工业上的用途远远超过它们的规模，但对全球气候产生了显着的延迟累积效应。

2a-型和 2b-型的共同之处在于，在这两种情况下，具有损害能力的行动者都会面临激励，鼓励在他们的情况下广泛的正常动机的行动者来处理导致损害的行动过程。如果只有一小部分可以驾驶汽车或砍掉一些树木的行动者选择这样做，全球变暖就不会成为问题。问题的出现只是因为许多行动者做出了这些选择。并且为了让*许多*行动者做出这些选择，必须通过具有广泛吸引力的激励（例如金钱，地位和便利性）来支持这些选择。同样，如果只有百万分之一的可以发动核第一次打击的行动者实际上会选择这样做，那么如果有少数行动者拥有这种能力，那就不会那么惊人了；但是，如果发起核打击得到强烈支持的激励措施强烈支持，那么它确实令人担忧，这些激励措施可以吸引正常动机的行动者（例如，一个人的对手先发制人的罢工动机）。这与 1-型弱势性形成对比，后者的问题源于破坏性能力的广泛扩散。只有具有相当不寻常价值的行动者才会以极大的代价和风险选择炸毁城市或释放世界末日的病原体；在这种情况下的麻烦是，如果有足够多的行动者拥有这样的能力，那么他们也具有世界末日动机的子集并非空洞。

0-型弱势性 Type-0 (“surprising strangelets”)

1942 年，曼哈顿科学家之一爱德华·泰勒（Edward Teller）发现，核爆炸会在地球历史上产生前所未有的温度，产生类似于太阳中心的温度，这可能会引发自我维持周围空气或水中的热核反应。[45] 洛杉矶奥拉莫斯实验室负责人罗伯特奥本海默立即认识到了泰勒关注的重要性。奥本海默通知了他的上级，并下令进一步计算调查可能性。这些计算表明不会发生大气点火。这种预测在 1945 年由 Trinity 测试确认，该测试涉及引爆世界上第一个核爆炸物。[46]

1954年，美国进行了另一次核试验，Castle Bravo试验，计划作为早期锂的秘密试验基于热核弹的设计。像铀一样，锂有两种重要的同位素：锂-6 和锂-7。在测试之前，核科学家计算出的产量为 6 兆吨（不确定性范围为 4-8 兆吨）。他们认为只有锂-6 会对反应有所贡献，但他们错了。锂-7 比锂-6 产生更多的能量，炸弹以15兆吨的产量引爆 —— 超过他们计算的两倍（相当于大约1,000个广岛）。意外强大的爆炸摧毁了大部分测试设备。放射性尘埃使顺风岛居民和日本渔船船员中毒，造成国际事件。

我们可以认为幸运的是，Castle Bravo 计算是不正确的，而不是计算三位一体测试是否会点燃大气。反事实上，如果大气已经被核爆炸点燃了，并且如果这个事实相对容易被忽视 - 让我们说这很容易忽视锂-7 在 Castle Bravo 测试中的贡献 - 那么人类的故事（以及所有地球生命的故事）将在 1945 年结束。我们可以将这种情景称为“Castle Bravissimo”。每当我们从发明的罐子中拉出球时，可能会发生意外破坏的可能性。通常，这种风险可以忽略不计，但在某些情况下，这种风险可能很大，尤其是当所涉及的技术产生某种新的自然扰动或引入历史上前所未有的条件时。这表明我们应该再添加一个类别，即技术命运的偶然文明破坏类型：

> 0-型弱势性：有一些技术带有隐藏的风险，以至于发现它时的默认结果是无意的文明破坏。[47]

---

[45] Rhodes (1986)

[46] 奥本海默委托的报告指出：“可以得出结论，本文的论点使得预期N + N反应可能传播是不合理的。 无限传播的可能性更小。 然而，论证的复杂性和缺乏令人满意的实验基础使得对该主题的进一步研究非常可取。“（Konopinski等人，1946）。

[47] 0-型可以被视为 1-型的限制性案例：它指的是一个弱势，该弱势需要零恶意行动者才能对结果进行文明破坏 —— 只有那些愿意继续使用技术的正常负责的行动者对新技术进行了普通的审查

然而，值得注意的是，“Castle Bravissimo”并不是 0-型弱势性的完美例证。假设经过仔细的计算表明，核爆炸有可能点燃大气和海洋，从而熄灭地球上的生命。进一步假设，已经知道要进一步解决问题并证明机会为零（或者机会是一个）将需要另外十年的细致研究。在这种情况下，目前还不清楚曼哈顿计划的领导人会做出什么决定。他们可能会认为人类最不希望继续发展核武器至少十年。[48] 另一方面，他们会担心德国可能会有一个先进的炸弹计划，希特勒可能不会拉断因为摧毁世界有 1%的风险。[49] 他们可能已经得出结论认为，测试核弹的风险是值得的，以减少纳粹德国结束核垄断的可能性。

在这个版本的“城堡Bravissimo”，文明被意外炸毁 —— 没有人试图造成破坏性的事件。然而，主要参与者被锁定在战略形势中，尽管存在风险，但仍鼓励他们继续前进。在这方面，该方案适合作为 2a-型弱势性；只是，它涉及的文明毁灭性的伤害是概率性的。当核技术成为可能时，强大的行动者在半无政府状态的默认条件下面临着以产生破坏文明危害的方式使用该技术的激励（这里采取风险外部性的形式）。[50]

因此，为了我们的诊断在 0-型弱势性中，我们要求满足更强的条件，而不仅仅是关键角色不打算破坏。我们规定“无意中”应该意味着不良结果源于运气不好，而非协调失败。在 0-型弱势性中，主要参与者即使得到充分协调，也会决定继续使用该技术，因为他们相信这些收益会超过成本 - 但这些都是错误的，并且成本会高于预期。[51]

因为“Castle Bravissimo”只是模糊地满足了这个标准（在原始的反事实中不清楚灾难会在多大程度上导致协调失败以及误判/运气不好的程度），引入 0-型弱势性的清洁示例可能很有用。因此，考虑一个“令人惊讶的奇怪的”情景，其中一些现代高能物理实验结果发起自我催化过程，其中普通物质转化为奇怪的物质，结果我们的星球被摧毁。这种情况及其变化，其中加速器实验产生稳定的黑洞或触发亚稳态真空状态的衰减，已在文献中进行了分析。[52] 这种结果确实非常令人惊讶，因为分析表明它们的机会完全可以忽略不计发生的。当然，如果运气不好，可能会发生可忽略不计的事件。但另外（在这种情况下更可能），分析可能有一个隐藏的缺陷，如 Castle Bravo 计算所做的；在这种情况下，机会可能并非如此微不足道。[53] [54]

---

[48] 如果有十年，为何不是永久

[49] 事实上，德国武装部长阿尔伯特·斯佩尔（Albert Speer）的一个说法表明，Werner Heisenberg 讨论了与希特勒发生连锁反应的可能性，这种可能性可能进一步挫伤了希特勒求得炸弹的热情（Rhodes 1986）。

[50] 这种 2a-型弱势性的真实版本，其中关键角色面临战略激励，采取行动，为文明创造不必要的风险，可能出现在发展机器超级智能的竞赛环境中。 在不利的情况下，竞争动态可能会让一个领先的开发商选择在安全之前启动他们自己的人工智能，或者将他们的领导放弃到愿意冒更大风险的其他开发人员（Armstrong 等人，2016）。

[51] 有关理性规划师如何在各种模型中平衡消费增长与安全的讨论，其中促进增长的创新也带来了引入减少寿命的创新的风险，请参阅 Jones（2016）。

[52] Jaffe et al. (2000); Tegmark and Bostrom (2005)

[53] Ord et al. (2010).

[54] 即使是“令人惊讶的奇怪”场景也可能会受到协调问题的困扰，尽管程度低于“Castle Bravo / Trinity测试”。 决定科学资助分配的人可能与提供资金的公众有不同的优先级。 例如，相对于保持低风险和为群众提供近期物质利益的价值而言，他们可能会更高地满足求知欲。 因此，在没有协调问题的情况下，主要代理问题可能会导致更多的粒子加速器资金。 国家之间的声望竞争 —— 可能部分被视为另一种协调失败 - 也可能是一般基础科学基金和特别是高能物理学的驱动力。

达成稳定 Achieving stabilization

VWH 的真相将是坏消息。但这并不意味着文明将遭到破坏。至少在原则上，即使存在弱势性，也有几种可能使世界稳定的潜在反应。回想一下，我们用黑球技术来定义假设，使得文明破坏极有可能取决于技术发展的持续性和半无政府状态的违约条件。因此，我们理论上可以考虑以下实现稳定的可能性：

1. 限制技术发展。
2. 确保不存在大量行动者，这些行动者代表了广泛和可识别的人类动机分布。
3. 建立极其有效的预防性警务。
4. 建立有效的全球治理。

我们将在后续章节中讨论（3）和（4）。这里我们考虑（1）和（2）。我们将争辩说，他们只有有限的承诺，作为防范潜在文明弱势性的方法。

技术放弃 Technological relinquishment

在一般形式中，技术放弃看起来非常没有希望。回想一下，我们广泛地解释了“技术”这个词；因此，完全阻止技术发展需要在世界各地接近停止发明活动的事情。这很难实现；如果能够做到这一点，那将是极其昂贵的 —— 以自身的方式构成存在主义的灾难。[55]

然而，普遍放弃科学和技术研究是一个不起作用的并不意味着有限的削减发明活动不是一项明智的政策。放弃特别危险的进步方向是有道理的。例如，回顾我们的“简单核武器”情景，劝阻研究用于铀浓缩的激光同位素分离是明智的。[56] 任何能够用更少的能源或更小的工业足迹生产武器级裂变材料的技术都将是合理的。侵蚀核扩散的重要障碍。很难看出核能价格略有下降会如何弥补。相反，如果以某种方式使浓缩铀变得更加困难和昂贵，世界可能会变得更好。我们理想在这个领域想要的不是进步而是技术倒退。

虽然有针对性的回归可能不会出现，但我们的目标可能是降低相对于保护技术进步速度的风险增加技术的进步速度。这是差别技术发展原则所表达的观点。在其最初的表述中，该原则侧重于存在风险；但我们可以更广泛地应用它，也包括具有“仅仅”破坏潜力的技术：

> *差别技术发展原则*：阻碍危险和有害技术的发展，特别是提高存在风险水平的技术；并加速有益技术的发展，特别是那些减少自然或其他技术带来的存在风险的技术。[57]

差别技术发展的原则与合理形式的技术决定论相一致。例如，即使规定所有可开发的技术都将被开发出来，它们在开发时仍然很重要。它们到达的顺序可能会产生重大影响 —— 理想情况下，保护性技术应该先于它们所保护的破坏性技术；或者，如果这是不可能的，那么最好将间隙最小化，以便其他对策（或运气）可以让我们渡过难关，直到可以获得更强大的保护。发明

[55] Namely, “permanent stagnation” (Bostrom 2013)
[56] Kemp (2012)
[57] Bostrom (2002).

的时机也影响了技术诞生的社会政治背景。例如，如果我们认为文明变得更有能力处理黑球的长期趋势，那么我们可能想要推迟最危险的技术发展，或者至少放弃加速它们。即使我们认为文明的破坏是不可避免的，但许多人宁愿在未来进一步发生，也许他们和他们的亲人无论如何都不再活着。[58]

差别技术发展并没有真正意义上的原始的创意的罐模型，每个球的颜色都是一个完全的惊喜。如果我们想在这个上下文中使用罐子隐喻，我们必须修改它。例如，我们可以假设球具有不同的纹理，并且纹理和颜色之间存在相关性，因此我们可以在从球中提取球之前获得关于球颜色的线索。使隐喻更逼真的另一种方法是想象一些球之间有绳子或松紧带，这样当我们拉上其中一个球时，我们就会拖拽其他几个球。据推测，罐模型是高度管状的，因为某些技术必须先出现才能达到其他技术（我们不太可能找到一个使用喷气式飞机和燧石轴的社会）。如果我们想象不仅只有一只手能够巧妙地探索罐子，那么这个比喻也会变得更加真实：相反，想象一群摇摇欲坠的探矿者为了金子，荣耀和引用而在他们怀里抱怨。

显而易见的是，正确实施不同的技术发展是一项艰巨的战略任务。[59] 然而，对于一个关心长期结果并且参与某些创造性企业的行动者（例如，作为研究人员，资助者，企业家，监管者，或立法者）值得做出尝试。无论如何，一些影响似乎相当明显 - 例如，不适用于激光同位素分离，不适用于生物武器，也不开发地形工程形式，使随机个体能够单方面对其进行大幅度的改造。地球的气候。在加速使能技术（例如DNA合成机器）之前要三思而后行 - 这将直接促进这种不祥的发展。[60] 但是提升技术主要是保护性的；例如，那些可以更有效地监测疾病爆发，或者更容易检测隐蔽的大规模杀伤性武器计划。

即使情况是所有可能的“坏”技术无论如何都必然会被开发出来，但购买一点时间仍然是有帮助的。[61] 然而，差别技术发展本身并没有为弱势性提供解决方案。持续很长时间 —— 那些充分保护性技术比破坏性对应物更难开发，或者即使在技术成熟时破坏也具有优势的地方。[62]

偏好修改 Preference modification

实现文明稳定的另一个理论上可行的方法是改变这样一个事实，即存在大量的行动者，这些行动者代表着广泛和可识别的人类动机分布。我们保留以后讨论干预措施，通过增加各种形式的

---

[58] 在灾难发生时活着的人们可能更喜欢它在它们出生之前就已经发生过，所以它们都将完全结束并且不会受到影响。 他们的偏好似乎遇到了一个非同一性的问题，因为如果在他们被怀孕之前发生了文明的破坏事件，他们几乎肯定不会存在（Parfit 1987）。

[59] Cf. Collingridge (1980).

[60] 更广泛地说，生物技术工具和技术的许多改进，使业余DIY生物黑客更容易完成以前只能由资源充足的专业研究实验室完成的工作，从这个角度来看是受到怀疑的。 令人质疑的是，DIY生物黑客（黑暗中发光植物？）的好处是否值得将生物工程转化为潜在风险或恶意目的的能力扩展到一系列相对无法解释的行动者。

[61] 反驳说“如果我不发展它，其他人会; 所以我不妨这样做“往往会忽略这样一个事实，即某个科学家或开发人员对新发现的预期时间至少有一些边际影响。 如果事实上科学家的努力对于发现或发明的时间没有任何影响，那么这些努力似乎是浪费时间和资源，并且应该因此而停止。 当某些技术能力可用时（例如，一个月），相对较小的转变在某些情况下可能很重要（例如，如果危险技术在开发和部署有效防御之前每月施加重大风险）

[62] “技术成熟度”是指实现能够提供一定程度的经济生产力和控制自然的能力，接近可实现的最大值（在充足的时间内）（Bostrom 2013）。

协调来减少独立行动者的有效数量。在这里，我们考虑修改偏好分布的可能性（在或多或少不变的行动者群体中）。

这种方法承诺的程度取决于我们所考虑的弱势性类型。

在 1-型弱势性的情况下，偏好修改看起来并不乐观，至少在没有非常有效的方法的情况下。如果世界上任何一个地方都有一个有动力去追求破坏性结果的人，那么考虑到一些第一类弱势性会导致文明的破坏。由于存在这种弱势性，减少世界末日残余中的人数无助于预防破坏，除非数量可以一直减少到零，这可能是完全不可行的。确实存在其他可能的 1-型弱势性，这些弱势性需要更大的世界末日残余才能发生文明破坏：例如，在“简单的核武器”这样的场景中，可能必须有来自世界末日的人在几百个城市中的每一个都有残余。但这仍然是一个很低的标准。很难想象干预 —— 在全球范围内彻底重新设计人性 —— 这足以耗尽世界末日的残余，以完全消除甚至大大降低1型弱势性的威胁。

请注意，将世界末日残差的大小减半的干预不会（至少不通过任何一阶效应）将 1-型弱势性的预期风险降低到接近任何数量。将世界末日残留量减半，将1型风险降低5%或10%将更为合理。原因在于对一些新的黑球技术的破坏性存在很大的不确定性，而且我们应该在对数空间（超过几个数量级）中具有相当均匀的先验，超过了世界末日残差的大小。文明破坏的顺序发生在一个 1-型弱势性的发生条件下。换句话说，根据正在开发的一些新技术，使普通人容易杀死至少一百万人，可能（大致）该技术可能使普通人杀死一百万人，十分百万人，一亿人，十亿人，或每个人活着。

尽管有这些考虑因素，但优先修改可能有助于在一系列授权参与者最初限于某些小的可定义子群体的情况下。一些黑球技术首次从罐子中出现时，可能难以使用并需要专门的设备。可能需要几年的时间才能使这种技术完善到一般人可以掌握它的程度。在这个早期阶段，一组有权力的行动者可能非常有限；例如，它可能仅由具有生物科学专业知识的个人在特定类型的实验室中工作。对这些实验室职位的申请人进行更密切的筛选，可以对破坏性个体在其出现的头几年内获得生物技术黑球的风险产生有意义的影响。[63]这种缓解可能提供了引入其他对策的机会提供更持久的稳定性，预期技术变得足够容易使用，并扩散到更广泛的人群。

对于 2a-型弱势性，授权的角色集合要小得多。通常我们在这里处理的是国家，也许还有一些特别强大的非国家行动者。在某些 2a-型情景中，该集合可能仅由两个超级大国或少数具有特殊能力的国家组成（例如，目前核武器就是这种情况）。因此，即使是一些强大的国家的偏好也在更加爱好和平的方向上转移，也可能非常有用。如果面临安全困境的行动者对彼此的态度与芬兰和瑞典之间的态度相似，那么“安全第一次打击”的情景将不那么令人担忧。对于由于一些技术突破而可能对强大的行动者产生的许多合理的激励措施，如果有关行动者有更多的寂寞倾向，那么非破坏性结果的前景将会大大增强。尽管这似乎难以实现，但似乎并不像说服世界末日残余中的几乎所有成员改变他们的性格一样困难。

---

[63] 自2001年“Amerithrax”事件以来，生物科学中的访问控制变得越来越重要。 在美国，处理危险病原体的机构有义务评估有权访问的员工的适用性，这些员工也会受到联邦机构的审查（2017年联邦选择代理计划），并建议向开发其生物安全基础设施的国家采用类似方法（ 生物安全和Biopreparedness 2017）。 现有制度存在两个缺点：第一，没有全球协调，所以不良行动者可以“逛逛”，以适应较为宽松的监管环境; 第二，重点仍然是获取生物材料（例如某些微生物的样本），而生物信息和技术越来越成为主要安全问题的对象（Lewis等人即将出版）。

最后，考虑 2b-型弱势性。回想一下，这样的弱势性需要“默认情况下”，许多行动者都会采取激励措施来采取一些破坏性的行动，这样的综合效应会加剧文明的破坏。因此，使用黑球技术的动机必须是掌握世界人口很大一部分的因素 - 经济收益可能是这种近乎普遍动机的主要例子。因此，想象一下几乎每个人都能获得的一些私人行动，它可以拯救每个人的年收入的一小部分 X，同时产生负外部性，这样如果世界上一半的人口采取行动，那么文明就会受到破坏。在 X = 0 时，我们可以假设很少有人会采取反社会行动。但是越大的 X 越大，那些屈服于诱惑的人口比例就越大。不幸的是，有可能导致至少一半人口采取行动的 X 值很小，可能不到 1%。[64] 虽然希望改变全球偏好的分布以便使人们更多的利他主义和提高 X 的价值，这似乎很难实现。（考虑一下已经在竞争心灵和思想的许多强大力量 —— 企业广告商，宗教组织，社会运动，教育系统等等。）即使世界上利他主义的数量急剧增加 —— 相应的，我们说， X 从 1% 增加到 2% —— 只能在相对狭窄的情景中防止灾难，即使用破坏性技术的私人利益在 1-2% 的范围内。私人收益超过 2% 的情景仍然会导致文明的破坏。

总而言之，修改由黑球发现破坏性赋予的一组行动者中的偏好分布可能是其他稳定手段的有用辅助，但它可能难以实施，并且最多只能提供非常局部的保护（除非我们假设世界范围内极端形式的人性重新设计）。[65]

一些具体的对策及其局限性 Some specific countermeasures and their limitations

除了影响科学技术进步的方向，或改变与销毁有关的偏好外，还有各种其他可能的对策可以减轻文明的弱势性。 例如，可以尝试：

- 防止危险信息传播
- 限制访问必要的材料，工具和基础设施
- 通过增加他们被抓住的机会来阻止潜在的恶人
- 要更加谨慎，做更多的风险评估工作
- 建立某种形式的监督和执行机制
- 有可能阻止企图进行破坏性行为

从我们之前的讨论和例子中可以清楚地看出，前四个不是一般的解决方案。防止信息传播很容易变得不可行。即使可以做到，也不会阻止危险信息被独立重新发现。审查制度似乎充其量只是权宜之计。[66] 限制对材料，工具和基础设施的使用是减轻某些（灰球）威胁的好方法，但它

---

[64] X值大大低于 1%似乎与大多数人给予全球慈善机构的数量一致。 然而，行为遗漏的区分可能会使人们愿意接受更大的个人牺牲，以便不为全球利益做出贡献，而不是为了贡献全球利益。

[65] 然而，请注意，偏好分布的正向变化 - 即使不足以通过简单地使某些个体行动者不选择破坏性选择来避免灾难 可能会产生重要的间接影响。 例如，如果许多人变得更加仁慈倾向，这可能会将社会转变为更加合作的均衡，支持更强大的基于治理的稳定方法，例如我们在下面讨论的方法（参见“道德提升”; Persson 和 Savulescu 2012）

[66] 在全球范围内，我们发现了一系列国家分类方案和信息控制系统。它们通常旨在保护军事和情报机密，或防止关于政权内部人员的尴尬事实暴露给公众，而不是规范科学或技术见解的传播。有一些例外，特别是在直接涉及国家安全的技术信息的情况下。例如，1951年美国的“发明保密法”赋予国防机构禁止授予专利的权力，并命令发明保密;虽然不寻求专利保护的发明家不受这些限制（Parker和Jacobs 2003）。核发明受1946年“原子能法”的“天生秘密”条款的约束，该条款宣布有关核武器的设计，开发和制造的所有信息 - 无论其来源如何，除非已经正式解密（同上）。 。其他法律工具，如出口管制，也被用于阻止科学信息的流动。多个美国政府机构阻止出版和使用20世纪70年代和80年代开发的强加密协议的

对其他类型的威胁无效 - 例如在经济中的许多地方需要必要的成分，或者在发现危险的想法时已经普遍存在的那些（例如“易核武器”情景中的玻璃，金属和电池）。遏制潜在的恶人是有道理的；但是对于充分破坏性的技术，即使每个犯罪者都肯定被抓住，世界末日残余的存在也会使威慑力不足。

更谨慎行事并进行更多风险评估也是一种薄弱而有限的策略。单方面决定更加谨慎的一个行动者对于2a-型弱势性可能没有多大帮助，对于Type-2b或Type-1中的任何一个弱势性基本上没有任何帮助。在0型-弱势性的情况下，如果关键角色更加谨慎，它可能会有所帮助 - 尽管只有当第一个谨慎的脚尖行动者没有跟随不谨慎的行动者获得相同的风险技术时（尽管世界已有）不知何故，在过渡期间，通过其他手段稳定了。）[67]至于风险评估，它只有在导致其他一些对策得以实施的情况下才能降低风险。[68]

列表监视中的最后一个对策确实指向更一般的解决方案。我们将在下一节“预防性警务”标题下讨论它。但我们已经注意到它本身并不足够。例如，考虑 2b-型弱势性，例如“更糟糕的全球变暖”。即使监督使一个国家能够完全执行其选择施加的任何环境法规，仍然存在使得足够多的州同意采用必要的法规的问题 - 这些法规很容易就不会发生。在 2a-型弱势性的情况下，监视的局限性更为明显，例如“安全第一次打击”，其中问题在于国家（或其他强有力的参与者）被强烈激励以执行破坏性行为。这些国家完全控制在自己境内发生的事情的能力并不能解决这个问题。可靠地解决涉及国际协调挑战的问题所需要的是有效的全球治理。

（不成功的）努力提供了一个值得注意的例子（Banisar 1999）。科学界在非常罕见的情况下进行了自愿的自我审查。

Leo Szilard在说服他的物理学家同事不要发表关于核裂变的方面（在曼哈顿计划开始之前和官方保密的开始之前）取得了一些部分成功，尽管他遇到了一些希望他们自己的工作出现的科学家的反对在期刊上或者认为开放是科学的神圣价值。
最近，有一些关于禽流感研究的科学自我审查尝试（Gronvall 2013）。在这种情况下，这些努力可能不仅是无效的，而且适得其反，因为公开辩论引发的争论是否应该公布某些结果会引起人们对这些结果的更多关注，而不是如果出版物没有得到反对就会得到，所以称为“Streisand效应”。

总的来说，科学自我审查的尝试似乎是相当半心半意和无效的。（我说出现了，因为在公开发布的事件中，事情是如何展开审判的。但真正成功的压制科学信息的尝试不一定会出现在公共记录中。）即使一些期刊编辑也同意关于如何处理造成信息危害的论文的标准，没有什么能阻止沮丧的作者将她的稿件发送到另一个标准较低的期刊或者将其发布在她的个人网页上。大多数科学界既没有文化，也没有激励措施，也没有安全和风险评估方面的专业知识，也没有有效处理信息危害所需的机构执法机制。科学精神就是这样：每一个球必须尽快从罐子中提取出来并立即向世界上的每个人展示;发生的越多，取得的进展就越大;你对此做出的贡献越多，你就越有科学家。黑球的可能性不会进入等式。

[67] 无论如何，目前还不清楚我们是否真的希望在整体上更加谨慎。可能需要（从各种评估角度）鼓励在可能存在极端全球缺点的情况下更加谨慎。然而，如果规范性过度冒险的原因是协调问题，那么在这些原因中行使自愿谨慎和克制的劝告可能不会非常有效：风险承担者在产生全球风险的同时获得一些私人利益（例如利润或声望）外部性。因此，在这种情况下，解决方案可能需要加强全球治理能力。

[68] 风险评估工作也可能通过产生信息危害来提高风险水平（Bostrom 2011）。

治理差距 Governance gaps

现在应该明确技术放弃，偏好修改和各种具体对策作为对潜在文明弱势性的反应的局限性。因此，在某种程度上，我们担心VWH可能是真的，我们必须考虑实现稳定的其余两种可能方式：
（a）建立极其有效的预防性警务能力。发展国家内部治理能力，以极高的可靠性，防止任何个人或小团体 - 包括无法阻止的人 - 不采取任何高度的行动非法；和
（b）建立强有力的全球治理能力。发展可靠地解决最严重的全球公共问题所需的国家间治理能力，并确保在重要安全利益受到威胁的国家（和其他强有力组织）之间进行强有力的合作 - 即使有非常强烈的动机要求协议或拒绝首先登录。

（a）和（b）反映的两个治理差距，一个是微观尺度，另一个是宏观尺度，是当代世界秩序的两个跟腱。只要他们保持不受保护，文明仍然容易受到潜在的技术黑球的影响，这种黑球可以使罢工直接进入那里。除非并且直到从罐子中出现这样的发现，否则很容易忽视我们的暴露程度。

在以下两节中，我们将讨论如何填补这些治理差距，以实现稳定潜在文明弱势性的一般能力。不言而喻，在寻求（a）和（b）方面取得进展时，存在很大困难，也存在非常严重的潜在缺点。在本文中，我们将很少谈论困难而几乎没有关于潜在缺点的部分 - 部分原因是这些已经相当广为人知并得到广泛认可。但是，我们强调，缺乏对（a）和（b）的论点的讨论不应被解释为这些论点薄弱或不指出重要问题的隐含断言。当然，在考虑全因素的评估中，必须考虑到它们。但是这样的评估超出了本贡献的范围，该贡献特别关注来自VWH的考虑因素。

预防性警务 Preventive policing

假设打开了Type-1弱势性。有人发现了导致大规模杀伤的一种非常简单的方法。有关发现的信息传播。必需的材料和仪器无处不在，不能迅速从循环中移除。当然，任何非国家行动者摧毁一个城市都是非常违法的，任何被捕的人都会受到严厉的处罚。但似乎有理由认为，一百万以上的人不属于不可摧毁的世界末日残余。虽然相对来说很小，如果每个这样的人创造了一个破坏城市的事件，那么绝对数量仍然太大，文明无法忍受。那么该怎么办？

如果我们突然发现自己处于这种状况，那么防止文明遭到破坏可能为时已晚。然而，有可能设想人类社会能够在如此挑战中幸存下来的情景 - 以及个人能够单独摧毁一个城市而不是整个世界的更加艰难的挑战。

稳定这些弱势性所需要的是一种极为发达的预防性警务能力。各国需要有能力密切监测其公民，使他们能够拦截任何开始准备大规模毁灭行为的人。

这种监视和拦截的可行性取决于场景的具体情况：破坏性地部署黑球技术需要多长时间？涉及的行动有多明显？他们可以区别于我们不想禁止的行为吗？但有理由认为，在最大程度上部署当前可用技术的国家可以稳定相当大的1类弱势性频谱。监测技术的预期进步将极大地扩展可实现的保护。

有关真正密集的监控级别的图片，请考虑以下小插曲：

> *高科技圆形监狱 Panopticon*
> 每个人都配备了一个“自由标签” - 与当今熟悉的更为有限的可穿戴监控设备相关，例如在一些国家用作监狱替代品的脚踝标签，许多警察部队所穿的防弹衣，口袋跟踪器和腕带一些家长用来跟踪他们的孩子，当然还有无处不在的手机（它被称为“个人跟踪设备，也可以用来打电话”）。[69]69自由标签稍微高级一些电器，戴在脖子上，并配有多向摄像头和麦克风。加密的视频和音频不断从设备上传到云端，并实时进行机器解码。 AI算法对佩戴者的活动，他的手部动作，附近的物体和其他情境线索进行分类。如果检测到可疑活动，则将馈送中继到几个爱国者监测站之一。这些是大型办公楼，每周7天，每天24小时为您服务。在那里，自由官员在几个屏幕上查看视频输入并听取耳机中的音频。然后，自由官员确定适当的动作，例如通过音频链接联系标签佩戴者以请求解释或请求更好的视图。自由官员还可以派遣检查员，警察快速反应部门或无人机进一步调查。在一小部分情况下，佩戴者在反复警告后拒绝停止被禁止的活动，可以进行逮捕或施加其他适当的处罚。公民不得移除自由标签，除非他们所处的环境中配备了足够的外部传感器（但不包括大多数室内空间和机动车辆）。该系统提供了相当复杂的隐私保护，例如亲密身体部位的自动模糊，并且它提供了编辑身份暴露数据（如面部和名称标签）的选项，并且仅在调查需要信息时才释放它。支持人工智能的机制和人类监督都密切监视自由官员的所有行动，以防止滥用。[70]

创建和运营高科技 圆形监狱 需要大量投资，但由于相机价格的下降，数据传输，存储和计算，以及基于人工智能的内容分析的快速进步，它可能很快变得技术上可行且价格合理。 例如，如果将一年内应用于一个人的成本降至 140 美元左右，那么可以持续监测整个世界人口，其成本不到世界 GDP 的 1%。 按照这个价格，该系统可以合理地代表净储蓄 —— 甚至将其用于防止文明规模的灾难 - 因为它可用于定期执法。 如果系统按照宣传的方式运作，几乎可以消除许多形式的犯罪，同时减少警务，法院，监狱和其他安全系统的成本。 它也可能在许多有益的文化习俗中产生增长，这些习俗目前受到缺乏社会信任的阻碍。

如果高科技 圆形监狱 的技术障碍正在迅速降低，那么其政治可行性如何呢？一种可能性是，即使没有对系统造成任何重大冲击，社会也会逐渐走向完全的社会透明度。收集和分析有关人员和物体的信息可能会变得越来越容易，并且可以证明完成这一过程非常方便，最终接近完全监视的某些事物变得现实 - 足够接近只有一个转向螺旋它可以变成高科技 圆形监狱。[71] 另一种可能性是，某些特定的 1-型弱势性足以引人注目，以吓唬各州采取极端措施，例如启动崩溃计划以创建通用监视。在没有完全普遍监测系统的情况下可以尝试的其他极端措施可能包括采取先发制人的监禁政策，比如每当一些不可靠的指标表明一些人试图破坏城市或更糟的可能性的概率 > 1%时。[72] 对这些政策的政治态度取决于许多因素，包括有关隐私和社会控制的文

---

[69] 奥威尔的名字当然是故意的，以提醒我们应用这种系统的各种方式。
[70] 实施细节仅供参考。 例如，类似的功能可以由混合现实眼镜而不是项链提供。 可以设计该设备的版本，其将为用户及其监视功能提供许多益处。 理论上，一些监测可能是众包的：当AI检测到可疑活动时，视频源被匿名化并发送给随机的100名公民，其职责是观察该源并投票决定是否需要进一步调查; 如果至少有10%的人认为这样做，那么（非匿名的）源会被转发给当局。
[71] 可以合理推动更具侵入性监控的“便利”的例子包括各种消费者应用和经济上有用或有利可图的监控（例如，用于广告定位，价格歧视等）; 能够防止引起公愤的各种事情，例如虐待儿童或小规模恐怖主义; 而且，特别是对威权政权来说，有能力压制政治反对派。
[72] 较温和的政策版本可能仅仅是禁止这些弱势嫌疑人获取产生破坏性影响所必需的设备和材料。 这可能足以取决于场景的细节。

化传统和规范；但它们还取决于文明弱势性的清晰程度。至少在有几个壮观的警告镜头的弱势性的情况下，可以非常清楚地看到风险。例如，在“轻松的核武器”情景中，在几个大城市被毁之后，公众可能会对一项政策提供强有力的支持，为了防止另一次袭击，这项政策将涉及为每一个聪明的绘图仪监禁一百名无辜的人。[73] 在这种情况下，高科技 圆形监狱 的创建可能会被广泛支持作为一个压倒性的紧急优先事项。但是，对于没有此类无可辩驳的证据之前或附带的弱势性，强有力的预防措施的意愿可能永远不会实现。

因此，通过无处不在的实时监视实现极其有效的预防性监管可能是稳定 1-型弱势性的必要条件。监控也与其他一些类型的弱势性相关，尽管不像 1-型弱势性那样集中。

在一个 2b-型弱势性中，不良结果是由大量独立行动者的共同行动引起的，这些行动者被激励去破坏性行为。但除非很难观察到破坏性行为，否则不需要加强监视或预防性警务来实现稳定。例如，在“更糟糕的全球变暖”中，个人行为被抢占并不是必要的。排放的危险程度需要一段时间才能积累，污染者可以追究责任；如果他们中的一些人穿过裂缝是可以忍受的。

然而，对于其他 2b-型弱势性，增强的监视和社会控制方法可能很重要。考虑一下“失控的暴徒”，一种暴徒形成杀死任何与之接触的人拒绝加入，并且变得越来越大，越来越强大。[74] 这种糟糕的社会平衡可以形成和传播的可能性，可行性一旦他们掌握了改革他们，他们确实对人类福利造成的损失，取决于可以通过技术创新改变的参数，可能更糟。即使在今天，许多国家仍在努力制服有组织犯罪。黑球发明 —— 也许是一些聪明的加密经济机制设计 —— 使得犯罪企业的社会影响更具可扩展性或更具破坏性，这可能会造成一种只有在国家拥有前所未有的监督和社会控制技术力量时才能稳定的弱势性。

至于 2a-型弱势性，问题源于国家权力或其他强大行动者所面临的激励，国内监督如何能够提供帮助尚不清楚。历史上，更强有力的社会控制手段甚至可能使国家间冲突恶化 —— 最血腥的国家间冲突依赖于现代国家的高效治理能力，例如税收、征兵和战争宣传。可以想象，改进监视可以间接地促进稳定 2a-型弱势性，例如通过改变社会文化动态或创造新的选择来使减少武器条约或非侵略条约更加可核查。但似乎同样合理的是，在没有可靠的解决国际争端的机制的情况下，加强国内监督和警务权力对 2a-型弱势性的净效应会在相反的方向上（即倾向于产生或加剧这种弱势性而不是稳定他们）。

全球治理 Global governance

再次考虑“安全第一次打击”：默认情况下可以使用黑球技术的国家面临着破坏性地使用它的强烈动机，即使对于没有国家这样做的每个人来说都会更好。最初的例子涉及核武器的反事实，但展望未来，我们可能会从生物武器的进步，或原子精确制造，或大量杀手无人机或人工智能的创造中获得这种黑球，或别的。然后，这组国家行动者面临集体行动问题。未能解决这一问题意味着文明在核世界末日或其他类似的灾难中遭到破坏。有理由认为，如果没有有效的全球治理，各州实际上无法解决这个问题。假设，我们这里面临的问题带来了特殊的挑战；然而，各州经常无法解决更容易的集体行动问题。人类的历史与战争的麻子一起被人们所覆盖。

[73] 高科技圆形监狱的部分实施可能会取代这种情况下的监禁，在这种情况下，只有那些长期列入“高度关注的人”的人才需要佩戴自由标签。
[74] Cf. Munz et al. (2009).

然而，通过有效的全球治理，解决方案变得微不足道：只是禁止所有国家破坏性地使用黑球技术。在“安全第一次打击”的情况下，最明显的做法是通过命令拆除所有核武器并建立检查制度，无论需要什么级别的侵入来保证没有人重建核能力。或者，全球治理机构本身可以保留核武库，作为对任何突破企图的缓冲。

为了应对 2a-型弱势性，文明需要的是实现全球协调的强大能力，特别是在国家行为具有极大外部性的问题上。有效的全球治理也将有助于那些第一类和第二类方案，其中一些国家不愿意采取必要的预防性警务措施，以便可靠地防止其境内的个人实施破坏性行为。

考虑一个足够强大的生物技术黑球，单个恶意使用可能会导致数百万人死于大流行，从而造成 1-型弱势性。如果即使是一个国家也未能建立持续监视和控制其公民所必需的机制（或任何其他必要的机制来防止恶意使用并具有几乎完美的可靠性），这将是不可接受的。一个拒绝执行必要保障措施的国家 —— 可能是因为它过分重视个人自由或者赋予公民宪法规定的隐私权 —— 将成为国际社会的违法成员。这样一个国家，即使其治理机构在其他方面表现令人钦佩，也类似于“失败的国家”，其内部缺乏控制使其成为海盗和国际恐怖分子的避风港（当然，在目前的情况下，风险很大）它会对世界其他地区施加的外部性将会更大。）其他州肯定会抱怨投诉。

类似的论点也适用于 2b-型弱势性，例如“更糟糕的全球变暖”情景，其中一些州倾向于搭乘其他人为减少排放而付出的代价高昂的努力。一个有效的全球治理机构可以迫使每个国家尽自己的力量。

因此，我们看到虽然一些可能的弱势性可以通过单独的预防性监管来稳定，而其他一些弱势性可以通过全球治理来稳定，但有一些弱势性需要两者兼顾。将需要极为有效的预防性警务，因为个人可以从事难以管制的活动，但必须对其进行有效监管，并且需要强有力的全球治理，因为各州可能有动力不对这些活动进行有效监管，即使他们拥有能力这样做。然而，综合起来，无处不在的监督驱动的预防性监管和有效的全球治理将足以稳定大多数弱势性，即使 VWH 属实，也可以安全地继续进行科学和技术发展。

讨论 Discussion

因此，全面监督和全球治理可以防范各种潜在的文明弱势性。这是支持实现这些条件的重要原因。这个原因的强度大致与脆弱世界假设为真的概率成正比。

不言而喻，一种能够实现前所未有的强烈监督形式的机制，或能够将意志强加于任何国家的全球治理机构，也可能产生不良后果。改善社会控制能力可以帮助专制政权保护自己免受反叛。无处不在的监视可以使霸权意识形态或不宽容的多数观点强加于生活的各个方面，防止生活方式不正常或不受欢迎的信仰的个人不愿透露姓名。如果人们相信他们所说和所做的一切，实际上是“记录在案”，他们可能会变得更加谨慎和平淡无奇，更贴近政治上正确的态度和行为的标准脚本，而不是胆敢说或做任何挑衅这可能会使他们成为愤怒暴民的目标，或者在他们的简历上留下不可磨灭的不合格标记。就全球治理而言，它可以减少有利形式的国家间竞争和多样性，创造一个单点失败的世界秩序：如果一个世界政府被一个充分有害的意识形态或特殊利益集团所俘获，它可能是游戏为了人类的政治进步，因为现任政权可能永远不会允许尝试和替代方案，以揭示有更好的方法。而且，与典型的州政府相比，甚至进一步脱离了个人和文化上有凝

聚力的“人民”，这样的机构可能被一些人视为不太合法，而且可能更容易受到官僚主义硬化或政治偏离等机构问题的影响。它可能在名义上受到的公共利益。

毫无疑问，除了稳定文明弱势性之外，更强有力的监督和全球治理可能会产生各种好的后果。[75] 更有效的社会控制方法可以减少犯罪并减轻对严厉刑事处罚的需求。它们可以营造一种信任气氛，使有益的新形式的社会互动和经济活动蓬勃发展。全球治理可以防止州际战争，包括不威胁文明破坏的战争，减少军事开支，促进贸易，解决各种全球环境和其他公地问题，平息民族主义仇恨和恐惧，随着时间的推移或许会增强国际大都会的意识团结。它也可能导致全球穷人的社会转移增加，有些人认为这是可取的。

很明显，在这些方向上都有支持和反对的重要论据。本文没有对这些论点的总体平衡做出判断。这里的野心更加有限：提供一个思考潜在的技术驱动的文明弱势性的框架，并指出，在一系列情景中稳定文明需要大大扩展预防性警务和全球治理的能力。是的，这种分析提供了另一个有利于发展这些能力的理由，这一理由似乎在最近关于相关问题的许多谈话中都没有发挥重要作用，例如关于政府监督的争论以及关于国际和超国家的拟议改革。[76] 当这一原因加入到组合中时，评估应该变得比在其他方面更加有利于以这些方式加强治理能力的政策。但是，这个增加的理由是否足以使总体平衡倾斜，将取决于本文范围之外的其他考虑因素。

值得强调的是，本文中的论点倾向于某些特定形式的治理能力强化。关于监督和预防性警务，VWH 关注点特别指出治理能力的可取性，这使得极其可靠地抑制非常大的绝大多数人口（以及权力加权的国内利益相关者）非常不赞同的活动成为可能）。只有在它们有助于创造这种特定能力的情况下，它才能为其他形式的治理加强提供支持。同样，在全球治理方面，基于 VWH 的论证支持能够可靠地解决非常高风险的国际协调问题的发展中国家的制度，而未能达成解决方案的问题将导致文明的破坏。这将包括有能力防止大国冲突、压制大规模杀伤性武器中的军备竞赛、规范发展竞赛和部署潜在的黑球技术，并成功地管理最严重的公地悲剧。它不需要包括使各国在许多其他问题上合作的能力，也不一定包括仅使用完全合法手段实现必要稳定的能力。虽然这些能力可能因其他原因而具有吸引力，但它们并不会因为严肃对待 VWH 而立即成为迫切需要。例如，就 VWH 而言，理论上如果必要的全球治理能力通过一个超级大国的崛起到足够的主导地位而在一个足够严峻的紧急情况下单方面地赋予其能力而成为理论上是令人满意的。对世界其他地区实施稳定计划。

我们仍需要讨论的一个重要问题是时间问题。即使我们严重关注发明的罐子可能包含黑球，但如果我们认为以后可以采取这些措施，我们现在不必支持建立更强大的监督或全球治理。当假设的弱势性清楚地出现时。然后我们可以让世界继续它的甜蜜睡眠，在充满信心的期望，一旦警报响起，它将跳出床并采取必要的行动。但我们应该质疑这个计划是多么现实。

一些历史反思在这里很有用。在整个冷战期间，两个超级大国（以及整个北半球）生活在对核毁灭的持续恐惧中，这可能是在任何时候由于意外或由于某些危机失控而引发的。各方都接受了威胁的现实。仅仅通过摆脱所有或大多数核武器就可以大大减少这种风险（此举可以节省超过 10 万亿美元）。[77] [78] 然而，经过几十年的努力，只有实施了有限的核裁军和其他降低风险

---

[75] See also Re (2016) and Bostrom (2006); cf. Torres (2018)

[76] 例如，监视辩论往往侧重于个人隐私权益与公众对小规模恐怖袭击的安全需求之间的权衡。（即使是通常被认为是大型的恐怖事件，例如9/11袭击，根据本文使用的标准，也不是很小规模。）

[77] 施瓦茨（Schwartz，1998）认为，冷战期间的核军备竞赛仅占美国支出的 5.8 万亿美元（1996 年的美元），相当于 2018 年美元的 9.3 万亿美元。 这个估计非常全面，涵盖了燃料循环，武器，运载系统，退役等。如果我们增加其他国家的支出，我们可以得出结论，人类文明在开发和维持能力方面花费了超过 10 万亿美元。 用核武器摧毁自己。 大部分成本是在世界 GDP 大大低于今天的时期内发生的。

措施。事实上，核毁灭的威胁至今仍然存在。由于缺乏强有力的全球治理可以执行条约并迫使争议者接受妥协，世界迄今无法解决这一最明显的集体行动问题。[79]

但也许世界未能消除核战争风险的原因是风险不是很大？如果风险更高，人们可以争辩说，那么就可以找到解决全球治理问题的必要意愿。也许 - 虽然看起来似乎是一个相当不稳定的理由来维持文明的命运。我们应该注意到，虽然一种比核武器更危险的技术可能会刺激更大的意愿来克服实现稳定的障碍，但黑球的其他特性可能会使全球治理问题比冷战期间更具挑战性。我们已经在诸如“安全第一次打击”和“更糟糕的全球变暖”等情景中说明了这种可能性。我们看到技术集的某些属性如何能够为破坏性使用产生更强的激励，或者拒绝加入（或豁免）任何协议来遏制其有害应用。[80]

我们仍需要讨论的一个重要问题是时间问题。即使我们严重关注发明的罐子可能包含黑球，但如果我们认为以后可以采取这些措施，我们现在不必支持建立更强大的监督或全球治理。当假设的弱势性清楚地出现时。然后我们可以让世界继续它的甜蜜睡眠，在充满信心的期望，一旦警报响起，它将跳出床并采取必要的行动。但我们应该质疑这个计划是多么现实。

即使人们对最终达成协议感到乐观，时间问题仍然是一个严重问题。国际集体行动问题，即使在有限的领域内，也可以长期抵制解决方案，即使赌注很大且无可争议。需要时间来解释为什么需要安排并回答反对意见，时间谈判双方可接受的合作理念实例，时间来确定细节，以及时间建立实施所需的体制机制。在许多情况下，持有问题和国内反对可能会拖延数十年的进展；当一个顽固的国家准备加入时，另一个先前同意的人可能已经改变了主意。然而与此同时，所有人都清楚可见的弱势性与必须采取稳定措施的时间点之间的间隔时间可能很短。如果弱势性的性质留下了拒绝的余地，或者由于信息危害而无法广泛提供具体的解释，那么它甚至可能是消极的。这些考虑表明，一旦出现弱势性，依靠自发的临时国际合作来挽救这一天是有问题的。[81]

预防性警务的情况在某些方面是相似的，尽管我们看到监测技术的进步推动了更快，更强劲的趋势 - 提高国家监测和潜在控制其本国公民行动的能力，而不是有效的全球治理。如果我们看一下物理领域，至少也是如此。在数字信息领域，由于加密和匿名化工具的使用日益广泛，以及破坏性创新的频率使得网络空间的未来更难以预见，因此前景不太明朗。然而，物理空间中足够强大的能力也会扩散到数字领域的强大能力中。在高科技 圆形监狱 中，当局无需破解密码，因为他们可以直接观察用户输入计算机的所有内容以及屏幕上显示的所有内容。

人们可以采取这样的立场：我们不应该制定改进的监督和社会控制方法，除非并且直到明确地看到一个特定的文明弱势性- 一个看起来足够严重以证明牺牲某些类型的隐私以及无意中促进极权主义的噩梦。但与国际合作的情况一样，我们面临时机问题。高度复杂的监视和响应系统，如“高科技 圆形监狱”所描述的系统，无法在一夜之间变得完全可靠。实际上，从我们目前的起点来看，实施这样一个系统需要很多年，更不用说建立政治支持所需的时间。然而，可能需

更多的金额用于非核军事能力（仅在美国，2018 年就超过 20 万亿）。核支出有可能通过减少非核军事开支来节省资金。 核军事和非核军事开支都反映了人类文明未能解决全球协调问题。

[78] 从根本上而不是完全地，因为即使所有核武器都被拆除，也可能会产生新的核武器。

[79] 当然，全面核裁军协议也可能涉及一些关于常规力量和其他事项的规定，以免危及战略稳定。

[80] 人们可能会查看其他历史示例以获得更大的参考类。到目前为止，世界在应对全球变暖方面所做的努力并没有激发人们对其迅速处理更加困难的全球集体行动问题的能力的信心。另一方面，“蒙特利尔议定书”成功解决了臭氧消耗问题。

[81] 单边强加可能更快，但它要求某些行动者有能力将其意志强加于世界其他地方。如果一个行动者具有如此强大的权力优势，那么事实上（弱或潜在）全球治理的形式可能已经到位。

要这样一个系统的弱势性可能不会给我们提供很多预先警告。上周，顶级学术生物实验室可能在科学上发表了一篇文章；当你正在阅读这些文字时，世界各地的一位受欢迎的博主，对网页浏览的热烈追求，可能会上传一篇文章，解释一些聪明的方式，实验室的结果可以被任何人用来造成大规模破坏。

在这种情况下，可能需要立即开启强烈的社会控制。在不利的情况下，提前期可能短至数小时或数天。当弱势性清晰可见时，开始开发监控架构将为时已晚。如果要避免破坏，则需要事先确定稳定机制。

理论上可行的是提前开发侵入式监视和实时拦截的能力，但最初并不是将这些能力用于任何类似的能力。这将是满足稳定 1-型弱势性（以及需要对各个操作进行高度可靠监控的其他弱势性）的要求的一种方法。通过赋予人类文明极其有效的预防性监管的能力，我们将退出半无政府状态违约条件的一个方面。

无可否认，构建这样一个系统并使其处于待命状态将意味着实际采取强烈形式社会控制的一些缺点将会产生。特别是，它可能会使压迫性结果更有可能："问题在于，监管系统的建立是否危险地改变了政府控制方向的这种平衡.....我们可以想象在家中安装强制摄像机系统，仅通过授权激活，多年来一直严格遵守法律。问题在于，这样的监视体系一旦建立，就很难拆除，并且如果它落入了人们的手中 —— 无论是通过恐慌，恶意还是误导他们的信心，那么它就是一种过于强大的控制工具。自己秘密判断公共利益的能力 —— 会试图用它来反对我们。"[82]

建立一个交钥匙极权主义的体系意味着冒险，即使一个人不打算改变关键。

人们可以通过设计具有适当技术和制度保障的系统来尝试降低这种风险。例如，人们可以通过组织信息架构来实现"结构化透明度"系统，以防止权力集中，以便多个独立的利益相关者必须给予他们的许可才能使系统运行，并且只有特定的信息才是某些决策者合法需要的是向她提供的，并且在目的许可的情况下适当的修改和匿名化。通过一些创造性的机制设计，一些机器学习以及一些花哨的加密步法，可能没有实现监控系统的根本障碍，该监控系统在其官方功能中同时非常有效，但也有一定程度上不能被颠覆为替代用途。

在实践中实现这一目标的可能性当然是另一个问题，需要进一步探索。[83] 即使极权主义的重大风险不可避免地伴随着善意的监督项目，也不会追求这样的项目会增加极权主义的风险。风险相对较低的善意项目，在相对平静的时候开始，例如可以通过抢占在危机期间开始的不太善意和风险更高的项目来降低极权主义的风险。但即使存在一些净极权主义风险增加效应，也可能值得承担这种风险，以获得稳定文明以抵御新兴的 1-型威胁的一般能力（或者为了其他有益的监督和预防性警务可以带来）。

结论 Conclusion

本文介绍了一个视角，我们可以更容易地看到文明如何容易受到我们技术创造的某些类型的可能结果的影响 —— 我们从发明的罐子中绘制出一个隐喻的黑球，我们有能力提取而不是我们

[82] Sanchez (2013)

[83] 例如，一个善意的项目可能会在其实施中被颠覆；或者它可能会导致错误或机构设计缺陷，只有在正常运行一段时间后才会变得明显。即使系统本身按预期运行并且保持不受损坏，它也可能激发创建其他没有相同民主保障的监视系统。

开发了这种潜在弱势性的类型，并展示了其中一些是由于破坏变得太容易导致的，另一些是由于一些强大的国家行动者或大量弱势行动者所面临的激励变化。

我们还研究了各种可能的反应及其局限性。我们追溯了我们文明暴露于当代世界秩序的两个结构特征的根本原因：一方面，缺乏预防性警务能力，以极高的可靠性阻止个人或小团体采取非常违法行为；而另一方面，缺乏全球治理能力以可靠地解决最严重的国际协调问题，即使在默认情况下重要的国家利益激励国家叛逃。

对于潜在的文明弱势性的一般稳定 —— 在一个技术创新沿着广阔前沿迅速发展的世界中，并且存在大量具有各种人类可识别动机的行动者 —— 将要求消除这两个治理差距。在这样的时间到来之前，人类将仍然容易受到技术黑球的影响。

显然，这些反思证明有利于支持加强监督能力和预防性警务系统，并支持能够采取果断行动的全球治理制度（无论是基于单边霸权力量还是强大的多边机构）。但是，我们还没有确定这些事情是否值得考虑所有事情，因为这样做需要分析一些超出本文范围的其他强有力的考虑因素。

因为我们的主要目标是在宏观战略环境中设置一些标志，我们将讨论集中在一个相当抽象的层面，开发可以帮助我们定位自己的概念（关于长期结果和全球可取性），这在某种程度上独立于我们不同的当地情况的细节。在实践中，如果要努力稳定我们的文明以防止潜在的黑球，人们可能会发现谨慎地将重点放在部分解决方案和低悬的果实上。因此，不是直接尝试实现极其有效的预防性监管或强有力的全球治理，人们可能会尝试修补黑球似乎最有可能出现的特定领域。例如，可以通过开发更好的方法来跟踪关键材料和设备，以及监测实验室内的活动，加强对生物技术相关活动的监督。人们还可以收紧生物技术供应部门的客户知识规定，并扩大背景检查的使用范围，以便在某些实验室工作或参与某些类型的实验。人们可以改进举报人制度，并试图在全球范围内提高生物安全标准。人们还可以追求差别技术发展，例如加强生物武器公约和维护生物武器的全球禁忌。可以鼓励资助机构和道德审批委员会更广泛地了解特定工作的潜在后果，不仅关注实验室工作人员，测试动物和人类研究课题的风险，还关注希望发现的方法可能会降低生物恐怖分子的竞争力。主要是保护性的工作（如疾病暴发监测，公共卫生能力建设，空气过滤装置的改进）可以有所不同。

然而，在追求这些有限的目标时，应该记住，它们提供的保护仅涵盖特殊的情景子集，并且可能是暂时的。如果人们发现自己能够影响预防性警务能力或全球治理能力的宏观参数，那么应该考虑到这些领域的根本性变化可能是实现稳定我们的文明以抵御新兴技术弱势性的一般能力的唯一途径。

参考文献 References

Aviles, L. and Purcell, J. 2012. “The evolution of inbred social systems in spiders and other organisms: From short-term gains to long-term evolutionary dead ends?” Advances in the Study of Behavior 44: 99–133. Academic Press.
Armstrong, S., Bostrom, N., and Shulman, C. 2016. “Racing to the precipice: a model of artificial intelligence development.” AI & Society 31 (2): 201–206.

Badash, L. 2001. “Nuclear winter: Scientists in the political arena.” Physics in Perspective 3 (1):76–105.
Banisar, D. 1999. “Stopping science: The case of cryptography.” Health Matrix 9: 253.
Blinder, A. and Perlroth, N., 2018. “A cyberattack hobbles Atlanta, and security experts shudder.” New York Times.
Bolt, J., Inklaar, R., de Jong, H., and Luiten van Zanden, J. 2018. “Rebasing ‘Maddison’: new income comparisons and the shape of long-run economic development.” GGDC Research Memorandum 174. 34
Bostrom, N. 2002. “Existential risks-analyzing human extinction scenarios and related hazards.” Journal of Evolution and Technology 9.
Bostrom, N. 2006. “What is a singleton.” Linguistic and Philosophical Investigations 5 (2): 48–54.
Bostrom, N. 2008. “Why I want to be a posthuman when I grow up.” In Medical Enhancement and Posthumanity, edited by Bert Gordijn and Ruth Chadwick, 107–37. New York: Springer.
Bostrom, N. and Ćirković, M. M. eds. 2011. Global Catastrophic Risks. Oxford University Press.
Bostrom, N. 2011. “Information hazards: A typology of potential harms from knowledge.” Review of Contemporary Philosophy 10: 44–79.
Bostrom, N. 2013. “Existential risk prevention as global priority.” Global Policy 4 (1): 15–31.
Bostrom, N., Douglas, T., and Sandberg, A. 2016. “The unilateralists curse and the case for a principle of conformity.” Social Epistemology 30 (4).
Brooks, S. G. 1999. “The globalization of production and the changing benefits of conquest.” Journal of Conflict Resolution 43 (5): 646–670.
Brower, M. 1989. “Targeting Soviet mobile missiles: Prospects and implications.” Survival 31 (5): 433–445.
Centre for Biosecurity and Biopreparedness. 2017. An efficient and practical approach to Biosecurity. Copenhagen, Denmark. Retrieved November 4, 2018. Available at https://www.biosecurity.dk/fileadmin/user_upload/PDF_FILER/Biosecurity_book/An_efficient_and_Practical_approach_to_Biosecurity_web1.pdf.
Collingridge, D. 1980. The social control of technology. New York: St. Martin's Press.
Crafts, N. and Fearon, P. 2010. “Lessons from the 1930s Great Depression.” Oxford Review of Economic Policy 26 (3): 285–317.
Day, E. H., Hua, X., and Bromham, L. 2016. “Is specialization an evolutionary dead end? Testing for differences in speciation, extinction and trait transition rates across diverse phylogenies of specialists and generalists.” Journal of Evolutionary Biology 29 (6): 1257–1267.
Diamond, J. 2005. Collapse: How societies choose to fail or succeed. Penguin.
Danzig, R. 2011. Aum Shinrikyo: Insights into how terrorists develop biological and chemical weapons. Center for a New American Security.
Drexler, K. E. 1986. Engines of Creation. Anchor.
Duprex, W. P., Fouchier, R. A., Imperiale, M. J., Lipsitch, M., and Relman, D. A. 2015. “Gain-of-function experiments: time for a real debate.” Nature Reviews Microbiology 13 (1): 58.
Ellsberg, D. 2017. The Doomsday Machine: Confessions of a Nuclear War Planner. Bloomsbury Publishing USA.

Fauci, A. S., Nabel, G. J., and Collins, F. S. 2011. “A flu virus risk worth taking.” The Washington Post.
Federal Select Agent Program. 2017. Suitability Assessment Program Guidance. Retrieved November 4, 2018. Available at https://www.selectagents.gov/resources/Suitability_Guidance.pdf.
Field, C. B. et al. 2014. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Working Group II Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Ford, C. A. and Rosenberg, D. A. 2005. “The Naval Intelligence Underpinnings of Reagan's 35 Maritime Strategy.” Journal of Strategic Studies 28 (2): 379–409.
Franck, J., Hughes, D. J., Nickson, J. J., Rabinowitch, E., Seaborg, G. T., Stearns, J. C., and Szilard, L. 1945. “Report of the committee on political and social problems.” Manhattan Project “Metallurgical Laboratory.”
Friedlingstein, P. et al. 2014. “Persistent growth of CO 2 emissions and implications for reaching climate targets.” Nature Geoscience 7 (10).
Gartzke, E. 2007. “The capitalist peace.” American Journal of Political Science 51 (1).
Gartzke, E. and Rohner, D. 2011. “The political economy of imperialism, decolonization and development.” British Journal of Political Science 41 (3).
Greenberg, A. 2012. This Machine Kills Secrets: How Wikileaks, Hacktivists, and Cypherpunks Are Freeing the World’s Information. Random House.
Gronvall, G. K. 2013. “H5n1: A case study for dual-use research.” New York, NY: Council on Foreign Relations.
Holloway, D. 1994. Stalin and the bomb: the Soviet Union and atomic energy, 1939-1956. Yale University Press.
Igic, B. and Busch, J. W. 2013. “Is self-fertilization an evolutionary dead end?” New Phytologist 198 (2).
Jaffe, R. L., Busza, W., Wilczek, F., and Sandweiss, J. 2000. “Review of speculative “disaster
scenarios” at RHIC.” Reviews of Modern Physics 72 (4): 1125.
Jones, C. I. 2016. “Life and Growth.” Journal of Political Economy 124 (2): 539–578.
Konopinski, E. J., Marvin, C., and Teller, E. 1946. “Ignition of the atmosphere with nuclear bombs.” Report LA-602. Los Alamos, NM: Los Alamos Laboratory.
Lechner, S. 2017. “Anarchy in International Relations.” Oxford Research Encyclopedia of International Studies. Oxford University Press.
Lewis, G. et al. (forthcoming). “Information hazards in biotechnology.” Risk Analysis.
Liberman, P. 1993. “The Spoils of Conquest.” International Security 18 (2).
Lieber, K. A. and Press, D. G. 2017. “The new era of counterforce: technological change and the future of nuclear deterrence.” International Security 41 (4).
Lundgren, C. 2013. “What are the odds? assessing the probability of a nuclear war.” The Nonproliferation Review 20 (2).
Mowatt-Larssen, R. and Allison, G. T. 2010. Al Qaeda Weapons of Mass Destruction threat: hype or reality? Belfer Center for Science and International Affairs.
Morland, H. 1979. “The h-bomb secret: How we got it; why were telling it.” The Progressive 43 (11).
Mueller, J. 2009. Atomic obsession: nuclear alarmism from Hiroshima to al-Qaeda. Oxford University Press.

Munz, P., Hudea, I., Imad, J., and Smith, R. J. 2009, “When zombies attack!: mathematical modelling of an outbreak of zombie infection.” Infectious Disease Modelling Research Progress 4.
Norris, R. S. and Kristensen, H. M. 2010. “Global nuclear weapons inventories, 1945-2010.” Bulletin of the Atomic Scientists 66 (4). 36
Olson, K. B. 1999. “Aum shinrikyo: once and future threat?” Emerging Infectious Diseases 5 (4).
Ord, T., Hillerbrand, R., and Sandberg, A. 2010. “Probing the improbable: methodological challenges for risks with low probabilities and high stakes.” Journal of Risk Research 13 (2).
Parfit, D. 1987. Reasons and Persons. Clarendon Press.
Persson, I. and Savulescu, J. 2012. Unfit for the future: The need for moral enhancement. Oxford University Press.
Re, R. M. 2016. “Imagining Perfect Surveillance.” UCLA Law Review Discourse 64.
Rhodes, R. 1986. The making of the atomic bomb. Simon and Schuster.
Rindskopf, E. P. and Jacobs, L. G. 2003. “Government controls of information and scientific inquiry.” Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science 1 (2): 83–95.
Sagan, S. D. 1995. The limits of safety: Organizations, accidents, and nuclear weapons. Princeton University Press.
Sagan, S. D. and Waltz, K. N. 2013. The spread of nuclear weapons: an enduring debate. WW Norton New York.
Sanchez, J. 2013. “A reply to Epstein and Pilon on NSA’s metadata program.” Cato at Liberty. Schelling, T. C. 1960. The Strategy of Conflict. Harvard.
Schlosser, E. 2013. Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety. Penguin Press.
Schwartz, S. I. Atomic audit: the costs and consequences of US nuclear weapons since 1940. Brookings Institution Press.
Sharp, P. A. 2005. “1918 flu and responsible science.” Science 310 (5745).
Shindell, D. T. 2014. “Inhomogeneous forcing and transient climate sensitivity.” Nature Climate Change 4 (4).
Sokoski, H. D. and Tertrais, B. 2013. Nuclear Weapons Security Crisis: What Does History Teach? Nonproliferation Policy Education Center.
Stevenson, J. 2008. Thinking beyond the unthinkable: Harnessing doom from the Cold War to the War on Terror. Penguin.
Stocker, T. F. et al. 2014. Climate Change 2013: The Physical Science Basis. Contribution of
Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Swire, P. 2005. The Declining Half-life of Secrets and the Future of Signals Intelligence. New America.
Tegmark, M. and Bostrom, N. 2005. “Astrophysics: Is a doomsday catastrophe likely?” Nature 438 (7069).
Torres, P. 2018. “Superintelligence and the Future of Governance: On Prioritizing the Control
Problem at the End of History.” In Artificial Intelligence Safety and Security, edited by Roman Yampolskiy. Chapman and Hall/CRC.