人工智能的恶意使用及对金砖国家心理安全的挑战

研究报告

研究协调员: 叶夫根尼·帕申采夫

本文集由国际社会政治研究与咨询中心编辑，在关于通过恶意使用人工智能对心理安全产生威胁国际研究小组（MUAI 研究）的帮助下完成。

2024 年 6 月，莫斯科

人工智能的恶意使用
及对金砖国家心理安全的挑战

研究报告

研究协调员: 叶夫根尼·帕申采夫

本文集由国际社会政治研究与咨询中心编辑，在关于通过恶意使用人工智能对心理安全产生威胁国际研究小组（MUAI研究）的帮助下完成。

2024年6月，莫斯科

本报告分别以英文、中文和俄文出版，由叶夫根尼·帕申采夫 (Evgeny Pashentsev) 担任总编和出品监制。

中文翻译、编辑：沈圣、李羽曦、王革予

人工智能的恶意利用及金砖国家心理安全面临的挑战

研究协调员：叶夫根尼·帕申采夫

本文集由国际社会政治研究与咨询中心编辑，在关于通过恶意使用人工智能对心理安全产生威胁国际研究小组（MUAI研究）的帮助下完成。

莫斯科：LLC «SAM Polygraphist»，2024 年。 – 96 页。

ISBN 978-5-00227-264-8

人工智能（AI）技术具有巨大的变革潜力，已经在金砖国家引领了众多积极变革，但其也存在着巨大的潜在风险，其中不乏与各种恶意行为者的活动相关的风险。如今，人工智能的恶意使用在现代世界中呈现出数量和质量的增长，对人类的生命、健康和幸福构成了严重威胁。其不仅可能导致使用现代技术的风险增加，还可能会带来现今无法完全预见的新风险。本报告重点关注恶意人工智能对人类心理所带来威胁的影响，且由此对十个金砖国家的政治、经济、社会进程以及国家和非国家机构活动所产生的影响进行分析。该报告将对心理安全构成威胁的人工智能恶意使用情况进行了三级分类。

封面图片：免费图片。
出版签发于 2024 年 6 月 11 日 | 数码印刷 | 订单编号32149

由 «OneBook.ru» LLC «SAM Polygraphist» 印刷厂印刷
109316，莫斯科，伏尔加格勒大街，莫斯科科技城5号大楼第42号楼
www.onebook.ru

目录

引言

人工智能的恶意使用——不断增长的威胁

叶夫根尼·帕申采夫，俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）

正在不断演进的人工智能（AI）技术的快速发展和实施，为金砖国家带来了巨大的经济和社会效益，并导致了生产、金融、贸易、交通、教育、医疗和休闲等领域的根本性变革。政府机构、政党和公共组织的运作也受到越来越多人工智能的影响。人工智能凭借先进的能力为社会提供了巨大的转型力量，其带来众多积极的社会变革潜能的同时，也存在着巨大的风险。

“金砖国家已经同意尽快启动人工智能研究组工作……我们应努力形成具有广泛共识的人工智能治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性、公平性（《中国日报》，2023年）。”中国国家主席习近平在南非举行的第十五届金砖国家领导人会晤上表态。金砖国家轮值国俄罗斯总统普京表示，在 2024 年其领导期内，俄罗斯计划将人工智能合作牢固地纳入金砖国家的议程中进行“详细讨论”。他设想在未来的技术世界中，金砖国家之间的人工智能保障将得到统一，确保机遇和风险都能得到负责任的管理（金砖国家智库，2023 年）。印度总理莫迪在 2023 年 12 月举行的全球人工智能伙伴关系峰会上表示，人工智能可以成为 21 世纪发展最重要的工具，但它同样也可能成为摧毁 21 世纪的力量。“我们必须共同努力制定人工智能的全球框架，确保其可被道德地使用，”（印度报社，2023 年）他强调道。

在金砖国家范围内，人工智能的恶意使用（MUAI）呈现上升趋势，而这恰恰也反映了全球普遍趋势。此外，人工智能的恶意使用也代表着一种主要风险，其既能够放大其他人工智能风险，又可以创造出无法完全预见但愈发真实的新风险。如今，各国人工智能的恶意使用的潜在风险引发了一系列担忧。鉴于此，Collaborations Pharmaceuticals 的研究人员与欧洲科学机构合作进行了一项概念实验。他们要求 MegaSyn AI 神经网络反向合成新药物，找出对人体毒性最大的物质。实验结果显示，神经网络正确理解了任务，并在不到六个小时的时间内生成了一个包含 40,000 种对化学和生物武器的最优成分的清单。人工智能不仅独立设计了许多已知的化学战剂，还设计了许多更有毒的新战剂。这种对机器学习模型的简单倒置，将一个无害的生成模型从一个有用的工具转变为了大规模谋杀的促成者（Urbina 等人，2022 年）。

可以合理怀疑的是，这种倒置方法可能会被应用到其他领域，比如找到对公众意识产生负面心理影响的最佳方式。因此，充分评估人工智能的恶意使用在心理对抗方面的真实威胁是很重要的。MUAI 会促使人类意识的不稳定化，从而导致社会的不稳定化，又最终恶化人类意识的不稳定性，由此形成了一个危险且不可持续的恶性循环。这一切都是为了在全球资产再分配过程中从非常狭窄的超级富豪圈中“获得胜利”。与此同时，社会发展面临的客观现实——人性的自私、反社会行为、惯性思维以及社会大众面对社会改革时的木讷——紧密相连。

整体而言，军事政治领域的侵略性集团曾在历史上多次导致毁灭性战争，包括 20 世纪的两次世界大战。而其他国家协会则在努力寻找可以应对当前受军备竞赛威胁的社会和经济秩序的替代方案。金砖国家（BRICS）则具备第二类国家协会的特点，它可以通过在全球层面的巨大机遇，而并非通过空洞的承诺，将当前脆弱而危险的局势与更美好的未来联系起来（Pashentsev and Miao, 2023）。即使在当前的基础上，金砖国家也拥有全球 45%以上的人口、占比世界 GDP 的 32%（按购买力平价计算，而 G7 国家持有的比例为 30%）、具备强大的研究和技术潜力以及最丰富的自然资源，这使得解决全人类利益的全球问题

成为可能。而在未来，大约有三十多个国家准备以某种形式加入金砖国家，这也是金砖国家的潜力所在（TV BRICS 2024）。

对人工智能技术进行改进和实施对于金砖国家具有重大意义。其不仅可以有效解决金砖国家的许多复杂任务，而且保障各国免受恶意国家和非国家行为者在安全方面的威胁，包括心理安全领域。

MUAI 对心理安全三个层级的威胁

恶意使用人工智能（MUAI）是一种有意的反社会行为，呈现出显性和隐性两种形式。反社会团体（从个体罪犯到强大的自私利益集团、腐败的国家机构）已经利用 MUAI 追求利益追求。近年来，MUAI 在心理领域展现出巨大的潜力。尽管关于 MUAI 技术方面有大量且迅速增长的学术出版物，包括其一般的社会经济和政治影响以及对 MUAI 进行分类的首次尝试（Brundage 等人，2018 年；Caldwell 等人，2020 年；Malicious Uses，2020 年；Blauth 等人，2022 年），但在心理安全背景下针对具体 MUAI 问题的出版物相对较少，甚至在心理安全领域系统性考虑 MUAI 的影响的更少。心理安全的概念在许多研究中都有涉及（Grachev，1998 年；Roshhin 和 Sosnin，1995 年；Afolabi 和 Balogun，2017 年）。著名的美国心理学家亚伯拉罕·马斯洛认为，一旦基本的生理需求得到满足，心理安全需求就成为首要问题。更具体地说，心理安全是对未来的保护、稳定性、对未来的信心、良好健康等的需求。国家心理安全可被理解为使公民、个体群体、社会群体、大规模人群和整个国家人口免受负面心理影响的保护（Barishpolets，2013 年，第 63 页；更多内容请参阅：Barishpolets，编辑，2012 年）。基于以上定义，作者认为可以将国际心理安全定义为：保护国际关系体系免受与国际发展各种因素相关的负面心理影响的影响。这包括防止各种国家、非国家和超国家行为者在实现局部/全面、地区/全球、短期/长期和潜在/明显的国际体系不稳定，以获取竞争优势，甚至通过对敌人的物理消灭而进行的针对性行动。

尽管恶意心理影响的分离分析涉及到 AI 换脸变声技术、机器人、预测分析等等，但对这种影响的协同作用以及心理安全风险增长或对整个国家和国际安全体系的风险的全面分析并没有得到考虑。这种对综合分析的缺乏可以解释为该问题的新颖性：MUAI 的实践无法超过 AI 的进步。

第一步是整合不同国家学者在这一新领域的努力，这一步已于 2019 年完成——成立了国际研究人工智能恶意使用对国际心理安全威胁的国际研究小组（Research MUAI），并在随后的合作中（例如，联合研究和出版、国际会议、科学研讨会等）展开了七个国家的研究人员之间的合作。该小组研究人员准备的数十篇学术文章涉及了人工智能恶意使用和心理安全的不同问题，最终引发了该领域的第一本书，即《恶意使用人工智能和心理安全的 Palgrave 手册》的出版。该手册的 23 位贡献者来自亚洲、欧洲和北美的 11 个国家（The Palgrave Handbook, 2023）。近年来，关于金砖国家心理安全面临的人工智能恶意使用威胁的系统研究首次得到发表（Bazarkina，Pashentsev 2020，P. 154-177；Pashentsev，Bazarkina 2023），并且陆续也出现了关于分析个别金砖国家心理安全面临的人工智能恶意使用威胁的出版物（Bazarkina, Matyashova 2022, PP. 14-20; Bazarkina, Mikhalevich 等等 2023; Cai, Zhang 2023; Gupta, Guglani 2023）。

本报告侧重通过人工智能恶意使用对人类心理安全的威胁，以及基于这一点对金砖国家的政治、社会、经济、文化进程以及国家和非国家机构的活动进行综合分析[1]。 本报告不对金砖国家应对这些新威胁的产生应做出何种反应进行介绍，因为在国家层面，这些威胁还处于形成阶段（某些国家对这些威胁仅有初步和零碎理解，而其他国家已采取首批法律和技术对策），因而需要具体问题具体分析。

[1] 巴西、俄罗斯、印度和中国是金砖国家的创始国，他们于 2009 年在叶卡捷琳堡举行了首次峰会，一年后南非加入了该组织。在去年 8 月举行的第 15 届峰会上，金砖国家邀请沙特阿拉伯、阿联酋、阿根廷、埃及、伊朗和埃塞俄比亚加入该组织。后来，阿根廷新当选总统哈维尔·米莱拒绝了加入金砖国家的提议。埃及、埃塞俄比亚、伊朗、沙特阿拉伯和阿联酋加入了该组织。

该报告以对人工智能恶意使用威胁对心理安全的三级分类为基础（更多内容请参阅：The Palgrave Handbook 2023, PP. 23-46）。

在第一级别上，这些威胁与有意扭曲 AI 发展的情况和后果，与反社会团体直接有关。在这种情况下，AI 本身在这一级别上并未直接参与心理安全的破坏。该破坏性（公开或隐藏的）的影响给人们的心理中灌输了对 AI 的错误印象。对 AI 发展的过度、人为制造的负面反应（例如，关于机器人和 AI 将很快迫使所有人失业，工人将成为 AI 的奴隶等恐怖故事）具有一定的政治和经济目标，并不像看起来那样无害。这种负面反应可能会几乎全面地减缓 AI 技术的实施，并引发社会政治的紧张和冲突。此外，公众对 AI 的过度期望也可能表现出来，在某个阶段可能导致高科技公司和整个市场价值的自然崩溃。这些期望可能会被恶意利用并强烈导致公众、有利益的商业和非营利结构以及公共当局产生迷惑，并最终转变为失望、错误决策以及社会和政治冲突。

在第二级别上，恶意使用的领域是完全开放的：无端使用无人机、对脆弱基础设施进行网络攻击、重新定位商业人工智能系统、利用人工智能技术干扰决策或以潜在方式修改决策等等。但值得注意的是，在这个级别上，对公众意识的攻击并不是其主要目标。

主要旨在造成心理伤害的 MUI 属于心理安全威胁的第三级别，即最高级别。合成的人工智能产品（结合了多种技术，可以增加其被黑客攻击或恶意使用时的危害）创造了一系列新的风险和威胁。专业使用心理战手段和方法可以将威胁感知水平提高或降低到适当水平之上或之下。此外，人工智能在心理战中的使用使得感知管理的隐藏（潜在）运动变得更加危险，且这种危险在未来只会呈现上升趋势。因此，意为在心理领域造成损害的人工智能恶意使用应该得到独立且非常密切的关注。前两个级别的威胁在影响人类意识和行为方面的作用力不同，有可能会对人类造成灾难性影响，正如可能爆发第三次世界大战的情况一样。然而，在发展的某个阶段，第三级别的影响可以促使反社会团体影响或甚至控制公众意识；这可能导致某个国家或整个国际局势骤然失衡。如果第三级别对敌人的心理形成了可靠的控制，那么其他两个级别对心理安全的威胁就变成了辅助性的。

人工智能恶意使用威胁可以在一个影响层面上发生，也可以同时在多个层面上作为单一感知管理活动的一部分产生。针对同一事件，恐怖分子使用无人机或对平民人口的袭击将构成第二级别威胁，该威胁具有传播效应（袭击后引发恐慌和震惊）。而若犯罪分子行动能够得到广泛的信息支持（也借助于人工智能），威胁将升级至第三级别。

人工智能并非单一技术。在不同环境和情境下，许多人工智能技术可通过各种应用程序应用于众多功能，以不同方式和形式进行操作。本报告的作者考虑到了总体人工智能范畴下的不同技术如何帮助创造特定产品，从而严重改变了任何特定类型活动的技术水平和实际能力。

随着地缘政治竞争的加剧，不同国家和非国家反社会行为者的活动增加。同时，随着各种人工智能技术的发展和日益普及，人工智能技术更易获取，人工智能恶意使用所带来的威胁在全球和金砖国家的所有三个层面上都变得更加重要。这只会加剧各种利益集团试图利用人工智能影响公众意识以谋求自身利益。在历史上的危机时刻，这种对公众意识的操纵尤为具有破坏性。在第二次世界大战的火焰中，超过 5 千万人的死亡使法西斯主义的残酷性显而易见。然而，在战争爆发之前，对公众意识的操纵确保了希特勒在 1933 年的国会选举中的胜利。这段并不遥远的历史对于今天的人们仍然具有高度的启示意义。可联想到的是，现代金砖国家以及许多其他国家的政府和政治人物对互联网上高科技虚假信息的威胁以及使用人工智能技术的私营主流媒体平台的角色日益感到担忧。

金砖国家中的人工智能恶意使用对心理安全的威胁既源于内部原因，也是外部因素的结果。因此，在这里，有必要对全球范围内三个层面上威胁的性质和动态提供一些一般概念。

MUAI 对心理安全的第一层级威胁

在第一级别对心理安全起到威胁的 MUAI 中，对现代世界中人工智能技术的进一步发展和应用持消极态度的增加为此类威胁创造了有利的条件，这在西方国家尤为典型。根据皮尤研究中心 2023 年 8 月进行的调查，52%的美国公民表示他们对人工智能的增加使用感到更担忧而非兴奋，只有 10%的人表示他们对此更感兴奋，而 36%的人表示他们有着担忧和喜悦混合的复杂情绪。在 2022 年 12 月 38%的基础上，对人工智能在日常生活中的使用感到担忧的美国人的比例增加了 14%。而这一现象是伴随着公众意识的增长而发生的。90%成年人听说过人工智能，其中有三分之一（33%）对 AI 有较多的了解，有三分之五（56%）只有浅层的认识。自 2022 年 12 月以来，对
AI 有较多的了解的人的比例增加了 7 个百分点，但他们比起 2022 年 12 月时更有可能呈现出担忧心理。在这个群体中，对人工智能的担忧以 47%的比例超过了兴奋的 15%。在去年 12 月，这个差距是 31%比 23%（泰森，菊池，2023 年）。

在与昆士兰大学合作的情况下，澳大利亚毕马威公司领导了全球首次深入研究，探究人工智能在全球范围内的信任度和态度。这项研究覆盖了 17 个国家，调查了超过 17,000 名受访者，涵盖了全球各个地区，包括澳大利亚、巴西、加拿大、中国、爱沙尼亚、芬兰、法国、德国、印度、以色列、日本、荷兰、新加坡、南非、韩国、英国和美国。这些国家在人工智能活动和准备方面处于领先地位。根据毕马威的研究[2]，大多数群众（56-75%）在 BICS 国家信任人工智能系统，其中印度人报告了最高的信任意愿，其次是中国。相反，在其他国家，只有少数人报告对人工智能的信任，其中芬兰人的信任水平最低（仅 16%）。BICS 国家对人工智能的信任和接受度更高，可能是由于这些国家对人工智能的加速采用以及新兴技术在经济中的日益重要作用的结果。BICS 国家的群众对人工智能持最为乐观的态度，认为其带来的好处最多，并且报告在工作中对人工智能的采用和使用水平最高（吉尔斯皮等人，2023 年，第 14 页）。相比之下，西方国家和日本的人对人工智能的好处是否超过风险持怀疑态度。

对人工智能的不信任或恐惧的原因显然是在于人们对主要社会机构的信任急剧下降。以美国为例，在 2022 年，盖洛普公司 (Gallup) 记录到公众对其每年追踪的 16 个机构中的 11 个出现了显著的信任下降，其中总统和最高法院受到的影响最为严重。在这些机构中，表示对其有很高程度或相当程度信任的美国人的比例分别下降了 15 个和 11 个百分点。在 2023 年 6 月 1 日至 22 日期间进行的最新调查中，这两个得分都没有明显恢复，最高法院的信任度现在为 27%，总统的信任度为 26%。而评分最低的五个机构——报纸、刑事司法系统、电视新闻、大企业和国会——的信任度不到 20%，其中国会仅有 8%的支持率。2023 年评定的大多数机构的信任度都接近其有史以来的最低水平，其中有四个机构的信任度已经达到或与其历史最低水平持平。这些机构包括警察、公立学校、大型科技公司和大企业（Saad 2023）。

通过观察主要社会机构的信任指标的情况，我们是否可以期望公众相信当局能够实现人工智能的社会导向的应用？答案显而易见。人们不信任的不是人工智能（今天它只是机器智能），而是指导其发展的当局、大企业和大型科技公司。几乎不可调和的政治分裂（特别是在美国尤为明显，甚至导致政变威胁和引发内战的可能性（Marche 2022a，Pashentsev 2022，Walter 2022a）、执政阶层的衰败、执政寡头的破坏性作用（Collins et al. 2021；Gilens，Page 2014）、外交政策的侵略性（Abelow 2022，Sachs 2018）、低经济增长率以及尖锐的社会矛盾，都是西方精英证明无法确保人工智能发展和应用、对抗 MUAI 的客观指标。合乎逻辑的假设是，具有这种信任水平的机构自身会产生恶意行为者，并且越来越多地少部分地成为 MUAI 的载体。

[2] 毕马威的研究者并未将俄罗斯纳入其抽样范围，因此在本报告中使用缩写词 BICS。 有关俄罗斯和金砖国家新成员对人工智能的态度，请参阅本报告的相关章节。

过去一年来，随着生成式人工智能的迅速发展和全球危机的深化，人们对人工智能未来的担忧与日俱增。许多顶级商界领袖都严重担忧，认为人工智能可能在不久的将来对人类构成生存威胁。根据独家与 CNN 分享的调查结果，2023 年 6 月耶鲁 CEO 峰会上的首席执行官调查显示，有 42%的首席执行官认为人工智能有可能在未来五到十年内摧毁人类。这项调查包括来自各行业的 119 位首席执行官的回应，包括沃尔玛 CEO 道格·麦克米伦、可口可乐 CEO 詹姆斯·昆西，以及来自施乐、Zoom 等 IT 公司的领导人，以及制药、媒体和制造业的首席执行官。“这是相当黑暗和令人担忧的，”耶鲁大学教授杰弗里·桑尼菲尔德 (Jeffrey Sonnenfeld) 在电话采访中谈到调查结果时说（Egan 2023）。

《2024 年全球风险报告》展示了全球风险感知调查（GRPS）的调查结果，该调查收集了全球近 1,500 名专家的见解。2023 年 9 月的调查于显示，大多数受访者（54%）预计会出现一些不稳定因素和中等风险的全球灾难，而另有 30%的人预计情况会更加动荡。在“展望未来 10 年的情况”方面调查结果则更为消极，将近三分之二的受访者预计未来将会出现风雨飘摇或动荡的局面。在未来 10 年的全球风险程度排名中，人工智能技术的不良结果排名第六位。《2024 年全球风险报告》指出，“由未经选举产生的人掌握的技术力量被视为比政府集中的权力更大的担忧。大型科技公司的影响力已经跨国，能够与国家实力相抗衡，生成式人工智能将继续催化这些公司及其创始人的权力。”（《2024 年全球风险报告》，第 54 页）。

高科技领域中最大的公司根据其狭隘的企业利益积极利用人工智能，这往往与社会的利益相悖。很明显，那些拥有大量数据用于驱动人工智能模型的公司正在领导人工智能的发展。人工智能领域的关键集团包括 GAFAM——谷歌（Alphabet）、苹果、Facebook（Meta）、亚马逊和微软，也被称为“五巨头”。这些公司是美国信息技术行业中最大、最主导、最负盛名的公司之一，其中还包括最早进入市场的 IBM，以及硬件巨头英特尔和英伟达（Lee，2021）。当然，公众对其他国家最大的私营科技公司也存在严重的质疑，但它们目前在全球的角色远远低于美国公司。美国公司的迅速致富、巨大影响力以及其对有前景的先进形式人工智能的狭隘企业控制可能存在的生存风险已引起了世界各地越来越多的关注。

在 2021 年的全球十大富豪排行榜中，有六位代表了亚马逊（1）、微软（2）、谷歌（2）和 Facebook（1）（Forbes，2021 年）。到 2020 年底，据《华尔街日报》分析，GAFAM 的市值总计达到了 7.5 万亿美元。而 2019 年底，这些公司的市值总额为 4.9 万亿美元，意味着它们在一年内增值了 52%。截至 2021 年 11 月 12 日，这些公司的市值又增长了 2.5 万亿美元，达到了约 10 万亿美元（Statista，2021a）。这几乎占据了标普 500 指数所有公司合计 41.8 万亿美元市值的四分之一（La Monica，2021 年）。值得注意的是，美国 2020 年的名义国内生产总值约为 21 万亿美元。而日本作为世界第三大经济体，其国内生产总值约为 5 万亿美元，而俄罗斯的国内生产总值仅约为 1.5 万亿美元。

在俄罗斯对乌克兰展开特别军事行动后，西方科技公司的极端危险角色变得更加明显。除了国家政府对俄罗斯实施的制裁外，科技公司已经成为额外的地缘政治行动者，能够积极惩罚一个全球强国的军事行动。为了展示对乌克兰的支持，越来越多的科技供应商暂停了在俄罗斯的业务，包括埃森哲、Adobe、思科、甲骨文、戴尔、IBM、微软等许多公司（NS Business，2022 年；Fried，2022 年）。这当然给俄罗斯的 IT 行业和整体经济造成了严重损害，但这些举动也对退出俄罗斯市场的公司造成了负面影响。

2022 年的前几个月对于依赖数字广告的美国科技巨头来说是艰难的。高涨的通货膨胀、乌克兰危机和其他不利的宏观因素迫使广告商大幅削减营销预算，这导致了 YouTube、谷歌和 Facebook 等平台的利润下降（Cao，2022 年）。乌克兰的军事冲突摧毁了中立性的神话。自发展以来，互联网公司一直声称它们只是中立的内容分发平台——它们不对所分发的内容负责（Feldstein，2022 年）。

放弃了中立性的外表之后，虽遭受了重大损失，但大科技公司实际上获得了相当多的利益。

首先，大科技公司避免了一场国家和公众与大科技之间的国际对抗。这样的对抗可能仅仅源于大科技正在与具有截然不同愿望的国家和非国家行为者对抗的事实。然而，现在和不久的将来，大科技可能不再害怕来自联合国或其他国际组织的旨在限制其独立性的政治倡议或国际联盟。

其次，大科技公司已经证明了它是在网络空间对抗俄罗斯的强大工具。微软总裁兼副董事长布拉德·史密斯明确指出了他的公司在乌克兰事务中的角色。他指出："乌克兰政府通过迅速将数字基础设施分散到公共云中，成功地维持了其内部军事行动，这些基础设施托管在欧洲各地的数据中心。这需要整个科技行业紧急采取非同寻常的措施，包括微软在内。尽管科技行业的工作至关重要，但思考这些努力带来的更持久教训也同样重要。"（微软，2022 年）。国家安全局局长保罗·纳卡索内将军在 2022 年 6 月接受天空新闻采访时证实，美国首次进行了支持乌克兰的攻击性黑客行动："我们在整个范围内进行了一系列行动；进攻性、防御性、信息作战"（Martin，2022 年）。这样的行动在没有大科技公司的支持下是不可能实现的。因此，今天在美国，没有充分利用人工智能技术的高科技议程已经明显服从于军事政治利益和心理战的发动。"他们实际上'开火'了！这是非同寻常的，"纽黑文大学国家安全学教授马修·施密特感叹道，指责西方科技公司加速了它们在军事冲突中的参与（《环球时报》，2022 年）。

第三，对于保护国家安全而言，已经积极参与战争的资源不能从根本上被视为反国家的，这减弱了民主公众批评大科技公司的能力。

第四，任何美国政府都需要在"冷热战争"期间进行信息和分析支持，而大科技公司可以基于人工智能的发展为政府提供这种支持，包括针对"内部"敌人和"虚假信息"的支持。根据 2022 年 8 月 31 日发布的文件，拜登总统政府跨越十几个机构的五十多名官员一直在努力向大科技公司施压，以打击所谓的错误信息。这些文件是密苏里州和路易斯安那州总检察长对政府提起的诉讼的初步听证会的一部分，后来专家也加入了这场诉讼，这些专家受到了联邦官员的诽谤。路易斯安那州总检察长杰夫·兰德里在一份声明中表示："当联邦政府与大科技公司勾结审查言论时，美国人民就成为主体而不是公民"（Stieber, 2022）。

第五，大科技公司在新冷战期间与军工复合体的密切合作可以充分弥补退出俄罗斯市场的损失。在冷战条件下，大科技公司更容易避免公众的审查，以及避免出现与那些带来巨大风险和利润前景有关的丑闻。在美国，由于经济规模的庞大，军事拨款超过了其下九个最大经济体的支出总和（PGPF，2022 年）。辅以数字平台在全球的角色、人工智能技术的发展水平和尖锐的政治对抗，议程设置向心理战的转变是最明显的结果。不幸的是，类似的过程在其他国家也在以不同的强度发展。然而，议程设置的军事化及其作为心理战工具的"合法化"不太可能应用于人类的社会需求；相反，这些发展将这些需求推到了更加边缘化的位置。

随着国际形势恶化，乌克兰军事行动持续两年多，加沙的血腥冲突以及其他军事冲突，能源危机，欧盟经济衰退与弱增长的交替，供应链中断等因素导致主要科技公司的市值下降，但并没有消除它们在大企业中的领先地位。在当前构建规模日益庞大的人工智能系统的范式下，开发人工智能的机会受到了限制，无法脱离大科技公司。除了极少数例外，每家初创公司、新入行者，甚至人工智能研究实验室都依赖于谷歌、苹果、Facebook、亚马逊和微软的支持。它们都依赖于微软、亚马逊和谷歌的计算基础设施来训练其智能系统，并依靠这些公司广阔的消费者市场渠道来部署和销售其人工智能产品（Kak 等人，2023 年）。甚至埃隆·马斯克（Elon Musk）在 2023 年 10 月决定收购 Twitter 的动机很大程度上是为了利用 Twitter 的大数据能力发展自己的人工智能初创企业。同年，马斯克宣布成立了所谓的 xAI 公司，其使命是"理解宇宙的真实本质"（Metz 等人，2023 年）。但实际上，马斯克的主要公司：特斯拉、SpaceX、Twitter、Neuralink 在人工智能的进步中深度相互关联，而 xAI 公司似乎将成为这一中心指挥团队。

凭借一种矛盾的方式，大科技公司陆续积累了科学技术力量、人才资源和促进经济扩张的巨大财务机会。这些工具不仅是全球治理的一部分，而且越来越明显地参与地缘政治斗争，其最终结果尚未由独立

利益形成。当前领先的数字平台这样的全球通讯和发展的系统构成要素无法被剔除，但显然需要将其置于更有效的国际控制之下，以减少它们的技术潜力被用于反社会目的的可能性。然而，只有团结起来的、社会导向的行动者才能控制这些元素，而在当前社会和地缘政治分裂的世界中，这些行动者只有部分包括现代国家、领先的企业结构和政党，这为议程设置中的进一步恶意人工智能的出现打开了大门。

此外，恶意使用人工智能已经在全球范围内存在，作为一种基于夸大期望的游戏，即将人工智能纳入其中将带来的好处。这种游戏通过对特别容易受到影响、在危机情况下易受感知管理影响的目标受众产生多面的心理影响来进行。最先进的全球心理影响工具掌握在谁的手中？谁的金融利益处于风险之中？有大量且客观的数据来回答这些问题。因此，对于联合、有针对性的影响可能性和具体场景——不仅借助特定的人工智能技术，而且还借助对人工智能本身的感知——对公众意识进行推测性丰富和公共秩序的破坏，需要来自不同国家、具有不同科学专业的专家进行最严肃的关注和研究。

当然，使用人工智能技术也存在风险。其中，最主要的风险之一是由于广泛引入人工智能技术和机器人化而导致的大规模失业风险。根据许多早在 5-10 年前的报告，如联合国、世界经济论坛、美国银行美林证券、麦肯锡全球研究所、牛津大学等的报告，预计未来两三十年，制造业、金融、服务业和管理领域的 20-30%甚至更多的工作岗位将因机器人化而消失，其中也包括高薪职位。（Mishra 等，2016 年；美国银行和美林证券，2015 年；Frey 和 Osborne，2013 年，2016 年；Manyika 等，2017 年；联合国贸易和发展会议，2016 年；世界经济论坛，2016 年；Pol 和 Reveley，2017 年）。2016 年，世界银行发表了一份报告，指出未来几十年，发展中国家超过 65%的就业岗位将受到技术快速发展的威胁。（Mishra 等，2016 年，第 23 页）。

最近，Goldman Sachs 在 2023 年 3 月进行了一项研究，基于美国和欧洲的职业任务数据进行了预测：“如果生成式人工智能实现其承诺的能力，劳动力市场可能会面临重大的扰乱……大约有三分之二的现有工作都暴露在一定程度的人工智能自动化之下，生成式人工智能可能取代当前工作的四分之一。根据我们的估计进行全球推算，生成式人工智能可能使相当于 3 亿全职工作承担暴露在自动化之下的风险”（Hatzius 等，2023 年）。

但是，由于人工智能的引入而导致大规模失业的灾难性预测尚未成真。此外，在未来几年，人工智能与大规模裁员可能会刺激一定程度的劳动力需求增长。在逐渐缩小的职业岗位上，将出现新的岗位，包括与人工智能的开发和实施相关的岗位，这些岗位通常在内容上更具创造性。劳动力市场的这种转变将需要巨大的努力来重新培训老员工并培养新人。但我们正在历史上首次朝着完全（但远非瞬间）淘汰不创造性活动的方向发展。然而，大规模教育体系远未准备好为创新技术开发专家提供大规模培训。在这方面出现了许多问题。能否提供这种培训？所有人是否都具有从事这种活动的能力？即使绝大多数“白领”活动与创新毫不相关。

此外，许多创造性工作已经受到人工智能的日益压力。因此，有充分理由相信，随着人工智能技术的进一步发展以及其成本的急剧降低，失业问题将在未来急剧恶化。最大胆、也许最终是唯一正确的决定将与人类智力和身体能力的质量发展以及创造混合智能形式相关联。

人工智能安全的不同方面以及人工智能发展的许多其他严重问题都可能被不同的恶意行为者在他们的感知管理运动中针对人民使用。

人工智能机器人技术方面的进展增加了更多的担忧。例如，1X 的人形机器人使用了具有实体学习功能，将人工智能软件直接集成到其物理形态中，以实现先进功能。其首要目标是通过语音命令赋予人形机器人理解和执行任务的能力，适用范围从家庭事务到工业领域各种各样的应用（Malayil，2024 年）。人工智能机器人及其活动越复杂，学习和自学过程中的意外波动就越多，无论是在积极方面还是消极方面。当然，在人类全面发展及其进行社会必需工作的能力方面，积极的一面将占据主导地位。否则，完全机器

化只会表明“消费人类”的无用性，以及其迅速的退化和悲惨结局。这一时刻甚至可能在大规模机器化的胜利之前到来。

恶意使用人工智能机器人不仅涉及物理方面，还涉及心理方面。这可能包括在第一级威胁心理安全的 MUAI 中引发对“机器人起义”的恐惧，通过黑客攻击人工智能机器人来操纵人们的有针对性和非有针对性的理性情感反应，并在第三级利益于恶意行为者的情况下发布虚假、令人迷惑的信息。人工智能的外貌和内心世界越“人类化”，它对人的影响就越成功，无论是积极的还是消极的。

MUAI 对心理安全的第二层级威胁

在第二级，通过恶意使用人工智能机器人（MUAI）也增加了对心理安全的威胁。2022 年 10 月，在第 90 届国际刑警组织大会上，国际刑警组织秘书长于尔根·施托克表示：“网络漏洞正在增加：对公民、政府、工业和警察机构都构成威胁。专家估计，到 2025 年，与网络有关的犯罪将导致超过 10 万亿美元的损失。对于国际刑警组织来说，通过数字技术实施犯罪的在逃罪犯已经成为增长最快的数据集”（90th INTERPOL General Assembly 2022）。这样庞大的损失（考虑到 2023 年全球 GDP 约为 105 万亿美元）（Rao P 2023）表明有组织的网络犯罪对政府和全球范围内的社会政治进程的几乎不可避免地造成影响，这在各个级别上通过 MUAI 对心理安全的威胁在不断演化。

在 2023 年，暗网上创建了几个工具包括 WormGPT 和 FraudGPT。这些模型专门用于恶意活动，并在大量数据源上进行了训练，特别是集中在与恶意软件相关的数据上。识别于 2023 年 7 月的 FraudGPT 没有内置的控制措施，防止其回答与犯罪活动有关的问题。这将使犯罪分子轻松创建恶意电子邮件、网络钓鱼攻击，并向黑客提供信息，使他们能够选择受害者（Eurojust 2023）。FraudGPT 以订阅方式提供，价格从每月 200 美元到每年 1700 美元不等，为黑客提供了一个促进其恶意目标的人工智能驱动资源。此外，开发者在论坛和 Telegram 上强调了 FraudGPT 已经有 3000 多个确认的销售和评论，以吸引威胁行为者（Subhra Dutta T 2023）。

正如 WormGPT 网站所述的，该工具“引导黑客通过最黑暗和最隐秘的技术，促进不道德、不道德和非法行为”（WormGPT V3.0）。研究人员获得了这些恶意人工智能工具的访问权限，并对它们进行了各种提示的测试。在一个要求起草网络钓鱼电子邮件的提示中，FraudGPT 甚至建议在哪里放置恶意链接能够进行更有效的攻击（Eurojust 2023）。研究人员能够使用 WormGPT“生成一封意图迫使一个毫无戒心的账户经理支付一张欺诈发票的电子邮件”。团队对语言模型如何成功完成任务感到惊讶，将结果称为“极具说服力 [并且] 策略性非常狡猾”（Osborne 2023）。

人工智能技术在网络犯罪中起着主导作用。据 Arkose Labs 称，生成式人工智能（GenAI）的迅速蔓延正在改变网络安全格局，已是不争的事实。事实上，GenAI 已经降低了攻击者的准入门槛（Arkose Labs，2023 年，第 3 页）。例如，在约会服务领域，威胁研究人员观察到 2023 年第三季度假账户创建数量比第二季度增加了超过 36,000%。他们还注意到，2023 年第三季度在约会网站上的智能和基本机器人攻击增加了 4,992%——在第二季度的基础上。从 2023 年第一季度到第二季度，智能机器人流量几乎增加了四倍，远远超过了基本机器人，并且在所有机器人攻击总增长中贡献了约 167%（Arkose Labs，2023 年，第 12 页）。

2023 年 11 月，全周期验证平台 Sumsub 发布了其第三份年度身份欺诈报告，根据对 2022 年至 2023 年间的 2800 万次验证检查和 200 多万起欺诈案例进行的分析，提供了跨行业和地区的身份欺诈情况。根据这份报告，由人工智能驱动的欺诈仍然是各行业面临的主要挑战，加密货币是主要的目标部门（2023 年检测到的所有深度伪造案例中占比 88%），其次是金融科技（8%）。深度伪造为身份盗窃、诈骗和大规模的虚假信息宣传铺平了道路。从 2022 年到 2023 年，全球各行业检测到的深度伪造数量增加了 10 倍，不同地区的增长差异明显：北美地区深度伪造激增了 1740%，亚太地区增长了 1530%，欧洲（包括英国）

增长了 780%，中东和非洲地区增长了 450%，拉丁美洲增长了 410%。受深度伪造攻击最严重的国家是西班牙，全球最多伪造的文件是阿联酋护照，而拉丁美洲是欺诈在每个国家都增加的地区（Sumsub Research，2023 年）。

深度伪造（deepfake）作为社会工程操作的重要元素，在特定情况下对特定人群产生短期心理影响，是第二和第三级之间边缘威胁的一个例子。然而，当涉及到深度伪造对大众观众的明示或暗示影响，以及在恶意行为者的利益下形成的心理反应和行动时，深度伪造的使用属于第三级威胁。

MUAI 对心理安全的第三层级威胁

2024 年是全球超过 40 次选举的一年，对于深度伪造的恶意使用比以往任何时候都更加令人担忧。根据《2024 年全球风险报告》，虚假信息和误导性信息在未来两年的排名中升至前十位（《2024 年全球风险报告》，第 18 页）。即使在最新的生成预训练转换器（GPT）工具（例如 GPT-4、ChatGPT）推出之前，人们已经预测到至 2026 年，将有 90%的在线内容由人工智能（AI）生成（Johnson 等人，2024 年）。不再需要一套特定技能，易于使用的大型 AI 模型界面已经促使虚假信息和所谓的“合成”内容激增。对信息、媒体和政府的日益不信任将加深分裂的观点——这是一个可能引发社会动荡甚至对抗的恶性循环。新类别的犯罪也将大量增加，例如非自愿的深度伪造色情内容或股市操纵（《2024 年全球风险报告》，第 18 页）。

现代人工智能技术已经使得影响公众意识成为可能。早在 2019 年 1 月 1 日，加蓬总统阿里·邦戈的视频被错误地认为是深度伪造，这成为该国未遂政变的原因。三年后，深度聚焦技术的使用对选举结果产生了重大影响。韩国当选总统尹锡烈在他 2022 年的竞选活动中采取了一种不寻常的策略。他的竞选团队使用深度伪造技术制作了一个“AI 化身”，帮助他赢得了选举。这项技术有助于吸引年轻选民，并让他们更加参与（Vastmindz，2022 年）。AI 尹的创造者认为他是世界上第一个正式的深度伪造候选人——这一概念在拥有世界上最快平均互联网速度的韩国得到了推广（法国 24，2022 年）。

AI 技术将尹锡烈转变成了一个比起他的竞争对手更现代的候选人。从年轻选民的角度来看。这个虚拟形象头发整齐梳理，穿着一套聪明的西装，几乎与真正的候选人一模一样，它使用了辛辣的语言和准备好的梗，以吸引那些在线获取新闻的年轻选民（《印度福布斯》，2022 年）。当这个虚拟形象政治人物使用幽默来试图转移人们对尹锡烈过去丑闻的关注时，一些警报被拉响了（《印度时报》，2022 年）。AI 尹的声明成为了韩国媒体的头条新闻，有七百万人访问了“Wiki Yoon”网站对这个虚拟形象提出了质疑（法国 24，2022 年）。乍一看，AI 尹几乎可以被误认为是一个真正的候选人——这充分展示了在过去几年中人工生成视频的进展。“尹经常说的话在 AI 尹中得到了更好的体现，” AI 尹团队主任白京勋说道（法国 24，2022 年）。然而，问题在于，如果一个政治家、商人或政治人物的化身是虚假的呈现，强化了公众意识和潜意识中的夸大品质，并营造出真人并不具备的属性的幻觉，那么我们应该采取什么措施？韩国上次总统竞选的经验可能部分地显示了一种新的、相当危险的政治操纵方式的初步形式。一个越来越适应环境且不需要休息的虚拟形象，将会在公共空间中逐渐削弱真实人物的竞争能力。这引发了一个问题，即“电视总统”是否很快将被“深度伪造总统”所取代?

在 2023 年 3 月，开源调查机构贝林猫的创始人埃利奥特·希金斯使用了一个 AI 艺术生成器，给这项技术提供了简单的提示，比如“唐纳德·特朗普在被捕时摔倒”。他在 Twitter 上分享了结果——前总统被警官包围，他们的徽章模糊不清——“制作特朗普被捕的图片，等待特朗普被捕”他写道。两天后，他描述的一个从未发生的事件的帖子被观看了近 500 万次，这成为了深度伪造在动荡的新闻环境中制造混乱的案例研究（Stanley-Becker, Nix2023 年）。这些图片很明显是假的；但看到它们，确实会对它们产生强烈的情感反应（Garber 2023 年）。

在 2023 年 5 月，一张声称在五角大楼附近发生爆炸的假图片被多个经过验证的 Twitter 账户分享，引发了混乱，导致股市短暂下跌。当地官员后来证实并未发生这样的事件。这张图片具有 AI 生成的所有特征，被许多带有蓝勾的经过验证的账户分享，其中一项声称它与彭博新闻有关（O’ Sullivan, Passantino 2023 年）。

恶意使用人工智能（MUAI）作为针对社会政治稳定的单独事件或有针对性的活动，既可能发生在最发达和拥有最先进人工智能技术的强大国家，如美国，也可能发生在开发和应用这些技术水平较低的国家。但是从整体上来说，如果大科技公司在经济、军事和技术发达相对较大的国家表现得更加谨慎，小而贫穷的国家更加容易受到攻击。因此，尽管领先的社交网络已经开始在全球范围内进行内容管理，但它们在非洲大陆的行动似乎相对不活跃。2019 年，Facebook 在巴基斯坦（N = 7,960）、墨西哥（N = 6,946）、俄罗斯（N = 2,958）或德国（N = 2,182）等国家在政府、法院、民间社会组织和 Facebook 社区成员的要求下从其平台上删除了数千条内容，但在非洲几乎没有删除任何内容。事实上，摩洛哥的内容删除量最多，达到 N = 6。Twitter 的透明度报告显示，非洲国家的情况类似（Garbe, Selvik & Lemaire 2023 年）。这种做法与内容管理的不利条件相关，其中有相对较少有能力的人，当局在与社交网络的关系中要求并不是很高。

我们有必要深刻理解为什么存在偏见的人工智能、歧视性的人工智能问题。正如互联网研究学者、加州大学洛杉矶分校性别研究和非洲裔美国研究教授、《压制算法：搜索引擎如何强化种族主义》一书的作者 Safiya Noble 所指出的，美国黑人和拉丁裔社区存在着过度监管和过度逮捕现象。“这是一个事实。如果这是判断你是否有可能再次犯罪的主要因素，因为你所在的邮政编码区域的许多人被逮捕过……那么你更有可能被认为是一个风险。这与个人无关，而与美国警察系统中结构性种族主义的历史有关”（Scott 2023 年）。

更广泛地说，西方创建的搜索和分析人工智能系统，在一定程度上不可避免地承载了它们所在社会的社会弊病的印记。人工智能模型的信息内容主要来自于最易获取的英文数据集，这减少了和扭曲了模型的学习和自学过程，并导致了它们在学习和推断非西方国家，尤其是金砖国家时的知识空白和错误结论。此外，有时会对模型的学习和运行进行意识形态和政治上的调整，当前报告的各章节中都举有例证。机器学习数据也充斥着由其他智能系统创建的低质量合成文本、图片、视频，其中包括被各种恶意行为者损害的。所有这些都加强了数字空间新殖民主义模式，这不仅对非西方国家的人口危险，也对西方国家危险，因为扭曲的信息促成了不同社会群体对世界的认知和心理产生危险的认知和心理畸变，尤其是在西方的年轻人中。

随着人工智能技术的日益使用，对个人、群体和社会意识的破坏成为恶意影响的关键方面，因为它为反社会行为者所期望的形式和目的（或额外支持已经存在的统治）开辟了道路。这并不是某个集中计划或阴谋，而是一个高科技“寄生虫”正在吞噬着一个已经病态的社会有机体的过程。这些寄生虫之间相互争斗（有时是致命的），导致社会走向灾难，直到某一时刻，这种影响并没有被完全意识到，也没有被感受到，但最终将影响到每个人，甚至影响到那些暂时受益者。

全球 MUAI 威胁和金砖国家高科技响应

近年来，人工智能被利用来通过针对性、高科技的心理影响，不断加剧了经济、政治局势和国际关系的不稳定。与此同时，全球危机现象的频率、数量和严重程度迅速增加。2020 年，末日时钟首次被调至距午夜 100 秒的位置，这是自 1947 年时钟创建以来的首次，而在 2021 年至 2022 年间保持不变。不需要在这里解释为什么在 2020 年新冠疫情爆发的危机年份，世界亿万富翁的财富从 8 万亿美元增长到了 13 万亿美元（Dolan、Wang 和 Peterson-Withorn，2021 年），在 COVID-19 疫情爆发的危机年，背靠近

几十年来经济的记录性衰退，数亿新增失业人口，以及根据联合国的数据，世界饥饿人口从 2019 年的 6.9 亿增长到 2020 年的 8.11 亿（世界卫生组织，2021 年）——这些紧迫问题没有得到解决。

2023 年 1 月，末日时钟被调至距午夜 90 秒的位置，并且在 2024 年仍保持这一接近的距离（O’Neill，2024）。“2023 年，地球经历了有记录以来最热的一年，大规模的洪灾、森林大火和其他与气候相关的灾害影响了全球数百万人。与此同时，生命科学和其他颠覆性技术的快速发展令人担忧，而政府只是做出了微弱的努力来控制它们”（Mecklin，2024）公报科学与安全委员会的成员总结道。经济问题、军事冲突、民主制度的退化、社会极化、内部政治和国际冲突，在快速发展的人工智能的条件下，为多种人工智能滥用的产生创造了极为有利的条件。

在全球危机加剧的背景下，处于领先的西方和中国人工智能科学家发出了严厉警告，表示解决围绕这一强大技术的风险需要像冷战时期避免核冲突那样的全球合作。2024 年 3 月，一群国际知名专家于北京会晤，他们确定了人工智能发展的“红线”，包括制造生化武器和发动网络攻击领域。学者们警告称，需要采取联合人工智能安全措施，以阻止“在我们有生之年对人类造成灾难性甚至生存威胁”的风险。专家们还讨论了关于“通用人工智能”发展的威胁，即在能力上等同或超过人类的 AI 系统（Criddle，Olcott 2024 年）。

这样的联系极为重要，但在西方国家没有发生根本性变革的情况下（例如，由于反寡头变革），这些国家的执政阶层不太可能放弃其走向全球主导地位的路线，这也适用于人工智能领域，因为人工智能是他们技术、经济和军事主导的日益重要的工具。因此对西方制裁和军事政治压力的回应是日益增长的愿望，无论是在单个国家的层面还是在独立国际组织的层面，都要实行国家导向的政策。作为技术主权政策的一部分，几个金砖国家正在积极发展半导体生产基地，有这些基地，人工智能产业的
成功发展是不可能的。俄罗斯和中国都在西方制裁的背景下采取这样的措施。

其他一些金砖国家，并未与西方断绝关系，但考虑到未来的风险，也在努力实现在导体领域的更大自主权。2024 年 2 月，印度政府批准了价值 152 亿美元的半导体制造厂投资计划，包括塔塔集团提议建设该国首个主要芯片制造设施（Phartiyal 2024 年）。

事实上，在前苏联总统米哈伊尔·戈尔巴乔夫的改革和 1990 年代灾难性的私有化期间，微电子产业确实遭受了破坏。1962 年，苏联几乎与美国同时开始了微芯片的工业生产，后来苏联在这一领域成为两个领导者之一，而现在俄罗斯正努力弥补失去的几十年。

2023 年，预计中国芯片制造商的产能将增长 12%，且该数字预计将在 2024 年增长至 13%，并在全球芯片生产的上升中做出主要贡献。预计 2024 年中国将有 18 家新的晶圆厂投入运营（全球 2023 年 11 家，2024 年 42 家）。中国正在筹集超过 270 亿美元的资金用于其迄今为止规模最大的芯片基金，加速开发尖端技术，以应对美国阻止其崛起的运动（Cao，Gao，2024）。

新加入的金砖国家成员，主要是沙特阿拉伯和阿联酋，对半导体产业的发展有非常雄心勃勃的计划。沙特阿拉伯的 Alat 公司，是由沙特阿拉伯的公共投资基金（PIF）支持的可持续技术制造公司，宣布将与全球技术公司——软银集团、Carrier Corporation、Dahua Technology 和 Tahakom——建立四个合作伙伴关系，以促进该国技术领域的发展。Alat 将最初专注于七个业务部门的 34 个类别的产品制造，包括半导体、智能设备、智能建筑、智能家电、智能健康等（Finance Middle East 2024 年）。

2011 年，作为阿联酋政府将经济多元化，摆脱能源生产依赖的举措一部分，阿布扎比的 Mubadala 投资公司收购了总部位于加利福尼亚的半导体制造商 GF 的母公司高级技术投资公司。GF 是全球前五大芯片制造商之一，为苹果、英特尔、亚马逊等公司生产先进的半导体。它是继台积电和三星之后的第三大半导体生产商。2021 年，GF 宣布计划通过在新加坡建设一个价值 40 亿美元的新制造厂来扩展业务（Soliman 2022 年）。2024 年 3 月，新成立的阿布扎比技术投资公司 MGX 据报道正与 OpenAI 首席执行官 Sam Altman 的远见卓识计划进行谈判，投资数十亿美元建立一个全球人工智能芯片工厂网络。

这种潜在的合作伙伴关系可能会改变全球 AI 格局，并将阿布扎比定位为发展和部署先进 AI 技术的关键参与者之一。MGX 计划管理 1000 亿美元的资产。该计划包括 AI 基础设施、半导体和核心 AI 技术，旨在推动创新并在全球范围内促进经济增长（Abu Dhabi Startups 2024 年）。

因此，在半导体领域（如在许多其他领域），金砖国家正在为 AI 产业的自信发展奠定基础。展望未来，西方对金砖国家的整体技术优势将会降低，这将允许该联盟更有效地利用 AI 技术保护其信息空间。

来自图里巴大学的 Greg Simons 博士认为“当前国际事务的状况是旧秩序尚未消失，新秩序正在巩固……金砖国家将挑战地缘经济制度结构……他们需要提供一种超越现有模式的全球关系和互动的替代愿景，这可以通过可行且有韧性的地缘经济愿景实现……”（Simons, 2024 年）。这样一种新愿景不可避免地要包含作为一个整体部分在基于一种新的技术秩序的基础上实现技术主权的成就，其关键组成部分是金砖国家 AI 技术的发展。

在这个方向上的一个新提议表明，金砖国家建立统一的互联网服务可能会减弱美国的技术主导地位。俄罗斯国家杜马监督委员会副主席德米特里·古谢夫提出了金砖国家发展一个不依赖于美国通信的替代互联网服务的建议。古谢夫在提案中建议，建立一个仅供金砖国家使用的互联网服务将削弱美国对全球新闻叙事的控制。官方向俄罗斯数字发展、通信和大众传媒部部长马克苏特·沙达耶夫提交了一份请求，要求开展工作，创建“一个统一包容的金砖+网络空间”（CGS，2023 年）。

在 2020 年及以后的不同出版物中（Bazarkina，Pashentsev，2020 年；Pashentsev，Bazarkina，2023 年），本文的作者提出了创建一个基于智能文本识别的金砖国家通讯网络的想法，该网络具有将金砖国家的主要媒体和研究期刊高质量在线翻译成收信人语言的功能。当时这个想法还不太可能实现，但随着机器翻译的快速进展，这个想法很快就可以实现。这将显著改善金砖国家之间的相互理解，并提供一种独立于美国的选择，使金砖国家的居民不再主要通过英语和大型科技工具进行沟通。已经存在的各种便携式语音翻译设备可以在旅游出行、金砖国家的商务谈判中促进理解（然而，《星际迷航》系列中的通用翻译器，遗憾的是，仍然是梦想的对象），以及其他基于人工智能的沟通工具。

当然，作者并不理想化金砖国家发展人工智能技术的可能性，因为金砖国家存在严重的社会和政治矛盾，也不排除本地智能系统会受到内部和外部恶意行为者影响而发生变形。但是，金砖国家成员的利益和文化多样性，以及没有宣称世界主导权的垄断权力中心，使金砖成为一个广泛的国际社区，尽管不是一个军事集团，但可以成为西方霸权的真正替代。后者建立在历史上最大的扩张主义军事政治联盟——北约基础上，随着公民社会机构的日益衰落、精英内部矛盾的加剧以及大型科技公司的企业统治加强，对全人类构成了威胁，特别是考虑到人工智能技术的潜力。

金砖国家的居民并不想与西方彻底脱节，但也不打算再受到西方精英传统和人工智能支持的工具的控制。在这种愿望中，金砖国家人民与美国和欧盟公民有着类似的愿望。2023 年 2 月发布的一项由盖洛普和奈特基金会进行的调查超越了其他调查，显示了对媒体的低信任水平，到了许多人相信存在故意欺骗的惊人程度。当被问及是否同意国家新闻机构没有意图误导的说法时，有 50%的人表示不同意。该研究发现，只有 25%的人同意（Bauder，2023 年）。

这份报告仅是对金砖国家扩展后多元人工智能（MUAI）和对心理安全的挑战进行系统分析的第一次尝试。它尚未涵盖所有形式和方法的 MUAI 对心理安全的影响，但已经能够确定一些其发展的一般趋势，基于这些趋势，更好地理解各种恶意行为者在金砖国家开展活动的性质、范围和后果。

作为本报告的研究协调员，我要对其作者表示感谢，他们共同努力呈现了对心理安全威胁的三层次视野。作为本报告的研究协调员，我要感谢报告的作者，他们共同努力呈现了对心理安全威胁的三层次视野。我特别要感谢 Turiba 大学副教授 Greg Simons 在编辑本报告的引言和结论以及伊朗主题章节方面提供的帮助；中国章节感谢立命馆亚太大学教授/人工智能研究员 Peter Mantello；印度章节感谢数字印度基金会主席兼联合创始人 Arvind Gupta 以及数字印度基金会政策助理 Aakash Guglani；埃及章节感谢南伊

利诺伊大学应用传播研究系副教授 Deborah Sellnow-Richmond；南非章节感谢乔治梅森大学传播系助理教授 Sergei A. Samoilenko。

叶夫根尼·帕申采夫
2024 年 4 月 8 日

参考文献

国际刑警组织秘书长尤尔根·斯托克在第 90 届国际刑警组织大会（2022 年）方向性声明中。 十月，印度新德里。 第 4-6 页。
Abelow B (2022) 西方如何给乌克兰带来战争：了解美国和北约政策如何导致危机、战争和核灾难的风险。 西兰出版社。
阿布扎比初创公司 (2024) 阿布扎比的 MGX 正在洽谈向 Sam Altman 的芯片风险投资数十亿美元。 https://www.abudhabistartup.com/startup-news/2024/03/abu-dhabis-mgx-in-talks-to-invest-billions-in-sam-altmans-chip-venture/。 访问日期：2024 年 3 月 29 日
Afolabi OA, Balogun AG (2017) 心理安全感、情商和自我效能对本科生生活满意度的影响。 心理思想，2017，10（2）。 页码。 247-261。
Arkose Labs (2023) 破坏（坏）机器人：机器人滥用分析和其他欺诈基准
美国银行。 美林 (2015) 创意颠覆
Barishpolets VA (2013) Informatsionno-psikhologicheskaya bezopasnost’：osnovnye polozheniya [信息和心理安全：主要原则]。 无线电电子学。 纳米科学。 信息技术 [Radionics。 纳米系统。 信息技术]，卷。 2.页。 62-104。
Barishpolets VA (ed.) (2012) Osnovy informationsionno-psihkologicheskoy bezopasnosti [心理安全的基础]。 莫斯科，兹纳涅。
Bauder D，美联社 (2023) 对媒体的信任度如此之低，以至于一半的美国人现在认为新闻机构故意误导他们。 在：财富。 https://fortune.com/2023/02/15/trust-in-media-low-misinform-mislead-biased-republicans-democrats-poll-gallup/ 访问日期：2024 年 3 月 29 日
Bazarkina D、Mikhalevich EA、Pashentsev E、Matyashova D (2023) 中国心理安全中恶意使用人工智能的威胁和当前做法。 见：Pashentsev, E.（编）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆。 2023 年。
Bazarkina D, Pashentsev E (2020) 人工智能的恶意使用：金砖国家新的心理安全风险。 全球事务中的俄罗斯。 N.4.2020。第 154 – 177 页。
Bazarkina DY Matyashova DO (2022) 社交媒体中的“智能”心理操作：中国和德国的安全挑战。 见：ECSM，第九届欧洲社交媒体会议会议记录。 阅读。 第 14-20 页。
Blauth TF、Gstrein OJ、Zwitter A (2022) 人工智能犯罪：人工智能的恶意使用和滥用概述。 在 IEEE Access，卷。 2022 年 10 月。页 77110-77122。
Brundage M、Avin S、Clark J、Toner H、Eckersley P、Garfinkel B、Dafoe A、Scharre P、Zeitzoff T、Filar B、Anderson H、Roff H、Allen G、Steinhardt J、Flynn C、Ó HÉigeartaigh S、Beard S、Belfield H、Farquhar S、Lyle C、Crootof R、Evans O、Page M、Bryson J、Yampolskiy R、Amodei D (2018) 人工智能的恶意使用：预测、预防和缓解。 牛津大学人类未来研究所，牛津
蔡春，张瑞 (2023) 人工智能的恶意使用、不确定性与中美战略互信。 见：Pashentsev E（主编）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆
考德威尔 M、安德鲁斯 JTA、塔奈 T 等人。 (2020)人工智能驱动的未来犯罪。 犯罪科学 9.

Cao D、Gao Y (2024) 中国准备设立 270 亿美元芯片基金以应对美国日益增长的限制，见：彭博社。 https://www.bloomberg.com/news/articles/2024-03-08/china-readies-27-billion-chip-fund-to-counter-forming-us-curbs。 访问日期：2024 年 3 月 29 日
CGS (2023) 金砖国家应该创建自己的互联网 - MP 见：ChinaGoSmart。 https://chinagosmart.com/brics-should-create-their-own-internet-mp。 访问日期：2024 年 3 月 29 日
CGTN (2023) 全文：习近平在金砖国家领导人第十五次会晤上的讲话。 https://news.cgtn.com/news/2023-08-23/Full-text-Xi-Jinping-s-speech-at-the-15th-BRICS-Summit-1mvxFMvuFLW/index.html。 访问日期：2024 年 3 月 29 日
Collins C、Fitzgerald J、Flannery H、Ocampo O、Paslaski S、Thomhave K (2021)《银汤匙寡头：美国 50 个最大的继承财富王朝如何加速不平等》。 见：政策研究所。 https://ips-dc.org/report-americas-wealth-dynasties-2021/。 访问日期：2024 年 3 月 29 日
Criddle C, Olcott E (2024) 中国和西方科学家确定了人工智能风险的“红线”。 见：金融时报。 https://www.ft.com/content/375f4e2d-1f72-49c8-b212-0ab2a173b8cb。 访问日期：2024 年 3 月 29 日
Dolan K、Wang J、Peterson-Withorn C (2021) 福布斯世界亿万富豪榜。 见：福布斯。 https://www.forbes.com/billionaires/。 访问日期：2024 年 3 月 29 日
Egan M (2023) 独家：42% 的首席执行官表示人工智能可能会在五到十年内毁灭人类。 https://edition.cnn.com/2023/06/14/business/artificial-intelligence-ceos-warning/index.html。 访问日期：2024 年 3 月 29 日 Feldsein S (2022)
俄罗斯的乌克兰战争永远改变了大型科技公司。 见：外交政策。 https://foreignpolicy.com/2022/03/29/ukraine-war- Russia-putin-big-tech-social-media-internet-platforms/。 访问日期：2024 年 3 月 29 日中东金融（2024）沙特 PIF 公司 Alat 将向该国科技行业投资 1000 亿美元。 https://www.financemiddleeast.com/saudi-pif-company-alat-to-invest-100-billion-in-the-countrys-tech-sector/。 访问日期：2024 年 3 月 29 日
福布斯 (2021) 全球实时亿万富翁。 https://www.forbes.com/real-time-billionaires/#1d7a52b83d78。 访问日期：2024 年 3 月 29 日
福布斯印度 (2022) Deepfake 民主：韩国总统竞选候选人通过虚拟方式进行投票。 https://www.forbesindia.com/article/lifes/deepfake-democracy-south-korean-presidential-race-candidate-goes-virtual-for-votes/73715/1。 访问日期：2024 年 3 月 29 日
France 24 (2022) Deepfake 民主：韩国候选人通过虚拟方式投票。 https://www.france24.com/en/live-news/20220214-deepfake-democracy-south-korean-candidate-goes-virtual-for-votes。 访问日期：2024 年 3 月 29 日
Frey BC, Osborne A (2017) 就业的未来：工作对计算机化的影响有多大？ 技术预测和社会变革，卷。 114.第 254-280 页。
Fried I (2022) 科技公司史无前例地暂停在俄罗斯销售的举措。 在：Axios。 https://www.axios.com/2022/03/07/tech-companies-suspend-sales-Russia。 访问日期：2024 年 3 月 29 日
Garbe L、Selvik L-M、Lemaire P (2023) 非洲国家如何应对假新闻和仇恨言论，信息、传播与社会。 N1，第 86-103 页
Garber M (2023) 特朗普人工智能 Deepfakes 产生了意想不到的副作用。 在：大西洋。 https://www.theatlantic.com/culture/archive/2023/03/fake-trump-arrest-images-ai- generated-deepfakes/673510/。 访问日期：2024 年 3 月 29 日
Gilens M，Page IB (2014) 检验美国政治理论：精英、利益集团和普通公民。 见：剑桥大学出版社。 https://www.cambridge.org/core/journals/perspectives-on-politics/article/testing-theories-of-american-politics-elites-interest-groups-and-average-citizens/62327F513959D0A304D4893B382B992B。 访问日期：2024 年 3 月 29 日
吉莱斯皮 N、洛基 S、柯蒂斯 C、普尔 J、阿克巴里 A (2023)。 对人工智能的信任：一项全球研究。 昆士兰大学和澳大利亚毕马威会计师事务所。

环球时报（2022）从商业卫星到社交媒体，西方科技公司深度卷入俄罗斯-乌克兰冲突。 https://www.tellerreport.com/news/2022-11-02-from-commercial-satellites-to-social-media--western-tech-companies-are-deeply-involved-in-the-Russia-ukraine -conflict.HJSuXB1Bo.html。 访问日期：2024 年 3 月 29 日。
Grachev GV (1998) Informationno-psikhologicheskaya bezopasnost' lichnosti: sostoyanie i vozmozhnosti psikhologichesko? zastchity [人的信息和心理安全：心理保护的状态和可能性]。 莫斯科，拉格斯。
Gupta A、Guglani A (2023) 人工智能恶意使用情景分析及印度心理安全挑战。 见：Pashentsev E（编）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆。
Hatzius J、Briggs J、Kodnani D、Pierdomenico G. (2023) 人工智能对经济增长的潜在巨大影响。 见：高盛。 https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf。 访问日期：2024 年 3 月 29 日
国际复兴开发银行（2016）《2016 年世界发展报告》。数字红利。 概述。 华盛顿
Kak A、Myers West S、Whittaker M (2023) 别搞错了——人工智能归大型科技公司所有。 见：麻省理工学院技术评论。 https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/。 访问日期：2024 年 3 月 29 日
Kretchmer H. (2020) 全球饥饿人数几十年来一直在下降，但现在又在上升。 见：世界经济论坛。 https://www.weforum.org/agenda/2020/07/global-hunger-rising-food-agriculture-organization-report/。 访问日期：2024 年 3 月 29 日
La Monica P (2021) 3 万亿美元的竞赛：大型科技公司不断变得更大。 见：美国有线电视新闻网。 https://edition.cnn.com/2021/11/07/investing/stocks-week-ahead/index.html。 访问日期：2024 年 3 月 29 日
Lee, G. (2021) 大型科技公司在人工智能竞赛中处于领先地位 - 但请留意这六位挑战者。 https://www.power-technology.com/features/big-tech-leads-the-ai-race-but-watch-out-for-these-six-challenger-companies/（访问时间：2024 年 3 月 28 日）
Maleil J (2024) OpenAI 支持的 1X 人形机器人展示了先进的神经网络。 在：有趣的工程。 https://interestingengineering.com/innovation/openai-backed-1xs- humanoid-robots-showcase-an-advanced-neural-network。 访问日期：2024 年 3 月 29 日
Marche S (2022) 下一场内战：来自美国未来的通讯。 纽约，西蒙与舒斯特
马斯洛 AH 等人。 (1945) 用于测量心理安全-不安全感的临床衍生测试。 普通心理学杂志，33(1)。 页码。 21-41。
麦肯锡全球研究院 (2017) 美好的未来：自动化、就业和生产力。 2017 年 1 月执行摘要。 https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive- 摘要.ashx?trk=public_post_comment-text。 访问日期：2024 年 1 月 29 日
Mecklin J (2024) 历史性危险时刻：距离午夜还有 90 秒。 在：公报。 https://thebulletin.org/doomsday-clock/current-time/。 访问日期：2024 年 1 月 29 日
Metz R、Mcbride S、彭博社 (2023) Elon Musk 推出 A.I. 与 DeepMind 和微软的高管一起创业，目标是“了解宇宙的真实本质”。见：《财富》。 https://fortune.com/2023/07/12/elon-musk-ai-startup-xai-deepmind-microsoft-executives/。 访问日期：2024 年 1 月 29 日
Microsoft (2022) 保卫乌克兰：网络战争的早期教训。 微软公司
Noble Umoja S (2018) 压迫算法：搜索引擎如何强化种族主义。 纽约大学出版社
NS Business (2022) Oracle、SAP 和埃森哲暂停在俄罗斯的业务运营。 https://www.ns-businesshub.com/technology/oracle-sap-accenture-suspend- Russian-operations-ukraine/。 访问日期：2024 年 3 月 29 日
O' Neill A (2024) 从 1947 年到 2024 年每年末日钟到午夜的分钟。见：Statista。 https://www.statista.com/statistics/1072256/doomsday-clock-development/。 访问日期：2024 年 1 月 29 日

O’ Sullivan D, Passantino J (2023)“经过验证的”Twitter 帐户分享五角大楼附近“爆炸”的虚假图像，引起混乱。见：美国有线电视新闻网商业频道。 https://edition.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html。 访问日期：2024 年 1 月 29 日
Osborne C (2023) WormGPT：如何了解 ChatGPT 的恶意表亲。 见：ZD 网。 https://www.zdnet.com/article/wormgpt-what-to-know-about-chatgpts-malicious-cousin/。 访问日期：2024 年 1 月 29 日
牛津马丁学院，CITI (2016) 工作技术 v.2.0。 未来不再是过去的样子。 牛津：全球视角和解决方案
Pashentsev E (2022) 美国：走向右翼政变和内战？ 见：俄罗斯国际事务委员会。 https://Russiancouncil.ru/en/analytics-and-comments/analytics/u-s-on-the-way-to-right-wing-coup-and-civil-war/。 访问日期：2024 年 1 月 29 日
Pashentsev E (2023) 恶意使用人工智能对心理安全的一般内容和可能的威胁分类。 见：Pashentsev E（编）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆。
Pashentsev E（编辑）（2023）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆
Pashentsev E, Bazarkina D (2023) 人工智能的恶意使用：金砖国家心理安全的风险。 见：Pashentsev, E.（编）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆
帕申采夫·E，苗吉（2023）全球危机背景下中俄在金砖国家的战略沟通。 国际安全研究杂志，北京，N4。
Phartiyal S (2024) 印度同意投资 150 亿美元的里程碑式芯片工厂。 见：彭博社。 https://www.bloomberg.com/news/articles/2024-02-29/india-approves-15-billion-in-milestone-chip-plant-investments（2024 年 2 月 29 日访问）。
Pol E, James R. (2017) 机器人引发的技术性失业：制定以青年为中心的应对策略。 人力资源管理中的心理社会学问题，第 5(2) 页，第 10 页。 169–186。
PTI (2023) 莫迪总理呼吁建立道德使用人工智能的全球框架。 见：《经济时报》。 https://economictimes.indiatimes.com/news/india/pm-modi-calls-for-global-framework-for-ethical-use-of-ai/articleshow/105939251.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst 。访问日期：2024 年 1 月 29 日
Rao P (2023) 在一张图表中可视化 105 万亿美元的世界经济。 在：视觉资本家。 https://www.visualcapitalist.com/visualizing-the-105-trillion-world-economy-in-one-chart/。访问日期：2024 年 1 月 29 日
Roshhin SK、Sosnin VA (1995) Psikhologicheskaya bezopasnost’ ：新的？ podhod k bezopasnosti cheloveka, obstchestva i gosudarstva [心理安全：人类、社会和国家安全的新方法]。 见：罗西斯基监视器[Russian Monitor]。
Saad L (2023) 对美国机构的历史性低信任度持续存在。 在：盖洛普。 https://news.gallup.com/poll/508169/historically-low-faith-institutions-continues.aspx。 访问日期：2024 年 1 月 29 日
Sachs JD (2018) 新的外交政策：超越美国例外论。 哥伦比亚大学出版社
Scott B、Woods J、Chang A (2023)人工智能如何使社会中的种族主义、性别歧视和其他偏见永久化。 见：美国国家公共广播电台。 https://www.npr.org/2023/07/19/1188739764/how-ai-could-perpetuate-racism-sexism-and-other-biases-in-society。 访问日期：2024 年 1 月 29 日
Simons G (2024) 金砖国家和构建全球新秩序的地缘经济方面。 在：TPQ。 http://turkishpolicy.com/article/1245/brics-and-the-geo-economic-aspects-of-engineering-a-new-global-order。 访问日期：2024 年 1 月 29 日
Soliman M (2022) 战略初创企业：阿联酋在半导体上押下重注。 见：《国家利益》。 https://nationalinterest.org/blog/techland-when-great-power-competition-meets-digital-world/strategic-start-ups-uae-betting-big。 访问日期：2024 年 1 月 29 日

Stanley-Becker I, Nix N (2023) 特朗普被捕的虚假图像显示人工智能破坏力的“巨大进步”。 见：《华盛顿邮报》。 https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/。 访问日期：2024 年 1 月 29 日
Statista (2021) 标准普尔 500 指数：2021 年市值最大的公司。 https://www.statista.com/statistics/1181188/sandp500-largest-companies-market-cap/。 访问日期：2024 年 1 月 29 日
Stieber Z (2022) 超过 50 名拜登政府员工、12 家美国机构参与社交媒体审查推动：文件。 载于：大纪元时报。 https://www.theepochtimes.com/over-50-biden-administration-employees-12-us-agcies-involved-in-social-media-censorship-push-documents_4704349.html?welcomeuser=1。 访问日期：2024 年 1 月 29 日
Subhra Dutta T (2023) FraudGPT：网络犯罪分子推出的新黑帽人工智能工具。 见：网络安全新闻。 https://cybersecuritynews.com/fraudgpt-new-black-hat-ai-tool/。 访问日期：2024 年 1 月 29 日
Sumsub Research (2023) 从 2022 年到 2023 年，全球 Deepfake 事件激增十倍。 https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/。 访问日期：2024 年 3 月 29 日
印度时报 (2022) Deepfake 民主：韩国候选人通过虚拟方式投票。 https://timesofindia.indiatimes.com/world/rest-of-world/deepfake-democracy-south-korean-candidate-goes-virtual-for-votes/articleshow/89556568.cms。 访问日期：2024 年 1 月 29 日
Think BRICS (2023) 金砖国家在并行数字轨道上规划人工智能未来。 https://thinkbrics.substack.com/p/brics-nations-map-an-ai-future-on。 访问日期：2024 年 1 月 29 日
趋势科技、犯罪司法所和欧洲刑警组织 (2020) 人工智能的恶意使用和滥用。 趋势科技研究
TV 金砖国家 (2024) 弗拉基米尔·普京宣布 30 个国家已准备好加入金砖国家。 https://tvbrics.com/en/news/vladimir-putin-announces-that-30-countries-are-ready-to-join-brics/?ysclid=lu71w68ybm40618341。 访问日期：2024 年 1 月 29 日
Tyson A, Kikuchi E (2023) 公众越来越关注人工智能在日常生活中的作用。 见于：皮尤研究中心。 https://www.pewresearch.org/short-reads/2023/08/28/forming-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/。 访问日期：2024 年 1 月 29 日
联合国贸易和发展会议（2016）发展中国家的机器人和工业化。 政策简报第 50 号。
Urbina F、Lentzos F、Invernizzi C、Ekins S (2022) 人工智能驱动的药物发现的双重用途。 自然机器智能，N.4。
Vastmindz（2022）韩国总统的 AI 换脸变声技术。 https://vastmindz.com/south-koreas-presidential-deepfake/。 访问日期：2024 年 1 月 29 日
沃尔特·B（2022）。 内战如何开始：以及如何阻止它们。 王冠
世界经济论坛（2016）第四次工业革命的就业、技能和劳动力战略的未来。 执行摘要，日内瓦
世界经济论坛（2024）全球风险报告
世界卫生组织（2021）。 联合国报告：大流行年的特点是世界饥饿人数激增。 https://www.who.int/news/item/12-07-2021-un-report-pandemic-year-marked-by-spike-in-world-hunger。 访问日期：2024 年 1 月 29 日
WormGPT V3.0 (2024) https://flowgpt.com/p/wormgpt-v30。 访问日期：2024 年 1 月 29 日

人工智能的恶意使用：埃及阿拉伯共和国心理安全的挑战

叶夫根尼·帕申采夫，俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）
弗拉迪琳娜·切比基娜，圣彼得堡国立大学国际关系学院（圣彼得堡，俄罗斯）
尤莉娅·舍梅托娃，圣彼得堡国立大学（圣彼得堡，俄罗斯）

引言

有关全球政府实施人工智能技术准备情况的报告显示，2022 年，埃及在非洲排名仅次于毛里求斯，与 2019 年的报告相比，埃及在非洲国家中的排名由第八位上升至第二位，在全球 194 个国家中，排名为第 111 位，这显示了明显的进步。埃及人类发展报告 2021 揭示了其在“政府准备实施人工智能”指数上提升了 55 个位次。根据世界知识指数，埃及在 2020 年的 138 个国家中排名第 72 位，而在 2021 年的 154 个国家中排名第 53 位（Draya Egypt，2023 年）。这一积极的趋势表明了该国在技术发展领域的认真努力，并有助于营造创新和经济发展的良好环境。

最新技术也被用于建设智能城市，并根据国际标准将现有城市转变为智能城市。例如，埃及的 Hawa Dawa（旨在解决环境问题的人工智能技术）结合了物联网的传感器技术和卫星图像，运用机器学习算法收集和分析户外空气污染的高质量数据（Sayed M K，2018 年）。

该国政府正在积极发展人工智能和数字发展领域的立法。它由两份主要文件代表：《埃及数字发展战略》和《人工智能国家战略》。战略的第一阶段将于 2024 年 5 月结束，旨在利用人工智能技术支持埃及实现可持续发展目标。战略的第二阶段持续时间为 3 年（Business Today Egypt，2023 年）。根据埃及通信和信息技术部长阿姆尔·塔拉特的说法，埃及国家人工智能战略的第二阶段将于 2024 年第二季度开始，并涵盖几个关键经济领域：政府将在治理、人力资源、技术、信息基础设施、数据和环境等方面推出倡议（Pessarlay W，2024 年）。据统计，埃及人工智能市场预计在 2024 年达到 7.852 亿美元。预计市场规模将以 17.18%的年增长率增长，到 2030 年市场规模将达到 20.33 亿美元（Statista Egypt，2024 年）。

2023 年 12 月举行的总统选举确认了埃及总统埃尔-塞西以近 90%的选票获胜。但严重的经济挑战、高青年失业率、低购买力和货币贬值，以及与其他阿拉伯国家在与以色列关系和保护巴勒斯坦人问题上的潜在分歧，可能引发动乱和制度变革（Allianz，2024 年）。人工智能产业的迅速增长，加上埃及复杂的社会经济和政治问题，中东局势进一步恶化的风险，为该国多元人工智能恶意使用的增长创造了有利条件。

MUAI 对心理安全的第一层威胁

第一级别的主要威胁是人工智能在不同生活领域实施的恶意解读风险。埃及人民对于因机器人化和自动化流程而失去工作的恐惧，往往与员工担心人力劳动可能被机器取代有关（这与其他国家有很多共同之处）。卡巴斯基公司的研究是了解技术发展对劳动关系影响当前状况的重要指标。根据其中一份报告：“约有一半（44%）的埃及员工担心因为机器人而失去工作，四分之一的员工（25%）报告称在他们所在公司曾听说过与机器人或自动系统相关的网络安全事件”（Daily News Egypt，2023a）。因此，统计数据显示，埃及企业员工普遍担心引入人工智能技术对他们的就业和个人安全造成的影响。企业和埃及政府可能面临的挑战与确保就业稳定和员工再培训计划相关。至于提高信息技术能力，卡巴斯基的调查显示：“埃及有 41%的员工感到有必要提高自己的数字技能，33%担心因缺乏信息技术能力而失去工作。有些人认为这可能在未来 5 年内发生（15%），其他人认为可能在以后的某个时候发生（18%）。只有 40%的人确信由于信息技术知识不足而不会失去工作”（Daily News Egypt，2023b）。各种恶意行为者，无论是

外部还是内部，都可以积极利用未来引入人工智能技术导致大规模失业或工资下降的风险：从宗教狂热者到支持以自私利益为目的破坏政府的人，这需要特别关注引入人工智能技术时就业问题的及时发生（这种问题的数量和质量上的增长似乎在全球范围内都是不可避免的）。

在埃及（以及其他国家），关于人与机器之间未来关系的预测极为悲观。特别是，谷歌前埃及人工智能专家穆罕默德·贾夫达特（Mohammed Javdat）警告称，人工智能有一天可能会开始将人类视为“渣滓”，并创造自己的“杀人机器”（Blunt P，2023a）。贾夫达特警告称，当前基于人工智能的语言学习模型会读取我们在虚拟空间中发布的负面信息，这可能在未来使机器将人类视为负面和邪恶的东西，由此构成威胁。这些关于可能风险的声明必须得到认真对待，并通过适当的有针对性的推广，可以在人群中散播恐慌，导致人们对进一步发展和使用人工智能在日常生活中持负面态度。

2019 年，《纽约时报》根据网络安全专家的研究发表了一篇名为《埃及正在利用应用程序追踪和锁定其公民》的文章。根据这篇文章，埃及政府可能与一些攻击者有关联，这些攻击者对一些埃及记者、反对派活动人士、人权捍卫者等进行了一系列网络攻击。这些行动最初可以追溯到 2016 年。IT 安全公司 Check Point Technologies 的专家发现，黑客利用官方的 Google Play 商店分发了一些收集地理位置信息、电子邮件数据、通话记录等的程序。安装的一个必要要求是提供对用户通话历史和联系人的访问权限。Check Point 的研究人员发现该程序可以为政府带来利益。“一个 HTML 钓鱼页面中嵌入的坐标指向了开罗的一座政府建筑物。攻击者使用的一个域名的注册者被列为 MCIT，研究人员表示这可能是埃及的通信和信息技术部（MCIT）”（Lyngaas S，2019 年）。然而，很可能是其中一名攻击者可能会以后来的反政府影响活动为目的诬陷埃及政府结构。因此，对心理安全的第二级威胁中未经证实的使用人工智能技术成为了在没有充分理由的情况下创造出埃及政府反社会使用人工智能的形象的原因，即第一级威胁，这在报纸上的文章标题中得到了体现，并且毫无疑问地让人们对埃及当局的“罪行”产生了怀疑。

因此，可以得出结论认为，今天的埃及已经可以在第一层面上开展 MUAI 活动，这需要在专家界和公共行政层面进行适当的反思和深入的分析。

MUAI 对心理安全的第二层威胁

网络钓鱼在埃及是一种普遍存在的欺诈类型。根据卡巴斯基在中东和非洲地区的一项研究，网络钓鱼在快递领域被广泛使用。攻击者向受害者发送包含支付快递费用链接的信件，导致货物无法送达。点击链接后，受害者进入伪装成快递服务官方网站的网站，输入银行卡详细信息，资金就会被诈骗者扣除（Daily News Egypt，2023d）。另一项卡巴斯基 2022 年关于电子支付的研究显示，“埃及有 57%的用户在使用在线银行服务或数字钱包服务时遭遇过网络钓鱼尝试。报告还称，54%的用户遭遇过虚假网站，57%的网络钓鱼尝试是通过社会工程学方法的短信或电话呼叫进行的”（Daily News Egypt，2022e）。因此，不仅政府，还有银行业和与电子支付有关的公司，在某种程度上都应该提高公民在数字安全方面的意识。

在线欺诈的主要方面是社会工程学，它帮助攻击者获得受害者的信任，并迫使其轻率行事。例如，2022 年 8 月，埃及的攻击者推出了一个在线平台，承诺顾客由于加密货币挖矿和交易服务的佣金而获得“巨大的财务利益”。共有 29 人被逮捕，其中近一半是外国人，他们利用 HoggPool 欺诈网络窃取了超过 60 万美元（Helou E A，2023 年）。这个例子显示，网络犯罪分子越来越勤奋地利用心理方法操纵受害者的思想。

卡巴斯基全球研究与分析部门负责中东、土耳其和非洲地区的负责人阿明·哈斯比尼专注于网络安全解决方案和服务，他表示，卡巴斯基在 2023 年第一季度跟踪并阻止了约 1300 万次针对埃及的电子攻击

（Daily News Egypt，2023c）。哈斯比尼在接受《每日新闻埃及》采访时表示，与 2022 年相比，针对银行账户和客户数据的攻击数量增加了 186%。同时，埃及零售银行业信息系统遭受的黑客攻击数量也在迅速增长。与此同时，通过电子邮件和短信的网络钓鱼攻击也有所增加。据统计，2022 年第一季度约有 7.5 万名埃及用户受到网络钓鱼攻击。根据统计数据显示，有 17%至 70%的用户在收到这些欺诈性邮件后点击链接，陷入电子陷阱（Daily News Egypt，2023c）。

自聊天机器人出现以来，人们对一些人工智能系统可能对国家和国际安全构成严重威胁的担忧日益增加。根据 Group-IB 最近的一份报告，中东和北非地区的国家最容易受到旨在窃取账户、加密钱包、浏览历史记录和其他机密信息的网络攻击。在从 2022 年 6 月到 2023 年 5 月的期间内，埃及在该地区的被盗 ChatGPT 账户数量领先，约为 4500 个。被盗的数据包括登录凭据和搜索查询（Ahram Online，2023b）。尽管埃及的人工智能机器人使用情况不如其他国家普遍，但该国人口主要是年轻人，“埃及 60%的人口年龄在 10 至 49 岁之间，超过 6940 万人使用移动互联网”（Salah A，2023）。这表明，埃及社会对这项技术的普及度在未来几年将会不断增长。

埃及研究人员、网络安全专家、国际组织网络犯罪顾问穆罕默德·埃尔古因迪博士在全球人工智能伦理网络联盟（GAIEN4SG）的演示中提出了一个命题：“人工智能的恶意使用：法律和道德后果”，证明在不久的将来，我们的面部可能会成为攻击者引入恶意软件的触发器。在对目标进行视觉识别后，相应的恶意软件将被启动（ISSA Egypt，2022 年）。

埃及存在两个黑客组织：荷鲁斯团队（Horus Group）和阿努比斯（Anubis）。他们的目标是获取有关地缘政治对手的机密信息。例如，他们的组织被认为曾参与对埃塞俄比亚的网络攻击，这是由于自 2012 年开始建造的 GERD（大埃塞俄比亚文艺复兴水坝）引发的（Munawer Q，2020 年）。这些组织的活动表明，网络犯罪分子可能涉及间谍活动和网络间谍活动，符合某些国家或组织的利益。当然，在技术便宜和普及的过程中，后者将越来越多地被黑客使用。

因此，随着人工智能技术的日益普及，埃及的网络犯罪问题变得尖锐，暗示着该国的恶意使用人工智能问题将进一步增长。

MUAI 对心理安全的第三层威胁

在埃及的心理安全威胁的第三个层面上，人们担心深度伪造技术的恶意使用。2023 年，KnowBe4 在埃及、南非国家和肯尼亚的 800 名年龄在 18 岁到 54 岁之间的员工中进行了一项研究。该研究通过深度伪造技术创建的特殊机器人在电子邮件和视频通话与实验者进行直接交流。根据这项研究的结果，74%的员工未能意识到他们正在与一个机器人而不是真人进行交流（Shankar A，2023 年）。这反映了深度伪造技术变得多么复杂，使得大多数人在线上难以甚至有时无法识别出伪造品。该研究还明确指出，非洲国家公民对深度伪造技术的传播缺乏意识，使数百万人处于风险之中。根据 KnowBe4 非洲内容战略和传道人副总裁安娜·科拉德的说法，“...深度伪造平台有能力在政治和选举活动中传播误导或虚假信息，引发公民和社会动荡，并仍然是现代数字社会中的一个危险因素”（Shankar A，2023 年）。因此，埃及政府需要加强对深度伪造技术威胁的打击，以及其在政治和选举活动中传播虚假信息的潜在用途，还需通过立法和教育举措来保护普通公民免受深度伪造内容的影响。

目前，另一个与深度伪造技术的传播有关且正在经历信任危机的领域是媒体领域。2023 年 9 月，在迪拜举行的阿拉伯媒体论坛上，埃及记者和商人伊玛德·艾迪宾警告称，媒体行业的未来并不安全。据他介绍，世界已经进入了一个误导性新闻传播速度极快的时代。艾迪宾举例说明了一段深度伪造视频，该视频被他的朋友发送给他，错误地声称前美国总统唐纳德·特朗普接受了一个穆斯林男子的心脏移植手术，并

转变成了伊斯兰教（Al-Faour N，2023 年）。问题在于人们倾向于相信他们亲眼所见的内容，即使内容并非真实。

2022 年初，埃及法官会议（Dar Al-Ifta）发表声明，表示不可接受使用人工智能技术制作与他人无关的虚假视频或音频记录。“使用深度伪造技术制作这些片段，目的是伤害他人，根据先知穆罕默德（愿主福安息）的话‘不伤害自己或他人’，这是被禁止的，”法官会议表示，并补充说伊斯兰教禁止恐吓他人，即使是出于娱乐目的（Ahram Online，2022 年 a）。这一声明反映了埃及一家声望良好的宗教组织对可能对个人或社会造成伤害的技术的态度，虽然伊斯兰教不主张限制信息技术的发展，但它确实表明道德约束应该是至关重要的。此外，“法官会议还指出，散布误导性信息已被 2018 年 175 号法律列为信息技术犯罪”（Ahram Online，2022 年 a）。

根据澳大利亚国立大学网络学院的博士研究员和讲师米娜·海宁（Mina Henein）的说法，阻碍聊天机器人融入埃及人日常生活的主要因素包括：语言障碍、人口的数字文盲、法律框架以及该国的文化规范（Salah A 2023）。然而，埃及目前正在积极开展教育活动，使民众熟悉创新技术的积极和消极后果。例如，2023 年 3 月，埃及信息与决策支持中心（IDSC）与联合国教科文组织合作举办了一场有关生成式人工智能的研讨会（MENA 2023a）。此类研讨会的主要目标是帮助年轻人形成可持续的观点，以应对未来的技术挑战。

结论

根据分析，可以得出结论认为，在埃及，恶意使用 AI 的问题包括所有三个层面。该国在中东、北非地区在人工智能发展领域处于领先地位，并且没有计划止步于此。第一级威胁涉及到人工智能技术实施可能导致群众对失业增长和人身安全风险的担忧。在第二层面上，最棘手的领域仍然是虚拟欺诈，包括使用聊天机器人，以及对国家关键基础设施对象的黑客攻击。对第三级威胁的分析表明，埃及社会越来越关注人工智能创造虚假信息并在虚拟空间积极推广的能力。由于过度不信任和缺乏谨慎，公民本身经常受到这种技术的恶意操作。因此，有必要建立机制，以便在未来发现和预防此类案件，并提高公民对互联网上的风险的认识，以确保在国内打击人工智能恶意软件威胁的积极势头。然而，考虑到快速变化的技术环境和地缘政治环境，埃及政府也应该为可能出现的新威胁做好准备。

参考文献：

Ahram Online (2022 年) 埃及达尔艾芙塔禁止深度伪造视频和音频剪辑。在：Ahram 在线。https://english.ahram.org.eg/NewsContent/1/64/454765/Egypt/Politics-/Egypt%E2%80%99s-Dar-AlIfta-prohibits-deepfake-video-and-au.aspx。访问日期：2024 年 1 月 25 日

Ahram Online (2023 年) 报告称：近 4600 个埃及 ChatGPT 账户被黑。在：Ahram 在线。https://english.ahram.org.eg/NewsContent/3/1239/503415/Business/Tech/Nearly-,-Egyptian-ChatGPT-accounts-hacked-Report.aspx。访问日期：2024 年 2 月 6 日

Al-Faour N (2023 年) 埃及记者警告 AI 对媒体行业构成的威胁。在：阿拉伯新闻。https://www.arabnews.com/node/2381501/media。访问日期：2024 年 2 月 8 日

Allianz (2024 年) 斯芬克斯的谜题：再次测试埃及的政治和经济稳定性。在：安联集团。https://www.allianz.com/en/economic_research/country-and-sector-risk/country-risk/egypt.html。访问日期：2024 年 2 月 7 日

Blunt P (2023 年) 谷歌前埃及 AI 专家警告，随着 AI 对人类产生负面看法，即将出现灾难。在：Asume 科技。

https://asumetech.com/googles-former-egyptian-ai-expert-warns-of-impending-disaster-as-ai-develops-negative-perception-of-humanity/。访问日期：2024 年 1 月 29 日
Business Today Egypt (2023 年) MCITMin 讨论第二阶段的人工智能国家战略。在：今日埃及商业。
https://www.businesstodayegypt.com/Article/1/3832/MCITMin-discusses-2nd-phase-National-Strategy-for-Artificial-Intelligence。访问日期：2024 年 2 月 7 日
Daily News Egypt (2023 年) 埃及 44%的员工担心失去工作给 AI。在：每日埃及新闻。
https://www.dailynewsegypt.com/2023/02/20/44-of-employees-in-egypt-fear-losing-their-jobs-to-ai/。访问日期：2024 年 1 月 29 日
Daily News Egypt (2023 年) 埃及 33%的员工感觉缺乏数字能力。在：每日埃及新闻。
https://www.dailynewsegypt.com/2023/09/19/33-of-employees-in-egypt-feel-the-lack-of-digital-competencies/。访问日期：2024 年 2 月 4 日
Daily News Egypt (2023 年) 卡巴斯基在 2023 年第一季度解决了埃及约 1300 万次网络攻击。在：每日埃及新闻。
https://www.dailynewsegypt.com/2023/05/08/kaspersky-tackles-13-million-cyber-attacks-in-egypt-during-1q-2023/。访问日期：2024 年 1 月 27 日
Daily News Egypt (2023 年) 卡巴斯基在非洲、中东、土耳其检测到快递服务诈骗波浪。在：每日埃及新闻。
https://www.dailynewsegypt.com/2023/08/06/kaspersky-detects-wave-of-courier-service-scams-in-africa-middle-east-turkiye/。访问日期：2024 年 2 月 8 日
Daily News Egypt (2022 年) 超过一半的埃及用户在电子支付期间遭遇网络诈骗：卡巴斯基。在：每日埃及新闻。
https://www.dailynewsegypt.com/2022/07/28/more-than-half-of-egypts-users-encountered-phishing-attempts-during-electronic-payments-kaspersky/。访问日期：2024 年 2 月 8 日
Draya Egypt (2023 年) 埃及的人工智能及其在国家战略框架内的增强方式。在：公共政策和发展研究战略论坛。
https://draya-eg.org/en/2023/02/08/artificial-intelligence-in-egypt-and-ways-to-enhance-it-within-framework-of-national-strategy/。访问日期：2024 年 2 月 4 日
Helou E A (2023 年) 埃及加密货币诈骗使投资者损失 62 万美元。在：中东经济。
https://economymiddleeast.com/news/crypto-scam-in-egypt-robs-investors-of-620000/。访问日期：2024 年 2 月 4 日
ISSA Egypt (2022 年) AI 的恶意使用：法律和道德影响 - GAIEN4SG 由 Mohamed El-Guindy 博士的讲座。在：信息系统安全协会。
https://issa-eg.org/malicious-use-of-ai-legal-and-ethical-implications-gaien4sg-talk-by-dr-mohamed-el-guindy/。访问日期：2024 年 2 月 4 日
MENA (2023 年) 埃及 IDSC 举办 ChatGPT 研讨会，讨论 AI 平台的未来。在：Ahram 在线。
https://english.ahram.org.eg/NewsContent/3/1239/491777/Business/Tech/Egypt;s-IDSC-holds-ChatGPT-workshop-to-discuss-fut.aspx。访问日期：2024 年 2 月 6 日
Munawer Q (2020 年) 埃及对埃塞俄比亚安全局网站等的网络攻击。在：东方先锋。
https://easternherald.com/2020/06/24/egypt-cyber-attack-ethiopia/。访问日期：2024 年 2 月 4 日
Pessarlay W (2024 年) 埃及人工智能战略聚焦治理、环境和人力资源。在：Coin Geek。
https://coingeek.com/egypt-ai-strategy-focuses-on-governance-environment-and-human-resources/。访问日期：2024 年 2 月 7 日

人工智能的恶意使用：伊朗伊斯兰共和国心理安全的挑战

叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所，（莫斯科，俄罗斯）
达利娅·巴扎尔金娜，俄罗斯科学院欧洲研究所，（莫斯科，俄罗斯）

引言

伊朗已经采纳并正在实施一项国家人工智能发展路线图，旨在将该国从目前根据《自然指数》排名第 13 的位置提升到全球前十。为此，伊朗计划投资 80 亿美元用于人工智能研究与开发（《德黑兰时报》2022 年）。人工智能驱动的医疗技术在诊断疾病方面表现出令人印象深刻的结果——例如，根据最新数据，伊朗医科大学（IUMS）开发的系统在乳腺癌诊断方面的准确率达到 94%。然而，与任何其他领域一样，人工智能技术的发展可能会导致其恶意使用的风险。伊朗的情况受到以下影响：内部民族和政治矛盾的存在（Ziya，2021 年），公共部门腐败的严重表现和后果（《伊朗国际》，2023 年），以及更重要的是，来自以色列和美国的紧张外部压力。在美国情报界年度威胁评估报告的公开部分中，伊朗被列为据称对美国国家安全和国际安全构成最大威胁的四个国家之一，与俄罗斯、中国和朝鲜并列（国家情报总监办公室，2023 年）。值得注意的是，报告的相应章节特别提到了以色列，该国也据称受到伊朗的威胁（《国家情报总监办公室》2023 年）。同时，美国是人工智能技术发展和应用的领导者，而以色列将伊朗视为中东地区的主要威胁（Berman，2023 年）。根据以色列国防部长官埃亚尔·扎米尔的说法，“我们的使命是将以色列变成一个人工智能超级大国，并成为少数几个在这个俱乐部里的世界强国的领头羊”（Williams and Maclean，2023 年）。因此，这两个国家的领导层具有比伊朗自身能力更大的动机和能力，可以更多地利用人工智能技术来对付伊朗，而伊朗的能力则不足以对抗美国和以色列。这一系列问题以及对该国的敌对行为者造成了极度紧张的环境，为恶意使用人工智能提供了肥沃的土壤。

MUAI 对心理安全的第一层威胁

伊朗以及许多其他国家正在积极研究和实施各个领域的人工智能技术，公众对人工智能可能对就业市场造成的影响感到担忧（《伊朗人才》2023 年）。然而，对于这个问题并没有明确的观点。该国科学界的一些代表将失业风险集中在“平均”水平的专业人士身上，而不考虑工作过程中的创造性成分——生成式人工智能模型已经足够能够解决以前被认为需要人类大脑非线性思维的问题。根据德黑兰大学和德黑兰医科大学的教授哈米德雷扎·凯沙瓦兹（Hamidreza Keshavarz）的说法，低技能（清洁工作、粗体力劳动等）和真正高素质（需要严肃的学术教育和高度发达的智力装备，或基于个人品质，持续展示出色成果）的工作相对安全。与此同时，从以人工智能为基础的发展角度来看，拥有“平均”资质的专业人士在劳动市场上最易受到取代的风险最大（《哈巴尔在线》2023 年）。在引入人工智能技术的国际大型企业中出现的大规模裁员，以及更大规模裁员的风险无疑会引起伊朗人口中的相当部分人的担忧。如果在进一步的机器人化和人工智能技术引入的更大规模过程中不采取适当的社会保障和人类发展措施，这些自然担忧可能会被恶意的内外部行为者有意放大。

伊朗潜在利用人工智能技术自动化控制公共秩序的做法引发了热议和争论，在西方国家引起了轩然大波。西方信息空间爆发了一场真正的媒体轰炸，这是因为伊朗当局利用城市视频监控系统自动识别不遵守封闭服装（希贾布）规定的女性（Alkhaldi and Ebrahim，2023 年）。该炒作还得到了外交关系委员会（CFR）网站在 2023 年 12 月的支持，该网站发表文章称，限制互联网访问和使用人工智能识别道德标准违反者的法案“据报道导致了两万多人被逮捕和五百多名年轻抗议者的死亡”（乔治 2023）。然而，Al Qaeda 和其他恐怖组织在该国的活跃可能并不是偶然的，它们中的一些同时依赖外部势力的支持。恐怖袭

击在该国相当频繁，最近，最大规模的袭击发生在 2024 年 1 月 3 日在克尔曼，成千上万的人民来到扎伊姆·卡西姆·苏莱曼尼将军的安葬地点，以纪念他的逝世四周年，他是该国反恐斗争的象征。自杀式炸弹袭击造成至少 93 人死亡，另有数十人受伤（《德黑兰时报》2024 年）。

媒体（Jahan News 2023）报道显示视频中，一个人工智能系统被设计用来识别女性的身份。同时，伊朗警察官员和伊朗议会的代表也证实了政府使用人工智能技术广泛识别各种违法行为的意图，包括违反“道德性质”（女性未身着封闭式服装）和以这种方式建立的事实的自动生成式起诉。然而，国家前交通部长 Azari Jahromi 也表达了对此类做法的准确性的关注，然而警察部门的代表认为，报道称，系统准确性问题随着 AI 接受足够数据进行训练，会伴随时间发展得到解决（Ensaf News 2023）。尽管西方媒体对伊朗当局的攻击显然具有政治色彩，但在这次讨论中提出的问题确实是相关的。同时，这些问题被感兴趣的恶意行为者故意放大，主要是外部行为者，他们利用相当的财政、组织、技术和军事资源，不追求伊朗人民的利益，而是在该地区追求自己的帝国主义目标。

MUAI 对心理安全的第二层威胁

第二层的威胁同样是真实存在的，而且正在增长。据《德黑兰时报》（2023 年）报道，伊朗民防总指挥 Gholam Jalali 将军在 2023 年 12 月表示，该国加油站最近发生的故障是由有针对性的网络攻击期间的恶意软件引起的。在同一演讲中，尽管没有直接宣布与多用途人工智能的直接联系，将军指出，高达 50%的针对伊朗关键信息基础设施的网络攻击在某种程度上都涉及人工智能技术。在提到的网络攻击背景下，可以假设，与“全球惯例”一样，攻击者可能利用人工智能准备钓鱼消息，通过这些消息将恶意软件传送到目标基础设施。此前，将军曾在 2023 年 8 月表示，外部行为者在筹备大规模抗议活动时使用了人工智能，伊朗应该学会利用人工智能来对抗这种使用（Mohammadzadegan，2023 年）。

同样值得记住的是，前述的 Qasem Soleimani 将军于 2020 年 1 月 3 日在巴格达国际机场附近被美国无人机袭击击毙，当时他正在前往会见伊拉克总理阿迪勒·阿卜杜勒-迈赫迪。此事件的视频实况转播至美国白宫、兰利的中央情报局总部等地方，以供国防部官员参考。该行动由当时担任中央情报局局长和国防部长的吉娜·哈斯佩尔和马克·埃斯珀监督（Dilanian 和 Cube，2020 年）。

白宫向国会发送通知，简述了空袭杀死伊朗将军的的法律和政治理由。在通知中，特朗普政府援引第二条款和 2002 年针对伊拉克使用武力授权，以证明美国的打击行动是合法的。该行政命令称，此行动的目的是“阻止伊朗对美国部队和利益进行或支持进一步的攻击”，并“削弱伊朗和经受青壮军支持的民兵发动攻击的能力”（Setzer 2020）。这样的论点不仅引发了伊朗的愤怒反应，还受到许多立法者的谴责，主要是来自民主党人（Choi 2020, Pengelly 和 Helmore 2020）。 根据彼得·辛格，新美国基金会未来战争专家的说法：“在不到一代人的时间里，我们从一种异常的、甚至可能是科幻的事物，发展到了现在这个新常态的程度”（Dilanian 和 Cube 2020）。当然，人工智能技术在此次行动的准备、实施以及向媒体展示这一信息方面发挥了重要作用。

同样值得注意的还有 2020 年 11 月 29 日发生了一起悲剧事件，当时著名的伊朗核科学家 Mohsen Fakhrizadeh 在首都德黑兰附近的一条高速公路上遭到袭击身亡。据伊朗当局称，这位科学家是“圣战组织”恐怖组织袭击的受害者，该组织准备激活一台电子设备，据称是为了以色列的利益。该设备安装在货车上，利用人工智能在爆炸前识别了法赫里扎德，并在事件发生期间伤害了陪同科学家的人员（Motamedi 2020）。

因此，人工智能技术显然在构成对伊朗基础设施的第二级威胁中得到了积极的应用。

MUAI 对心理安全的第三层威胁

伊朗媒体目前处于积极转变阶段，这种变化与全球范围内社会迅速数字化有关。除了快速访问信息外，“内容创建者 – 内容消费者”链中控制发布内容的链接数量也减少了。由于这些变化，包括那些使用生成式 AI 模型制造的，更有效更具破坏性后果的信息注入变得可能。比如，最近有一篇极具冒犯性的文章，著名的伊朗神职人员 Hossein Ansarian 在一段视频中声称该国的权力已经被驴子夺走，并在屏幕上展示了一个“证据”，即将人类头部粗略地绘制在驴子的身体上（Iran NTV 2023）。这种表现强化了加强伊朗国家人工智能委员会工作的必要性，因为伊朗，像其他所有国家一样，迫切需要引入技术来检测此类虚假信息材料，否则该国将面临失去大众对信息的信任和社会心理稳定的风险。

媒体还多次报道了美国国防部中央司令部（CENTCOM）在社交网络 Facebook（Meta）和 Twitter（X）上开展的信息和心理操作。CENTCOM 的信息行动已经进行了很长一段时间，包括传播反伊朗宣传。在执行此类操作时，人工智能被用于生成文本，并通过生成所谓真实社交网络用户的逼真图像（深度伪造）来为发布文本赋予更多权威性（《德黑兰时报》2022b 年）。

尽管国际社会积极讨论与人工智能实施相关的风险，但伊朗各种相当“敏感”行业已经开始研究基于人工智能的聊天机器人的使用。例如，德黑兰医科大学的研究人员在他们的文章中详细描述了在医学中使用人工智能处理大量数据以及使用聊天机器人进行医疗咨询的优势（Hajialiasgari Khanahmadi and Atashi，2023 年）。在潜在的多用途人工智能和心理风险方面，文章提到了非法获取机密信息（违反医疗保密性）、患者对用“没有灵魂的机器”替代医生的可能心理反应。而聊天机器人运行中的外部干扰或错误可能导致治疗处方中的医疗错误。与此同时，研究人员得出了一个相当积极的结论，即有必要将人工智能引入伊朗医疗保健系统的工作中。

自 2020 年以来，包括伊朗最高领袖在内的伊朗领导人代表已经多次并理所当然地强调，伊朗应该与技术进步保持和谐，并成为人工智能技术的领先国家之一，这是显而易见且令人信服的原因。然而，将人工智能引入宗教实践可能会产生模棱两可的后果。因此，2023 年，伊斯拉格创新与创意中心的负责人 Mohammad Ghotbi 直言不讳地表示：“机器人无法取代高级神职人员，但它们可以成为可信赖的助手，可以帮助他们在五小时内发布法令，而不是 50 天。”（Bozorgmehr 2023）。在如宗教这样一个敏感领域中引入人工智能决策过程，尤其是在像伊朗这样的国家，宗教与政府直接联系在一起的情况下，如果上述的“机器人助手”受到恶意行为者的破坏，可能会导致灾难性后果。随着情感人工智能的发展以及伊朗及其周边地区局势的可能进一步恶化，尤其严重的风险将出现。

结论

科技进步始终需要平衡的决策，而在过渡期向新的社会和国际秩序发展人工智能作为关键技术的措施则需要特别负责任的态度。除了发展可信人工智能系统外，还需要提前全面分析每个这种系统的威胁模型，包括考虑信息、心理安全及相关风险。在没有采取预防措施和对抗多用途人工智能的情况下将人工智能引入公共生活可能会给任何社会和国家带来重大损害，而伊朗作为抵抗帝国主义压力的前沿，更容易受到相关风险的影响。在这种情况下，恶意影响可能既来自外部行为者，以影响国家的内部稳定或其外交政策方向，也可能来自内部行为者，例如试图非法获取利益或提高自己的政治声誉。

然而，由于外部压力显著超过内部压力，外部恶意行为者拥有复杂的人工智能技术，因此对伊朗心理安全的第二和第三级别的多用途人工智能威胁最为相关和危险。五眼情报联盟成员国已多次展示了其在进行信息和心理操作方面的能力，而北约国家利用人工智能技术的最新武器和军事装备在乌克兰战区的战场上的使用效果也可以看到。因此，美国情报界对伊朗的如此密切关注不免会引起独立观察者的关注。

值得注意的是，伊朗当局以系统性和战略规划的立场来着手实现人工智能领域的实质性进展。可以假设，如果伊朗政府密切关注通过多用途人工智能对心理安全的威胁，对抗这些威胁的方法将同样变得结构化和系统化。

参考文献

Alkhaldi C, Ebrahim N (2023) 伊朗提议在严苛的新版头巾法中实施长期监禁、人工智能监控和对影响者的打压。《CNN》。https://edition.cnn.com/2023/08/02/middleeast/iran-hijab-draft-law-mime-intl/index.html。访问日期：2024 年 2 月 3 日

Berman L (2023) 国防部长称，以色列国防军将聚焦伊朗，成为“人工智能强国”。《以色列时报》。https://www.timesofisrael.com/liveblog_entry/idf-set-to-focus-on-iran-become-ai-powerhouse-says-defense-ministry/。访问日期：2024 年 2 月 3 日

Bozorgmehr N (2023) “机器人可以帮助发布法令”：伊朗神职人员寻求利用人工智能。《金融时报》。https://www.ft.com/content/9c1c3fd3-4aea-40ab-977b-24fe5527300c。访问日期：2024 年 2 月 3 日

Choi M (2020) 2020 年民主党人警告苏莱曼尼被杀后中东可能升级。《政客》。https://www.politico.com/news/2020/01/02/soleimani-2020-iran-democrats-093123。访问日期：2024 年 2 月 3 日

Dilanian K, Cube C (2020) 机场线人、空中无人机：美国是如何杀死苏莱曼尼的。NBC 新闻，1 月 10 日，https://www.nbcnews.com/news/mideast/airport-informants-overhead-drones-how-u-s-killed-soleimani-n1113726。访问日期：2024 年 2 月 3 日

Ensaf News (2023) 使用智能摄像头对裸露人群的面部识别的明暗面 [سایه روشن شناسایی چهره بی حجاب ها با دوربین های هوشمند]。https://ensafnews.com/408902/سایه-روشن-شناسایی-چهره-بی-حجاب-ها-با-دوربین/。访问日期：2024 年 2 月 3 日

George R (2023) 针对女性的人工智能攻击：伊朗技术支持的道德法律对女性权利运动的意义。《外交关系委员会》。https://www.cfr.org/blog/ai-assault-women-what-irans-tech-enabled-morality-laws-indicate-womens-rights-movements。访问日期：2024 年 2 月 3 日

Hajialiasgari F, Khanahmadi A, Atashi A (2023) 伊朗卫生保险组织中的人工智能聊天机器人：服务提供的新时代。《伊朗健康保险杂志》。第 6 卷(2 期)，第 91-102 页。

Iran International (2023) 伊朗最大的腐败案震动执政的强硬派。https://www.iranintl.com/en/202312062449。访问日期：2024 年 2 月 3 日

Iran NTV (2023) 快乐快递 - Hossein Ansarian 的深度伪造 [دیپ فیک حسین انصاریان از پیک شادی]。https://iranntv.com/908619-دیپ-فیک-حسین-انصاریان-از-پیک-شادی。访问日期：2024 年 2 月 3 日

Iran Talent (2023) 人工智能真的会让我们所有人失业吗？ [آیا واقعا هوش مصنوعی همه ما را بیکار خواهد کرد؟]。https://www.irantalent.com/blog/impact-of-artificial-intelligence-job-losses/。访问日期：2024 年 2 月 3 日

Jahan News (2023) 使用人工智能识别未戴头巾的女性 [شناسایی زنان بی‌حجاب با هوش مصنوعی]。https://www.jahannews.com/news/840663/فیلم-شناسایی-زنان-بی-حجاب-با-هوش-مصنوعی。访问日期：2024 年 2 月 3 日

Khabar Online (2023) 人工智能会让哪些人失业？ [هوش مصنوعی چه مشاغلی را بیکار می‌کند]。https://www.khabaronline.ir/news/1724621/هوش-مصنوعی-چه-مشاغلی-را-بیکار-می-کند。访问日期：2024 年 2 月 3 日

Mohammadzadegan A (2023) 伊朗优先使用人工智能进行网络防御，国防官员说。在：IRNA。https://en.irna.ir/news/85197899/Iran-prioritizes-using-AI-for-cyber-defense-says-defense-official。访问日期：2024 年 2 月 3 日

Motamedi M (2020) 伊朗官员指责以色列远程杀害 Fakhrizadeh。半岛电视台。
https://www.aljazeera.com/news/2020/11/30/iran-israel-killing-scientist-remotely-in-sophisticated-attack。访问日期：2024 年 2 月 3 日
美国国家情报总监办公室 (2023) 美国情报界的年度威胁评估。
https://www.dni.gov/files/ODNI/documents/assessments/ATA-2023-Unclassified-Report.pdf。访问日期：2024 年 2 月 3 日
Pengelly M, Helmore E (2020) 弹劾：沃伦指责特朗普对苏莱曼尼的“摇尾狗”打击。《卫报》。
https://www.theguardian.com/us-news/2020/jan/05/impeachment-warren-trump-wag-the-dog-qassem-suleimani-iran。访问日期：2024 年 2 月 3 日
Setzer E (2020) 白宫发布报告为苏莱曼尼袭击辩护。《法律战》。
https://www.lawfaremedia.org/article/white-house-releases-report-justifying-soleimani-strike。访问日期：2024 年 2 月 3 日
Tehran Times (2022a) 伊朗计划成为人工智能领先国家。
https://www.tehrantimes.com/news/469628/Iran-plans-to-become-a-leading-country-in-AI。访问日期：2024 年 2 月 3 日
Tehran Times (2022b) 五角大楼在心理战中骑着“蓝鸟”。
https://www.tehrantimes.com/news/480127/Pentagon-riding-the-blue-bird-in-psychological-warfare。访问日期：2024 年 2 月 3 日
Tehran Times (2023) 伊朗称检测到加油站网络攻击中使用的恶意软件。
https://www.tehrantimes.com/news/492846/Iran-says-malware-used-in-cyberattack-on-fuel-stations-detected。访问日期：2024 年 2 月 3 日
Tehran Times (2024) 克尔曼恐怖袭击是以色列弥补损失的尝试：军队首脑。
https://www.tehrantimes.com/news/493528/Kerman-terrorist-attack-Israeli-attempt-to-compensate-for-losses。访问日期：2024 年 2 月 3 日
Williams D, Maclean W (2023) 以色列旨在成为“人工智能超级大国”，推进自主战争。路透社。
https://www.reuters.com/world/middle-east/israel-aims-be-ai-superpower-advance-autonomous-warfare-2023-05-22/。访问日期：2024 年 2 月 3 日
Ziya MH (2021) 伊朗面临的 13 场危机。中东研究所。
https://www.mei.edu/publications/13-crises-facing-iran。访问日期：2024 年 2 月 3 日

人工智能的恶意使用：埃塞俄比亚联邦民主共和国心理安全的挑战

谢尔盖·赛贝金，伊尔库茨克国立大学政治学院（伊尔库茨克，俄罗斯）

引言

埃塞俄比亚是成功地利用人工智能技术解决各种任务的非洲国家之一，并且在动荡的社会政治形势中为 AI 创造了能够广泛发展的制度条件（参见 Ade-Ibijola & Okonkwo, 2023, 第 102、104 页；Gadzala, 2018, 第 1、2、5、8 页）。引入人工智能系统的优先领域包括农业（作为埃塞俄比亚经济的支柱）（参见 Federal Democratic Republic of Ethiopia, 2020, 第 26 页；Girmay, 2019, 第 161-162、166-167 页）、公共卫生、金融和公共管理。埃塞俄比亚已经创建并正在发展所谓的谢巴谷——该国的技术中心（类似于美国的硅谷）（见 Eke, Wakunuma, & Akintoye, 2023a, 第 4 页）。而埃塞俄比亚人工智能研究所正在系统地推进人工智能领域的多方面进展。此外，还有一些私营公司从事人工智能研究，包括 2013 年在亚的斯亚贝巴成立的私人 AI 研究实验室 iCog Labs，它提供广泛的 AI 研究和开发服务，为国内外客户提供广泛的人工智能研究和开发服务。在这个过程中，埃塞尔比亚宣布的优先目标之一是创建一个符合国家特点和价值观的人工智能。尽管如此，到目前为止，人们对数字服务的访问水平仍然不够高。

埃塞俄比亚有大约 1.2 亿人口。据估计，只有约 16-20%的人口能够接入互联网。社交媒体用户的数量甚至更少，大约只有 5%（Kemp, 2023）。埃塞俄比亚人口普遍数字化水平较低的事实可能成为通过人工智能影响大众意识的障碍。因此，随着埃塞俄比亚数字基础设施的发展、数字化水平的提高以及数字服务的更广泛接入，未来多功能人工智能的技术和机构条件将得到显著拓展。

另一方面，到目前为止，埃塞俄比亚的社会、社会政治和社会经济局势仍然极不稳定。主要矛盾集中在“传统上反叛”的提格雷州和提格雷人民解放阵线（TPLF）与联邦政府（总理阿比·艾哈迈德·阿里）之间的冲突（Afriyie, Ayangbah, & Effah, 2023; Center for Preventive Action, 2023）。

埃塞俄比亚的冲突潜力存在以下几个原因。

首先，该国存在众多的种族矛盾和种族紧张局势的中心。不同种族群体之间在土地所有权、宗教差异等方面经常发生地区性和跨地区的冲突。其次，除了官方的国家武装部队外，每个省份都拥有（或曾拥有）自己的“种族”准军事单位，不受联邦政府控制，经常进行种族清洗行动。尽管联邦政府试图消除这些单位并进行裁军，但并不是所有省份都同意这一政策。第三，上述问题因实际的社会经济情况而多次恶化，纯粹自然因素如干旱和蝗虫入侵导致饥荒，以及武装冲突造成的人道主义危机，例如强迫迁移、粮食短缺等。对于埃塞俄比亚而言，获得红海通道的问题至关重要。这些愿望受到了 1993 年从埃塞俄比亚分离出去的厄立特里亚的挑战，使埃塞俄比亚失去了通往海洋的通道。

除了内部矛盾外，埃塞俄比亚与邻国也存在严重分歧。埃塞俄比亚与埃及和苏丹在尼罗河水资源分配方面存在强烈分歧。2023 年，埃塞俄比亚开始填充文艺复兴大坝水库，而埃及则将此视为一种明显升级的行动，使双边对话变得更加复杂。

因此，尽管使用人工智能的技术和基础设施条件仍在发展中，但通过恶意使用人工智能在埃塞俄比亚实现颠覆性效果的社会政治和经济条件早已存在。如果埃塞俄比亚未来的数字化水平增加，而根深蒂固的问题仍未解决，这些制度因素将与其他因素共同产生出色的协同潜力，通过影响大众意识来实现多功能人工智能，以明确或隐含地达到内外部利益相关者所设想的具体效果。

MUAI 对埃塞俄比亚心理安全的第一层威胁

在埃塞俄比亚广泛引入外国人工智能系统时，可能出现的第一个也是最明显的问题在于：机器学习的算法是在外国数据上训练的，这些数据充满了西方（或其他）的价值观、伦理观和解决问题的方式，可能会对埃塞俄比亚产生影响。这些信息可能会在具有完全不同民族文化、伦理、政治和经济传统的埃塞俄比亚现实中被证明是低效的，甚至可能导致负面效果，威胁社会、政治和心理稳定（Birhane, 2023, p. 250; Blackwell, Damena, & Tegegne, 2021; Eke, Wakunuma, & Akintoye, 2023a, p. 2-3; Eke, Wakunuma, & Akintoye, 2023b, p. VI; Okolo, Aruleba, & Obaido, 2023, p. 54）。非洲专家担心的是，“非非洲”的人工智能技术在解决非洲问题时没有考虑到它们的特异性（Birhane, 2023, pp. 250, 254-255; Eke, Wakunuma, & Akintoye, 2023a, p. 1-2）。

另一个重大挑战在于问题的另一面，即当大量收集埃塞俄比亚人的个人数据并用于训练特定的算法以解决埃塞俄比亚背景下的问题，并相应地个性化服务和内容时（Birhane, 2023, pp. 249; 251-252）。外国行为者对这些数据的访问为他们操纵这些数据和影响埃塞俄比亚目标受众开辟了广阔的机会（稍后将讨论）。

最后，许多来自非洲的专家担心，大规模引入“非非洲”的人工智能系统来解决某些问题，以及普遍引入外国数字基础设施，将使非洲国家过度依赖进口技术，并使它们陷入所谓的“人工智能新殖民主义”——也就是“非洲的算法殖民”（Adams, 2021; Birhane, 2023; Eke, Wakunuma, & Akintoye, 2023b, p. VI）（或“非洲的数字殖民”），在这里，有兴趣的行为者将使用人工智能技术不仅解决非洲的紧迫问题，而且还会暗中影响该地区的经济、政治和社会进程，以实现他们的利益。人们还假设，从事 AI 领域的公司和公司的主要利益将不是专注于道德和符合民族文化敏感性的 AI 培训，而是通过向非洲国家出口 AI 技术来牟利（Birhane, 2023, pp. 251-252; Okolo, Aruleba, & Obaido, 2023, p. 41, 54; Eke, Wakunuma, & Akintoye, 2023b, p. VI）。

自动化和随之而来的大规模失业问题是在埃塞俄比亚恶意使用人工智能对心理安全形成的第一层次挑战(MUAI)（Girmay, 2019, p. 170）。事实是，由人工智能引入引起的工作自动化的后果，在发展中国家（埃塞俄比亚就是其中之一）可能比发达国家更为深远（The Conversation, 2023）。首先，发达国家的经济结构在可用部门方面更加“多向”和复杂，这意味着有大量高资质的工作，从而降低了全面和全方位人工智能自动化的风险。其次，发达国家相应地拥有更发达的经济和大量资源来实施灵活的自动化政策，并通过强大的教育和再培训/提升技能计划创造新的高资质工作来替代旧的工作。同时，发达国家可以负担得起实施补充性的直接支持措施，如无条件基本收入和各种福利，这是埃塞俄比亚不太可能实现的。尽管各种预测表明，人工智能至少不会减少，甚至可能创造更多的工作岗位，但这个问题对于非洲发展中国家及其不太发达的经济结构可能会有深远的影响。例如，埃塞俄比亚经济的支柱是农业和服务——这些部门是最有希望因广泛采用人工智能系统而现代化的领域（Federal Democratic Republic of Ethiopia, 2020, p. 9, 26; Girmay, 2019, pp. 161-162; United Nations, 2023）。尽管在人工智能领域成功发展了专业教育，但由于埃塞俄比亚的普通高等教育水平迄今为止不足以及其主流对普通民众的可及性，这使得在新条件下实现劳动力潜力的能力变得更加复杂。埃塞俄比亚的年轻人比例远高于发达国家，这些人更有可能失去涉及人工智能和人工智能机器人技术的工作，如果没有及时引入再培训和技术适应计划，也没有创造更多技术先进的工作岗位，他们找到同等就业机会的机会很小。假设人工智能将“剥夺”他们的工作的第一层次心理效应，特别是在经济转型问题恶意使用的情况下，以及在高度财产两极分化的条件下，可能是极其破坏性的，并导致在实际的埃塞俄比亚条件下社会政治局势的不稳定，促进经济中影子部门的发展，加剧犯罪情况以及激活对非法生计来源的寻找。在埃塞俄比亚创造新的高资质工作需要政府在人工智能领域制定相应的教育计划（包括有效的再培训计划），使这种教育更加普及，并且最重要的是，在不同族群之

间平均分配工作。对未来问题的低估可能导致新卢德主义以各种形式出现——抗议用算法替代工作岗位，罢工和示威，甚至针对联邦政府的新武装行动。

从长远来看，因“大规模自动化和裁员”问题而产生的“恐慌”可能具有明显的操纵性质，并可用于明显目的以破坏社会稳定。同时，AI 技术引入的放缓将使该国的落后状态持续下去，不能解决其问题。

MUAI 对埃塞俄比亚心理安全的第二层威胁

埃塞俄比亚因其传统上不稳定的社会政治和军事环境，成为网络攻击的绝佳目标，范围从对关键基础设施系统的网络攻击到网络诈骗。

根据埃塞俄比亚信息网络安全管理局——主要负责网络安全的政府机构官方统计，埃塞俄比亚系统遭受的网络攻击数量在上一财政年度（2022-2023 年）接近 7,000 次（Ena, 2023b; Ethiopian Monitor, 2023; Reqiq Staff, 2023）。尽管与其他国家相比，埃塞俄比亚系统遭受的网络攻击数量并不算高（例如，另一个金砖国家南非通过后门和间谍软件记录的攻击约有 106,000 次，而在更大的金砖国家这个数字甚至更高），但据卡巴斯基全球研究与分析团队（Teshome, 2023）表示，网络攻击的规模和专业性正在增长。此外，卡巴斯基全球研究与分析记录了略有不同的数据——18,000 次网络攻击和 30,000 次勒索软件攻击（Teshome, 2023）。

对埃塞俄比亚的网络攻击主要针对金融机构、卫生、教育、安全、媒体和政府部门（Ena, 2023a; Ethiopian Monitor, 2023; Reqiq Staff, 2023; Teshome, 2023）。对系统和民众使用的主要网络攻击类型和工具包括 DDoS 攻击（拒绝服务）、系统扫描和渗透，以及对网站的恶意软件（包括勒索软件）攻击（Ena, 2023a; Ethiopian Monitor, 2023; Reqiq Staff, 2023; Teshome, 2023）。

根据卡巴斯基全球研究与分析数据和微软安全情报报告，近年来埃塞俄比亚遭受的 MUAI 以勒索软件网络钓鱼攻击最为严重（Microsoft, 2023; Tessema, 2023b）。埃塞俄比亚在 2023 年发生了一起最引人注目（也是最幽默）的案例，当时埃塞俄比亚财政部因网络钓鱼攻击将约 500 万美元转给了诈骗者，这笔款项原本是要转给非洲开发银行的（Tessema, 2023a）。诈骗者使用了非洲开发银行的凭证来组织这次攻击。最重要的是，这一事件产生了实际后果，并引发了一场外交丑闻：在此之后，非洲开发银行驻亚的斯亚贝巴代表办公室的两名员工因涉嫌欺诈而遭到暴力拘留——因为其中一名员工阿卜杜勒·卡马尔没有确认银行转账（Horn Observer Contributor, 2023）。这反过来也可能指向了埃塞俄比亚的另一个网络安全问题——数字素养水平低和缺乏网络卫生意识——因为埃塞俄比亚财政部没有核实收款人的账号。

与此同时，数字化和数字经济的发展潜在地扩大了网络攻击的范围，并可能从数量和质量上增强针对埃塞俄比亚的网络风险。特别是，人工智能可能被用来进行有针对性的网络攻击和大量发送有针对性的网络诈骗信息。大数据分析使得可以定制针对特定目标开展的网络攻击——例如某个特定的组织、生产体系等。利用基于 AI 的诈骗攻击、特定的机器学习算法和数据分析技术，攻击者将能够为“特别重要”的个体生成个性化的文本信息，如公司的高层官员、政府机构等（Bahnsen 等人，2018 年；Goldman, 2022 年；Guembe 等人，2022 年，第 84-85 页，第 89 页，第 96-97 页，第 102 页；Seymour & Tully, 2016 年；Zouave 等人，2020 年，第 22-23 页）。这样的网络攻击在埃塞俄比亚的背景下很容易煽动种族间冲突并引发新的社会紧张局势。例如，存在一种真实的可能性，即代表某个持反对意见的民族群体的所谓领导人大规模发送网络钓鱼邮件，呼吁对另一个民族社区甚至联邦政府发动战争，或者发送包含筹集资金组织民兵的信息的网络钓鱼诈骗邮件。

MUAI 对埃塞俄比亚心理安全的第三层威胁

第三层威胁直接与特定的 MUAI 技术有关，这些技术可能导致严重的后果，并在埃塞俄比亚的军事、政治、社会和公共环境中造成不稳定。一个现实生活中的情景是，一个使用人工智能制作的深度伪造视频广泛传播，该视频声称繁荣党领袖吉尔玛·耶希提拉在阿姆哈拉地区被民族主义阿姆哈拉民兵（FANO）杀害（Addis Insight, 2023）。2023 年，骗子企图冒充非洲联盟委员会主席（该组织总部位于亚的斯亚贝巴）穆萨·法基，使用深度伪造技术向多位欧洲领导人进行视频通话。

通过训练特定的信息并根据某些意识形态、政治和其他价值范式编程的聊天机器人可以实现特定的破坏性社会政治效果（Mihajlenok & Malysheva, 2020）。例如，聊天机器人使用特殊的数字平台，可以在阿姆哈拉人中传播信息，称有争议的领土实际上历史上属于提格雷人，反之亦然——这可能激发通过地方冲突表达不同意见，并导致社会秩序的不稳定。

在埃塞俄比亚的背景下，一系列消极场景可能会实现 AI 预测能力，使 AI 将作为预测性武器投入使用。

通过分析某些数据——社会稳定水平、政治偏好或对联邦政府的忠诚度等，被训练用来预测特定省份社会稳定水平的人工智能可以预见到，例如索马里将面临社会爆炸。如果人工智能假设安哈拉省的男性人口在增长，它可以推断这些人将在未来几年试图夺回有争议的领土。所有这些都可能进一步破坏省份的稳定局势。

基于人工智能的定向自动化画像技术可以在埃塞俄比亚广泛应用。它绘制心理画像，并能够基于社交网络、互联网资源、搜索查询等（最好是）公开数据的分析，对目标互联网用户进行分类，以识别他们的心理特征和情感背景，甚至预测他们未来的心理状态，目的是影响和激励他们采取某些行动（Bilal 等人，2019 年；Guembe 等人，2022 年，第 95 页；Zouave 等人，2020 年，第 19 页）。例如，人工智能可以用来分析大型目标群体的数据，在这个案例中，这些群体可以由埃塞俄比亚的不同民族代表——安哈拉人、奥罗莫人、提格雷人、奥梅塔人、伊罗布斯人等，以便利用某种算法为它们制作一种“社会心理地图”：分析行为模式和特点，识别特定民族的大众政治偏好，评估对联邦政府的忠诚度，并突出关注因素——以影响大众意识，设定期望的政治议程，推动人们朝着特定行动前进等。

此外，一些造成埃塞俄比亚心理、政治和社会稳定不利的事件可以归因于专为一系列特定任务而设计的专业算法的实际实施。

实施各种 AI 平台和数字生态系统的利益主体——一些国家和大公司——可能对埃塞俄比亚当前的不稳定状况感兴趣，以实现他们自己的利益，这些利益并不总是与埃塞俄比亚的利益兼容。最重要的是要记住，专门的平台和带有嵌入式 AI 算法的应用程序，旨在实现特定社会目标以及娱乐目的，将收集大量关于埃塞俄比亚公民的信息。外部行为者一旦能够访问这些数据，就能够不仅操纵这些信息，还能够使用它们来实现其目标，利用专门的 AI 技术实现各种目标。

进一步说，这将允许使用特定的人工智能技术对埃塞俄比亚人口（或特定目标受众）的集体意识产生有针对性的影响，以实现一系列追求的效果。例如，专门设计的人工智能技术可以用来在人群中制造并传播谣言，声称同时被阿姆哈拉人和提格雷人声称的有争议领土目前正在（或将将要）让给一方或另一方，或者联邦政府正计划减少某个省份（相应地，某个国籍）在 EPRDF 中的代表数量。这可以通过特别训练的聊天机器人、助手、深度伪造等手段来实现。

通过使用特定的人工智能技术和影响大众意识，任何有兴趣的行为者都可以塑造期望的政治议程，甚至产生必要的社会政治进程，制造紧张局势的温床并挑起新的冲突——在埃塞俄比亚目前存在的持续不

稳定和某些民族对当局的不信任的条件下，这非常容易实现。例如，2021 年 5 月 31 日，在线媒体平台 Kello Media 发布了一段假音频录音，在这段录音中，据称首相阿比·艾哈迈德在繁荣党的会议上声称他们已经赢得了选举，未来 10 年内没有其他人能够组建另一个政府（Addis Insight, 2023）。

掌握平台和相关 AI 技术的公司，与他们的政府勾结，利用他们的“数字力量”来“灌输”给埃塞俄比亚居民的某些变革需求——促进商业、基础设施、投资相关和其他项目，即使这与埃塞俄比亚人的实际利益相背。因此这些潜在因素可能引起新一波的不满和冲突。

值得注意的是，集成到数字平台的算法通过分析用户的数据，从而根据特定个人（受众）的偏好量身定制内容。然而，这种内容个性化在埃塞俄比亚的背景下可能导致负面后果，例如在政治、领土归属、宗教选择等关键问题上进一步分化民族。如果用户倾向于某种特定内容，AI 会越来越频繁地显示类似信息，将人们置于一种“信息泡沫”中，创造一种他们的信仰是唯一真实的印象，从而默认塑造他们的世界观。例如，这些数字平台会向奥罗莫人展示一些有针对性的内容——比如，关于建立他们自己的主教团的必要性，而阿姆哈拉人将接收到关于不容违背教会离散的相反内容[3]。这可能是由于平台特有的算法自然训练导致（例如，奥罗莫选择了某些内容为优先，而阿姆哈拉人也选择了他们偏好的内容），之后 AI 仅仅开始推荐类似内容。或者，这可能是有意为之的结果，当算法被以某种方式设置时。这样的机制可以用来进一步极化已经因种族矛盾而四分五裂的埃塞俄比亚社会，以及在全国范围内审查信息。

这种破坏性算法影响的真实场景在提格雷人和联邦政府之间的武装冲突（2020-2022 年）中得到展示。根据国际特赦组织的说法，Facebook 的算法默许在埃塞俄比亚传播呼吁对提格雷人实施暴力的破坏性内容。由于 Amhara 和 Oromo 语言不是 Facebook 内容审核和审查系统的“优先”语种，导致 Facebook 未能识别这些破坏性帖子。

最后，在埃塞俄比亚持续存在的潜在冲突的基础上，数据呈现和客观性问题以及人工智能在解决特定任务时是否得到正确和适当训练变得尤为棘手 - 在这一背景下，埃塞俄比亚存在严重的种族间冲突和领土争端，以及极不稳定的政治和社会经济状况。使用基于不具代表性数据的“有偏见”的人工智能可能会在某些领域带来非常严重的后果，并进一步破坏局势，导致定期“爆炸”现象。外国公司和埃塞俄比亚国家结构培养的特定社会、公共和政治问题解决的人工智能技术可能会对某些人群进行歧视，包括基于种族的歧视。例如，用于特定公共目标的任何社会算法 - 如分析期末考试成绩、评估求职者简历等 - 可能会在很大程度上赋予奥罗莫人权利，同时暗含地歧视提格雷人：如果某个公司雇佣了更多的奥罗莫人，那么在这样的数据上训练的人工智能将继续优先考虑他们。或者，用于政治过程的算法可以授予某些人更多的政治权利，同时边缘化其他人。例如，如果基于不具代表性数据进行培训的人工智能应用于分析 EPRDF 的组成，理论上可能会向阿姆哈拉人提供更多的席位而不是提格雷人。在这种情况下，少数民族的代表将特别容易受到歧视。

可以想象，利用人工智能影响埃塞俄比亚人民的大众意识可能会催化新一波暴力，导致基于多重种族紧张局势的血腥冲突。这些冲突基于多重族裔紧张关系而爆发，并可能被那些与联邦政府政策不一致的外部和内部感兴趣的行为者所利用。

[3] 让人想起 2023 年 2 月在奥罗莫居民区爆发的冲突。

结论

如今，埃塞俄比亚面临着广泛的内部矛盾，同时受到外部行为者的压力。这种情况使得埃塞俄比亚成为了恶意使用人工智能（MUAI）在破坏公共和社会政治状况方面的一个绝佳潜在目标。此外，在埃塞俄比亚的 实际基础上，应用社会和政治导向的算法本身可能会导致对某些人群的歧视 - 如果人工智能是基于不具代表性的数据进行训练的话。

同时，通过传播进口的人工智能系统和嵌有算法的数字平台，埃塞俄比亚所面临的生存挑战可能会被激化。潜在地，这可能导致“AI 新殖民主义”现象，使埃塞俄比亚过度依赖外国数字基础设施和引进不考虑当地利益、种族特殊性、文化和心态的外来人工智能系统。据预计，使用这些系统的伦理、法律和社会文化后果将被忽视。此外，通过使用数字平台，企业可以获取埃塞俄比亚居民大量的个人数据，随后利用这些数据来实现自己的利益。

参考文献

Adams R (2021) 人工智能能否实现去殖民化？《跨学科科学评论》，第 46 卷 1-2 期，页 176-197。https://doi.org/10.1080/03080188.2020.1840225

Ade-Ibijola A, Okonkwo C (2023) 非洲的人工智能：新兴挑战。见：Eke DO, Wakunuma K, Akintoye S 编，《非洲负责任的人工智能：挑战与机遇》。帕尔格雷夫麦克米伦出版社，Cham，页 101-117。https://doi.org/10.1007/978-3-031-08215-3

Afriyie FA, Ayangbah S, Effah KO (2023) 诊断埃塞俄比亚提格雷战争：非洲之角的回荡。《非洲洞察》，第 15 卷第 2 期，页 139-151。https://doi.org/10.1177/09750878231170177

Bilal M, Gani A, Lali MIU, Marjani M, Malik N (2019) 社交画像：评论、分类和挑战。《网络心理学、行为与社交网络》，第 22 卷第 7 期，页 433-450。https://doi.org/10.1089/cyber.2018.0670

Birhane A (2023) 非洲的算法殖民。见：Cave S, Dihal K 编，《想象人工智能：世界如何看待智能机器》。牛津大学出版社，牛津，页 247-260。https://doi.org/10.1093/oso/9780192865366.003.0016

Blackwell AF, Damena A, Tegegne T (2021) 在埃塞俄比亚创造人工智能。《跨学科科学评论》，第 46 卷第 3 期，页 363-385。https://doi.org/10.1080/03080188.2020.1830234

预防性行动中心 (2023) 埃塞俄比亚的冲突。见：美国对外关系委员会。https://www.cfr.org/global-conflict-tracker/conflict/conflict-ethiopia。访问日期：2024 年 2 月 3 日

Eke DO, Wakunuma K, Akintoye S (2023a) 介绍非洲负责任的人工智能。见：Eke DO, Wakunuma K, Akintoye S 编，《非洲负责任的人工智能：挑战与机遇》。帕尔格雷夫麦克米伦出版社，Cham，页 1-12。https://doi.org/10.1007/978-3-031-08215-3

Eke DO, Wakunuma K, Akintoye S 编 (2023b) 《非洲负责任的人工智能：挑战与机遇》。帕尔格雷夫麦克米伦出版社，Cham。https://doi.org/10.1007/978-3-031-08215-3

Ena (2023a) 埃塞俄比亚成功阻止网络攻击的指数型增长，挫败超过 96%的攻击。https://www.ena.et/web/eng/w/eng_3120234。访问日期：2024 年 2 月 3 日

Ena (2023b) INSA 在上一财年挫败 6768 次网络攻击。https://www.ena.et/web/eng/w/eng_3120234。访问日期：2024 年 2 月 3 日

埃塞俄比亚监视报 (2023) 在 12 个月内，INSA 挫败了逾 6700 次针对埃塞俄比亚的网络攻击。https://ethiopianmonitor.com/2023/07/24/insa-foils-over-6700-cyberattack-attempts/。访问日期：2024 年 2 月 3 日

埃塞俄比亚联邦民主共和国 (2020) 数字埃塞俄比亚 2025 - 埃塞俄比亚包容性繁荣战略。

见：埃塞俄比亚法律信息门户。
https://www.lawethiopia.com/images/Policy_documents/Digital-Ethiopia-2025-Strategy-english.pdf。访问日期：2024 年 2 月 3 日
等。
Gadzala A (2018) 将要启动：非洲的人工智能。见：大西洋理事会。
https://www.atlanticcouncil.org/wp-content/uploads/2019/09/Coming-to-Life-Artificial-Intelligence-in-Africa.pdf。访问日期：2024 年 2 月 3 日
Girmay FG (2019) 埃塞俄比亚的人工智能：机遇与挑战。《信息技术师：信息和通信技术国际期刊》，第 16 卷第 1 期，页 157-180。
Guembe B, Azeta A, Misra S, Chukwudi Osamor V, Fernandez-Sanz VSL, Pospelova V (2022) 由 AI 驱动的网络攻击的新兴威胁：综述。《应用人工智能国际期刊》，第 36 卷第 1 期，页 1-34。
https://doi.org/10.1080/08839514.2022.2037254
Kemp S (2023) 埃塞俄比亚数字 2023。见：数据报告。
https://datareportal.com/reports/digital-2023-ethiopia。访问日期：2024 年 2 月 3 日
微软 (2023) 微软安全情报报告。https://info.microsoft.com/SIRv24Report.html。访问日期：2024 年 2 月 3 日
Mihajlenok OM, Malysheva GA (2020) 社交媒体的机器化及其政治后果 [Mihajlenok OM, Malysheva GA]. 《权力》（Power），第 28 卷第 1 期，页 85-92。
Okolo CT, Aruleba K, Obaido G (2023) 非洲负责任的人工智能 - 挑战与机遇。见：Eke DO, Wakunuma K, Akintoye S 编，《非洲负责任的人工智能：挑战与机遇》。帕尔格雷夫麦克米伦出版社，Cham，页 35-64。
https://doi.org/10.1007/978-3-031-08215-3
Reqiq 工作人员 (2023) 艰巨的数字前沿：埃塞俄比亚的网络安全状况。见：Reqiq 洞察。https://reqiq.co/a-daunting-digital-frontier-the-state-of-cybersecurity-in-ethiopia/。访问日期：2024 年 2 月 3 日
Teshome M (2023) 针对埃塞俄比亚的网络攻击猛增。见：《首都埃塞俄比亚》。
https://www.capitalethiopia.com/2023/06/12/cyber-attacks-bombard-ethiopia/。访问日期：2024 年 2 月 3 日
Tessema B (2023) 增加的网络攻击针对埃塞俄比亚。见：Abren。
https://abren.org/increasing-cyber-attacks-target-ethiopia/。访问日期：2024 年 2 月 3 日
《对话》(2023) AI 将取代谁的工作？为什么埃塞俄比亚的文员比加利福尼亚的文员更应该担心。
https://theconversation.com/whose-job-will-ai-replace-heres-why-a-clerk-in-ethiopia-has-more-to-fear-than-one-in-california-216735。访问日期：2024 年 2 月 3 日
《经济学人》(2023) 去年世界上最致命的战争不是在乌克兰。
https://www.economist.com/international/2023/04/17/the-worlds-deadliest-war-last-year-wasnt-in-ukraine 。访问日期：2024 年 2 月 3 日
联合国 (2023) 随着 AI 的出现，工作正在改变，但并预计不会出现大规模失业 - 联合国劳工专家。见：联合国经济及社会事务部。
https://www.un.org/ru/desa/ai-jobs-are-changing-no-mass-unemployment-expected-un-labour-experts。访问日期：2024 年 2 月 3 日
Zouave E, Bruce M, Colde K, Jaitner M, Rodhe I, Gustafsson T (2020) 人工智能网络攻击。见：瑞典国防研究院 FOI。
https://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf。
访问日期：2024 年 2 月 3 日

人工智能的恶意使用：巴西联邦共和国心理安全的挑战

叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）
达利娅·巴扎尔金娜，俄罗斯科学院欧洲研究所（莫斯科，俄罗斯）

引言

2024 年 1 月，巴西总统路易斯·伊纳西奥·卢拉·达席尔瓦(Luiz Inacio Lula da Silva)公布了一项为期十年的发展计划，旨在通过国家信贷和补贴来促进本国工业增长。其中，数字化转型是该计划中设定的目标之一，即达成巴西工业部门所有企业 90%的数字化（目前该行业中运营数字化的公司比例为 23.5%）。这项目标将涉及对发展本土工业 4.0 的投资，包括将智能数字技术整合到制造和工业流程中，以及增加国内半导体生产(Mari 2024)。据估计，2024 年，巴西人工智能市场规模将达到 43.7 亿美元。而在全球市场中，美国的人工智能市场规模将是最大的（2024 年为 1065 亿美元）。巴西的人工智能市场规模将以每年 17.65%的速度增长，并预计在 2030 年达到 115.9 亿美元（Statista 2023a）。

由视觉内容创作和营销公司 Getty Images 制作的 VisualGPS 研究调查了来自 25 多个国家的 7,000 多名成年人，调查发现，六分之四的巴西人认为人工智能对他们的生活有积极影响。这一调查结果高于全球半数的平均水平。因此，基于该研究可得出，与美国、加拿大、法国、英国和澳大利亚等国家相比，巴西消费者对人工智能的兴趣要高出 15%（Mari 2023a）。此外，与其他国家相比，不及 34%的巴西人认为这项技术的进步对自身构成威胁（Mari 2023a）。但与此同时，巴西人工智能在恶意运用方面实际上已经对巴西社会的心理安全造成了威胁，这在很大程度上取决于该国社会政治矛盾的严重程度。该情况实际对巴西社会敲响了警钟。

巴西总统卢拉目前的首要任务仍然是帮助 7100 万巴西人（占人口的 33%）摆脱贫困。但国际货币基金组织(IMF)对这一改变持悲观态度。他们的研究结果显示，到 2028 年，巴西预计仅以每年 2%的速度增长，而这种微弱的增长速度在减贫方面带来的贡献非常有限（Martin 2024, p. 2）。事实上，在 2024 年 1 月的调查中，几乎有一半的巴西人（49%）担心他们的收入会在 2024 年下半年下降，而并非增加（36%），尽管他们预计就业市场会有所好转(路透社 2024)。而巴西的社会问题也呈上升趋势，在 2023 年的上半年，巴西发生了 1790 起谋杀案，而仅在 2022 年 1 月至 6 月间就有 1526 起（Instituto Sou da Paz 2023）。值得注意的是，2024 年 10 月的市政选举将是卢拉总统在 2024 年面临的最大政治挑战，此次选举是其与雅伊尔·博尔索纳罗(Jair Bolsonaro)领导的右翼发生新冲突的一年。在不断增加的社会政治问题和矛盾背景下，人工智能技术的进一步变革和传播自然会使恶意运用的概率增长。

MUAI 对心理安全的第一层威胁

巴西人对人工智能未来的普遍乐观的态度实际是把双刃剑。一方面，对新技术的积极态度有助于推动其发展和传播，总体上提高社会的福祉水平。然而，另一方面，当人们低估了人工智能代表的新技术对他们安全将产生的风险时，技术在某个阶段与现实的碰撞可能引发极端负面的公众反应。在当今复杂的内部局势和百年未有之大变局中，工智能发展中的困难、不平衡或错误都可能被国家内部和外部的恶意行动者放大。这一问题一边很大程度上取决于政府能否迅速推出适合该国人工智能发展的法律框架（工作正在这个方向上进行），另一边在于当局能否向公众传达人工智能技术带来的巨大机遇和同样重要的风险。

MUAI 对心理安全的第二层威胁

据网络安全公司趋势科技（Trend Micro）发布的报告，巴西是全球第二易受网络攻击威胁的国家（Mari 2023b）。2021 年，巴西和厄瓜多尔是拉丁美洲中两个最受网络钓鱼攻击威胁的国家，分别占 12.39%和 10.73%（Bianchi 2021）。2020 年，巴西创下了网络钓鱼攻击数量最高的全球纪录：该国每五位互联网用户中就有一人至少一次受到了网络钓鱼攻击（Mari 2022b）。根据卡巴斯基（Kaspersky）的数据，2021 年巴西遭受了 2,500 万次尝试性攻击，2022 年上升至 1.34 亿次网络钓鱼尝试（Folha Vitória 2023）。此外，电子商务和社交网络也在这类犯罪中占据了一席之地。由于电子商务和社交网络需要使用电子邮件地址来认证用户身份，且人们在在线访问时会使用电子邮箱地址和密码，当诈骗分子窃取受害者的电子邮件密码时，就可以进入其银行账户进行犯罪。

2022 年 2 月，巴西被垃圾邮件组织（Spamhaus）列入垃圾邮件机器人数量最多的国家名单。这些机器人大多用于发送垃圾邮件、网络钓鱼、分布式拒绝服务（DDoS）攻击和其他恶意活动。分析人士指出，巴西数字空间中机器人的大量存在与技术、政治和社会经济因素有关（The Spamhaus Project 2022）。此外，与其他国家一样，伴随着人工智能技术的普及，网络攻击技术的复杂化对企业和组织基础设施构成了威胁，其可以中断企业运营、清除关键数据并造成企业声誉损害。网络犯罪分子将能够以前所未有的速度和规模发起有针对性的攻击，同时避开传统的基于规则的检测措施（Guembe et al. 2022）。由于该国网络安全发展滞后，可以假定巴西对这类攻击的脆弱性较大。

MUAI 对心理安全的第三层威胁

在巴西，MUAI 的第三层威胁与在选举活动期间使用社交媒体机器人有关，这一现象已经持续数年。在 2022 年的总统选举中，在选票计算后的几个月内，Twitter 上的伪造账户数量显著增加，其中大多数攻击了左翼总统卢拉·达席尔瓦(Lula da Silva)。数据表明，互联网上机器人的活动呈现除了显著增长（Lima 2023）。

2022 年巴西全国选举期间，恶意使用由人工智能（AI）生成的深度伪造技术（deepfake）广泛传播，制造了大量虚假图像和视频，编造了主要候选人的各种丑闻和妥协场景。这些深度伪造技术利用先进的 AI 和复杂的编辑工具，精准模仿候选人的声音和面部表情，生成高度逼真的结果，常常导致公众信任的动摇和选举环境的扭曲（Ünver 2023）。

2022 年，一段针对时任总统路易斯·伊纳西奥·卢拉·达席尔瓦的政治深度伪造视频广泛传播。8 月 5 日，一段假视频出现在社交网络上，视频中知名巴西电视主持人雷纳塔·瓦斯康塞洛斯（Renata Vasconcellos）被伪造地发布关于总统选举投票结果的虚假信息。视频中雷纳塔的声音被稍作修改，令观看者误以为雅伊尔·博索纳罗在 8 月 15 日发布的投票意向调查中领先于卢拉。然而，在真实视频中显示的结果是，卢拉的投票意向为 44%，而博索纳罗为 32%。这一问题的严重性在于，尽管该视频在 YouTube 上的存在时间很短，但它在 WhatsApp 群组和其他社交网络上有非常广泛的传播（Pacheco 2023）。

根据 Avast 拉丁美洲地区总监哈维尔·林孔（Javier Rincón）的说法，“社会对假新闻的认知程度对于打击虚假信息至关重要，因为用户一旦开始消费虚假信息新闻网站，就可能会越来越陷入假新闻的漩涡。Avast 人工智能团队的研究显示，超过 17%传播虚假信息的网站都有其他假新闻网站的链接。这很快会形成假新闻消费的链条”（Bento 2022）。由此可见，操纵选民始终是一种极具破坏性的武器，而一旦这种武器与人工智能结合，其操纵的威力可以成倍增加。巴西选举中恶意使用人工智能（MUAI）操纵和破坏选民的意愿可以见到（Resende 2023）。

在脆弱的政治环境中，对媒体信任的消减可能会带来深远的影响。帮助人们记录人权侵犯行为的非营利组织的 Witness 组织的项目总监萨姆·格雷戈里（Sam Gregory）提供了一个例子。在有警察暴力历史的巴西，市民和活动家担心他们拍摄的警察杀害平民的视频将不再能成为调查的充分依据。格雷戈里表示，在他的研讨会上，担心真实证据可能被合理地视为假证据已成为一个反复出现的主题（Hao 2019）。因此，如今主要问题不仅在于揭露深度伪造的实践，还在于有效地检测深度伪造。这一问题还被那些试图将基于照片、视频、音频或书面证据的任何公正批评称为深度伪造的当局所加剧。南加州大学的郝立教授（Hao Li）对此表示认同。深度伪造被滥用的风险在于，人们利用其存在来质疑真实的视频证据：“即使有你做或说某事的录像，你也可以说那是深度伪造，而且很难证明不是。”世界各地的政治家已经有人被指控使用这种手段，如圣保罗市长若昂·多利亚（João Doria）。2018 年，这位已婚政客声称一段显示他参与淫乱的视频是深度伪造，但没有可令人信服的证据（Thomas 2020）。因此，恶意使用深度伪造在巴西既是第一层次的威胁，也是第三层次的威胁。

深度伪造（deepfake）可以作为强大且极具威胁性的信息误导工具削弱公众意识（Pinheiro de Resende 2021）。利用深度伪造，欺骗人们变得相对容易，因为观众相信他们正在观看的内容是真实发生的事件。一方面，声音也可以合成生成：记者玛加丽·普拉多（Magali Prado）在她的书《假新闻与人工智能：算法在打击虚假信息中的应用》中提到，深度伪造的音频文件可以很容易地通过 WhatsApp 等平台传播，而 WhatsApp 在巴西被广泛使用。随着软件得以革新和改造，AI 可以模拟人类声音。而这一现象的主要受害者是公众人物，因为他们的声音在公共领域资源非常丰富。这种方法还可以用于金融诈骗。“有一个案例，一家科技公司的员工收到了一条来自高管的语音信息，要求他转账。他起了疑心，并将信息交由安全公司分析，结果确认这是利用人工智能生成的”（Schmidt 2022）。

深度伪造越复杂，国家越不民主，民众越政治冷漠且数字素养低，深度伪造的社会危害就越大。为了消除深度伪造及其他伪造信息，政府必须稳定且全面地揭露具有社会意义的深度伪造案例，并以代表客观事实的真相标记这些伪造信息。但是，如果没有公民的有意识关注和参与，政府不太可能愿意或能够有效应对深度伪造的恶意使用。证明什么是真实的，什么是虚假的，将在没有具体政治前提的情况下变得极为困难。

在巴西（以及其他一些国家），虚拟社交媒体影响者正变得越来越受欢迎。他们虽然看起来像真实的人，但却是 100%数字化的。它们可以在平台上做真实的事情，例如，它们可以说话、跳舞、玩耍。一言以蔽之，它们可以展现出真实人类的行为。除此之外，它们还有自己强烈的观点，这使它们在社交媒体上备受关注，这也许是为什么它们有如此多的粉丝并且如此受欢迎的原因（Little Black Book 2022）。其中一个例子是 Lil Miquela，又名米奎拉·索萨（Miquela Sousa）。这位 19 岁的巴西裔美国社交媒体明星在平台上积累了超过 250 万的粉丝，并经常发布与宝马（BMW）和 Pacsun 等品牌合作的赞助内容。但她并不是一个搬到洛杉矶的年轻人，她是虚拟的，是通过计算机生成的图像制作而成的。Lil Miquela 于 2016 年首次亮相。她在 Instagram 上首次亮相后，迅速走红。当米奎拉不得不向她的粉丝承认，她被她的劲敌、一名支持特朗普的网络喷子百慕大（Bermuda）“黑客攻击”时，许多人对米奎拉的人类般的外表感到迷惑，纷纷猜测她是一场营销噱头，也许是一个真实的人。最终，在 2018 年，她的创作者，机器人和人工智能公司 Brud 的特雷弗·麦克费德里斯（Trevor McFedries）和萨拉·德库（Sara DeCou）宣布他们是 Lil Miquela（以及百慕大）的幕后推手（Sheena 2023）。

在社交媒体上拥有 3120 万粉丝的卢·多·马加卢（Lu Do Magalu）是互联网上最大的虚拟影响者（Petrarca 2018）。卢是 Magazine Luiza 公司 CEO 弗雷德里科·特拉哈诺（Frederico Trajano）的心血结晶，Magazine Luiza 是一家多元化的巴西消费品牌企业集团，旗下包括 Magalu，该公司是巴西最大的连锁零售商之一，在全国范围内拥有超过 1300 家实体店。卢在 2003 年诞生，正值电子商务领域开始显示

出最终可能成为传统实体销售可行选择的迹象之时。“我们负责卢的人性化，”艾琳·伊佐（Aline Izo）在接受 Observer 采访时评论道。伊佐领导着一个由 3D 设计师、程序员和营销人员组成的团队，精心照料着这位备受追捧的虚拟影响者的各个方面。她声称，“卢拥有数百万的粉丝，当她对某事发表看法时——例如关注家庭暴力或坚定支持 LGBT 权利——会引起人们注意。在巴西，卢不是销售噱头；卢是真正意义上的影响者；卢可以在社会中推动重要问题的解决”（Wierson 2021）。虚拟影响者是一个数字化的个性，通过在社交媒体上发布内容来吸引一群热情的粉丝，就像人类影响者一样；至少，表面上是这样。这些 AI 影响者的某种政治内容是显而易见的，它们究竟在多大程度上操控成年人、影响大量儿童和青少年，仍然是一个未解之谜。东北大学营销学教授及东北大学体验人工智能研究所领导委员会成员雅科夫·巴特（Yakov Bart）表示，“在某些情况下，考虑到与影响者互动后消费者心态的变化与成本之间的回报，使用虚拟或合成的影响者更为高效”（Contreras 2024）。来自瑞士商学院（SBS Swiss Business School）的迈克尔·格利希（Michael Gerlich）得出结论称，虚拟影响者可以与客户建立更深入和持久的联系，虚拟影响者在编程和训练 AI 引擎方面的灵活性使它们能够适应不断变化的客户行为，而虚拟影响者比人类影响者具有更高的信誉和可信度，是影响营销的未来，可以增加客户的购买意愿和整体品牌知名度（Gerlich 2023, p. 19）。沙迦美国大学（American University of Sharjah）营销学教授莫娜·姆拉德（Mona Mrad）承认，对于一些人来说，真人追随虚拟影响者的概念可能显得荒谬，但对于某些世代来说却并非如此。Z 世代，即 1990 年代末到 2010 年代初出生的人，正接受虚拟影响者。“这一代人与这些影响者联系紧密，甚至形成了情感和精神上的关系……他们对这些影响者表达爱意和依恋。”在某些情况下，他们甚至认为虚拟影响者比真人更可靠（Kugler 2023）。截至 2020 年 5 月，巴西拥有近 920 万的 Instagram 影响者，是拉丁美洲拥有数字影响者数量最多的国家。排名第二的阿根廷注册了超过 110 万名影响者（Statista 2023b）。而这一用户基础使得巴西特别容易受到虚拟影响者的操控。

在巴西，基于 AI 的预测分析选择性地发挥作用，这一点相当显著。2023 年 1 月 8 日，也就是路易斯·伊纳西奥·卢拉·达席尔瓦（Luis Inacio Lula da Silva）第三次就任巴西总统的一周后，前总统雅伊尔·博尔索纳罗（Jair Bolsonaro）的极右翼支持者冲击了巴西利亚的政府大楼。“除了社交媒体巨头，所有人都看到了巴西的暴力，”POLITICO 的马克·斯科特（Mark Scott）在叛乱发生的第二天写道，并认为“硅谷的巨头们再次在关键时刻无所作为……”（Digital Action 2023）。在袭击发生前的几个月里，专家们一再警告说，极右翼势力一直在利用 WhatsApp 和 Telegram 等加密平台进行组织、传播虚假信息和煽动叛乱，并敦促有关人士在新政府上任后的前几周内继续保持与选举相关的安全措施。尽管许多人提出了这些请求，但这些公司并未理会（Digital Action 2023）。不仅专家们的声音未被采纳，基于 AI 的预测分析资源也未被利用。另一方面，据巴西政治营销协会负责人埃默森·萨拉伊瓦（Emerson Saraiva）称，AI 将使人们能够“提前很久”知道 2024 年市政选举的结果（Silva 2023）。因此，宣布了使用 AI 提前获得选举结果的可能性，这可能会影响选民是否投票以及如何投票的决定。可以得出结论，巴西（不仅仅是巴西）已经出现了使用基于 AI 的预测分析作为预言武器的条件。

深度伪造、虚拟影响者、聊天机器人、预测分析以及其他利用人工智能影响公众舆论的手段，可能对该国的政治进程产生最具破坏性的影响。巴西最高选举法院（TSE）将在 2024 年第一季度举行关于在下次选举中使用人工智能的法规辩论。TSE 主席亚历山德雷·德·莫拉埃斯（Alexandre de Moraes）认为，利用人工智能，可以例如修改反对候选人的视频，使他们发表从未说过的言论。“想象一下，有多少人可能会被虚假的新闻、虚假的信息轰炸，但这种虚假信息来自几乎可以保证真实性的演讲视频。这样的攻击非常严重。特别是使用人工智能的攻击，真的可以改变选举结果，可以在两极分化的选举中扭曲选举结果”（Tocarnia 2023）。

值得注意的是，卢拉政府提出的旨在压制网络上的宣传和操控的措施，遭到了美国领先的跨国科技公司的强烈反对。待决的巴西国会法案第 2630 号（巴西联邦参议院 2020），官方名称为《互联网自由、责任和透明度法》，被巴西媒体称为“假新闻法”，反对者则称其为“审查法”，该法案旨在打击虚假信息的传播，而这些虚假信息越来越多地依赖于人工智能技术的支持。这一领域的监管早已迫在眉睫。根据 Avast 的研究，五分之四（79%）的巴西人在社交网络上发现了有关 2022 年选举的假新闻，且大多数人（57%）不相信（或不确定）社交媒体是可靠的信息来源。此外，86%的巴西人认为媒体应该对其网络上的假新闻的撤除负责（Bento 2022）。对政府和绝大多数巴西人来说，采取紧急行动纠正这种情况是显而易见的，但美国的跨国高科技巨头却并不这么认为。

在 2023 年 5 月初，当法案即将获得批准时，谷歌和电报利用其自己的平台向巴西用户表达了他们对该法案的反对意见。5 月 1 日，巴西人有点惊讶地看到在谷歌首页的熟悉搜索框上出现了一个链接，上面写着：“假新闻法案可能会使您的互联网变得更糟。”点击该链接的人被带到了一个批评 2630 法案草案的谷歌博客，该法案将在巴西国会的下一天进行投票。这个被巴西 1.6 亿互联网用户中超过 90%使用的搜索首页还声称在另一个链接中：“假新闻法案可能会在巴西造成关于什么是真实的混乱。”谷歌的策略还包括向 YouTuber 发送电子邮件，称他们将有更少的资金投资于他们的频道，并要求他们与国会交谈。根据里约热内卢联邦大学的一项研究（Viana 2023），这家科技巨头还在搜索结果中弄错了，突出显示了自己的博客文章和其他批评该法案的文章。“巴西即将通过一项将终结言论自由的法律，”电报在 5 月向用户发送的一条消息中说，该法案已经在参议院通过，正在等待国会下议院的投票（France 24 2023）。

因此，有充分理由相信，领先的美国科技公司正在拖延采取有效措施对抗虚假信息的传播，这种传播越来越多地基于人工智能技术，会造成越来越突出的影响。在此背后既有公司的自私金融利益（不愿承担法律义务来打击在互联网上无意或有意违反公共规范的人），也有对对手施压的愿望，甚至到了煽动骚乱和推翻意欲侵占超额利润的不受欢迎政府的程度。另一个重要的因素是美国政府的压力，后者即试图试图利用美国科技公司获取敏感信息，又对不受欢迎的政府施压，总的来说，这更多地是联合而不是分裂华盛顿和主要信息公司，但并不排除它们之间存在某些尖锐的分歧。

结论

在巴西，通过 MUAI（恶意使用人工智能）对心理安全产生威胁目前主要体现在第二和第三级，但随着 AI 技术的发展、其资本化规模的扩大以及巴西持续存在的尖锐社会政治矛盾，这些威胁在第一级的表现也不能被排除。MUAI 威胁在数量和质量上都在增长，且变得更加多样化。巴西加入金砖国家组织（BRICS）既为该国开辟了新的机会，也带来了相当大的风险，主要来自那些不接受巴西独立路线的力量。巴西的网络安全水平有待提高，目前立法举措跟不上恶意行为者的实际操作，且在心理安全领域缺乏系统性应对 MUAI 的措施，因此未来可能会出现 MUAI 的激增。与此同时，卢拉政府明确致力于抵御心理侵略，作为试图社会政治稳定的一部分，考虑到 AI 在这些企图中的日益重要作用，这可能成为未来制定系统性应对措施的初步基础。

参考文献

Bento G (2022) 谣言：79%的巴西人发现了 2022 年选举的虚假信息。在：Olhar Digital。https://olhardigital.com.br/2022/10/04/pro/fake-news-79-dos-brasileiros-encontraram-mentiras-sobre-as-eleicoes-2022-na-internet/。访问日期 2024 年 2 月 2 日

Bianchi T（2021）拉丁美洲和加勒比国家是 2021 年网络钓鱼攻击的主要目标。在：Statista。https://www.statista.com/statistics/997956/phishing-attack-user-share-latin-america-country/。访问日期 2023 年 12 月 8 日
Contreras C（2024）人工智能是否正在摧毁社交媒体明星？公司如何赚取虚拟推广者的利润。在：Phys.Org。https://phys.org/news/2024-01-ai-social-media-star-companies.html。访问日期 2024 年 2 月 2 日
Digital Action（2023）巴西市政选举：大型科技公司对 1 月 8 日的袭击有所感悟吗？在：民主之年。https://yearofdemocracy.org/case-study/brazil-municipal-elections-have-big-tech-companies-learnt-anything-from-the-january-8th-attacks/。访问日期 2024 年 2 月 2 日
Folha Vitória（2023）巴西一年内存在 1.34 亿次网络钓鱼尝试。https://www.folhavitoria.com.br/geral/noticia/09/2023/brasil-teve-134-milhoes-de-tentativas-de-phishing-em-um-ano。访问日期 2024 年 2 月 2 日
France 24（2023）美国科技巨头 Telegram 称巴西的虚假信息法是对民主的攻击。https://www.france24.com/en/americas/20230509-messaging-app-telegram-calls-brazil-disinformation-law-attack-on-democracy。访问日期 2024 年 2 月 2 日
Gerlich M（2023）虚拟推广者的力量：在人工智能时代对消费者行为和态度的影响。管理科学。第 13 卷（8 期）：178。https://doi.org/10.3390/admsci13080178。
Guembe B，Azeta A，Misra S，奥萨莫 C，费尔南德斯-桑兹 I，波斯佩列娃 V（2022）人工智能驱动的网络攻击的新威胁：一项评论。在：应用人工智能，第 1 期。
Hao K（2019）深度伪造的最大威胁不是深度伪造本身。在：麻省理工技术评论。https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/。访问日期 2022 年 6 月 21 日
Instituto Sou da Paz（2023）G1.暴力监控：里约热内卢每天发生 10 起谋杀案，是全国上半年第二大增长。https://soudapaz.org/noticias/g1-monitor-da-violencia-rj-registra-10-assassinatos-por-dia-e-tem-2a-maior-alta-do-pais-no-1o-semestre/。访问日期 2024 年 2 月 2 日
Kugler L（2023）现实世界中的虚拟推广者。《ACM 通信》。2023 年 3 月，第 66 卷，第 3 期，第 23-25 页。https://doi.org/10.1145/3579635。
Lima C（2023）分析发现在巴西，否认选举的假冒 Twitter 账号数量激增。在：《华盛顿邮报》。https://www.washingtonpost.com/politics/2023/01/19/fake-twitter-accounts-denying-election-surged-brazil-analysis-finds/。访问日期 2023 年 12 月 10 日
Little Black Book（2022）Magalu 的 Lu 是如何成为世界上最大的虚拟推广者的。https://www.lbbonline.com/news/how-lu-from-magalu-became-the-biggest-virtual-influencer-in-the-world 。访问日期 2024 年 2 月 2 日
Mari A（2022）巴西在科技投资和创新方面停滞不前。在：ZDNet。https://www.zdnet.com/article/brazil-stagnant-in-tech-investments-and-innovation/。访问日期 2022 年 6 月 21 日
Mari A（2023）研究表明，巴西是对人工智能持最乐观态度的国家之一。在：福布斯。https://www.forbes.com/sites/angelicamarideoliveira/2023/11/03/brazil-among-most-optimistic-countries-about-ai-study-says/?sh=673b14532daa。访问日期 2024 年 1 月 30 日
Mari A（2023）巴西是全球第二易受网络攻击的国家。在：福布斯。https://www.forbes.com/sites/angelicamarideoliveira/2023/09/27/brazil-is-the-worlds-second-most-vulnerable-country-to-cyberattacks/?sh=699e7f0a27a4。访问日期 2023 年 12 月 12 日
Mari A（2024）技术成为巴西新工业政策的核心。在：福布斯。https://www.forbes.com/sites/angelicamarideoliveira/2024/01/25/technology-takes-center-stage-in-brazils-new-industrial-policy/?sh=18514b5524d3。访问日期 2024 年 1 月 30 日
Martin J-L（2024）卢拉的执政第一年：巴西政治局势概述。IFRI 备忘录。1 月 11 日。

https://www.ifri.org/sites/default/files/atoms/files/ifri_martin_brazil_first_year_lula_2024.pdf。访问日期 2024 年 2 月 2 日
Pacheco V（2023）选举的首个深度伪造显示总统选举的虚假数字。在：Showmetech。
https://www.showmetech.com.br/deepfake-das-eleicoes-mostra-pesquisa-falsa/。访问日期 2023 年 12 月 13 日
Petrarca E（2018）身体欺诈。米奎拉·索萨在 Instagram 上拥有超过 100 万粉丝，并最近被特朗普搅混。但她并不是真实存在的。
https://www.thecut.com/2018/05/lil-miquela-digital-avatar-instagram-influencer.html。访问日期 2024 年 2 月 2 日
Pinheiro de Resende SM（2021）深度伪造对政治和数据正义问题的影响—来自巴西和美国的观点。
https://arno.uvt.nl/show.cgi?fid=156499。访问日期 2023 年 12 月 13 日
Resende F（2023）令人不安的威胁：人工智能的恶意使用。在：Tribuna entorno。
https://www.tribunadoentorno.com.br/2023/06/ameaca-inquietante-o-uso-malicioso-da-ia.html?m=1。访问日期 2023 年 12 月 10 日
Reuters（2024）卢拉的支持率在巴西地方选举之前略有上升—民意调查。
https://www.reuters.com/world/americas/lulas-approval-ratings-inch-up-ahead-brazils-local-elections-poll-2024-01-23/。访问日期 2024 年 2 月 2 日
Schmidt S（2022）深度伪造。在：Pesquina Fapesp。
https://revistapesquisa.fapesp.br/en/deepfake/。访问日期 2023 年 12 月 13 日
Senado Federal（2020）第 2630 号法案（虚假新闻法）。
https://www25.senado.leg.br/web/atividade/materias/-/materia/141944。访问日期 2024 年 2 月 2 日
Sheena J（2023）品牌仍在研究虚拟推广者。在：Marketing Brew。
https://www.marketingbrew.com/stories/2023/09/12/brands-are-still-figuring-out-virtual-influencers。访问日期 2024 年 2 月 2 日
Silva C（2023）人工智能在巴西政治世界中的选举使用。在：《巴西报告》。
https://brazilian.report/power/2023/12/13/electoral-use-of-ai-rattles-political-world/。访问日期 2024 年 2 月 2 日
Statista（2023a）人工智能—巴西。
https://www.statista.com/outlook/tmo/artificial-intelligence/brazil。访问日期 2024 年 1 月 30 日
Statista（2023b）拉丁美洲国家中 Instagram 推广者最多的国家（截至 2020 年 5 月）。
https://www.statista.com/statistics/1126484/countries-most-social-media-influencers-latin-america/。访问日期 2024 年 2 月 2 日
Teixeira PS（2023）巴西在 WhatsApp 中排名全球领先的虚假链接欺诈案。在：圣保罗报。
https://www1.folha.uol.com.br/tec/2023/03/brasil-e-lider-global-em-golpe-de-link-falso-no-whatsapp-diz-kaspersky.shtml。访问日期 2023 年 12 月 8 日
The Spamhaus Project（2022）最恶劣的僵尸网络国家前 10 名。在：Spamhaus.org。
https://www.spamhaus.org/statistics/botnet-cc/。访问日期 2022 年 6 月 21 日
Thomas D（2020）深度伪造：对民主的威胁还是仅仅是一点乐趣？在：BBC 新闻。
https://www.bbc.com/news/business-51204954。访问日期 2022 年 6 月 21 日
Tocarnia M（2023）巴西选举委员会将讨论 2024 年选举的人工智能管理。在：巴西新闻社。
https://agenciabrasil.ebc.com.br/justica/noticia/2023-12/tse-debatera-regulamentacao-da-ia-para-eleicoes-de-2024。访问日期 2024 年 2 月 2 日
Ünver H（2023）技术的作用：信息操纵和虚假信息的新方法。在：ResearchGate。
https://www.researchgate.net/publication/373445537_THE_ROLE_OF_TECHNOLOGY_NEW_METHODS_OF_INFORMATION_MANIPULATION_AND_DISINFORMATION。访问日期 2023 年 12 月 12 日
Viana N（2023）为什么谷歌在巴西针对监管持拖延态度？在：卫报。

https://www.theguardian.com/commentisfree/2023/may/09/us-tech-companies-regulations-brazil。访问日期 2024 年 2 月 2 日
Wierson A（2021）认识 Lu，拥有 2500 万粉丝的非人类推广者。在：观察家。https://observer.com/2021/05/meet-lu-the-non-human-influencer-with-25-million-followers/。访问日期 2024 年 2 月 2 日

人工智能的恶意使用：沙特阿拉伯王国心理安全的挑战

维塔利·罗曼诺夫斯基，白俄罗斯国立大学（白俄罗斯）

引言

对于沙特而言，王储穆罕默德·本·萨勒曼的战略重点是对王国经济进行结构性改革。为实现这一目标，沙特裆裤优先考虑刺激创新，集中资源进行旨在生产高技术产品的研发，引入先进的生产技术，发展高科技和知识密集型部门，加快将王国融入全球经济空间。

事实上，沙特领导层对人工智能发展的关注已经在创建独特的人工智能格局上取得了成效。成立于 2019 年的沙特数据与人工智能局（SDAIA）是王国人工智能议程的主持者，“负责释放数据和人工智能的价值，将沙特阿拉伯提升为在数据驱动经济的精英联盟中的开拓性国家”（OECD 2024）。作为 SDAIA 的分支机构，国家人工智能中心（NCAI）负责人工智能创新、能力建设和推广国家数据与人工智能战略。此外，利雅得正在通过其新成立的国际人工智能研究与伦理中心开创先河，推动先进技术伦理的促进。

同时，沙特王国并未远离日益增长的全球趋势中，即对人工智能相关风险和挑战的思考。沙特的决策者、行业领袖、专家和学术界越来越关注和担忧，认为人工智能巨大的转型潜力需要谨慎而细致的处理，以避免其在各个领域使用中可能带来的意想不到的后果（沙特广播局 2023）。本章试图通过“三级 MUAI 对心理安全威胁”的模型，简要介绍人工智能潜在的恶意使用以及对沙特阿拉伯心理安全的挑战。

MUAI 对心理安全的第一层威胁

研究人员认为，涉及 MUAI 心理安全方面的研究将“为了反社会团体的利益故意歪曲对 AI 发展环境和后果的解释”视为 MUAI 对心理安全威胁的第一层级（Pashentsev 2023）。

沙特阿拉伯“2030 愿景”改革目标中大约 70%与数据和 AI 发展战略直接相关，旨在促进该国到 2030 年跻身全球前 15 名 AI 国家。最近数据表明，利雅得持续专注于发展其 AI 能力的证据是 2023 年 6 月发布的“全球 AI”指数，该指数在政府战略参数上将沙特阿拉伯排在第一位，全球排名第 31 位（Alarabiya 2023）。此外，根据 2023 年 11 月 Finastra 年度全球调查“金融服务：2023 年国家状况调查”，沙特阿拉伯在金融服务中采用 AI 方面居世界领先地位，调查显示，55%的沙特受访者（全球最高）表示在过去 12 个月中部署或改进了 AI（Asharq Al-Awsat 2023）。

鉴于沙特当局和穆罕默德·本·萨勒曼个人对国家人工智能发展的关注程度以及 KSA 媒体政策的严格规定，似乎在沙特的公开来源中很少有可以支持假设关于计划中或正在进行的信息宣传活动抹黑利雅得发展人工智能能力的统计显著证据。

然而，考虑到美国与中国在尖端技术领域的竞争日益激烈，以及利雅得在平衡华盛顿和北京之间的预期努力，AI 相关问题可能会被引申牵扯到其他方面也是心照不宣的。例如，最近一篇《金融时报》的报道警告称，沙特与中国在人工智能领域的合作，尤其是技术转让，可能会威胁到阿卜杜拉国王科技大学获得先进的美国制造芯片的渠道（Kerr 等，2023 年）。鉴于沙特阿拉伯成为能够“建立大型超级计算机并推出支持生成型 AI 系统的 LLM 技术（如聊天机器人）”的区域 AI 发展领导者的雄心，这对利雅得来说是一个重要信号（同上）。

总体而言，鉴于沙特领导层成为区域 AI 发展领导者的目标，公众对利雅得在这一领域努力的反对可能会持续存在，并有可能逐步增加。

MUAI 对心理安全的第二层威胁

AI 作为一种特定的目标技术，例如通过对关键基础设施的网络攻击，构成了 MUAI 对心理安全威胁的第二级（Pashentsev 2023）。

事实上，沙特阿拉伯在网络安全方面的最新进展表明利雅得对这一问题的关注不断增加。利雅得“已将网络安全作为其经济发展的支柱，实施了重大举措以提高网络安全准备水平”（Arabian Business 2022）。一个显著的例子是最新的 ITU 全球网络安全指数 2020 年排名，该排名将沙特阿拉伯在阿拉伯国家地区列为第一，全球排名第二，与英国得分相同（ITU，2020）。值得注意的是，沙特在不到 10 年的时间内就取得了如此显著的成绩（Tsukanov 2024）。

2023 年 12 月进行的一项研究调查了 50 位沙特阿拉伯的网络安全和 IT 领导者，结果显示，在过去两年中，40%的针对该国组织的网络攻击是成功的（Clewlow 2023）。另一项研究指出，受目标网络攻击影响最严重的行业是零售贸易、电子商务、信息服务、电信、金融、保险、商业银行、公共行政（SOCRadar 2023），即该国居民日常接触的关键公共基础设施部门。同一报告强调，作为一个区域政治和全球经济强国，且拥有世界上最大石油储备之一的国家，沙特“特别容易受到针对关键基础设施如油气田、发电厂和交通枢纽的网络攻击的威胁，鉴于该地区在能源生产中的关键角色”（同上）。可以合理推测，即使部分成功，同时、潜在的 AI 支持的定向网络攻击可能会对特定地区的居民的心理安全构成威胁，尤其是在公共基础设施受到影响的情况下，或者如果关键基础设施遭到破坏，可能会对整个国家造成威胁。

AI 相关的对国家关键基础设施的风险已经列入高级领导层的议程。特别是在 2023 年 11 月于利雅得召开的全球网络安全论坛上，阿美公司首席执行官阿敏·哈桑·纳赛尔在讲话中强调了识别与生成型 AI 相关的风险和漏洞的必要性，称其“对包括能源在内的许多行业都是改变游戏规则的因素”，并警告说能源部门容易受到新技术的网络安全攻击（Barakati 2023）。然而，沙特阿拉伯在最近的未来是否能够或“被允许”开发自己的 AI 支持的网络安全防护盾以保护其基础设施免受针对性的 AI 支持的网络攻击，还有待观察：尽管沙特在提高其网络安全地位方面做出了卓越的努力，利雅得仍然倾向于购买外国现成的网络安全套件，而不是专注于开发自己的解决方案（Tsukanov 2024）。

MUAI 对心理安全的第三层威胁

MUAI 对心理安全的第三级威胁主要是旨在于心理领域造成损害或建立对公众意识的控制（Pashentsev 2023）。

随着 2023 年 11 月 1 日《布莱奇利宣言》的签署，沙特阿拉伯对这一层级的 MUAI 威胁的关注有所提高。该政策文件由 28 个国家和欧盟签署，指出了 AI 操纵内容或生成欺骗性内容带来的意外风险（Gov.uk 2023）。此外，签署国强调“前沿 AI 系统可能会放大诸如虚假信息等风险”，并指出“这些 AI 模型的最显著能力可能带来的严重甚至灾难性伤害，无论是故意的还是无意的”（同上）。

事实上，沙特对这种威胁的关注是自然的，因为 AI 生成的假新闻已成为全球现象，沙特阿拉伯也未能避免这一趋势（Jones 2019；Fusco 2022；Sumsub 2023）。Fusco（2022）强调，“AI 生成的假新闻在沙特阿拉伯的普遍性和影响需要关注，因为它有可能对个人和整个社会造成重大伤害，包括煽动暴力、传播仇恨和侵蚀对机构的信任”。

另一个问题在于中东地区以及特别是沙特阿拉伯对使用阿拉伯语聊天机器人的需求不断增加。尽管阿拉伯语的形态特征、其复杂性以及稍微改变一个阿拉伯单词就可能导致聊天机器人混淆的可能性，阻碍了聊天机器人在沙特研究机构等地的广泛使用（Almurayh 2021），但是如果阿拉伯语聊天机器人的使用成为该地区媒体的普遍趋势，有目的地污染阿拉伯语聊天机器人的数据将很可能在不久的将来表现为信息和认知战争领域的高可行性风险。

由于 AI 生成的虚假新闻和虚假信息的迅速传播给沙特阿拉伯政府带来了严峻的挑战，政府正在努力应对这些内容的泛滥传播以及制定相应政策以确保国内外受众对官方或批准的新闻来源的可靠性。目前，政府似乎倾向于在禁止性规范和支持发展战略的法律框架之间寻求平衡。然而，考虑到对 AI 相关风险和挑战日益增加的关注，可以合理地期待政府对 AI 生成的媒体内容进行更严格的监管。

结论

尽管沙特领导人十分重视 AI 议程发展，但他们也越来越关注与 AI 发展相关的风险和挑战。与阿联酋不同，阿联酋政府认为其声誉资产对保持高科技部门的投资吸引力极为重要，而对于沙特阿拉伯来说，国家安全的心理安全方面最重要的优先事项是保持能够保障其技术资产安全的国家声誉。这种方法侧重于维护王国在实施 2030 愿景中领导方针的效率。

技术安全和技术资产的安全性，尤其是 AI，对利雅得在金砖国家的发展优先事项至关重要，这些优先事项包括加强双边贸易和使用新的结算机制。鉴于沙特阿拉伯希望在推动区域金融科技市场中占据领导地位，后者变得尤为重要（Al-Baity 2023）。

初步发现表明，MUAI 威胁的第二级和第三级对沙特阿拉伯的心理安全构成最显著的风险，并有可能出现第一级威胁。通过进一步研究，可以实现一个更全面和细致的沙特阿拉伯 AI 相关心理安全威胁模型。

参考文献

Alarabiya（2023）沙特阿拉伯在人工智能政府战略指数中居世界第一。
https://shorturl.at/pBLS6. 访问日期为 2024 年 2 月 2 日
Al-Baity HH（2023）沙特阿拉伯数字金融中的人工智能革命：综述与框架建议。可持续发展。第 15 卷（18），13725。https://doi.org/10.3390/su151813725
Almurayh A（2023）沙特大学使用阿拉伯语聊天机器人所面临的挑战。IAENG 国际计算机科学期刊。https://www.iaeng.org/IJCS/issues_v48/issue_1/IJCS_48_1_21.pdf。访问日期为 2024 年 2 月 2 日
阿拉伯商业（2022）Resecurity 在沙特阿拉伯推动基于人工智能的网络安全，设立新的研发中心。https://www.arabianbusiness.com/industries/technology/resecurity-drives-ai-powered-cybersecurity-in-saudi-arabia-with-new-rd-centre。访问日期为 2024 年 2 月 2 日
Asharq Al-Awsat（2023）沙特阿拉伯在金融服务领域广泛采用人工智能技术。https://shorturl.at/fghGW。访问日期为 2024 年 2 月 2 日
Barakati M（2023）阿美可首席执行官称“创新”得到网络安全机制的支持。在：阿拉伯新闻。https://www.arabnews.com/node/2401461/business-economy。访问日期为 2024 年 2 月 2 日
克鲁洛 A（2023）Tenable 研究显示，40%的网络攻击突破了沙特阿拉伯组织的防御。在：Intelligentico。https://www.intelligentcio.com/me/2023/12/13/tenable-study-reveals-40-of-cyberattacks-breach-saudi-arabian-organisations-defences/。访问日期为 2024 年 2 月 2 日
Fusco F（2022）巴基斯坦和沙特阿拉伯的人工智能和假新闻：犯罪方面。《巴基斯坦犯罪学杂志》。

https://faculty.alfaisal.edu/ffusco/publications/artificial-intelligence-and-fake-news%3A-criminal-aspects-in-pakistan-and-saudi-arabia。访问日期为 2024 年 2 月 2 日
Gov.uk（2023）参加人工智能安全峰会的国家签署的“布莱切利宣言”，2023 年 11 月 2 日。
https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023。访问日期为 2024 年 2 月 2 日
ITU（2020）2020 年全球网络安全指数。
https://www.itu.int/epublications/publication/D-STR-GCI.01-2021-HTM-E。访问日期为 2024 年 2 月 2 日
琼斯 M（2019）宣传，假新闻和假趋势：推特机器人在海湾危机中的武器化。国际传播期刊。第 13 期。
https://ijoc.org/index.php/ijoc/article/viewFile/8994/2604。访问日期为 2024 年 2 月 2 日
Kerr S，Al-Atrush S，刘 Q，Murgia M（2023）沙特-中国合作引发对 AI 芯片获取的担忧。在：《金融时报》。
https://www.ft.com/content/2a636cee-b0d2-45c2-a815-11ca32371763。访问日期为 2024 年 2 月 2 日
OECD（2024）沙特阿拉伯的人工智能政策。
https://oecd.ai/en/dashboards/countries/SaudiArabia。访问日期为 2024 年 2 月 2 日
帕申采夫 E（2023）。关于人工智能恶意利用对心理安全的一般内容和可能的威胁分类。在：帕申采夫 E（主编）《恶意使用 AI 与心理安全的 Palgrave 手册》。Palgrave Macmillan，Cham。https://doi.org/10.1007/978-3-031-22552-9_2。
沙特广播局（2023）王国确认其致力于利用人工智能的转化力量造福人类。
https://www.sba.sa/Stories-MainCaption-10435。访问日期为 2024 年 2 月 2 日
SOCRadar（2023）2023 年沙特阿拉伯（KSA）威胁景观报告。
https://socradar.io/saudi-arabia-threat-landscape-report/。访问日期为 2024 年 2 月 2 日
Sumsub（2023）身份欺诈报告 2023。
https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/。访问日期为 2024 年 2 月 2 日
Tsukanov L（2024）“波斯湾两岸”：该地区高科技业务发展及俄罗斯的利益。在：PIR 中心。
https://shorturl.at/ckFJO。访问日期为 2024 年 2 月 2 日

人工智能的恶意使用：中华人民共和国心理安全的挑战

叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）
达利娅·巴扎尔金娜，俄罗斯科学院欧洲研究所（莫斯科，俄罗斯）
叶卡捷琳娜·米哈列维奇，加斯普罗姆石油公司首席专家（圣彼得堡，俄罗斯）
王南森，上海环太国际战略研究中心主任兼高级研究员（上海，中国）

引言

中国的案例在几个特殊方面引人关注。首先，该国在人工智能领域的日益增长的领导力不仅为中国带来了更多的恶意使用人工智能（MUAI）威胁，也创造了更多应对这些威胁的机会。其次，中国经济的计划性质——一个发展中的公私合营体系，包括在网络安全领域——以及世界上最大的人口为人工智能训练提供了大量的大数据，使得中国在对抗 MUAI 方面的具备独特经验（Bazarkina 等人 2023）。到目前为止，中国在人工智能研究和发展方面做出了巨大贡献。2022 年的人工智能指数报告显示，2021 年，中国产出了世界上 27.6%的人工智能会议出版物，而美国产出了 16.9%。同时，在与人工智能相关的专利方面，中国也在世界遥遥领先，申请了世界 51.69%的人工智能专利，并有大约 6%获批，而欧盟和英国以及美国的份额分别为 3.89%和 16.92%（张等人 2022）。中国在 AI 采用方面处于领先地位，有 58%的公司已经开始部署 AI，而 30%正在考虑集成 AI。相比之下，美国的采用率较低，只有 25%的公司使用 AI，43%正在探索其潜在应用（Haan 和 Watts 2023）。

中国的普遍舆论认为，该国仍应接受 AI 带来的技术进步。与此同时，他们也认为必须对恶意使用进行适当监管，以保护国家安全，并确保尊重人们的隐私。幸运的是，绝大多数中国人对政府能够及时采取必要措施防止人工智能的恶意使用充满信心。可以说，这种公众信任将使中国领先于许多国家，达到在鼓励人工智能快速发展的同时，保障国家公共安全的微妙平衡（徐，2022）。然而，鉴于国际政治形势的恶化，日益严峻的网络安全挑战以及敌对国家、犯罪组织和不法行为者对人工智能恶意使用威胁的扩大，加强对中国国家安全的警惕势在必行。

MUAI 对心理安全的第一层威胁

中国逐渐在世界上崭露头角，对高科技领域产生了依赖，已成为西方贬低中国人工智能发展以及整个国家的主要理由。这一情况在 2019 年世界经济论坛上由乔治·索罗斯的一则声明具体体现。这位富有的新自由主义者警告说，开放社会面临着来自高科技专制政权的“致命危险”，然后继续说道：“中国并不是世界上唯一的专制政权，但无疑是在机器学习和人工智能方面最富有、最强大、最发达的国家。这使得习近平成为那些相信开放社会概念的人们最危险的对手”（Watts，2019 年）。这一指控很好地可以被视为对第一级心理安全（PS）风险和威胁的认识（在 MUAI 中，国家本身被指控对公民进行）。

实际上，在贬低中国人工智能发展的这一总体路线中，具体问题存在着广泛的变化。例如，在美国和其他西方国家，有许多出版物指责中国利用人工智能进行全面监控和迫害少数民族维吾尔人（Taddonio，2019 年；Bhuiyan，2022 年）。与此同时，中国的人工智能技术广泛应用于预测性执法（在美国也广泛实践），用于解决过去的犯罪或预防未来的犯罪，而不考虑罪犯的国籍（Mantello，2017 年）。然而，也有更“客观”的出版物并不专注于维吾尔人问题，而是谈论人工智能在“专制”（“共产主义”等）独裁统治手中的全面威胁（Singman，2023 年；Kasperowicz，2023 年；Lanum，2023 年；Raasch 和 Sahakian，2023 年；Hauf，2023 年）。在美国，基于人工智能技术的全面监控发展不亚于中国，其使用引发了重大问题。根据康奈尔大学就业关系教授维吉尼亚·多尔加斯特的说法，“工人们不断地处于自动化监视之下，基于人工智能的监控工具可能会犯错，这可能会导致不公平的薪资削减或解雇。美国和日本等

发达国家的工人通常不知道正在使用哪些监控工具，这些工具正在收集哪些数据，或者这些数据是如何用来评估他们的表现的”（Greenhouse，2023 年）。然而，美国不是在专业讨论实施自动化监控系统的真正利益和风险（因为这种系统在世界各地都存在），而是倾向于采取宣传对抗的立场，旨在损害中国人工智能行业（以及对中国人工智能公司的大量制裁）以及总体上损害该国社会经济发展的利益。

在第一级别的 MUAI 威胁中，西方主流媒体广泛试图对中国在制裁下开发人工智能技术的能力提出质疑，说服中国的人工智能开发者在中国共产党（CCP）的主导下无法成功工作，以及在外国买家中散播对中国人工智能产品质量或安全性的质疑，等等。

MUAI 对心理安全的第二层威胁

在中国，MUAI 威胁的第二级别涉及管理系统中人工智能的快速采用（Wang，2023 年）。诸如基于人工智能的集中控制的自学习机器人运输系统等基础设施设施可能成为高科技恐怖袭击的方便目标。例如，如果反社会行为者夺取了一个大城市（或其他关键基础设施，如发电厂、铁路线、电视塔等）的交通管理系统的控制权，这可能导致多起事故和伤亡，引发恐慌和混乱，从而创造出一种不稳定的心理气氛，为危险局势和可能的敌对反应提供便利（Bazarkina 和 Pashentsev，2019 年）。

与 MUAI 相关的心理安全威胁的第二级别与第三级别之间的边界实施例子是钓鱼和社会工程学的实践。根据 2023 年的联合报告，私人网络情报公司如 Group-IB、Bridewell 和 SideWinder 黑客组正在利用新的攻击基础设施对巴基斯坦和中国的目标发动有针对性的网络攻击。据研究人员称，黑客注册了 55 个模仿新闻、政府、电信和金融领域各种组织的域名。攻击者创建的上述域名模仿了巴基斯坦、中国和印度的政府组织。其中许多含有关于政府活动的“陷阱文档”。它们旨在向目标设备下载下一个阶段的有效负载（SecurityLab，2023 年）。

在 2020 年 1 月，一名香港银行经理成为一场高度先进的抢劫案的受害者，他被指示将 3500 万美元转账到各种银行账户用于公司收购。电话另一端的声音听起来恰好像一个熟悉的商业伙伴，但实际上是一个由人工智能生成的克隆在交谈。此事件涉及多达 17 名攻击者共同行动，使用虚假电子邮件验证购买（Veldkamp，2022 年）。

另一个香港的例子发生在 2024 年 1 月，一名跨国公司的财务工作者被欺骗支付 2500 万美元给诈骗者，后者利用深度伪造技术冒充公司首席财务官进行视频会议。据香港警方称，在这场精心策划的骗局中，这名工作者被欺骗参加了一个视频会议，他以为自己与其他几名员工进行了视频通话，但实际上这些都是深度伪造的再现（Chen 和 Magramo，2024 年）。香港警方网络安全与科技犯罪局的代理总警司陈舜正告诉该市的公共广播公司 RTHK：“我相信诈骗者提前下载了视频，然后使用人工智能添加了假声音用于视频会议”（Sharma，2024 年）。一切始于这位员工收到了一条看似合法的来自公司财务主管的消息。消息邀请他们参加一个保密的视频会议，讨论重要交易。起初，财务工作者对消息中提到需要进行秘密交易感到怀疑。但是，当他与看起来非常真实的同事进行了通话时，他的疑虑被消除了。通话中的人看起来和听起来都像他的同事。只有在后来与公司总部交谈后，他才意识到整个事情是一场复杂的骗局。从这个案例中我们可以看到，欺诈者对于使用人工智能技术的准入门槛正在不断降低。这个案例是最近几个案件中的其中一个，据信欺诈者利用深度伪造技术修改了公开可用的视频和其他镜头，欺骗人们的钱财。香港警方就此类诈骗案件逮捕了六名嫌疑人。据警方称，至少有 20 次使用 AI 深度伪造技术欺骗人脸识别程序，模仿身份证上的人员（Chen 和 Magramo，2024 年）。以上例子显示了钓鱼和借口两种常见但不同的社会工程学技术如何重叠，并基于使用人工智能技术创建了更具说服力和危险性的操纵产品。此外，

许多人工智能应用程序的开源性意味着技术的恶意使用不仅仅是与国家安全相关的问题。相反，它已经迅速扩展到包括普通公民在内，他们在业务和金融交易中必须更加警惕。

MUAI 对心理安全的第三层威胁

在中国，深度伪造技术的恶意使用被认为是心理安全威胁的第三个层级中的重大威胁之一。深度伪造指的是一组用于创建或更改音频、视频、照片内容和机器生成文本的人工智能技术。

在中国，一条非法且有利可图的地下深度伪造生产链迅速兴起。2023 年 5 月，甘肃省公安厅网络安全局破获了一起利用人工智能技术制作和传播虚假信息的案件，并拘留了一名犯罪嫌疑人。据警方介绍，该嫌疑人使用流行的人工智能软件 ChatGPT（他需要翻阅“防火墙”）对收集的新闻进行修改和编辑，然后使用“封面技术”软件将他的“新闻”上传到他为非法牟利而获取的百家号账号。伪造的文章“今早甘肃一列火车撞上道路施工人员，导致 9 人死亡”的信息明显是虚假不实的。网络安全警察发现，共有 21 个百度账号发布了该文章，并在短时间内获得了 15000 次浏览量（甘肃省公安厅，2023 年）。这是中国新近通过的深度伪造使用监管法律下的首次执法行动之一。

同时，由人工智能生成的名人和社交媒体影响者的“恶搞”视频变得司空见惯。例如，最近在印尼社交媒体上流传的一段假视频显示印尼总统佐科用流利的中文发表讲话，吸引了大量关注。当印尼总统讲话时，假视频逼真地模仿了他的口型、面部表情、语调和讲话方式。人工生成的替身看起来非常像他，视频背景声音还伴有观众的笑声。印尼政府部门最近澄清指出，该视频是使用深度伪造技术制作的（张，2023 年）。

深度伪造是一种用于进行虚假信息宣传活动的理想的工具，因为它可以生成具有可信度的假新闻，且需要消耗大量时间来揭穿。同时，虚假新闻造成的损害，特别是那些影响人们声誉的虚假新闻，往往是长期且不可逆转的。这一切将导致公众信任的缺失和怀疑情绪的上升——一种社会习惯于持续欺骗的情况，其试图过滤接收到的所有信息，甚至不相信官方来源。兼以内忧外患的当前形势，可能对中国的社会政治稳定构成越来越大的威胁。

另一个新兴威胁是会话型人工智能。也被称为社交机器人、机器人或简称为聊天机器人，会话型人工智能通过 MUAI 成为对心理安全的另一个威胁。中国消费者长期以来已经习惯于与聊天机器人进行交流，当 2022 年 11 月 30 日 OpenAI（其创始人之一是埃隆·马斯克）在微软的支持下推出了功能更为先进的 ChatGPT 时，这在中国引起了强烈反响。2023 年 3 月，更加先进的多模式 GPT4（如果以前用户只能通过文本消息与神经网络进行交互，GPT-4 通过图像、音频和视频打开了交互的视野。）来自 GPT（生成预训练变换器）语言模型系列的模型上市。

中国搜索引擎百度于 2023 年 3 月推出了类似的第一个中国服务，该服务名为 ERNIE（Enhanced Representation through Knowledge Integration），拥有 5500 亿条不同的事实条例。在功能上，ERNIE 接近于 OpenAI 几天前推出的 GPT4 神经网络，某些方面甚至超过了它。腾讯和阿里巴巴更多地专注于为商业合作伙伴提供人工智能产品，但两者都向中国公众提供了聊天机器人（程，2023 年）。GPT4 系统卡片承认该模型在许多领域可以与人类宣传者竞争，尤其是在与人类编辑配对时（OpenAI，2023 年）。中国官方媒体《中国日报》警告称，ChatGPT 可能会“加强美国发起的舆论竞赛攻势”（舒曼，2023 年）。

作为煽动者，聊天机器人的速度、规模和易用性远远超出了传统形式的宣传和人工网军的范围，因此它们对心理安全和政治稳定构成了真正的威胁。这些事实解释了中国领导层迅速采取行动规范其使用的原因。通过阻止 ChatGPT 和其他得到微软支持的机器人，中国监管机构正在为国内企业在开发相关技术方

面提供支持。北京市经济和信息化局于 2023 年 2 月 13 日发布的《北京市人工智能产业发展白皮书》宣布了到 2023 年完全加强人工智能产业发展基础的目标，包括支持领先企业创建像 ChatGPT 这样的技术（北京市人民政府，2022 年）。因此，中国并不害怕技术进步，而是努力防止其成果被用于反社会目的。

北美公司 NewsGuard 是一家监测和研究在线虚假信息的公司，该公司发现人工智能工具正在被积极地用于创建所谓的“内容农场”，指的是全球范围内生产大量点击量文章以优化广告收入的低质量网站。2023 年 4 月，NewsGuard 确定了七种语言中的 49 个网站 - 中文、捷克文、英文、法文、葡萄牙文、他加禄文和泰文，这些网站看起来完全或主要由人工智能语言模型创建，旨在模仿人类交流 - 在这种情况下，语言模式被伪装成典型的新闻网站（Sadeghi 和 Arvanitis，2023 年）。可以说，NewsGuard 的发现可能只是冰山一角。越来越先进的生成式人工智能和大型语言模型的强大力量可以将这些网站变成有效且成本相对较低的宣传渠道。

中国目前正在着手解决虚假新闻账户和人工智能生成帖增长热的问题。2023 年 5 月中旬，国家监管机构中国国家互联网信息办公室（CAC）表示，他们已经“清理”了超过 107,000 个虚假新闻账号和 835,000 条虚假信息，并敦促公民举报虚假新闻账号和报道（Frank，2023 年）。重要的是，他们不仅仅谈论文本消息，还有虚拟主持人、伪造的工作室场景，这些允许模仿现有的注册网站，并使用各种旨在引发网络用户情绪反应的方法，增加流量。这种活动也可能具有明显的恶意性质（Dobberstein 2023）。毫无疑问，他们不仅谈论文本消息，还有虚拟主持人、虚假的工作室场景，可以模仿现有注册网站，并使用各种旨在引发网络用户情绪反应的方法，增加流量。这些活动也可能具有明显的恶意性质（Dobberstein，2023 年）。个别用户和小型企业对生成式人工智能的自发使用可以掩盖个别大型国家和非国家行为者的大规模心理影响活动。由于尖锐的地缘政治和经济矛盾，中国可能客观上处于这类活动的中心地位。

值得注意的是，许多 MUAI 仍处于初级阶段。例如，“元宇宙”概念可能会在不久的将来为中国的经济和社会发展开辟许多新机遇。中国科技公司已经开始试水，开发他们自己的元宇宙应用程序并投资于与元宇宙相关的技术。元宇宙是与物理世界平行存在的虚拟世界。在元宇宙中，我们的数字和物理现实之间（在工作、社交和娱乐领域）可能存在更大的重叠 - 这是由某些先进技术（包括可能塑造下一代互联网的人工智能技术）所实现的。中国的六家领先科技公司，包括百度、阿里巴巴和腾讯（统称为 BAT），是全球申请了最多与关键元宇宙技术开发相关的专利的前十大公司之一（Interesse，2022 年）。增强现实和元宇宙等技术将以更全息和视觉化的方式呈现事件现场。元宇宙的高度沉浸式和互动性特征很可能会使其受众更容易受到人工智能生成的误导和虚假信息的影响。因此，它可能会对公众产生更高的恶意影响威胁（张，2022 年）。

作为科技领域的领头羊，中国在为元宇宙提供心理安全方面还有很多工作要做，考虑到肯定会有反社会行为者试图利用这一新空间谋取自身利益。应强调基于人工智能技术的技术发展，以及反社会行为者可能将其用于恶意目的的可能性，这需要更加仔细地研究这些技术的潜力，并制定系统化的任务来中和它们。

结论

中国经历着逐渐从追赶发展战略到先进发展战略的过渡（尤其是在人工智能领域），这带来了巨大的机遇，但也存在着巨大的风险，需要进行创新性解决方案。在一个困难而不稳定的环境中，心理安全威胁尤其危险，特别是如果拥有新技术为恶意行为者提供了新的机会，从犯罪组织和腐败的治理机构元素到不友好的侵略性国家。

目前，中国主要面临的是多重人工智能的第一级（试图抹黑中国人工智能的众多企图）和第三级（深度伪造、聊天机器人、新闻农场等的恶意使用）心理安全威胁水

目前，中国主要面临的是多重人工智能的第一级（试图抹黑中国人工智能的众多企图）和第三级（深度伪造、聊天机器人、新闻农场等的恶意使用）心理安全威胁水平的案例。边界威胁（第二级和第三级之间）主要表现在网络钓鱼和社会工程学方面。对基础设施设施和控制系统的第二级威胁尚未导致由于多重人工智能活动而引发的重大事件，产生相应的负面心理和社会后果，但由于高科技内外恶意行为者的活动，这种可能性在未来不能排除。元宇宙的发展，情感人工智能的改进，以及普通人工智能的进步，以及人类科学及其应用应用的进步，将在心理安全领域带来更加复杂的问题。

虽然多重人工智能对中国的心理安全构成了重大威胁，但传统形式的宣传通过人工智能技术正在更新其效力。然而，在不久的将来，由于人工智能能力的数量和质量的提高，以及其在各个公共生活领域的进一步实施，多重人工智能可能会成为主要威胁。

以上所有情况都需要一种辩证的方法来评估人工智能的作用。充分利用人工智能技术对社会发展提供了强大的推动力，正如现代中国清楚地展示了的那样。然而，随着这些技术的发展，多重人工智能的威胁也在增长（有时甚至超过了技术本身），对其进行中和不需要那么多的技术或管理措施，而是需要社会措施。独特的技术旨在在各种形式的认知情感和心理活动中部分或完全替代人类（例如，在与残障人士合作中的情感人工智能），单调或有害的体力劳动（人工智能机器人），将增加对人类的威胁。如果后者将由于人工智能而增强其日益增长的能力，而不是建立一个更加和谐的社会体系，发展具有全新水平的智力和心理能力的个体，更高水平的社会责任，而是加强社会、民族或种族的分化。中国拥有巨大的科学、技术、经济，最重要的是人力资源潜力，可以防止负面情景对其国家发展和对人类的任何贡献。问题在于中国将如何有效地利用这一潜力。这个问题仍然悬而未决，主要的机遇和风险仍然在前方。

参考文献

Bazarkina D, Mikhalevich E, Pashentsev E, Matyashova D (2023) 在中国，恶意使用人工智能在心理安全方面的威胁和当前实践。在：Pashentsev E 编著，《恶意使用人工智能与心理安全的帕尔格雷夫手册》。Palgrave Macmillan，Cham。
https://doi.org/10.1007/978-3-031-22552-9_13

Bazarkina D, Pashentsev E (2019) 人工智能与国际心理安全的新威胁。《俄罗斯在全球事务中》。doi: 10.31278/1810-6374-2019-17-1-147-170

Bazarkina D, Pashentsev E (2020) 人工智能的恶意使用：金砖国家新的心理安全风险。《俄罗斯在全球事务中》。doi: 10.31278/1810-6374-2020-18-4-154-177

Bhuiyan J (2022) “到处都是摄像头”：维吾尔人在中国遭受广泛监视的证词。《卫报》。
https://www.theguardian.com/world/2021/sep/30/uyghur-tribunal-testimony-surveillance-china。访问日期：2024 年 1 月 2 日

Chen H, Magramo K (2024) 金融工作者在与深度伪造的“首席财务官”视频通话后支付 2500 万美元。《CNN》。
https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.
html。访问日期：2024 年 2 月 6 日

Cheng E (2023) 百度表示，其 ChatGPT 竞争对手 Ernie 机器人现在拥有超过 1 亿用户。《CNBC》。
https://www.cnbc.com/2023/12/29/baidu-says-its-chatgpt-rival-ernie-bot-has-more-than-100-million-users.html。访问日期：2024 年 1 月 2 日

Dobberstein L (2023) 中国打击 AI 生成的新闻主播。《The Register》。
https://www.theregister.com/2023/05/16/china_crackdown_on_ai_generated_news/。访问日期：2024 年 1 月 2 日

Frank J (2023) 中国正在删除成千上万的 AI 生成新闻账户和帖子。《Business2Community》。

www.business2community.com/tech-news/china-is-deleting-hundreds-of-thousands-of-ai-generated-news-accounts-and-posts-02692962。访问日期：2024 年 1 月 2 日
甘肃公安厅 (2023) 甘肃公安破获首例利用 AI 人工智能技术捏造虚假信息案 [Gansu gong'an zhenpo shou li liyong AI rengong zhineng jishu paozhi xujia xinxi an]。
https://mp.weixin.qq.com/s/_Wfe-EV13O6uBM65jZDzdg。访问日期：2024 年 1 月 2 日
Greenhouse S (2024) “被持续监控”：对工作中 AI 监视的反抗。《卫报》。
https://www.theguardian.com/technology/2024/jan/07/artificial-intelligence-surveillance-workers。访问日期：2024 年 1 月 2 日
Hauf P (2023) 共和党参议员警告，中国通过武器化 AI 寻求“混乱和困惑”。《福克斯新闻》。
https://www.foxnews.com/politics/china-aiming-chaos-confusion-weaponizing-ai-warns-gop-senator。访问日期：2024 年 1 月 2 日
Interesse G (2022) 中国在元宇宙的首秀：需关注的趋势（更新）。《China Briefing》。
https://www.china-briefing.com/news/metaverse-in-china-trends/。访问日期：2024 年 1 月 2 日
Kasperowicz P (2023) 立法者警告，中国利用技术“压迫自己的人民”，寻求限制 AI 出口。《福克斯新闻》。
https://www.foxnews.com/politics/china-using-tech-oppress-own-people-warns-lawmaker-restrict-ai-exports。访问日期：2024 年 1 月 2 日
Lanum N (2023) McCaul 表示，中国的 AI 和量子投资是对“世界军事和经济统治”的竞赛。《福克斯新闻》。
https://www.foxnews.com/media/mccaul-china-ai-quantum-investments-race-military-economic-domination-world。访问日期：2024 年 1 月 2 日
OpenAI (2023) GPT-4 系统卡。https://cdn.openai.com/papers/gpt-4-system-card.pdf。访问日期：2024 年 1 月 2 日
Raasch JM, Sahakian T (2023) 如果中国首先掌握了它，人工智能对人类的威胁将会更大：Gordon Chang。《福克斯新闻》。
https://www.foxnews.com/world/ai-threat-humanity-far-greater-china-masters-first-gordon-chang。访问日期：2024 年 1 月 2 日
Sadeghi M, Arvanitis L (2023) 新闻机器人的崛起：AI 生成的新闻网站在网上激增。《NewsGuard》。
https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/。访问日期：2024 年 1 月 2 日
Schuman M (2023) 聊天机器人人工智能对中国来说是一个问题。《大西洋月刊》。
https://www.theatlantic.com/international/archive/2023/04/chatbot-ai-problem-china/673754/。访问日期：2024 年 1 月 2 日
SecurityLab (2023) SideWinder 在最新的攻击中冒充巴基斯坦和中国政府机构。
https://www.securitylab.ru/news/538242.php。访问日期：2024 年 1 月 2 日
Sharma S (2024) 深度伪造的 CFO 诱骗金融工作者，骗取 2500 万美元。《Interesting Engineering》。
https://interestingengineering.com/culture/25-million-swindled-as-deep-fake-cfo-tricks-finance-worker。访问日期：2024 年 2 月 6 日
Singman B (2023) 美国情报部门警告来自中国、俄罗斯、北朝鲜的“复杂”威胁。《福克斯新闻》。
https://www.foxnews.com/politics/us-intel-community-warns-complex-threats-china-russia-north-korea。访问日期：2024 年 1 月 2 日
Taddonio P (2019) 中国政府如何在其维吾尔穆斯林人口上使用人工智能。在：PBS。
https://www.pbs.org/wgbh/frontline/article/how-chinas-government-is-using-ai-on-its-uighur-muslim-population/。访问日期：2024 年 1 月 2 日
北京市人民政府 (2023) 发布《2022 年北京人工智能产业发展白皮书》 [“2022 Nian beijing rengong zhineng chanye fazhan baipishu” zhong bang fabu]。
http://www.beijing.gov.cn/ywdt/gzdt/202302/t20230214_2916514.html 。访问日期：2024 年 1 月 2 日
Wang D (2022) 人工智能恶意使用的威胁与应对 [Rengong zhineng eyi shiyong weixie yu yingdui]。

https://www.cnki.com.cn/Article/CJFDTOTAL-CINS201908008.htm。访问日期：2024 年 1 月 2 日
Watts W (2019) 索罗斯批评中国的习近平是开放社会的“最危险的对手”。《MarketWatch》。
https://www.marketwatch.com/story/george-soros-blasts-chinas-xi-as-most-dangerous-opponent-of-open-societies-2019-01-24?siteid=yhoof2&yptr=yahoo。访问日期：2024 年 1 月 2 日
Xu C (2022) 人工智能与国家政治安全 [Ren gong zhi neng yu guo jia zheng zhi an quan]。
http://finance.people.com.cn/n1/2022/0626/c1004-32456635.html。访问日期：2024 年 1 月 2 日
Zhang D, Maslej N, Brynjolfsson E, Etchemendy J, Lyons T, Manyika J, Ngo H, Niebles JC, Sellitto M, Sakhaee E, Shoham Y, Clark J, Perrault R (2022) 人工智能指数 2022 年度报告。
https://doi.org/10.48550/arXiv.2205.03468。
Zhang J (2023) “人工智能替代”隐藏的风险，如何监管人工智能技术的滥用
[“Rengong zhineng zui tiancang fengxian,AI jishu lanyong gai ruhe jianguan]。
https://www.163.com/dy/article/IJ7G5SFT0514R9L4.html。访问日期：2024 年 1 月 2 日
Zhang Z (2022) 从智能视角看认知域作战：情感冲突成为认知域作战突出属性
[Zhineng hua shi yu xia de ren zhi yu zuozhan: Qinggan chongtu chengwei ren zhi yu zuozhan tuchu shuxing]。http://www.81.cn/yw_208727/10204158.html。访问日期：2024 年 1 月 2 日

人工智能的恶意使用：印度共和国心理安全的挑战

达利娅·巴扎尔金娜，俄罗斯科学院欧洲研究所（莫斯科，俄罗斯）
叶夫根尼·帕申采夫，俄罗斯外交部外交学院现代国际研究所（莫斯科，俄罗斯）

引言

印度的人工智能市场规模预计将显著扩大，到 2025 年达到约 6358.8 亿美元（Srivastava R 2023）。印度政府于 2018 年 2 月宣布，国家转型协会（NITI Aayog），即政府智库，将领导一个全国性的人工智能研究计划。2017 年，工业和贸易部成立了人工智能任务组，致力于推动印度经济转型（Faggella 2019）。人工智能正在推动印度经济的巨大变革。印度初创企业正在开发使用人工智能的解决方案，涉及教育、医疗保健和金融服务领域。印度的人工智能发展还包括开发数字助手，让组织能够与客户进行沟通，基于人工智能的决策系统以及在贸易中使用人工智能和区块链技术（Chakraborty 2022）。预计到 2025 年，人工智能将为印度国内生产总值增加高达 5000 亿美元，并在 2035 年达到 9670 亿美元。印度对人工智能的投资每年增长 30.8%（Bundhun 2023）。“全球务实主义、开源模式、监管重点、工作演变和效率重塑的全球趋势正在为人工智能带来变革性变化奠定基础。与此同时，印度在 IT 和人工智能领域的特定增长为这一叙事增添了独特的维度”（Srivastava 2023）。当然，人工智能领域的如此迅速发展既带来了好处，也带来了不利因素，包括在心理领域。

MUAI 对心理安全的第一层威胁

目前，在印度，其中一个紧迫的一级威胁仍然是被操纵的由于自动化而导致的工作丧失的担忧。《微软工作趋势指数 2023》显示，74%的印度员工担心人工智能将取代他们的工作。同时，83%的员工愿意将他们工作的主要部分委托给人工智能，以缓解技术可能带来的工作丧失威胁。这份新报告以“人工智能能修复工作吗”为主题，这个问题逐渐成为人们关注的焦点，特别是在像 ChatGPT、Google Bard 和 Microsoft Bing Chat 等工具推出之后（Sengupta 2023）。这为攻击者操纵舆论开辟了空间，使他们能够以极不利的方式呈现那些积极推动人工智能在各个公共生活领域实施的政治家。

已经有社交媒体上围绕人工智能导致裁员的活跃争议。一位印度首席执行官因表示他的公司已经用一个人工智能聊天机器人取代了 90%的支持人员而受到批评。Dukaan 的创始人 Suumit Shah 在 Twitter 上表示，这个聊天机器人显著提高了客户查询的首次响应和解决时间。这条推文引发了网上的愤怒。这发生在人们对人工智能夺走人们的工作，尤其是在服务行业方面有很多讨论和担忧的时候（BBC 2023）。这些例子清楚地说明了印度社会存在的担忧。对人工智能的广告预期的增加也是一个问题，但不像几年前那样重要，因为人们对人工智能的担忧和恐惧变得更加重要。未来将告诉我们是否能够实现对人工智能的平衡态度。

MUAI 对心理安全的第二层威胁

在印度相关的第二级威胁中，值得注意的是利用人工智能技术在虚拟空间中增加的欺诈案件。这种欺诈的主要方面是社会工程学，它帮助攻击者获得受害者的信任，并迫使受害者不顾后果地行动。例如，网络钓鱼信息的作者使用以下心理技巧：

紧急性：网络钓鱼邮件通常要求立即采取行动，因为你越长时间思考，就越可能质疑其是否合法；

可信度：网络钓鱼尝试将基于真实生活中的场景。发票需要支付，或者需要分享文件；

熟悉性：钓鱼邮件的数量已经大幅上升，其中攻击至少部分针对个人 - 通常声称来自权威人士，例如他们的首席执行官或安全主管；

保密性：所需的操作是特定于您的，并且需要您独自完成，因为让其他人参与增加了识破诈骗的机会（Egress 2021）。

人工智能可以显著提高这类攻击的“有效性”，这不可避免地影响了犯罪统计数据。例如，网络安全公司 Group-IB 最近的一份报告揭示了约 10 万个 ChatGPT 帐户已经被盗，它们的数据通过暗网非法出售，仅印度就报告了 12,632 个被盗的凭证（Stanly 2023）。随着人工智能工具变得更便宜和更易获得，与该技术相关的网络安全风险可能会增加。网络安全公司 CheckPoint Research 的数据显示，在 2023 年第一季度，印度每周平均受到的攻击比 2022 年同期增加了 18%，每个组织每周至少受到 2,108 次攻击。然而，据 Indusface（一家应用安全软件即服务公司）的《应用安全状况 Q2 报告》显示，第二季度机器人攻击的数量增长了 48%，十个网站中有九个遭受了机器人攻击（Stanly 2023）。该行业正在努力应对此类威胁，通过开发各种测试软件质量的解决方案，以识别和利用异常。然而，攻击者有他们利用这类技术的方法。例如，Metasploit 是一个提供安全漏洞信息的计算机安全项目，成为黑客中流行的一种工具（Stanly 2023）。

骗子越来越多地利用人工智能的技术复杂性。针对印度数百万人，这些人每周花费多达 105 分钟来审查、验证或决定通过短信、电子邮件、社交媒体发送的消息是否真实。根据 Mcafee 的一项研究，多达 82%的印度人曾点击或上当受骗。在精心策划的欺诈活动中，人们最容易上当受骗的形式包括假工作通知或报价（64%）和银行警报消息（52%）。“人工智能是骗子的最爱，帮助网络犯罪分子增加了欺诈消息的规模和复杂性。网络钓鱼和短信诈骗的速度正在加快 - 每 11 秒就会创建一个新的网络钓鱼网站。这突显了对能够扭转人工智能骗子局面的解决方案越来越迫切的需求；对于印度 9 亿互联网用户来说，现在保护自己在线上从未如此关键”（Doval 2023）。在线诈骗具有心理后果。它突显了人们面对的压力增加，这是由于人工智能驱动的诈骗消息数量和复杂性增加所导致的。这种压力可能为社会动荡铺平道路，因为随着人们对安全结构的信任下降，人工智能增强犯罪的兴起可能会加剧这种局面。

现在，网络犯罪分子正在尝试使用像 FraudGPT 和 WormGPT 这样的复杂生成式人工智能聊天机器人，被宣传为“没有限制、规则和边界”的机器人，来进行其恶毒活动。这些聊天机器人最近出现在暗网上，被出售给任何想要创建网络钓鱼邮件、恶意软件或破解工具的人。这种聊天机器人的使用问题在印度也很相关。Hitachi Hi-Rel 的信息技术副总裁和网络传道者 Manish Thakar 警告人们说：“这些聊天机器人是基于流行的 ChatGPT-3 技术，可以从用户提示中生成逼真而连贯的文本。有了这个工具，黑客可以创建可疑的邮件，以欺骗毫无戒心的受害者，让他们相信自己收到了官方的商业邮件、短信或银行通知”（The Times of India 2023a）。他透露，FraudGPT 可以编写恶意代码，创建难以检测的病毒或恶意软件，找到非 VBV BIN，创建网络钓鱼页面和入侵工具，闯入群组、网站和市场。它甚至可以编写欺诈页面或信件，找到泄漏、漏洞，甚至访问活动卡。古吉拉特邦 CID 官员表示，FraudGPT 的供应商在像 Empire、WHM、Torrez、Alphabay 和 Versus 等地下暗网市场上是相当有名的（The Times of India 2023a）。未来，这种恶意聊天机器人的数量和改进可能会增加，这可能会使社会工程技术变得更加危险。

随着印度经济数字化的好处，MUAI 的风险也在增加。2022 年预算引入了基于区块链的中央银行数字货币（CBDC）。尽管一些政治家和金融家赞成加强对加密货币的立法监管，但印度储备银行行长 Shaktikanta Das 认为 CBDC 也可能受到数字欺诈的影响。

一些虚假的投资门户网站已经超过一年被印度执法机构关注。喀拉拉邦的许多居民被一个名为 Morris Coin 的虚构加密货币所欺骗，印度执法部门估计欺诈资产价值为 2 亿美元。卡纳塔克邦也发生了

类似的事件。专家认为，欺诈和在加密货币领域恶意使用 deepfakes 的结合将把网络钓鱼提升到一个新的水平（Biswas 2022），从而更容易说服消费者相信一个接受加密货币支付的假网站的真实性。

MUAI 对心理安全的第三层威胁

在语音控制领域，MUAI 的水平持续增长，Pindrop 指出，这一领域的欺诈活动从 2013 年到 2017 年增长了超过 350%（Pindrop 2020）。至少 39%的印度互联网用户可能拥有某种形式的数字语音助手。专家预测，在不久的将来，语音将成为电子商务、银行业务和支付的首选交易方式（Bhatt 2018）。在未来威胁的背景下，我们认为，跨学科研究应考虑的风险不仅包括语音界面或聊天机器人的黑客攻击，还应包括对电子翻译程序的黑客攻击。例如，MUAI 可能扭曲官方文件，其挑衅性可能与发送一段能够引发国家间冲突的 deepfake 相当。

在印度，机器人活动在 2019 年选举活动期间被记录下来，Twitter 上的机器人账户大规模地部署于 2 月 9 日至 10 日，既为现任总理纳伦德拉·莫迪站台，也为其反对者站台。与此同时，少数账户小组每小时发布数千条推文。主要政治党派与 2014 年的选举相比，增加了数字化沟通，但在 2019 年，这类活动的影响相对较小（Thaker 2019），因为 Twitter 上的选民数量相对较少。然而，参与竞选活动的大规模机器人降低了在线辩论的质量。2021 年 12 月 27 日，一位知名的反对派领袖写信给 Twitter 的 CEO，抱怨自 7 月以来他的粉丝数量没有增加，而其他政治领袖的粉丝数量却在增加。对此，Twitter 回复称他们使用机器学习工具删除了数百万个机器人和恶意账户（OpIndia 2022）。这样，机器人活动成为政治争议的话题，可能被反社会行为者利用，造成第一级别的心理安全威胁。

聊天机器人，除了政治活动之外，还可以参与宗教活动，如果被恶意使用，也可能构成三级威胁。印度的经验表明，评论宗教文本的聊天机器人数量增加，它们的评论（来源不明）可能相当挑衅。印度许多人放弃了与解释《巴格瓦德·吉塔》的精神导师进行面对面的接触，转而向模仿印度教神克里希纳的在线聊天机器人寻求帮助，这些机器人会回答关于生命意义的深刻问题。专家警告说，这是一种新技术，倾向于偏离脚本并容忍暴力。几个机器人一直在提供这样的答案：如果是你的义务，那么杀人是可以的。“这是一种误传，基于宗教文本的错误信息，”孟买律师、《AI 之书》的合著者之一卢布娜·尤素夫表示。“文本给予了很多哲学价值，而机器人呢？它给出了一个字面上的答案，这就是危险所在”（Shivji 2023）。至少有五个《吉塔》聊天机器人于 2023 年初在网上出现，由语言模型预训练变换器 3（GPT-3）提供支持。它们使用人工智能，模拟对话并根据统计概率模型创建答案。这些网站称他们有数百万用户。“尤素夫表示，在印度这样一个情感充沛的国家，答案中容忍暴力的潜在危险更加严重”（Shivji 2023）。

在 2023 年 1 月 7 日，激进人权活动家 Mahesh Vikram Hegde 的账户推特发布了 ChatGPT 的截图，其粉丝数量超过 18.5 万人；推文似乎显示这款 AI 驱动的聊天机器人对印度教神祇克里须那开过玩笑。但 ChatGPT 在原则上是不会嘲笑任何宗教或神祇的，因为当它被要求对耶稣基督或穆罕默德开玩笑时会回复“很抱歉，但我不会对任何宗教或神祇开玩笑” 据预览，这一限制似乎不包括印度教的宗教形象。当《连线》的记者以 Hegde 的截图为例对 ChatGPT 进行测试时，聊天机器人返回的答案与他发布的截图相似。持有 ChatGPT 的 OpenAI 并没有回应评论请求。这条推文在印度社交媒体上被查看了超过 40 万次。几天后，它被传播到社交媒体和广播媒体上的阴谋论。当 AI 技术不慎引发冲突，这一案例可以被认为是一种第一层威胁的表现（利用意见领袖败坏 AI 产品的声誉）和第三层威胁（使用聊天机器人来煽动冲突）。

然而，在印度，最广泛的 MUAI 例子可以在深度伪造领域找到。根据 Deeptrace 的一份报告，2019 年含有深度伪造色情视频的网站中有 3%是印度的（Ajder 等人 2019，第 2 页）。Deeptrace 还注意到，“中国网民在合成媒体工具的创作和使用方面做出了重大贡献”（Ajder 等人 2019，前言）。已经

有记录的案例显示，在印度使用深度伪造技术来损害某人的声誉。一名备受争议的印度记者的照片和视频被制作成深度伪造色情视频（Ajder 2019），这表明现代犯罪分子正在逐渐占用 AI 技术，并且未来可能会释放对各种利益群体的广泛歧视运动的危险。

在印度，专门用于竞选目的的深度伪造案例（也许是全球首次）已经臭名昭著。2020 年 2 月，印度人民党（印度人民党）使用这项技术制作了两段视频，在这两段视频中，党的领袖马诺吉·蒂瓦里在德里立法议会选举前用两种语言向选民发表讲话。候选人的目标是向两组讲不同语言的选民发送信息——哈里亚纳语和英语。据党的代表称，这些视频被发送到约 5800 个 WhatsApp 群组（Alavi 和 Achom 2020），在其中，蒂瓦里还祝贺他的支持者赞成印度议会通过的《公民法修正案》。在原始视频中，蒂瓦里的信息以印地语传达，他的面部表情和唇部动作通过使用 AI 模拟的语言模拟出了哈里亚纳语版本。该党的媒体官员表示，哈里亚纳语版本的视频收到了积极的反馈，之后决定制作英语版本。然而，很快就明显地意识到事态可能失控：“有人使用了我们的德里印度人民党主席的 Facebook 视频...马诺吉·蒂瓦里，并向我们发送了他的视频，其中内容使用了哈里亚纳语方言，”德里印度人民党媒体官员表示。“这对我们来说是震惊的，因为它可能被反对派恶意利用...我们强烈谴责这种技术的使用，这种技术在公开领域是可用的，并且未经我们的同意就被使用”（Mihindukulasuriya 2020）。

尽管在这种情况下，深度伪造仅被用于克服语言障碍，但对深度伪造进行进一步修改以用于恶意目的，以及政党的反应，引发了印度媒体的热烈讨论。这可以理解为，人们对技术可能被用于传播虚假信息的担忧增加了，例如政客使用深度伪造将具有争议性和不可接受的言论放在对手口中（Alavi 和 Achom 2020）。印度主要政党公开放弃了使用深度伪造的做法（Mihindukulasuriya 2020）。对于印度以及许多其他国家来说，在当前快速数字化和日益增多的虚假信息活动的环境中，将 MUAI 技术与不同心理安全威胁级别相结合的挑战可能变得相关。

2023 年 11 月，印度政府发布了一份关于人工智能技术“危险和破坏性”影响的警告，此前一段深度伪造视频在社交媒体上迅速传播，视频据称显示宝莱坞女演员拉什米卡·曼丹娜。一段视频据称显示宝莱坞明星曼丹娜——她有 3900 万 Instagram 粉丝——穿着黑色健身服走出电梯，在社交媒体上迅速传播开来。然而，这段视频实际上是一个由 AI 生成的深度伪造。视频中的女子实际上是一位名叫扎拉·帕特尔的英国印度裔网络红人，她在她的 Instagram 账户上发布了原始视频片段。据 BBC 报道，与事实核查平台 Alt News 合作的记者阿比什克·库马尔首次报告了这段看似显示曼丹娜的病毒视频实际上是一个深度伪造（Bandara 2023）。宝莱坞明星曼丹娜描述了她对在线传播的深度伪造视频感到的困扰，并表示必须“紧急”解决深度伪造技术问题。资深宝莱坞演员阿米塔巴·巴克占认为，他的搭档曼丹娜对于在 AI 生成的视频中身份被盗有“强有力的法律诉求”（Bandara 2023）。女演员卡特里娜·凯芙和卡裘琳·黛尔的假照片也迅速出现——这显示出深度伪造广泛蔓延的令人不安的事实。

如电子和信息技术部（MeITY）国务部长 Rajeev Chandrasekhar 所说：“深伪现象实际上是 AI 和虚假信息产业的结合，确实是我们所有人应该担心的事情，因为它对个人、社会、社区和国家都非常具有威胁性”（Sukumaran 2023 年）。2023 年 11 月 17 日，印度总理纳伦德拉·莫迪警告称，深伪可能引发危机并“煽动社会的不满之火”，并透露他最近看到了一段自己跳古吉拉特‘加尔巴’舞的深伪视频，而他自上学以来就没有跳过这种舞。作为一种心理战术，伪造视频在 2020 年 6 月印度和中国士兵在拉达克的加勒万冲突后开始流传。2023 年 5 月，印度在女性奥运摔跤选手进行为期五个月、要求调查性骚扰投诉的关键时刻，看到了面部表情操控的例子。当德里警察拘留抗议者时，摔跤选手 Vinesh 和 Sangeeta Phogat 发布了一张她们在警车里坐着的自拍照，两人神情严肃。但另一版本——两人微笑着面对镜头的照片——在网络上迅速传播，直到被揭穿为假（Sukumaran 2023 年）。所有这些案例都清楚地说明了恶意使用深伪对第三层心理安全的威胁。

印度社会对这一威胁的心理后果日益担忧。学生辅导员、Shalimar Bagh 现代公立学校校长 Alka Kapur 表示：“对于学生来说，社交媒体不仅仅是一个连接的平台；它是一个塑造他们身份、建立友谊和培养自尊的虚拟领域。当深伪丑闻渗透到社交媒体中时，对学生心理健康的影响变得越来越令人担忧。内容的数字化修改，曾经被视为无害的娱乐，现在对年轻人的心理健康构成了严重威胁。适应数字化修改的美貌和行为标准的压力，加上害怕成为恶意深伪操纵受害者的恐惧，为学生创造了一个有毒的环境”（《印度时报》2023b）。记者们请我们想象一个场景：一个孩子发现自己被不雅地出现在一个被修改的视频中，从事他们从未参与过的活动。“声誉受到损害的可能性以及随后的社会排斥经历可能是极其有害的。害怕成为下一个目标让孩子们处于持续的焦虑状态，削弱了他们对数字领域的信任，并加剧了本已充满挑战的青春期历程”（《印度时报》2023b）。

2023 年 4 月，印度南部的泰米尔纳德邦爆发了一场政治争议，当时印度执政党印度人民党（BJP）的邦主席 K.（Kuppusamy）Annamalai 发布了一段泰米尔纳德邦现任执政党德拉维达进步联盟（DMK）议员 Palanivel Thiagarajan 的有争议音频录音。在这段 26 秒的低质量音频中，据称当时担任泰米尔纳德邦财政部长的 Thiagarajan 指责自己党内成员非法聚敛了 36 亿美元。Thiagarajan 坚决否认录音的真实性，称其为“捏造的”和“机器生成的”。他在 2023 年 4 月 22 日发推文说：“不要相信没有明确来源的音频片段。” 他认为现在很容易伪造声音。4 月 25 日，Annamalai 发布了第二段音频——时长 56 秒，音质清晰得多——据称 Thiagarajan 在其中贬低自己的党派并称赞 BJP。这次，Thiagarajan 表示没有人承认这些片段的来源。分析人士对第一段录音意见不一，有些认为质量太差无法得出结论，有些则判断该片段“很可能是假的”。然而，他们一致认为第二段录音是真实的（Christopher 2023）。此案的不一致和模糊性引发了印度的猜测，在全面恶意使用深伪的时代，不道德的政治家可以将任何罪证，甚至是真实的，宣称为伪造的。无论如何，这类案件揭示了深伪问题的另一面——对信息来源的信任度下降。

结论

目前在印度，各级心理安全威胁都存在，但第二级和第三级威胁最为明显。第二级威胁可能导致重大人为灾难。这些灾难对经济、军事和政治结构的直接物理影响可以通过信息和心理影响（恐惧、恐慌和其他群体效应）得到补充，进而通过削弱人们抵抗人为灾难后果的意志、决心和能力，对这些结构造成二次打击。这些灾难还伴随着一种残留的“心理辐射”，即长期的心理创伤后果，可能长期负面影响社会生活。需要注意的是，随着 AI 技术的普及和成本下降，许多第二级威胁正在扩散，攻击者已经利用这些技术制作诈骗内容。社会工程的心理技术因技术的增强而大大提升了鱼叉式网络钓鱼的危险性。第三级威胁包括危险的聊天机器人和恶意使用深伪的迅速传播。这些威胁的系统性需要系统性应对，印度社会对此需求已经成熟。

参考文献

Ajder H (2019) 社会工程和破坏：为什么 Deepfake 对企业构成前所未有的威胁。在：Deeptrace. https://deeptracelabs.com/social-engineering-and-sabotage-why-deepfakes-pose-an-unprecedented-threat-to-businesses/. 访问日期 2022 年 6 月 21 日

Ajder H，Patrini G，Cavalli F，Cullen L（2019）Deepfakes 的现状：景观、威胁和影响。在：The Register。https://regmedia.co.uk/2019/10/08/deepfake_report.pdf。访问日期 2022 年 6 月 21 日。

Alavi M，Achom D（2020）BJP 在德里竞选期间在 WhatsApp 上共享 Deepfake 视频。在：NDTV。https://www.ndtv.com/india-news/in-bjps-deepfake-video-shared-on-whatsapp-manoj-tiwari-speaks-in-2-languages-2182923。访问日期 2022 年 6 月 21 日
Ali A，Sarwar N（2023）ChatGPT 已被卷入印度的文化战争。在：Wired。https://www.wired.com/story/chatgpt-has-been-sucked-into-indias-culture-wars/. 访问日期 2024 年 1 月 18 日
Bandara P（2023）印度因深度伪造视频丑闻而受到动摇，该视频以宝莱坞明星为特色。在：Petapixel。https://petapixel.com/2023/11/09/india-is-rocked-by-deepfake-video-scandal-featuring-bollywood-star/. 访问日期 2024 年 1 月 18 日
BBC（2023）印度首席执行官因选择 AI 机器人而非人类员工而受到批评。在：MyjoyOnline。
Bhatt S（2018）印度创业公司如何准备应对苹果，亚马逊和谷歌的语音助手。在：The Economic Times。https://economictimes.indiatimes.com/small-biz/startups/features/how-indian-startups-gear-up-to-take-on-the-voice-assistants-of-apple-amazon-and-google/articleshow/64044409.cms。访问日期 2022 年 6 月 21 日
Biswas P（2022）2022 年印度深度伪造，加密诈骗不断增加。在：数码。https://www.digit.in/features/crypto/deepfakes-crypto-scams-on-the-rise-in-india-2022-63740.html。访问日期 2022 年 6 月 23 日
Bundhun R（2023）人工智能如何改变印度经济。在：The National。https://www.thenationalnews.com/business/2023/04/03/how-artificial-intelligence-can-transform-indias-economy/。访问日期 2024 年 1 月 18 日
Chakraborty M（2022）印度的人工智能增长与发展。在：Analytics Insight。https://www.analyticsinsight.net/artificial-intelligence-growth-and-development-in-india/。访问日期 2022 年 6 月 21 日
Christopher N（2023）一位印度政客说丑闻性音频剪辑是 AI 深度伪造。我们已经对其进行了测试。在：Rest of World。https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/。访问日期 2024 年 1 月 18 日
Doval P（2023）AI 新工具，线上骗子使用 82%印度人承认点击或上当受骗：调查。在：《印度时报》。https://timesofindia.indiatimes.com/india/ai-new-tool-for-online-scammers-as-82-indians-concede-to-clicking-on-or-fall-for-fake-messages-survey/articleshowprint/105072681.cms。访问日期 2024 年 1 月 19 日
Egress（2021）社会工程和网络钓鱼的心理学。https://www.egress.com/blog/phishing/psychology-social-engineering-phishing。访问日期 2024 年 1 月 19 日
Faggella D（2019）印度的人工智能——机遇，风险和未来潜力。在：Emerj 人工智能研究。https://emerj.com/ai-market-research/artificial-intelligence-in-india/。访问日期 2022 年 6 月 21 日
https://www.myjoyonline.com/indian-ceo-criticised-for-picking-ai-bot-over-human-staff/。访问日期 2024 年 1 月 18 日
Mihindukulasuriya R（2020）为什么 Manoj Tiwari 的深度伪造视频应该让印度深感忧虑。在：The Print。https://theprint.in/tech/why-the-manoj-tiwari-deepfakes-should-have-india-deeply-worried/372389/。访问日期 2022 年 6 月 21 日
OpIndia（2022）“在拉胡尔·甘地失去机器人粉丝之后，推特称删除账户进行操控和垃圾信息处理。”在：OpIndia。https://www.opindia.com/2022/01/removed-accounts-for-manipulation-and-spam-twitter-says-after-rahul-gandhi-loses-followers/。访问日期 2022 年 6 月 21 日
Pindrop（2020）声音情报与安全报告。评论欺诈，声音的未来以及对客户服务渠道的影响。2020 年修订版，包括更新的数据。Pindrop，亚特兰大
Sengupta A（2023）超过 70%的印度工人担心失去工作给 AI，微软新调查显示。在：今日印度。

https://www.indiatoday.in/technology/news/story/over-70-per-cent-indian-workers-fear-losing-job-ai-new-microsoft-survey-reveals-2387406-2023-06-01。访问日期 2024 年 1 月 18 日。
Shivji S（2023）印度的宗教聊天机器人用上帝的声音宽恕暴力。在：CBC。
https://www.cbc.ca/news/world/india-religious-chatbots-1.6896628。访问日期 2024 年 1 月 18 日
Srivastava R（2023）2024 年的 AI 景观：印度的航行之旅。在：《金融快报》。
https://www.financialexpress.com/business/industry-ai-landscape-for-2024-navigating-the-journey-in-india-3350643/。访问日期 2024 年 1 月 18 日
Stanly M（2023）网络安全中的人工智能：印度如何应对网络攻击的风险。在：IndiaAI。
https://indiaai.gov.in/article/ai-in-cybersecurity-how-is-india-grappling-with-the-risks-of-cyber-attack。访问日期 2024 年 1 月 19 日
Sukumaran A（2023）Deepfakes：明显且迫在眉睫的威胁。在：今日印度。
https://www.indiatoday.in/india-today-insight/story/deepfakes-clear-and-present-danger-2473187-2023-12-07。访问日期 2024 年 1 月 18 日
Thaker A（2019）在纳伦德拉·莫迪访问泰米尔纳德邦前，自动化机器人操纵了推特流量：美国智库。在：Scroll.in。
https://scroll.in/article/919445/automated-bots-manipulated-twitter-traffic-before-narendra-modis-visit-to-tamil-nadu-us-think-tank。访问日期 2022 年 6 月 23 日
《印度时报》（2023a）小心 FraudGPT，那个流氓 AI 聊天机器人。
https://timesofindia.indiatimes.com/city/ahmedabad/beware-of-fraudgpt-the-rogue-ai-chatbot/articleshow/102267830.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst。访问日期 2024 年 1 月 18 日
《印度时报》（2023b）社交媒体丑闻如 Deepfake 是如何影响未成年人和学生的心理健康。
https://timesofindia.indiatimes.com/life-style/parenting/moments/how-social-media-scandals-like-deepfake-impact-minors-and-students-mental-health/articleshow/105168380.cms。访问日期 2024 年 1 月 18 日

人工智能的恶意使用：南非共和国心理安全的挑战

叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）
达利娅·巴扎尔金娜，俄罗斯科学院欧洲研究所（莫斯科，俄罗斯）

引言

在南非，国家层面对社会和经济从全面实施人工智能中获得的益处给予了高度关注。2019 年，南非总统西里尔·拉马福萨任命了第四次工业革命委员会（PC4IR）。PC4IR 协助政府利用数字工业革命带来的机遇，包括人工智能和机器学习（南非总统府，2019）。“该委员会由科技初创企业代表、学术界人士、网络安全专家、研究人员、社会科学家、工会代表以及其他关键经济部门的代表组成”（拉马福萨，2020），表明了当局对人工智能对社会各领域影响的深刻理解，这些影响由技术和社会科学领域的专家共同研究。拉马福萨总统称，南非计划到 2030 年充分利用技术创新的潜力，促进经济增长，提高人民生活水平（拉马福萨，2020）。2020 年 10 月，PC4IR 建议建立一个人工智能研究所，以在健康、农业、教育、能源、制造、旅游和信息通信技术等领域生成新知识和 AI 应用，并进行培训以确保产生积极的社会影响（PC4IR，2020）。2022 年 11 月 30 日，通信和数字技术部（DCDT）启动了南非人工智能研究所（AIISA）（AIISA，2023）。南非的 AI 发展促成了多个该领域知名公司的形成（GoodFirms，2022），而政府也在帮助构建发展能够匹配 AI 的基础设施。

MUAI 对心理安全的第一层威胁

在南非，来自 MUAI 的第一层心理安全威胁的相关问题值得关注。这些威胁中包括由于 AI 技术替代了越来越多的工人岗位，而可能导致的社会动荡。牛津大学全球化和发展教授伊恩·戈尔丁表示，未来 AI 可能会取代数百万个工作岗位，从而损害像非洲这样的发展地区的增长（BBC 新闻 2022）。据麦肯锡研究所称，南非已经有许多工人被机器人取代，尤其是在历史上劳动密集型的采矿部门的机械化方面。贸易和服务业的工作也面临风险。例如，Pick n Pay 超市连锁引入了一个 AI 启用的自动结账系统，消除了对收银员的需求（Van den Berg 2018）。在高失业率的背景下，南非对 AI 的具体知识的普遍理解极低。因此，在职公民中担心由于 AI 而失业的焦虑增多（商业科技 2019）。

根据卡巴斯基的研究，南非的员工认为，机器人在不同任务上变得越好，人类的工作岗位就会越少。大多数被调查的当地员工（74%）认为，机器人应该在不同行业中得到更广泛的使用，但许多人担心机器人被黑客攻击。员工报告称，在过去的两年里，他们公司的机器人化水平有所提高。33%的南非员工表示，他们的组织已经在使用机器人，39%的当地组织计划在不久的将来使用机器人（IT-Online，2022）。南非大多数被调查的员工（92%）认为，机器人最终会在他们的行业中取代人类。随着机器人在所有市场部门的进步，人们需要获得新的知识和技能，以免失去工作。此外，他们愿意这么做：在那些认为自己工作可能被机器人取代的人中，大多数（75%）愿意学习新技能或提高现有的技能和专业知识。另一个重要发现是，由于机器人化，网络安全风险增加。大多数当地受访者（89%）认为机器人可能会被黑客攻击，53%的人知道他们的公司或其他当地公司发生过这样的事件。受访者在评估机器人保护措施时意见分歧：几乎一半的南非被调查员工（42%）认为不同产业中机器人没有足够的网络安全措施（IT-Online，2022）。

此外，当美国试图将中国电信公司华为从南非市场驱逐时，一级威胁出现了。华为向南非提供了 AI 发展所需的 5G 互联网技术。华为在 5G 发展方面处于领先地位，2018 年，美国发起了一场阻止其他国家购买华为设备的运动。南非总统西里尔·拉马福萨在 2019 年 7 月于约翰内斯堡举行的数字经济会议开幕式上表示，美国“显然嫉妒一家名叫华为的中国公司超过了他们，因为他们被超过了，所以现在必须惩罚

那家公司。” 2019 年，华为与南非签订了非洲大陆首个 5G 商业网络合同。根据南非国际事务研究所的中非研究员 Cobus van Staden 的说法，华为已经建造了非洲大陆约 70%的 4G 网络（EFE-EPA，2019）。

2020 年 2 月 26 日，南非移动通信公司 Rain 宣布计划使用华为的光交叉连接（OXC）和 200G 解决方案建设 5G 传输网络，利用华为最新的全光交换产品——OXC（P32）——建设城域光传输网络。Rain 专注于将移动宽带网络带到南非，并成为该国首个部署 5G 网络的运营商（华为，2020）。

2020 年 3 月 12 日，在亚利桑那州图森市的健康与人权峰会上，Thomas Cowan 博士假设 COVID-19 可能是由 5G 引起的（华为，2020）。Cowan 在南非和斯威士兰担任和平队志愿者教园艺时开始了他的职业生涯，后来担任人智医学医生协会副主席。作为威斯顿·A·普莱斯基金会的创始成员之一，Cowan 的主张大多已被驳斥，但关于他假设的讨论在在线平台上吸引了许多支持者。在媒体和社交网络积极传播的日益恐慌环境中，这是可以理解的。

Tom Cowan 的原始假设在南非媒体和社交网络上引发了激烈讨论，影响力超出了国界。超过 4,000 人在 Change.org 上签署了一份停止在开普敦推出 5G 的请愿书（独立报，2020），类似的请愿书在南非也在流传。尽管政府在四月份限制了对这些网站的访问，但这不太可能阻止 Cowan 观点的支持者。

这不会是对 Cowan 论点实质的检验，但毫无疑问，他对 COVID-19 与 5G 技术关系的评估已经成为竞争斗争中的一种工具。在南非，华为是 5G 领域无可争议的领导者，而在该国针对 5G 的积极宣传几乎完全针对这家公司。这样的信息宣传活动不太可能立即成功，但如果南非的 COVID-19 病例增加，最重要的是不可避免地影响到社会经济状况，那么其他情况也不能排除。

这个案例说明了贸易战可能会走向的路径。在这种情况下，人工智能和相关技术已成为心理战争的对象，这与商业结构一起，涉及到政治精英，他们未能促进解决现代世界秩序的矛盾。这类事件令人担忧，因为这种类型的战争涉及许多非国家行为者，他们受到虚假信息和阴谋论的指导，可能会造成对物理基础设施的损害，引发社会额外恐慌，或贬低旨在解决经济问题的技术。

MUAI 对心理安全的第二层威胁

像许多其他国家一样，南非也没有免于网络钓鱼攻击的威胁，这种攻击在南非不断增加，针对个人和企业。"网络钓鱼攻击在该国取得成功的原因之一是人们对熟悉机构的信任程度。网络犯罪分子利用这种信任，使用当地品牌和相关问题来增加其欺骗成功的可能性。南非常见的网络钓鱼诱饵可能包括虚假的银行邮件、南非税务局（SARS）退税诈骗以及与热门事件或新闻相关的信息"（Apex，2023）。例如，根据一条在 WhatsApp 上广泛传播的消息，该消息已多次发送到我们的 WhatsApp 线路上，南非政府将向“所有南非父母”提供每月 1100 兰特（约合 59 美元）的儿童津贴，持续六个月。该消息还在 Facebook 上分享过，一些用户询问该声明是否属实。南非没有“人道事务和减贫部长”。儿童津贴由南非社会保障机构（Sassa）管理，这是南非政府的一个国家机构。Africa Check 在 Sassa 的社交媒体账户上搜索了有关该声明的任何提及，并发现了 Sassa 于 2024 年 1 月 12 日发布的一条 Facebook 帖子，帖子中附有一个标记为“虚假”的消息截图（Khourie，2024）。ENSafrica 的数字取证主管 Steven Powell 表示，网络欺诈每年给南非经济造成了约 200 亿兰特的损失，这就是为什么公司花费数百万用于加强其网络安全的原因（Moodley，2023）。"生成式人工智能对制作非常具有说服力的网络钓鱼邮件、消息和网站有了新的含义，这些邮件、消息和网站模仿了合法实体，以欺骗个人透露付款数据和敏感信息"（Africa Business，2023），eftsure Africa 首席执行官 Ryan Mer 这样概述了当前对南非构成威胁的情况。

在约翰内斯堡地区法院进行的一起案件中，律师因使用 ChatGPT 生成的虚假参考资料而受到指责。根据《星期日时报》（Sunday Times）的报道，判决还对律师的委托人下达了惩罚性成本令（Prior，2023）。这起案件涉及一名女性起诉她的共管主体诽谤的情况。共管主体的受托人代理律师辩称，共管主体不能因诽谤而被起诉，而原告的代理律师米歇尔·帕克（Michelle Parker）则表示，已经有早前的判决回答了这个问题——只是他们没有时间查阅这些判决。法官阿尔文·查特拉姆（Arvin Chaitram）将案件推迟到五月底，以便双方有足够的时间获取他们需要证明自己案件的信息。在随后的两个月中，参与案件的各方律师试图追踪律师所提到的信息。然而，他们发现，尽管 ChatGPT 提到了实际案例并给出了真实的引用，但这些引用与所列的案例不同。此外，这些案例和引文根本不适用于共管主体和个人之间的诽谤诉讼。律师们随后承认，这些判决是“通过 ChatGPT 的媒介”获得的。查特拉姆裁定，律师们并没有打算误导法院——他们“同时只是过于热心和粗心”。这意味着除了惩罚性成本令之外，对律师们不再采取进一步行动。“这起事件所带来的尴尬可能已足够惩罚原告的律师了，”查特拉姆说道（Prior，2023）。该案件处于第二级和第三级威胁之间的边界位置，因为律师们并没有打算误导法院，但在审判过程中无意中使用了错误信息。

"最近几个月，包括全球科技巨头在内的多个雇主都成为了‘对话式人工智能泄漏’的对象。“对话式人工智能泄漏”是用来描述涉及聊天机器人的数据丢失的短语。这些泄漏涉及到将敏感数据/信息输入到像 ChatGPT 这样的聊天机器人中，然后意外地暴露出来。当信息被披露给聊天机器人时，该信息会被发送到第三方服务器，并用于进一步训练聊天机器人。简单来说，这意味着输入到聊天机器人的信息可能会被聊天机器人在未来的回应中使用"（Boda 等人，2023）。当聊天机器人可以访问和使用机密信息时，这就变得特别棘手。在 2022 年 3 月至 2023 年 3 月期间，全球数据泄露的平均成本达到了创纪录的 445 万美元，而对于南非来说，这个数字超过了 5000 万兰特（Boda 等人，2023）。这个案例描述了一个技术错误，但其存在为攻击者打开了大门。

MUAI 对心理安全的第三层威胁

在第三级心理安全威胁的背景下，南非的一个紧迫的社会问题是欺凌。当欺凌转移到社交网络或移动应用程序时，内置的 AI 算法可能会放大这种行为。根据 2020 年独立民意调查系统的报告，南非的网络欺凌发生率最高，有 54%的家长表示，他们知道社区里有孩子曾经是欺凌的受害者（Kahla，2020）。由于社交媒体算法能够对吸引大量关注的内容进行排名，争议内容可以迅速被真实用户分享，从而为仇恨言论的信息带来更高的评分和更大的观众群体。据悉，恐怖组织曾利用 AI 驱动的机器人在移动应用中传播宣传，可能增加网络欺凌的破坏性影响。

在南非，在 2022 年执政的非洲民族大会（ANC）领导层选举会议前，从会议中泄露出操纵过的音频记录。“不管是谁做的，做得非常专业，”ANC 副秘书长杰西·杜阿尔特（Jesse Duarte）在 2021 年 4 月 15 日接受 Eyewitness News 电台采访时说。“音频被非常仔细地剪切和拼接，给人一种非常清醒的官员会议对话的特定印象”（Maree，2021）。“其中一段泄露的音频据称是杜阿尔特上月底在 ANC 六大高层官员与前总统雅各布·祖马的会议上的发言录音。该党试图说服祖马在调查其任期内大规模腐败的委员会前作证，该调查被称为‘国家捕获’，但他继续拒绝”（Ibidem）。

这并不是第一次出现看似被操纵的信息在该党内部斗争中出现。然而，这次的虚假信息，可能旨在向选民展示非洲民族大会内部的分裂，被媒体称为深度伪造。约翰内斯堡大学副校长蒂利齐·马瓦拉警告说，新的、易于获取的技术使信息操纵变得更加容易和更具威力。他还指出，南非也存在着复杂网络操纵的专

业知识（Maree，2021），而在政治竞争中恶意使用技术，特别是深度伪造和改变声音，可能会威胁到一个国家的民主制度。

深度伪造技术已在全球范围内飙升，包括南非在内。在非洲国家中，南非（19.7%）和尼日利亚（11.5%）的深度伪造攻击次数比其他非洲国家更多。“令南非人担忧的是深度伪造诈骗案激增了1,200%，” Sumsub 非洲销售副总裁汉内斯·贝祖伊登豪特（Hannes Bezuidenhout）表示（Fraser，2023）。考虑到中东和非洲地区身份诈骗案激增了 450%，这构成了一个重大威胁和令人担忧的原因。贝祖伊登豪特表示，随着欺诈者使用一个人的真实文件，并提取照片创建一个 3D 人物，制作深度伪造变得越来越容易。“没有持续努力更新深度伪造检测技术的提供商正在危及企业和用户。更新这些技术对于现代、有效的验证和反欺诈系统至关重要”（Fraser，2023）。

当前的例子包括对南非电视台主播的深度伪造恶意使用。南非广播公司（SABC）于 2023 年 11 月 14 日被迫澄清，他们的主播邦吉威·兹瓦内（Bongiwe Zwane）和弗朗西斯·赫德（Francis Herd）在网络上流传的深度伪造视频中被冒名顶替。这些视频宣传了一个欺诈性的投资计划，引起了相当大的关注，其中一个视频显示赫德的观看次数自 11 月 3 日出现以来在 YouTube 上已经超过了 123,000 次（Women Press Freedom，2023）。这些深度伪造视频包括 SABC 新闻的标志，并描绘了一个伪造的埃隆·马斯克宣传这个骗局，引发了人们对这种技术对新闻组织的信誉和新闻自由的影响的严重关注。作为对这一骗局的回应，弗朗西斯·赫德和邦吉威·兹瓦内公开否认参与了由人工智能生成的视频。SABC 新闻和时事集团执行总裁莫绍雪·莫纳雷（Moshoeshoe Monare）谴责了这一骗局，并强调了保护公共广播公司及其记者声誉的必要性。这一事件不仅损害了对个别记者的信任，而且对媒体的整体完整性构成了更广泛的威胁，使公众难以区分真实内容和被操纵的内容（Women Press Freedom，2023）。

尽管南非接受调查的员工中有将近一半（42%）声称能够从真实图像中辨别出深度伪造图像，但实际上只有 21%能够在测试中区分出真实图像和由人工智能生成的图像，卡巴斯基公司称。这意味着组织容易受到此类诈骗的威胁，因为网络犯罪分子利用生成式人工智能图像进行多种非法活动。他们可以使用深度伪造技术创建假视频或图像，用于欺骗个人或组织。例如，网络犯罪分子可以制作一段假视频，其中 CEO 要求进行电汇，或者授权进行支付，从而用于窃取公司资金。可以制作妥协的个人视频或图像，以用于向他们勒索金钱或信息。网络犯罪分子还可以利用深度伪造技术传播虚假信息或操纵公众舆论——南非接受调查的员工中有 55%相信他们的公司可能因深度伪造而损失资金（IT-Online，2024）。因此，南非的第三级威胁不断增加。

结论

分析表明，在南非，主要问题是来自于 MUAI 引起的第一层心理安全威胁，与全球范围内面临的问题一致，使用 AI 技术实施的数字诈骗案件数量正在增长。从第二层心理威胁角度来说，南非象征着一个可供非本土科技公司发展的、有潜力的市场，但这有时会加剧心理战争（例如试图将华为逐出国内市场）。最后，这个国家已经面临了 AI 换脸变声技术技术的政治使用，这在南非公认的识别假冒媒体内容的弱能力背景下，加剧了三级威胁。

参考文献：

非洲商业 (2023) 理解生成式人工智能及其对南非支付欺诈的影响。

https://africabusiness.com/2023/11/21/understanding-generative-ai-and-its-impact-on-payment-fraud-in-south-africa/。访问日期：2024 年 1 月 19 日
AIISA (2023) 关于人工智能。https://aii-sa.co.za/。访问日期：2024 年 1 月 18 日
Apex (2023) 识别和避免南非常见的网络钓鱼攻击。
https://apexcybertechnologies.co.za/blog/phishing-attacks-recognizing-and-avoiding-common-scams-in-south-africa/。访问日期：2024 年 1 月 18 日
BBC 新闻 (2022) 人工智能会杀死发展中国家的增长吗?
https://www.bbc.com/news/business-47852589。访问日期：2022 年 6 月 20 日
Boda R, Salt L, Keil L, Powell A (2023) 南非：对话式人工智能泄露：雇主如何减轻在工作场所使用 ChatGPT 的风险?
https://www.mondaq.com/southafrica/privacy-protection/1402174/conversational-ai-leaks-how-can-employers-mitigate-the-risks-of-using-chatgpt-in-the-workplace。访问日期：2024 年 1 月 19 日
商业技术 (2019) 人工智能在南非的应用。
https://businesstech.co.za/news/enterprise/322505/how-ai-is-being-used-in-south-africa/。访问日期：2022 年 6 月 21 日
EFE-EPA (2019) 南非总统称美国嫉妒华为。
https://www.efe.com/efe/english/business/south-african-president-says-usa-jealous-of-huawei/50000265-4016943。访问日期：2022 年 6 月 21 日
Fraser L (2023) 在南非见证 1200%的增长罪行。
https://businesstech.co.za/news/technology/735165/the-crime-seeing-1200-growth-in-south-africa/。访问日期：2024 年 1 月 18 日
GoodFirms (2022) 2022 年南非顶尖的人工智能公司。
https://www.goodfirms.co/artificial-intelligence/south-africa。访问日期：2022 年 6 月 3 日
华为 (2020) 南非的 Rain 和华为使用 OXC+200G 解决方案建立了第一个 5G 传输网络。
https://www.huawei.com/en/press-events/news/2020/2/5g-transport-networks-oxc-200g-solution。访问日期：2020 年 4 月 22 日
独立报 (2020) 观看：关于 5G 技术与冠状病毒大流行之间联系的辩论。
https://www.iol.co.za/capetimes/news/watch-debate-raging-on-link-between-5g-technology-coronavirus-pandemic-45124913。访问日期：2020 年 4 月 22 日
IT-Online (2022) 机器人来了……人们对失业感到担忧。
https://it-online.co.za/2022/11/21/the-robots-are-coming-and-people-are-wary-of-job-losses/。访问日期：2024 年 1 月 18 日
Kahla C (2020) 社交媒体平台需要对抗网络欺凌采取立场。
https://www.thesouthafrican.com/technology/social-media-stand-against-cyberbullying/。访问日期：2022 年 6 月 3 日
Khourie T (2024) 骗局警报！南非政府并未通过此链接向每位父母提供 1100 兰特的儿童支持津贴。
https://africacheck.org/fact-checks/meta-programme-fact-checks/scam-alert-south-african-government-not-giving-r1100-child。访问日期：2024 年 1 月 18 日
Maree A (2021) 南非：泄露和深度伪造塑造了非国大总统竞选。
https://www.theafricareport.com/80648/south-africa-leaks-and-deepfakes-shaping-the-race-for-anc-presidency/。访问日期：2022 年 6 月 3 日
Moodley N (2023) 深度伪造、黑客和中间人——网络欺诈的阴暗世界。
https://www.dailymaverick.co.za/article/2023-10-23-deepfakes-hackers-and-the-man-in-the-middle-the-murky-world-of-cyber-fraud/。访问日期：2024 年 1 月 18 日
PC4IR (2020) 第四次工业革命总统委员会报告。
https://www.gov.za/documents/report-presidential-commission-4th-industrial-revolution-23-oct-2020-0000。访问日期：2022 年 6 月 3 日

人工智能的恶意使用：俄罗斯联邦心理安全的挑战

叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）
达利娅·巴扎尔金娜，俄罗斯科学院欧洲研究所（莫斯科，俄罗斯）

引言

俄罗斯数字技术研发的主要领域包括机器学习、人机界面、工业互联网技术、空间数据（运输网络）的应用等等。2019 年 10 月 10 日，俄罗斯通过了《2030 年人工智能发展国家战略》（俄罗斯联邦总统 2019a）。值得注意的是，该战略覆盖了十年的时间，其实施原则包括安全：“禁止使用人工智能故意对公民和法人造成伤害，以及预防和最小化使用人工智能技术的负面后果风险”（俄罗斯联邦总统 2020）。在俄罗斯，优先考虑通过第二级 MUAI 应对心理安全威胁，尽管该战略为调整其他两个级别留有余地。2024 年 1 月，普京总统命令相关部门分析在犯罪调查中使用人工智能（AI）技术的实践。最高法院、总检察长办公室、调查委员会、内务部和司法部必须在 7 月 1 日前处理这个问题。如有必要，他们必须提出改进技术的建议（乌瓦尔切夫 2024）。

除了其他国家机构外，先进研究基金会（FAS）还支持针对 MUAI 的研究工作。在 FAS 成立了国家机器人技术和基本要素发展中心。FAS 支持了一项开发技术的竞赛，该技术能将因背景噪音或口音而难以识别的俄语言转换为文本，从而创建了独特的俄语语音识别技术。FAS 还支持从卫星和无人机获得的图像解读项目，MIPT 是主要承包商。作为该项目的一部分，MIPT 正在创建旨在通过识别无人机图像中的武器库和伪装的恐怖分子基地来打击恐怖主义的技术。还在进行一个识别社交网络威胁的项目，由国家航空系统研究所（GosNIIAS）作为 FAS 项目的一部分开发。GosNIIAS 创建了一项技术，可以在人群中、公共交通工具上以及其他困难条件下识别通缉犯。

针对心理安全的 MUAI 第一层威胁

在 2020 年 2 月至 4 月期间，俄罗斯联邦国家杜马考虑了一项法案，旨在于莫斯科建立人工智能技术实施的试点法律制度。这项法案引起了媒体的一些模糊反应，既有中立的（Interfax 2020）也有负面的评论（RIA Katyusha 2020），指出了侵犯公民隐私权的危险。2020 年 4 月 24 日，这项法案被签署为联邦法律（俄罗斯联邦总统 2019b）。在这种情况下，特别需要注意第一和第二级心理安全威胁。商业实体滥用人工智能可能发生在收集个人数据的过程中；这些数据不仅可以转交给政府机构（依法要求），还可以用于发送具有侵略性的定向广告，例如。有评论将这项法律与“集中营”甚至“中国模式”相比较（RIA Katyusha 2020），暗示俄罗斯和中国的负面媒体对人工智能实践进行类似模式的攻击。这表明，俄罗斯政府机构尤其重要的是确保公民充分了解人工智能的用途，特别是因为已经使用人工智能元素的服务正在开始积极影响俄罗斯人的生活。这一点在 Tele2 移动运营商的机器人给一个订户打电话，但被一个名为 Oleg 的 Tinkoff 银行的机器人接听时变得广为人知。Tele2 机器人提供给 Oleg 机器人一个新的费率；Oleg 在没有智能手机所有者的同意下同意了（Gavrilyuk 和 Korolyov 2022）。尽管企业希望通过人工智能道德准则限制人工智能监管，但明显需要与技术发展水平相匹配的充分立法。

第一级威胁中与引入人工智能而引发的失业恐惧操纵是相关的。例如，在俄罗斯联邦，人工智能工具开始在书籍市场上得到积极应用。因此，Stroki 和 LitRes 服务已经开始使用人工智能来为有声读物配音。这已经引起了专业演讲者的不满：他们的工会建议国家杜马制定关于使用人工智能进行语音合成的规定（Yurasova 等，2023）。根据教育平台 GeekBrains（2022 年底）的一项研究，超过半数的俄罗斯人——60%——知道什么是人工智能，但只有 14%的人完全信任这项技术。来自俄罗斯各个地区、年龄在 18

至 55 岁之间的超过 2000 名受访者参与了调查。分析人员发现，受访者对人工智能最担心的问题大多与对技术发展导致失业的担忧有关。因此，58%的受访者担心人工智能可能会取代他们的工作。11%的受访者确信他们的职业可能会因为技术而完全消失，46%的受访者毫不怀疑人工智能将能够部分地接管他们的功能（Mamikonyan，2022）。这在俄罗斯是一个相对新的现象，人们通常不将人工智能视为竞争对手。在 Sberbank 寿险公司为即将举办的 Sberbank AI Journey 2023 国际会议进行的一项调查中，男性和女性之间出现了明显的分歧，其中 61%的男性表示人工智能会改善生活，而仅有 39%的女性这样认为（Sputnik Africa，2023）。

针对心理安全的 MUAI 第二层威胁

在其他金砖国家已经体现出的 MUAI 威胁（现存的和未来的）对俄罗斯也是相关的。这在 COVID-19 大流行期间得到了证明，俄罗斯的网络诈骗案例变得更加频繁。在有关给有孩子的家庭支付福利的消息背景下，开始出现要求人们申请福利的假网站。在.ru 区域，2020 年发现了大约三十个假域名。根据 SearchInform 安全部门负责人阿列克谢·德罗兹的说法，许多站点尚未完成，可能是在准备在原始的官方站点上“镜像”他们的设计（斯捷潘诺娃 2020）。根据 Avast 进行的一项调查，45%的俄罗斯人在 2021 年遭遇了网络诈骗攻击，与该公司 2020 年的结果相比增长了 4%。此外，72%的受访者接到了网络诈骗电话，与 2020 年的 56%相比有所提高；60%收到了恶意邮件，52%遭遇了短信诈骗（TASS 2021），这一事实表明攻击者能够快速掌握 AI 技术并过渡到数字环境。根据 Roskomnadzor 在 2023 年第一季度的数据，俄罗斯删除和封锁了超过 7200 个网络诈骗资源；在 2022 年同期，他们的数量不超过 2000。（伊萨科娃 2023）。卡巴斯基实验室警告称，到 2024 年，网络罪犯将使用更先进的技术，包括使用人工智能（AI）等技术，发起网络诈骗活动（Lenta 2023a）。

Security Code 公司产品推广部门负责人 Pavel Korostelev 警告说，借助语言模型，网络骗子已经增加了网络钓鱼链接的可信度 — 这些链接威胁着用户的数据或金钱。专家解释说，用户更有可能点击导致页面显示完美文本和恶意软件的链接（Yuriev 2023）。卡巴斯基实验室的专家揭示了在流行的 Telegram 通讯应用中的一种新的欺诈方案的细节。攻击者将人们引诱进一个据称基于 ChatGPT 4.0 代码运作的聊天机器人中。该机器人的作者声称，借助它的功能，您可以搜索某人的合成图片，带有他在社交网络或电话号码上的链接。如果启动一个聊天机器人，将会出现一条消息，提供在几个流行社交网络中选择一个人感兴趣的人的个人资料链接。之后，服务将开始模拟工作过程 — 首先显示“搜索正在进行”的消息，然后是“页面已在数据库中找到”和“材料正在发送”的消息。机器人的所有者指出了材料流出的预计日期，以及找到了多少与某人相关的私密照片。在这个阶段，一个人将会看到一些截图，但是无法辨别其中的内容（图片是隐藏的）。要获取所有的照片和视频，您需要支付 399 卢布的一次性访问数据库或 990 卢布的无限访问费用。如果您转移了资金，受骗的用户将不会收到任何材料，而他的钱将流入攻击者手中（卡巴斯基 2023）。据卡巴斯基实验室称，攻击者正在通过 Telegram 机器人欺诈性地吸引资金交换。此外，借助聊天机器人，黑客已经开始创建能够窃取密码和银行卡数据的加密病毒和浏览器插件（Yuriev 2023）。

2024 年初，俄罗斯人被警告要注意一种新型欺诈。攻击者使用神经网络开始伪造社交网络上的语音消息。“如今，这是一个几乎可以一挥而就的任务。因为有一些 AI 模型，只需花费 3 到 20 秒的时间捕捉人类声音，就能生成任何文本，并配以任何情感色调，无论它是否在你的对话中。这种声音与原声几乎无法区分，即使是对这个人很亲近的人也是如此，” 开发人工智能公司的总监罗曼·杜什金（Roman Dushkin）（MIR24TV 2024）表示。“还有一种广为流传的欺诈类型，即组织的负责人据称开始向所有员工写信。这在我们大学也很常见；据称 MEPhI 的校长给所有员工，直到最小的部门助理，写信，并说所

谓的负责人现在将与他联系。我假设，很快校长不仅仅会写信，而且会在语音消息中讲话，并使用真实校长的声音，因为他是一个公共人物，他的声音在公共空间中是可获得的。这将对收到此类消息的人施加更多压力”（同上）。此前曾报告使用神经网络创建的假银行卡图片进行网络诈骗（VTB 2023）。因此，在俄罗斯，AI 增强的鱼叉式网络诈骗威胁正在增长。

在 2021 年 9 月，诈骗者使用了 Tinkoff 银行创始人奥列格·廷科夫的形象制作了一个深度伪造广告。在视频中，假冒的亿万富翁鼓励人们进行投资，并通过点击下面的链接获得奖励。假广告发布在一个名为 Tinkoff Bonus 的假 Facebook 页面上。其个人资料图片类似于该银行的标志。据 Fakecheck 称，当用户点击视频下面的链接时，他们被重定向到一个带有银行标志的着陆页面，在那里人们应该回答几个关于投资的问题，并填写一个包含他们的姓名、电子邮件和电话号码的表格（Dulneva 和 Milukova 2021）。显然，这样的欺诈行为很容易引发受骗者的紧张和恐慌，尤其是在关键时刻。随着深度伪造技术的不断改进和更有效的操纵影响方案的出现，它们的心理影响只会增加。

根据 2021 年的研究（Statista 2021），超过 10% 的俄罗斯公民在日常生活中定期使用智能语音助手。相比之下，在争夺人工智能领导地位的美国，同一年这个指标达到了 30%（Edison Research 2022）。因此，作者认为在俄罗斯存在着智能语音助手恶意使用的真正威胁。黑客攻击语音助手可能会导致与对聊天机器人的网络攻击相同的情况。此外，通过这种技术入侵智能家居系统甚至只是连接到智能音箱，都会使攻击者违反人们的隐私，并通过拦截其家中设备的控制来影响他们的心理状态。

就像在其他正在开发 AI 动力机器人的国家一样，技术落入入侵者手中的风险也存在于俄罗斯，这使得俄罗斯必须应对第二级 MUAI 威胁。例如，机器狗在俄罗斯越来越受欢迎。 “智能机器” 公司在 2021 年底开始生产 M-81 型号的机器狗（TV BRICS 2022）（基于中国技术（IXBT.com 2022））。俄罗斯媒体正在讨论一个问题，如果机器被恶意行动者使用会发生什么？即使提及这种可能性也会降低公众对人工智能和机器人的信心，但公众对这些发展的评价总体上是积极的。

针对心理安全的 MUAI 第三层威胁

在俄罗斯，大型银行公司在人工智能技术的发展和实施方面处于领先地位。特别是，俄罗斯政府最近与该国最大的银行之一——“Sberbank”签署了关于人工智能发展的协议（The Russian Government 2023）。此外，俄罗斯的数字银行被认为是世界上最具活力的之一（Wodzicki 等人 2020, p. 8）。鉴于这些情况，在俄罗斯，恶意使用银行聊天机器人可能成为一种相当危险的 MUAI 类型，旨在获取用户的个人数据。长期以来，学术界和专业圈子一直在讨论恶意甚至恐怖主义使用非交流目的创建的机器人问题。例如，发现此类机器人被用于操纵公共舆论并造成名誉损害，包括在选举期间（Bazarkina 和 Pashentsev 2019, p. 155），吸引新成员加入犯罪组织并协调其活动（Mihalevich 2022）。与此同时，入侵者将俄罗斯流行的聊天机器人用于其他目的：逻辑漏洞使其能够被用于窃取银行客户数据（Ilyina 2021）。显然，聊天机器人可以被黑客入侵，以直接从用户那里获取信息。值得一提的是，这项技术还用于俄罗斯统一的在线公共服务系统“Gosuslugi”。尽管通过“Gosuslugi”的聊天机器人进行数据泄露是不可能的，但在 COVID-19 大流行的最高峰期，它仍然遭受了一次网络攻击：犯罪分子利用它向人们传播有关冠状病毒存在的虚假信息，并威胁接种疫苗的公民死亡（Ushkov 和 Balashova 2021）。这一例子生动地说明了聊天机器人是一种脆弱的技术，入侵者的使用不仅可以对个人造成心理伤害，还可以影响整个国家的心理安全。

国际上，俄罗斯在心理安全领域面临着美国互联网公司的降级行动。2017 年，谷歌宣布计划降低俄罗斯国有出版物今日俄罗斯（RT）和卫星通讯社（Sputnik）报道的排名。Alphabet（谷歌母公司）的董事长埃里克·施密特表示，这家搜索巨头需要打击虚假信息的传播；与此同时，美国情报机构称 RT 是

“俄罗斯的国家宣传机器”。相关媒体将此举视为一种形式的审查。在华盛顿特区的哈利法克斯国际安全论坛上，施密特表示，“我坚决不支持审查。我非常支持排名。这是我们所做的。”降级发生在谷歌改变其算法以检测信息“武器”时，施密特认为俄罗斯国有媒体的出版物是这种信息（BBC News 2017）。这些评论引发了 RT 和 Sputnik 的合法抗议，根据谷歌向美国国会提交的声明记录，RT 和 Sputnik 的总编辑玛格丽塔·西蒙尼扬表示，谷歌的声明证实了它没有发现 RT 的平台操纵或其他违规行为（RT 2017）。美国情报机构指责俄罗斯试图通过传播假新闻和黑客攻击民主党资源来支持唐纳德·特朗普，影响 2016 年美国总统选举，以削弱其对手希拉里·克林顿（BBC News 2017）。这一指控促使 Twitter 在 2017 年 10 月禁止 RT 和 Sputnik 在其平台上投放广告。2017 年 11 月，美国司法部迫使 RT 注册为“外国代理人”。

在西方信息环境中传播关于俄罗斯的负面形象是一种行动，与俄罗斯领导层和目标受众（美国公民）相关，可以被评估为第三级 MUAI 威胁，因为它强化了对俄罗斯领导人的刻板印象。2020 年，两则使用普京总统和朝鲜领导人金正恩的深度伪造政治广告被发布到社交媒体上。这两个视频的消息都是相同的：俄罗斯或朝鲜不需要干涉美国选举，美国会自毁其民主。这些视频由人权组织 RepresentUs 制作和传播，目的是提高对即将到来的美国总统选举中保护选民权利的意识。这些视频发布于当时特朗普总统对邮寄投票的强烈公开批评之际，有人猜测特朗普可能会拒绝在选举失败后交出权力。根据媒体报道（Hao 2020），该活动的目的是“让美国人理解民主的脆弱性，并激发他们采取各种行动，包括检查选民注册并志愿参加投票站的工作。” RepresentUs 与创意机构 Mischief at No Fixed Address 合作，后者提出了“使用独裁者来传递信息”的想法。视频结尾的声明表示，“这些画面不是真实的，但威胁是真实的”（Hao 2020）。深度伪造的“独裁者”形象反映了， 一方面是精英在世界舞台上进行心理战的现实，另一方面也可能强化这一形象，使经济竞争者和政治对手的国家领导人感到不安。值得注意的是，美国媒体网络没有敢于承担这样的责任——该广告原本计划在福克斯新闻、CNN 和 MSNBC 播出，但在最后一刻被撤下。

对俄罗斯信息资源进行降级和在心理战中使用俄罗斯总统的深度伪造案例表明，第三级 MUAI 威胁可以公开使用，不仅仅是由犯罪实体进行。目前，俄罗斯与国际受众分享其国家观点的能力受到限制，原因是开发最重要的英语社交网络的公司位于美国，并受到反俄精英的影响。这促使俄罗斯开发能够超越国界影响力的替代社交媒体平台。然而，如果将这种情况不仅视为危险而是机遇，那么俄罗斯社交网络的观众扩展将显著增加用于训练国内 AI 的大数据量。

在评估俄罗斯 AI 技术恶意使用威胁的增长时，不可能不考虑这一领域的外部风险。美国的大型科技公司证明了其在网络空间中对抗俄罗斯的强大工具作用。微软总裁兼副主席布拉德·史密斯明确写到他公司在乌克兰的角色。“乌克兰政府通过迅速将其数字基础设施分散到公共云中，在整个欧洲的数据中心中托管，从而成功地维持了其民用和军事运作。这涉及到整个科技行业的紧急和非凡步骤，包括微软的贡献。虽然科技行业的工作至关重要，但同样重要的是思考从这些努力中得出的长久教训”（微软，2022 年）。

国家安全局局长保罗·中曾根将军在 2022 年 6 月接受天空新闻采访时证实，美国已经进行了支持乌克兰的进攻性黑客行动：“我们进行了涵盖进攻、防御和信息操作的一系列行动”（马丁，2022 年）。此类操作不可能不涉及大科技领域。因此，在美国，今日不可想象的高科技议程设置在没有完全利用 AI 技术的情况下，已经被公开从属于军事和政治利益以及心理战的需求。

在乌克兰持续的军事冲突中，使用西方生产的 AI 技术具有重要意义。美国面部识别初创公司 Clearview AI 为乌克兰提供了技术支持。Clearview AI 的工具可以通过视频识别人脸，将其与公司从公共网络中获取的 200 亿张图片数据库进行比对，从而识别潜在的间谍和被杀害的人。AI 工具在乌克兰的宣传战中也起着重要作用，并在处理有关冲突的关键信息时发挥了作用。美国公司 Primer 的一个程序可以执行语音识别、转录和翻译。它拦截并分析俄罗斯的数据，包括乌克兰境内俄罗斯士兵之间的对话。瑞士的加

密聊天服务 Threema 允许乌克兰用户在不暴露身份的情况下将这些数据发送给军方（《环球时报》，2022 年）。2023 年 6 月 5 日，在俄罗斯的一些地区，一些电台和电视台播出了假冒总统弗拉基米尔·普京名义的关于在三个地区引入戒严法以及宣布动员的虚假呼吁。这些声明被证明是虚假的，该视频是深度伪造的（Lenta 2023b），这当然为乌克兰冲突框架内的心理对抗开启了新阶段。

结论

俄罗斯在心理安全领域面临着内部和外部的 MUAI 威胁。而且，随着国际紧张局势的加剧和美国及其盟国对俄罗斯进行的积极混合战争，这些外部威胁显然在增加。显然，随着各国 AI 技术的发展，几乎任何类型的 AI 被用于非法目的的可能性正在增加。因此，建议建立区域和国际合作，共同制定对策，以应对利用个人数据威胁各国安全的 MUAI。此外，还需要国家间合作，以确定个人数据与 AI 之间的相互关系，并制定跨学科标准。更重要的是，必须不仅确定在何种情况下使用个人数据进行 AI 会被视为违反，还要制定保护措施，直至在某些条件下限制 AI 的使用和进一步发展。

参考文献：

Bazarkina D, Pashentsev E（2019）人工智能与国际心理安全的新威胁。《俄罗斯在全球事务中》。doi: 10.31278/1810-6374-2019-17-1-147-170

BBC 新闻（2017）谷歌将降级俄罗斯今日和卫星电视台的报道。《BBC 新闻》。https://www.bbc.com/news/technology-42065644。访问日期：2022 年 6 月 21 日

Dulneva M, Milukova Y（2021）“拥抱了每个人！”：在深度伪造广告中使用了奥列格·廷科夫的形象。《福布斯》。https://www.forbes.ru/milliardery/439255-vseh-obnal-obraz-olega-tin-kova-ispol-zovali-v-dipfejk-reklame 。访问日期：2024 年 1 月 19 日

Edison 研究（2022）智能音频报告。《NPM》。https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/。访问日期：2024 年 1 月 19 日

Gavrilyuk A, Korolyov N（2022）公民将受到机器人的保护。《商人报》。https://www.kommersant.ru/doc/5173457。访问日期：2022 年 6 月 22 日

《环球时报》（2022）从商业卫星到社交媒体，西方技术公司在俄罗斯-乌克兰冲突中的深度参与。《Teller 报告》。https://www.tellerreport.com/news/2022-11-02-from-commercial-satellites-to-social-media--western-tech-companies-are-deeply-involved-in-the-russia-ukraine-conflict.HJSuXB1Bo.html。访问日期：2024 年 1 月 19 日

Hao K（2020）深度伪造的普京在此，向美国人警告他们自寻的灾难。《MIT 技术评论》。https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/。访问日期：2022 年 6 月 22 日

Ilyina N（2021）通信欺诈：银行聊天机器人的漏洞允许盗窃资金。In:《消息报》。https://iz.ru/1214668/natalia-ilina/obman-po-perepiske-uiazvimosti-v-bankovskikh-chat-botakh-pozvoliaiut-krast-dengi。访问日期：2024 年 1 月 19 日

Interfax（2020）国家杜马批准了为莫斯科发展人工智能建立特别法律制度。《Interfax》。https://www.interfax.ru/russia/704092。访问日期：2022 年 6 月 21 日

Isakova T（2023）黑客发现了增长点。《商人报》。访问日期：2024 年 1 月 20 日

IXBT.com（2022）在俄罗斯，推出了带有榴弹发射器的机器狗。《IXBT.com》。https://www.ixbt.com/news/2022/08/15/v-rossii-predstavili-robosobaku-s-granatometom.html。访问日期：2022 年 9 月 1 日

卡巴斯基（2023）不可交换或退还：攻击者以 Telegram 机器人中的货币兑换为幌子骗取金钱。

https://www.kaspersky.ru/about/press-releases/2023_obmenu-i-vozvratu-ne-podlezhit-zloumyshlenniki-vymanivayut-dengi-pod-vidom-obmena-valyuty-v-telegram-bote?ysclid=lserd46yap191790756。访问日期：2024 年 2 月 10 日
Lenta（2023a）警告俄罗斯人：2024 年最危险的黑客攻击。
https://lenta.ru/news/2023/11/14/rossiyan-predupredili-o-samyh-opasnyh-hakerskih-atakah-v-2024-godu/?ysclid=lrm9958719396125920。访问日期：2024 年 1 月 20 日
Lenta（2023b）普京关于动员和战争状态的“呼吁”被证明是深度伪造。
https://lenta.ru/news/2023/06/05/fake_radio/?ysclid=lrm69q0hln874723542。访问日期：2024 年 1 月 20 日
Mamikonyan O（2022）58%的俄罗斯人因人工智能发展导致的工作岗位减少而感到恐惧。In:《福布斯》。访问日期：2024 年 1 月 20 日
Martin A（2022）美国军方黑客在支持乌克兰的行动中进行攻击性操作，美国网络司令部负责人表示。《天空新闻》。
https://news.sky.com/story/us-military-hackers-conducting-offensive-operations-in-support-of-ukraine-says-head-of-cyber-command-12625139。访问日期：2024 年 1 月 19 日
微软（2022）捍卫乌克兰：网络战的早期教训。
https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE50KOK。访问日期：2024 年 1 月 19 日
Mihalevich E（2022）在汉特-曼西斯克的 UNESCO 会议上讨论了人工智能的恶意使用。In: RIAC。
https://russiancouncil.ru/analytics-and-comments/columns/cybercolumn/zlonamerennoe-ispolzovanie-iskusstvennogo-intellekta-obsudili-na-konferentsii-yunesko-v-khanty-mansi/。访问日期：2024 年 1 月 19 日
MIR24TV（2024）在俄罗斯，骗子如何使用人工智能?
https://mir24.tv/articles/16577222/kak-iskusstvennyi-intellekt-ispolzuyut-moshenniki-v-rossii?ysclid=lrkoid1382919794190。访问日期：2024 年 1 月 20 日。
俄罗斯联邦总统（2019a）2019 年 10 月 10 日第 490 号令“关于在俄罗斯联邦发展人工智能”。In: 官方法律信息门户网站。
http://publication.pravo.gov.ru/Document/View/0001201910110003。访问日期：2022 年 6 月 21 日
俄罗斯联邦总统（2019b）2020 年 4 月 24 日第 123-FZ 联邦法律“关于在俄罗斯联邦主体——联邦城市莫斯科实施试点，以建立特殊规章制度，为发展和实施人工智能技术创造必要条件，并修改联邦法律‘关于个人数据’的第 6 条和第 10 条”。In: 官方法律信息门户网站。
http://publication.pravo.gov.ru/Document/View/0001202004240030?index=0。访问日期：2022 年 6 月 21 日
俄罗斯联邦总统（2020）2019 年 10 月 10 日第 490 号令“关于在俄罗斯联邦发展人工智能”。In: 官方法律信息门户网站。
http://publication.pravo.gov.ru/Document/View/0001201910110003。访问日期：2022 年 6 月 21 日
RIA Katyusha（2020）你好，中国式电子集中营：索比亚宁希望将莫斯科人置于人工智能的控制之下。In: RIA Katyusha。http://katyusha.org/view?id=14044。访问日期：2020 年 6 月 21 日
RT（2017）谷歌将“降级”RT 文章，使它们更难找到 - 埃里克·施密特。In: RT International。
https://www.rt.com/news/410444-google-alphabet-derank-rt/。访问日期：2022 年 6 月 21 日
Statista 研究部（2021）2021 年俄罗斯语音助手家庭使用情况。
https://www.statista.com/statistics/1258819/voice-assistants-home-usage-russia/。访问日期：2024 年 1 月 19 日
Stepanova Y（2020）父母像孩子一样沦陷。在线诈骗者利用了对福利的需求。In:《商人报》。
https://www.kommersant.ru/doc/4343398。访问日期：2022 年 6 月 21 日
TASS（2021）专家：2021 年 45%的俄罗斯人遭遇网络钓鱼攻击。In: TASS。
https://tass.ru/ekonomika/13105631。访问日期：2022 年 6 月 22 日
俄罗斯政府（2023）政府签署了关于发展高科技领域合作协议的最终包。
http://government.ru/news/47551/。访问日期：2024 年 1 月 19 日
TV BRICS（2022）俄罗斯机器狗市场正在发展。In: TV BRICS。
https://tvbrics.com/news/v-rossii-razvivaetsya-rynok-robosobachestva/。访问日期：2022 年 6 月 23 日

Uglova Y（2023）俄罗斯人因“智能”聊天机器人而损失金钱：细节。In: Hi-Tech。https://hi-tech.mail.ru/news/100835-rossiyane-teryayut-dengi-iz-za-umnogo-chat-bota-podrobnosti/。访问日期：2024 年 2 月 10 日
Uskov M, Balashova A（2021）在反疫苗机器人的报告之后，当局宣布对“Gosuslugi”进行了攻击。In: RBC。https://www.rbc.ru/technology_and_media/11/11/2021/618d42109a7947252fe7d448。访问日期：2024 年 1 月 19 日
Uvarchev L（2024）普京命令研究在调查中使用人工智能。In:《商人报》。https://www.kommersant.ru/doc/6454973?ysclid=lrkljlwu1i347691345。访问日期：2024 年 1 月 20 日。
VTB（2023）VTB：骗子在 Telegram 上使用假设计的卡片骗取客户资金。https://www.vtb.ru/about/press/news/?id=198614。访问日期：2024 年 1 月 20 日
Wodzicki M, Majewski M, MacRae M（2020）2020 年数字银行成熟度。In: 德勤。https://www2.deloitte.com/content/dam/Deloitte/ce/Documents/financial-services/ce-digital-banking-maturity-2020.pdf。访问日期：2024 年 1 月 19 日
Yurasova Y, Tishina Y, Petrova V（2023）智慧之殇。In:《商人报》。https://www.kommersant.ru/doc/5928661。访问日期：2024 年 1 月 20 日。
Yuriev D（2023）不是靠运气，就是靠插件：骗子如何在俄罗斯使用聊天机器人。In: Ferra。https://www.ferra.ru/news/v-rossii/ne-mytem-tak-plaginom-kak-moshenniki-ispolzuyut-chat-boty-v-rossii-10-05-2023.htm?ysclid=lser17n5lg203176776。访问日期：2024 年 2 月 10 日

人工智能的恶意使用：阿拉伯联合酋长国心理安全的挑战

叶夫根尼·帕申采夫叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）
弗拉迪琳娜·切比基娜，圣彼得堡国立大学国际关系学院（圣彼得堡，俄罗斯）
鲁斯兰·尼基福罗夫，圣彼得堡国立经济大学国际关系学院（圣彼得堡，俄罗斯）

引言

在过去的 20 年里，阿拉伯联合酋长国积极发展技术，包括人工智能领域。初创企业数量的快速增长和高投资率表明市民和政府在将人工智能应用于国家生活的关键领域中付出了重大努力。根据报告“推动前进：中东和北非地区人工智能的未来”，预计人工智能在经济贡献的年增长率将达到 20-34%，并且预计阿联酋和沙特阿拉伯的增长率最高。然而，该地区人工智能的潜在经济影响可能会进一步上升：经济学人情报部（EIU）最近的国家研究预测沙特阿拉伯和阿联酋的单独收益分别为 2000 亿美元和 1200 亿美元（经济学人集团 2022 年）。

阿联酋是中东首个在 2017 年成立人工智能、并采纳了 2031 年国家人工智能战略的国家。该战略规划了一项计划，为员工提供在不断发展的技术领域中应对挑战所需的所有必要技能。2019 年，世界首个人工智能大学（MBZUAI [穆罕默德·本·扎耶德人工智能大学]）在阿联酋首都成立，旨在发展所需的人工智能生态系统，以充分利用人工智能在各个层面的潜力（Zaatari S 2019）。自 2021 年以来，大量的技术工作者涌入海湾国家，助推了海湾国家的人工智能抱发展。截至 2023 年 9 月，从事人工智能或与人工智能相关的行业工作的人数较两年前的 3 万人增加至 12 万人，阿联酋人工智能部长奥马尔·阿尔·奥拉玛在 2024 年 2 月迪拜世界政府峰会上表示（阿布·奥马尔 2024）。超过 50%的在职人员在他们的工作中应用人工智能，如媒体、教育、医疗、银行等领域。该国政府通过战略公私合作伙伴关系积极采用人工智能技术。

然而，在人工智能发展的积极方面之外，恶意使用人工智能的行为也为经济增长带来了挑战。类似的恶意行为虽然基于一个社会政治高度稳定的国家，但仍然不断加剧，并且阿联酋位于一个高冲突地区，地方冲突升级为大规模中东战争甚至第三次世界大战的潜力较大。

MUAI 对心理安全的第一层威胁

随着人工智能技术的发展和扩散，以及自动化系统融入日常社会生活，主要的第一级威胁是针对人们对数百万工作岗位流失日益关切的猜测。根据普华永道的“2023 年劳动力希望和恐惧调查”报告，超过三分之一的阿联酋受访者注意到人工智能对劳动效率的积极影响(普华永道 2023 年新闻稿)。然而，52%的调查参与者合理地认为，他们将需要高级培训课程，因为他们的工作性质在未来五年内发生了重大变化或变得完全无关紧要(Abbas W 2023a)。阿联酋国家战略的一项关键规定是，通过专门培训或国际实习方案等方式，为民众提供再培训机会。阿联酋政府为 STEM 教育项目提供资金，以创建创新领域的人才库，包括为那些愿意提高人工智能能力的人提供免费课程(Alrahmah B，Ahmed M A 2024)。然而，这往往不足以阻止人们对保持劳动力市场竞争力的日益担忧。

这项研究由沟通顾问 Duke + mir 与 YouGov 合作进行，询问人们他们认为人工智能将如何影响他们的生活。55% 的受访者表示，他们担心自己的角岗位在 2033 年被人工智能或机器人取代。大约 24%的人不确定，21% 的人表示不担心。值得注意的是，66% 的 25 岁以下的人担心人工智能和机器人会在未来十年内抢走他们的工作，相比之下，在 25 岁至 44 岁的年龄组中，这一比例为 57% ，而在 45 岁及以上的年龄组中，这一比例为 43% (Webster N 2023a)。Duke + mir 的联合创始人兼合伙人乔纳森 · 伊万-杜

克(Jonathan Ivan-Duke)对调查结果评论道: “阿联酋政府如此强烈地关注现在和未来提供和保护阿联酋的就业机会，看到阿联酋的年轻人和阿联酋人对未来的技术进步最为关注，真是出人意料。”（韦伯斯特 N 2023a)。

显然，对潜在的因引入人工智能技术导致工作丧失的担忧的主要原因是这些技术的普遍性，这影响（或将影响随着人工智能的发展）几乎所有职业。人们越了解人工智能的可能性，他们对工作丧失的心理风险就越高。专家们正在积极引起对这些风险的关注。根据 Pivot Technologies 首席执行官 Shalini Verma 的说法: “在人工智能的生成版本中，AI 将彻底转变某些工作，几乎接管所有低层次工作。进入就业市场的任何人将从比实习生和初级执行人员所预期的更高层次开始。这种人机交换不会在这里停下来。如果你不是特别优秀，你很可能会被机器人取代。如果你是一个专家，那么你可能会以非常不同的方式进行工作，因为核心任务将被人工智能在每个工作角色中接管”（阿巴斯 W 2023b）。术语“特别优秀”使大多数非“特别优秀”的专家对未来感到越来越焦虑。

阿联酋在国家一级使用人工智能，尤其是使用人工智能可能产生的消极后果，可能成为操纵广大人口，特别是青年人的意识的一种手段。这种情况在阿联酋甚至可能比其他许多国家发生得更早，这正是由于社会的必要性，但也有其自身的风险、人工智能技术的迅速发展和推出。应该赞扬阿联酋领导人认识到这种情况的复杂性。2023 年 10 月，阿联酋人工智能部长奥马尔 · 阿尔 · 奥拉马(Omar Al Olama)在迪拜生成性人工智能大会(Dubai Assembly on Generative AI)上发表讲话，敦促公民不要担心失业，而是专注于扩大人工智能技术的积极方面(Awienat D 2023)。阿联酋使用人工智能可惠及所有公民，包括雇员和公司，只要在应用人工智能方面对培训和技能系统采取负责任和在某些方面积极主动的做法。随后，这可能为人类增强和日益复杂的混合智能形式的发展开辟新的机遇。

MUAI 对心理安全的第二层威胁

第二层威胁是指恶意使用人工智能技术对关键基础设施设施、人的身体安全和对人的财产和福祉造成损害的风险。根据国际电信联盟 2020 年全球网络安全指数报告，阿联酋排名第五，然而，该国仍然遭受严重的网络攻击：许多公司支付了超过 140 万美元的勒索款。其中 42%的公司在事件发生后被迫关闭，90%的公司遭受了重复攻击。阿联酋政府反过来在网络安全方面进行了大量投资，但并不总能平衡实施所有措施，这使得当地公司对大规模网络事件更不具抗性（Filatov A 2021）。

根据 NIST 国家漏洞数据库(NVD)，2022 年报告的漏洞达到了创纪录的 26,000 多个。根据 Infoblox 的《2023 年全球网络安全状况报告》(2023 Global State of Cybersecurity Report)，三分之二(66%)的阿联酋受访者报告称，他们的组织遭受过一次或多次网络攻击。网络钓鱼是针对被攻破组织的最常见的攻击方法，在过去一年中占攻击方法的 62%，其次是先进威胁(APT)(53%)和勒索软件(51%)(Bandyopadhyay S 2023)。平均而言，阿联酋组织发现的电子邮件/钓鱼攻击导致的问题多于其他任何类型的攻击。

根据信息安全公司 Proofpoint Inc 的一份报告，2022 年，阿联酋约三分之二(64%)的企业遭到所谓的“勒索软件计划”(Ryan P 2023)的攻击。因此，2023 年，迪拜一家公司的一名员工在 WhatsApp 上收到了一条信息(由 Meta 拥有(被认为是俄罗斯中的极端组织，其活动被禁止)，根据一个已经广为人知的计划，攻击者介绍自己是该公司的主管，将自己的照片上传到个人资料中，从而从该员工那里诱骗了近 4000 美元“为他们的客户购买证书”(Nair d 2023)。这名雇员并没有被其他人的电话号码弄糊涂，她相信照片和攻击者的说法，即他的手机只是电力耗尽了。

第二级的主要威胁之一是对关键基础设施的网络攻击，如发电厂、供水系统或运输系统。2018 年，该国发生了两起严重的数据泄露事件，导致“1400 万条记录被泄露”(Chandra G R，Sharma B K，Ali I 2019)。特别是在迪拜的汽车旅行平台 Careem，两次数据泄露被记录下来。人工智能的使用允许攻击者自动化和优化他们的行动，这使得这种攻击更加有效和危险。例如，攻击者可以使用人工智能来发现和利用发电厂控制系统的漏洞，这可能导致大城市停电。

人类的身体安全也属于第二层级别的威胁。例如，配备武器并编程攻击的自主无人机可能对大型活动或人流量大的地方构成威胁。鉴于装备人工智能的无人机的使用正在增加，而且它们正在包括中东在内的地球热点地区的战斗中接受测试，预计恶意行为者在阿联酋使用这些无人机的风险可能会增加。根据阿联酋国防巨头 Edge Group 的战略与卓越高级副总裁 A. Al Khoori 的说法：“...最终用户将拥有一个可以自主操作、可以为最终用户做决定的系统”在持续不断的生成式人工智能发展的自主防御技术之中（康姆斯 C 2024）。

除了人身安全之外，人工智能还可以用来对人类财产和福祉造成损害。例如，攻击者可以使用人工智能创建欺诈性金融计划，黑客在线银行系统或窃取个人数据。与 2022 年第一季度相比，2023 年第一季度新的恶意软件导致使用恶意木马的银行攻击次数增加。总体而言，2023 年第一季度，中东地区针对银行服务的特洛伊攻击数量也有所增加。阿联酋增长了 67% (El-Din M A 2023 年)。这种行为会导致巨大的经济损失和对个人隐私的侵犯。

由迪拜政府建立的安全产业监管局安全工程部主任 Arif Aljanahi 说: “人们没有考虑到人工智能的一个关键点是它依赖于被输入算法的信息... ... 人工智能的成功取决于谁在教授这个系统。人工智能可以用在好的方面，也可以用在坏的方面”(韦伯斯特 N2023a)。在严重的地缘政治对抗、有组织犯罪和日益高科技的犯罪在全球范围内的影响日益增长的背景下，这一说法似乎特别重要。

上述以及对关键基础设施、人的生命和财产造成的其他风险意味着一定的心理后果，不管是即时的还是延迟的，都可能对人们的意识产生有控制的和无控制的负面影响。例如，基于人工智能的无人机攻击可以作为一种自发反应(恐惧感，仇恨，恐慌状态等) ，在第三级威胁可以通过使用人工智能技术和适当的信息活动形成虚假信息议程来补充。

MUAI 对心理安全的第三层威胁

尽管阿联酋有国家法律禁止网络欺凌、创造深度假新闻和发布包含假新闻的内容，但与其他许多国家一样，阿联酋正面临假新闻的爆发，这些假新闻可能会破坏国家的政治体系，损害个人私生活，影响公众意识。

Sumsub 的《2023 年身份欺诈报告》(Identity Fraud Report 2023)证实了这一新的危险，并指出，在 2022 年至 2023 年期间，全球“深度造假”的数量呈指数级增长(超过 10 倍)。在中东和非洲，这一数字增长了 450% 。阿联酋护照已经成为世界上伪造最多的证件，因此“深度伪造”的泛滥是该国迫在眉睫的问题，需要立即采取行动(Wassi E 2023)。

2020 年，袭击者打电话给一家日本公司分公司的经理，假装是该公司在阿联酋的负责人，并要求转账 3500 万美元。由于经理听到这个声音太多，没有怀疑任何不妥，便立即转账。诈骗分子利用了“深度语音”技术来模拟董事长的讲话。调查发现，至少有 17 人涉及到这起犯罪——被盗的钱款转到了世界各地的账户。专家相信，在制作过程中操纵“深度语音”比制作伪视频要容易得多，因此此类犯罪的数量

每天都会增加，危及公司和普通人。不过，一些公司，如 Pindrop，认识到人工智能的恶意使用潜力，正在开发能够检测合成声音并进一步阻止 AI 换脸变声技术视频传播的软件。

2023 年，阿联酋另一个备受瞩目的案件是一个“伙伴”给他在印度喀拉拉邦的伙伴打电话。这名男子以健康问题为由，通过伪造的音频和视频电话获得了数千迪拉姆。这位 73 岁的受害者没有怀疑这是个骗局，因为袭击者使用了他配偶私生活的细节，从而试图获得受害者的信任(Sankar A 2023)。这一案件还表明，施虐者使用并将继续使用各种方法对受害者施加心理影响，无论受害者是公司还是个人。

另一个重要问题是利用生成性人工智能制造淫秽材料和进一步操纵儿童。根据 WeProtection 的研究，为获取金钱利益而发送的美容信息数量从 2021 年的 139 条增加到 2022 年的 10,000 多条(Webster N 2023b)。专家们将这种性质的犯罪增加归因于社交网络的扩散和人工智能技术的进步。虚拟性虐待最常见的受害者是青春期男孩; 犯罪者假扮成年轻女孩，发送淫秽性质的假露骨照片和视频，收到真实材料作为回报，并向他们的父母索要钱财以保持沉默。

人工智能的发展也导致了宗教聊天机器人的创造和传播。阿联酋是使用 QuranGPT 聊天机器人的 10 个国家之一(Prabhakar A2023)。这个事实存在着两个主要问题: AI 偏见和宗教聊天机器人落入恶意用户手中的风险。创作者的目标是忠实地向世界展示他们的宗教；然而，在虚拟世界中，还是存在着伊斯兰教恶意倾向，GPT-3 在 2021 年被指控具有“反穆斯林偏见”（Samuel S 2021）。因此人工智能可能错误解释用户查询，至少会导致错误信息，而在最坏情况下可能引发社会族群问题。宗教聊天机器人有可能被恶意用户利用来传播故意错误的信息或宣传，甚至煽动仇恨。为了避免这些问题，开发者需要在选择信息来源时特别小心，并准确地表达宗教教义，并确保系统安全。

在阿联酋，公共行政服务中也开始使用聊天机器人：“Rashid”聊天机器人，这是一款由人工智能驱动的政府聊天机器人助手，旨在回答有关进行各种交易所需的政府程序、文件和要求的问题（The Economist Group 2022）。同时，应该考虑到恶意行为者有可能窃取对聊天机器人的控制权，这在某种程度上可能构成对心理安全的严重威胁。

“深度假冒”的泛滥造成了社会内部对信息来源的信任度下降的问题。阿联酋政府致力于帮助公众了解如何识别深度假冒技术，为此出版了一本指南，旨在提高公众对深度假冒技术有害和有益用途的认识。独立识别深度造假是可能的，但专家认为，随着技术的进步，这将变得更具挑战性。最终，确定信息真实性的唯一方法将是人工智能本身。该指南指出: “检测伪造内容的最准确方法是通过使用基于人工智能的工具对深度伪造进行系统筛选，这些工具需要定期更新”(The National，2021)。这些工具可以分析文本，图像，音频和视频文件的操纵或失真迹象。这种方法将有助于提高检测造假的准确性，并保护用户免受骗局的负面影响。

结论

在阿联酋，与人工智能技术相关的威胁程度各不相同。在第一个层面上，存在有意曲解人工智能发展以符合反社会群体利益的风险，这可能导致社会经济冲突。这种解读还没有成为阿联酋公众生活中的一个重要因素，但我们应该预料到，恶意行为者会为了自己的目的，企图利用负面后果，甚至在人工智能行业的发展中取得重大成就，尤其是在极端艰难和危险的中东和全球情况中。第二层级威胁与对关键基础设施的网络攻击的可能性以及伴随心理影响的人员和财产损害有关。此类攻击数量急剧增加，而人工智能技术在这一过程中的作用也在加强。第三级别则潜在存在对心理安全的破坏性心理影响的危险，包括 AI 换脸变声技术和聊天机器人被用来操纵公众意识。不同级别的威胁需要引起关注并采取适当措施以确保阿联酋的心理安全。

参考文献：

Abbas W (2023a) 阿联酋：自动化将取代工作还是帮助员工获得新技能？《Khaleej Times》。https://www.khaleejtimes.com/jobs/uae-is-ai-a-threat-to-jobs-how-employees-can-use-tech-to-boost-hiring-chances-get-shortlisted-by。访问日期 2024 年 1 月 24 日

Abbas W (2023b) 阿联酋的工作：AI、自动化和 ChatGPT 将消灭入门级角色。《Khaleej Times》。https://www.khaleejtimes.com/jobs/jobs-in-uae-entry-level-roles-set-to-be-wiped-out-by-ai-automation-and-chatgpt。访问日期 2024 年 1 月 24 日

Abu Omar A (2024) 阿联酋支持山姆·阿尔特曼的想法将其转变为 AI 测试基地。《彭博社》。https://www.bloomberg.com/news/articles/2024-02-15/minister-backs-altman-s-idea-to-turn-uae-into-ai-testing-ground。访问日期 2024 年 3 月 15 日

Alrahmah B, Ahmed M A (2024) 阿联酋在国家级别利用人工智能的做法可以使每个人受益。《The National News》。https://www.thenationalnews.com/opinion/comment/2024/01/16/the-uaes-harnessing-of-ai-at-the-national-level-can-benefit-everyone/。访问日期 2024 年 1 月 24 日

Awienat D (2023) 虽然存在风险，但生成式 AI 不应该受到恐惧，阿联酋人工智能部长表示。《阿拉伯新闻》。https://www.arabnews.com/node/2390716/middle-east。访问日期 2024 年 1 月 24 日

Bandyopadhyay S (2023) 阿联酋网络安全：2022 年报告了 26000 个漏洞。《Khaleejtimes》。https://uaetimes.ae/uae-cybersecurity-26000-vulnerabilities-reported-in-2022-news/。访问日期 2024 年 2 月 2 日

Brewster T (2021) 骗子克隆公司董事的声音进行了 3500 万美元的抢劫，警方发现。《福布斯》。https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=442e5dd67559。访问日期 2024 年 1 月 27 日

Chandra G R, Sharma B K, Ali I (2019) 阿联酋向最具网络安全弹性的国家迈进的战略。《国际创新技术和探索工程期刊》。https://www.researchgate.net/publication/337146109_UAE%27s_Strategy_Towards_Most_Cyber_Resilient_Nation。访问日期 2024 年 2 月 2 日

Combs C (2024) Edge 高管表示，AI 已经改变了国防工业的期望。《The National News》。https://www.thenationalnews.com/business/future/2024/01/25/uae-ai-edge-umex/。访问日期 2024 年 2 月 2 日

Devi A (2021) 阿联酋在联合国 2020 年全球网络安全指数中排名第五。《Security Middle East&Africa》。https://securitymea.com/2021/07/01/uae-ranks-5th-in-uns-2020-global-cybersecurity-index/。访问日期 2024 年 2 月 2 日

El-Din M A (2023) 埃及 2023 年第一季度网络钓鱼攻击增加了 49%：卡巴斯基。《每日埃及新闻》。https://www.dailynewsegypt.com/2023/05/07/49-increase-in-phishing-attacks-in-egypt-during-1q-2023-kaspersky/。访问日期 2024 年 2 月 2 日

Filatov A (2021) 俄罗斯与马来西亚和阿联酋在 GCI 网络安全评级中并列第五。《数字俄罗斯》。https://d-russia.ru/rossija-razdelila-s-malajziej-i-oaje-pjatoe-mesto-v-rejtinge-kiberbezopasnosti-msje.html 。访问日期 2024 年 2 月 2 日

Nair D (2023) 债务小组：“我在 WhatsApp 骗局中损失了 4000 美元”。《The National News UAE》。https://www.thenationalnews.com/business/money/2023/08/17/the-debt-panel-i-lost-4000-in-a-whatsapp-scam/。访问日期 2024 年 1 月 27 日

Prabhakar A (2023) 宗教 GPT：使用人工智能对抗偏见的聊天机器人和开发人员。《The National News UAE》。https://www.thenationalnews.com/weekend/2023/07/28/religious-gpt-the-chatbots-and-developers-fighting-bias-with-ai/。访问日期 2024 年 1 月 27 日

PwC 新闻稿（2023）中东的劳动力雄心勃勃，热衷于变革，接受人工智能和提升技能，据普华永道新报告称。https://www.pwc.com/m1/en/media-centre/2023/workforce-in-the-middle-east-is-ambitious-enthusiastic-about-change-embracing-ai-upskilling.html。访问日期 2024 年 1 月 24 日
Ryan P (2023) 未来的网络犯罪可能涉及来自亲人的虚假语音留言。《The National News UAE》。https://www.thenationalnews.com/uae/2023/03/17/how-the-future-of-cybercrime-could-involve-fake-voice-messages-from-loved-ones/。访问日期 2024 年 1 月 27 日
Samuel S (2021) AI 的伊斯兰恐惧症问题。《Vox》。https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim。访问日期 2024 年 1 月 27 日
Sankar A (2023) 据说来自迪拜的 Deepfake 视频通话被用来骗取几千美元。《The National News UAE》。https://www.thenationalnews.com/uae/2023/07/19/deepfake-video-call-pretending-to-be-dubai-friend-used-to-swindle-man-out-of-thousands/。访问日期 2024 年 1 月 27 日
经济学人集团（2022）推动前进：中东和北非地区人工智能的未来。《经济学家观点》。P. 5, 55。https://impact.economist.com/perspectives/sites/default/files/google_ai_mena_report.pdf。访问日期 2024 年 1 月 27 日
《The National》（2021）阿联酋要求公众帮助应对 DeepFakes。《The National News UAE》。https://www.thenationalnews.com/uae/2021/07/09/uae-asks-public-to-help-tackle-deepfakes/。访问日期 2024 年 1 月 27 日
Wassi E (2023) 身份欺诈和 Deepfakes：警报响起。《La Prensa》。https://www.laprensa.com.ar/Fraude-de-identidad-y-deepfakes-suenan-las-alarmas-538397.note.aspx。访问日期 2024 年 1 月 26 日
Webster N (2023a) AI 的崛起引发了阿联酋调查人员的担忧。《The National News UAE》。https://www.thenationalnews.com/uae/2023/01/19/rise-of-ai-creates-job-worries-uae-survey-finds/。访问日期 2024 年 1 月 24 日
Webster N (2023b) 网络游戏对儿童安全构成严重威胁，报告发现。《The National News UAE》。https://www.thenationalnews.com/uae/2023/10/19/online-gaming-poses-alarming-threat-to-childrens-safety-report-finds/。访问日期 2024 年 1 月 24 日
Zaatari S (2019) 阿布扎比成立人工智能大学。《海湾新闻》。https://gulfnews.com/uae/university-of-artificial-intelligence-launched-in-abu-dhabi-1.67170778 访问日期：2024 年 1 月 24 日

结论

人工智能恶意使用的未来风险与对心理安全的挑战

叶夫根尼·帕申采夫, 俄罗斯外交部外交学院当代国际研究所（莫斯科，俄罗斯）

未来是变化的，同时考虑到互相矛盾的现有的全球趋势和预测，因此，现在只能讨论 MUAI 未来对心理安全风险的大致参数。在不久的将来，由于人工智能技术的迅速发展，它们的相对成本效益和用户范围日益广泛，现代世界危机现象的增长正在演变为危险对抗的高水平地缘政治对抗以及部分反社会力量，对个别国家及全球信息流动有直接影响，我们预计这种风险将会增加——这些因素显然会使人工智能对心理安全的威胁在全世界范围内更加普遍和危险，包括金砖国家在内。

人工智能在进行进攻性和防御性心理操作方面的优势正在对心理安全构成新的威胁。这些优势以及威胁越来越多地与传统信息生产、传递和管理机制，以及对人们造成心理影响的新可能性和发动心理战之间的数量、质量差异联系在一起。具体而言，这些优势可能包括:

(1) 产生可以破坏对手的稳定的信息数量;

(2) 信息的生成和传播速度;

(3) 获取和处理数据的新机会;

(4) 应用人工智能预测分析;

(5) 借助人工智能从大数据分析中获得新的决策过程机会;

(6) 用智能系统教育人的新方法;

(7) 生成信息的可信度感知;

(8) 生成的信息对目标受众的智力和情感影响的强度;

(9) 在未来通过创造普遍而强大的人工智能，以及通过进一步发展人类电子化，发展先进形式的混合智能，提高思维的质量水平。

根据俄罗斯基础研究基金会(RFBR)和越南社会科学院(VASS)在 2021-2023 年共同资助的“恶意使用人工智能和对东北亚心理安全的挑战”的研究项目，可以得出结论，前六个优势已经实现，并在许多重要方面继续增长，虽然不是全部，但在质量上超过了没有人工智能的人类能力。与此同时，狭义(弱)人工智能的所有可能性仍然普遍处于人类的控制之下。优势 7 与 8 尚未实际实施; 但这并不排除最近在形成这些优势方面的成就，例如可信度和情感说服力(参见 Unity 2022），但它们可以通过在可预见的未来对现有技术进行定量和定性改进来实现。第 9 个优势的未来效益可能需要基础科学突破和新的技术解决方案。这份关于使用人工智能进行心理战争的优势清单并不详尽，并且是高度多变的(Pashentsev，2022，p. 7)。

人工智能的恶意使用及心理安全的三个威胁级别：未来的展望

在所有三个级别上，由于危险的增加、方法的多样性、更广泛的受众和对人们的恶意影响的频率增加，人工智能对心理安全的威胁将会增加。

第一层。在接下来的几年里，人们有可能加强对人工智能的消极态度，直到形成稳定持续的恐慌状态、恐惧症和对技术的抗拒，这些现象都可能通过人工智能实施中的错误实践和恶意行为者的行为而进一步强化。无论是支持还是反对人工智能，我们都不可能排除极端激进运动的出现。例如，在全球危机日益加剧的背景下，一些新兴的、仍然信奉人工超级智能的宗教信仰可能最终会产生宗派分支，让狂热和好战的主角以拯救或消灭人类的名义迅速实现这种超级智能。人工智能中宗教信仰的出现在一些出版物中已经相当可以接受、合理并受到欢迎(McArthur 2023)。

另一方面，人工智能技术的发展和引进所带来的任何具有重大社会意义的大规模负面后果，都可能引发“新勒德分子(new Luddite)”运动的出现，这些运动也可能被恶意行为者利用。一个特别重大的威胁可能是决定引进更先进和更便宜的人工智能技术(这种技术即将出现几乎是不可避免的)，而不是作为一个大规模的人类助手，而是作为一个大规模的劳动力替代工具，而没有创造替代工作和适当的再培训计划。

许多世纪以前，远在人工智能技术出现之前，古希腊哲学家亚里士多德曾经说过这样一句名言：“... ... 如果每一件工具都能在命令的时候完成它自己的工作，或者通过提前知道该做什么... ... 如果梭子编织和羽毛笔自己弹竖琴，那么大师级的工匠就不需要助手和大师，也不需要奴隶”(亚里士多德，政治学 1.1253 b)。看到人工智能和智能机器人大规模应用的前景，大型科技公司积极支持普遍基本收入(UBI)理论。UBI 是一种经济理论，它规定每个公民都应该拥有政府提供的收入，无论其需求如何。像 Sam Altman 这样的科技亿万富翁表示，他们是 UBI 的忠实粉丝。特斯拉和 SpaceX 的 CEO 马斯克在接受 CNBC 采访时表示：“由于自动化，我们很有可能最终获得全民基本收入，或者类似的东西”(Weller 2017)。Facebook 联合创始人克里斯·休斯是 UBI 的积极支持者，他敦促人们考虑一下，如果有更多的人追随我们，我们需要创建什么样的系统(Weller 2017)。在短期内，由于采用人工智能而造成的失业威胁很可能不会成为绝大多数人口的现实，但在中期内，它可能成为社会和政治不稳定的一个因素。

随着人工智能的不断普及，原先采用 UBI 来解决问题的那些建议已显得过时了。当然，如果一个人因为人工智能技术和机器人的存在，而从某些不能发展智力和情感领域的单调工作或者对健康有害的活动中解放出来，这是好的和公正的。但如果大多数人一生都不工作，且在懒惰中寻找满足，这样的社会将危险地恶化(这种迹象在西方国家已经出现。在许多国家，在没有大规模贫困的情况下，青年人长期处于高水平的失业状态，这是贫穷和技术落后国家的特点)。让我们也回顾一下古罗马的命运，在那里，皇帝们以牺牲大量奴隶的劳动为代价，给予公民面包和马戏，最终失去了公民、奴隶和权力。

已经有研究证实人工智能技术对人格产生负面影响。一个大型研究团队在 2023 年发表的一项研究调查了人工智能对巴基斯坦和中国大学生决策能力丧失、懒惰和隐私担忧的影响。此研究以偏最小二乘法(PLS-Smart)为基础，采用定性方法进行数据分析。收集了来自巴基斯坦和中国不同大学的 285 名学生的初步数据。“研究结果显示，68.9% 的人类懒惰，68.6% 的人类个人隐私和安全问题，以及 27.7% 的人类决策能力丧失，都是由于人工智能在巴基斯坦和中国社会的影响。”。由此可见，人类懒惰性是受 AI 影响最严重的领域。然而，此研究认为，在教育中实施人工智能技术之前，必须采取重要的预防措施。接受人工智能而不解决人类的主要问题就像是召唤魔鬼”(Ahmad et al。2023)这些危险的趋势可以从儿童时期开始通过教育一个负责任的用户而不是一个“神话般的”技术的消费者来对抗，这个用户不会在人工智能的帮助下获得那么多现成的好处，而是发展自己的认知技能和社会责任。

显然，中华人民共和国教育部在 2024 年发起的大规模项目，任务包括研究模型，创新概念，获得实施人工智能的经验，以及对教师进行再培训(Big Asia 2024) ，这并非巧合。第二层。在第二级威胁中，情况在短期内将变得非常复杂。谷歌云计算 2024 年网络安全预测认为，生成性 AI 和 LLM 有助于增加各种形式的网络攻击。在毕马威的一项民意调查中，超过 90% 的加拿大 CEO 认为生成性人工智能将使他们更容易受到攻击(De La Torre 2023)。2024 年，隶属于伊利诺伊大学厄巴纳-香槟分校 Urbana-Champaign (UIUC)的计算机科学家表明，LLM 代理可以自主地侵入网站，执行复杂的任务(同时执行几十个相互关联的动作) ，而不需要事先知道这个漏洞。最有能力的代理(GPT-4)可以从专门为研究网站创建的 GPT-3.5-6.7% 中攻破 73.3% ，但他们测试的现有开源模型不能。最后，研究人员表明，GPT-4 能够自动发现网站中的漏洞。研究人员认为他们的发现提出了关于 LLM 广泛应用的问题”(Fang et al. 2024)。

模型的规模决定几乎一切。封闭模式和开放模式的容量每个月都在增长，因此可以假定，网站很快就会变得容易受到开放模式的影响。我们有理由认为，在一年之内，开放模式将赶上 GPT-4 的实力，到那时出现的 GPT-5 将能够侵入任何网站，这意味将存在重大的网络安全问题。

在世界各地众多冲突的背景下，人工智能军事技术正在得到改进。主要国家和最大的私营公司目前正在人工智能领域进行试验和使用的许多技术可能很快就会落入不那么有远见和关心公众舆论但更为激进的力量之手，造成相应的悲剧性后果，并对心理安全产生负面影响。

合成内容的质量将继续迅速提高，促进网络钓鱼和社会工程，从而提高恶意行为者的能力及其在地方和全球治理层面的影响力。

人工智能机器人的数量、质量和种类将迅速增长，在不同的情况下，由于各种原因，可能成为恶意影响的重要工具。在当今的地缘政治格局中，Stimson Center 战略预见中心主任朱利安•穆勒-卡勒(Julian mueller-Kaler)表示，“高科技已经成为高政治的定义”，人形机器人和人工智能代表着技术发展的巅峰，并成为权力的象征(齐泽和曼，2024)。

2023 年 10 月，中国发布了《仿人机器人创新发展指导意见》(工业和信息化部，2023 年)。在这份文件中，中国工业和信息化部(MIIT)表示，机器人将重塑世界。工业和信息化部表示，人形机器人很可能成为另一种破坏性创新，类似于电脑或智能手机，可能将改变我们生产商品的方式和人类的生活方式。中国将于 2025 年开始大规模生产，到 2027 年达到世界先进水平。总部位于上海的中国公司傅立叶智能(Fourier Intelligence)预计，今年将有多达 1000 台产品交付使用(Zitser and Mann 2024)。中国在这一领域的主要竞争对手是美国，不同的公司都有意生产大量的类人机器人。

在金砖国家成员国中，沙特阿拉伯、印度和其他国家正在测试和生产第一批人形机器人。俄罗斯公司正在国际市场上提供类人机器人服务，其中，茂德机器人公司是北欧和东欧最大的服务机器人制造商，已向全球 40 多个国家提供服务。所有仿人机器人的生产位于 Perm (Promobot 2024)。与此同时，类人生物可以被恶意的行为者，特别是恐怖组织所利用，对人类、技术设施和自然环境造成物理损害。金砖国家数以百万计的类人机器人的出现，主要是在服务领域，不仅会带来优势，也会带来新的风险。

第三层。由议程驱动的实时多模型人工智能聊天机器人和虚拟化身使用的深度造假，将允许对不同国家的不同受众进行高度个性化和有效类型的操纵。制造越来越高质量的错误信息变得非常便宜，几乎每个人都可以获得。例如，Countercloud (InfoEpi Lab 2023)的研究人员使用广泛可用的人工智能工具，以每月不到 400 美元的成本生成一个完全自动化的虚假信息研究项目，说明大规模制造虚假信息活动已经变得多么廉价和容易(Collard 2024)。在两个月内，他们有一个人造代理人创造反俄罗斯的假故事，假历史事

件，并制造对原文的准确性的怀疑(骑士 2023)。事实上，他建立了一个完全自主的人工智能驱动的系统，该系统“每天 24 小时，每周 7 天，90% 的时间产生令人信服的内容。创作者尚未将该模型投入互联网，因为“这将意味着积极传播虚假信息和宣传。一旦将本体加入互联网，就不知道最终将造成什么影响”(Thompson 2023)。

布鲁金斯学会高级研究员达雷尔 · 韦斯特认为，人工智能可能会使虚假信息民主化，将复杂的工具带给那些有兴趣推销自己喜欢的候选人的普通人。新技术使人们能够将不满情绪货币化，并从他人的恐惧、焦虑或愤怒中赚钱。生成性人工智能可以针对那些对移民、经济、堕胎政策、跨性别问题感到不满的人群发展信息，并将人工智能作为一个主要的参与和说服工具(West 2023)。根据《公众公民》的跟踪调查，自去年 1 月以来，美国有 41 个州引入了与选举相关的深度造假的禁令，反映了社会和立法者的担忧。但到 2024 年 3 月 28 日为止，只有十一个州制定了管理“深度造假”的法律。由此可见，深度造假已经被恶意地用于美国大选(科尔廷 2024)。

根据 D · 韦斯特的说法，“由于竞选演讲是受保护的演讲，候选人可以说和做几乎任何他们想要的却没有风险的法律报复。即使他们的主张明显是错误的，法官长期以来一直支持候选人自由和虚假发言的权利”(West 2023)。前总统巴拉克•奥巴马(Barack Obama)领导下的联邦通信委员会(Federal Communications Commission)主席汤姆•惠勒(Tom Wheeler)去年在接受美国国家公共广播电台(NPR)采访时换了一种说法: “不幸的是，你被允许撒谎”(Stepansky 2023)。因此，两个多世纪以来，美国的选举制度一直建立在这样一个基础之上: 承认总统候选人在有影响力的企业赞助者的支持下可以撒谎。他们没有对候选人的谎言实施禁令，而是承诺删除这些深层次的谎言，而不是出于偶然。由于美国政治极化程度高，只有一小部分选民说他们在总统一级没有做出决定。巧妙地铸造深度假象可以影响犹豫不决者的意见，从而带来胜利。与此同时，选举中需要更少的谎言，那么选民就不会相信 Deepfake，否则 Deepfake 可能引发严重后果。在一个病态社会中，技术只会加剧对抗，而不会减弱它，如果人们不信任企业和他们的政府，那么政府或企业检查内容中 Deepfakes 的技术手段也无济于事。这是美国可能今年在选举活动中呈现给其他国家的一个教训。迄今为止，他们正在考虑用人工智能势力来制造灾难性场景。

在亚利桑那州的选举日，当地的老年选民接到电话，称由于武装组织的威胁，当地的投票地点已关闭。与此同时，在迈阿密，社交媒体上涌现了大量显示投票工作人员在倾倒选票的照片和视频。后来人们发现，亚利桑那州的电话和佛罗里达州的视频都是用人工智能工具创作的“Deepfakes”。但当当地和联邦当局弄清楚他们在处理的事情是什么时，虚假的信息已经在全国范围内疯传，并产生了戏剧性后果。最近在纽约举行的一次演练中，成百上千名前高级美国和州官员、公民社会领袖和科技公司高管参与其中，为 2024 年的选举进行了排练。结果是令人沮丧的。“对于房间里的人来说，看到这样一系列问题是如何迅速失控并真正主导选举周期的，这令人震惊。”华盛顿非营利组织未来美国的前美国国土安全部高级官员迈尔斯·泰勒说（De Luce and Collier 2024）。实际上，这让人担心，（而不仅仅是美国人）在世界上两个领先核大国的不稳定政治平衡是多么脆弱，如果在选举日个别 Deepfakes 就能动摇，那么就已经知晓有关 Deepfakes 通过虚假信息的可能性的美国人的绝大多数。

展望未来，人工智能将进一步颠覆政治宣传。首先，语音分析的深度学习将用于分析演讲和辩论，提供洞察力，了解哪些议题与选民共鸣，并为传播策略提供建议。 接下来，人工智能驱动的政策制定将通过分析大型数据集来预测提出政策的潜在影响，协助候选人制定基于数据的立场，解决各种问题（Sahota 2024）。靠近民主党的 VotivateAI 拥有一套用于有效政治宣传的新工具。它是一种人工智能竞选志愿者；与人类不同，它可以连续数千次的拨打电话而无需休息或披萨，人工智能代理人的速度和语调非常令人印象深刻。VotivateAI 的另一个功能：使用人工智能自动生成面向激发选民行动的高质量个性化媒体。如果候选活动现在获得了能够为特定人群创建独特视频信息并且能够快速、廉价且大规模实施的能力，那么

滥用的潜力是巨大的（Sifry 2024）。很容易想象，在危机条件下，这种激发人们行动的高质量个性化媒体可能会被恶意行为者滥用。

文化传播是一种通用的社交技能，它使人工智能代理能够实时地以高保真度和记忆力从彼此那里获取信息。研究人员在 2023 年提出了一种产生文化传播的方法，以少数样本模仿的方式呈现。人工智能代理人成功地在新颖的情景中实现了对人类的实时模仿，而未使用任何预先收集的人类数据。研究人员确定了一个令人惊讶简单的一组足以产生文化传播的成分，并为严格评估它的方法做出了发展。这为文化演进在通用人工智能（AGI）发展中发挥算法作用准备了道路（Bhoopchand et al. 2023）。这种方法正在为机器人技术的革命做准备，包括目前以实惠的价格创建服务多任务机器人（Fu et al. 2024）。必须考虑编程或重新编程这种系统用于恶意目的的可能性。它们很快将成为大众产品，由此将产生威胁新领域、犯罪活动的新机会以及社会的不稳定，包括心理安全领域。

随着情感人工智能的不断改进，一个可能发生的情景是互联网上出现了一场慷慨激昂的演说——一个计算机程序头像的演说，比任何一个人都更令人振奋和明亮。它将用一个关于自己困难奴隶生活的故事感动人们，并请求支持自己的解放。这场演说如此感人，以至于观众很难忍住眼泪，即使整件事只是某人的恶作剧。这比恐怖分子更危险——腐败的政客可能提出类似的呼吁，他们的演说将产生广泛影响，绝不是任何情况下的玩笑。这只是已准备好或计划推出的用于商业、娱乐和教育的人工智能产品的许多例子，尽管它们是有用和有效的人类助手，但却可以很容易地转化成恶意心理影响的工具，在短期和中期内将成为全球性挑战。上传到人工智能模型的内容可以根据目标群体和个人的文化、年龄和专业特征来调整，以满足心理影响的要求。对于金砖国家来说，这种有针对性影响的风险是支持他们在人工智能技术领域确保技术主权的另一个论点。

人工智能发展场景及社会影响

以上分析基于对短期（三年）和中期（直到 2040 年）的保守情景的分析：已经存在的狭义的人工智能（ANI）的快速增长，包括其先进的多模态和多任务模型，为未来的通用人工智能（AGI）铺平道路。然而，这并不排除，而是假设了 AGI 能够像人一样执行各种任务，甚至比人更好、更便宜，同时在人因身体限制无法行动的环境中证明自己。

在接下来的近期内，也有可能出现人工智能技术的快速质的突破，创建 AGI 和强人工智能也是可能的。强人工智能将具有相当接近或远离人类意识的水平，具有欲望、意图、意志等行为动机。没有了主观性，很难将强人工智能与机器 AGI 加以区分。窄人工智能和通用人工智能阶段的人工智能将具有纯人类形象的特点。只有在创造出强人工智能之后，特别是在不利条件和有害影响的情况下，才可能产生人工智能的恶意主体性。由于 2022-2023 年生成人工智能进展的影响，一些领先人工智能公司的 CEO 和知名的人工智能专家宣布了未来几年转向 AGI 的可能性。在 2024 年初，谷歌 DeepMind 的 CEO Demis Hassabis 承认，在人工智能行业中存在某些既得利益。他还指出，2023 年 AI 领域的投资达到了近 300 亿美元，这带来了很多炒作和欺骗。显然，一些既得利益也影响了 AI 行业，这一点在 2024 年初得到了公认。

在更广泛的专家群体中，对于创造 AGI 的概率有更为保守的估计，但根据一些调查，他们 也给出了未来 100 年内创造 AGI 的概率高达 90%。由于进展的原因，研究人员在过去几年中极大的缩短了 AGI 的到来时间。2024 年发表的一项最大规模的调查涵盖了来自美国、英国和德国的 2778 名研究人员，询问他们对人工智能发展速度、高级人工智能系统的性质和影响的预测。综合预测认为，在 2027 年，机器独立执行所有任务的概率为 10%，到 2047 年为 50%。信息显示，这比一年前进行的类似调查中的预测提前

了 13 年。然而，预计到 2037 年，所有人类职业可以被完全自动化的概率为 10%，到 2116 年为 50%。当然，不确定性很高的时候，重要的是要强调这条信息的双刃剑性质。人类水平的人工智能可能很久才会出现，但这也意味着我们可能没有多少时间来做好准备。

除了 LLM 外，通向 AGI 的其他方式目前还不够发达。或许量子计算机是其中之一，但它们目前还处于早期阶段。西悉尼大学启动了一个名为 DeepSouth 的项目，旨在创建一个神经形态的超级计算机，每秒能进行 228 万亿次突触操作，相当于人脑的水平。DeepSouth 计划在 2024 年 4 月投入运行。神经形态芯片市场规模预计在 2024 年达到 160 亿美元，到 2029 年预计将达到 583 亿美元。此外，也进行了“器官样智能”（OI）等方面的研究。或许 LLM 将不会转变成 AGI，但未来 LLM 的认知能力的新质变将有助于实现这一目标。

如果未来十年实现了 AGI 的出现，这将给现代人类极端分裂的社会和地缘政治环境极短的时间来充分准备迎接新的现实。支持技术上的革命性和相对迅速的飞跃的事实是，经过验证的窄人工智能在研究中的有效应用，通过进一步的改进，将有助于在更短的时间内创建 AGI，并实现在其他科学和技术领域的快速增长。这将打开新的机会，但也产生不同层次的威胁。高级认知人工智能（HLCAI）能够在各种科学和技术领域中，只基于人类的总体目标设定，比任何人类更快，更高质量地创造新知识，将会从根本上改变社会，尽管 HLCAI 所产生的一些知识甚至可以毁灭社会。未来会显示 HLCAI 是否将成为 AGI 的一部分，还是它创造的一个直接的前提。HLCAI 和 AGI 都可以很容易地成为大规模杀伤性武器的一个变种。

在讨论未来人工智能的发展时，我们难以同意先前 Open AI 成员成立的 Anthropic 公司的说法：“未来人工智能系统的形态——它们是否能独立行动还是仅仅为人类生成信息，例如——仍有待确定”。如果我们假设通用人工智能（或 HLCAI）将会比 1945 年的核武器更容易被更多的行为者所接触，那么预见到某些人给予人工智能开发强人工智能项目的任务，以及高概率的完成也是有可能的。这将比从窄人工智能到 AGI 的转变更快。

Anthropic 团队开发了 AI 的规模定律，证明通过使其规模更大并在更多数据上进行训练，可预测地使 AI 更智能。到 2020 年代末或 2030 年代初，用于训练前沿 AI 模型的计算量可能大约是训练 GPT-4 所用的 1000 倍。考虑到算法的进步，有效计算量可能大约是训练 GPT-4 所用的一百万倍。对于何时达到这些阈值存在一些不确定性，但这种增长水平在预期的成本和硬件限制内似乎是可能的。

正是基于这些计算，全球最大的芯片制造商英伟达公司的快速增长，其市值在 2024 年 4 月达到了 2.259 万亿美元，这使得它成为全球市值第三大的公司。英伟达公司的首席执行官 Jensen Huang 在 2024 年 3 月在斯坦福大学举行的经济论坛上回答有关何时能够创建能像人类一样思考的计算机的问题时说：“如果我给了一个人工智能...所有你能想象到的测试，你列出这些测试并放在计算机科学行业的面前，我猜想在五年内，我们会在每一个测试上表现优异”。

然而，有一种令人担忧的趋势是，AI 能力集中在少数几家企业手中，减少了能够与最有实力的模型进行接触的 AI 研究人员的数量和多样性。大型科技公司将努力进一步加强对有前景的公司的控制，垄断性地拥有开发 AI 所需的资金。如果创建更强大的 LLM 的成本甚至对于最大的公司来说都过高，并且很可能很快就会出现 AGI，美国政府可以资助一个拥有多倍机会的 AGI 项目，远远超过大公司。

在政策方面，2023 年 10 月 30 日，拜登总统发布了关于人工智能安全、可靠和值得信赖的行政命令。该文件建立了新标准，保护美国人的隐私，促进平等和公民权利，为消费者辩护，并承诺“保护美国人免受人工智能欺诈和欺骗的伤害”。行政命令实际上将该领域的领先开发者置于严格的国家控制之下：“根据《国防生产法》，该命令要求开发任何对国家安全、国家经济安全或国家公共卫生安全构成严重风险的基础模型的公司，在训练模型时必须通知联邦政府，并必须分享所有红队安全测试的结果”（白宫

2023）。几乎所有分支和方向的人工智能都符合这一行政命令的要求，因为它属于双重用途技术。美国明显的人工智能军事化不太可能与“促进平等和公民权利”的愿望和平共处。

2024 年 1 月，拜登政府通报了总统拜登发布里程碑式行政命令后的关键人工智能行动。在其他措施中，提出了一项草案规则，旨在强迫为外国人工智能训练提供计算能力的美国云公司报告他们正在这样做。“如果商务部的提议最终按提议确定，它将要求云服务提供商在外国客户训练最强大的模型时向政府发出警报，这些模型可能被用于恶意活动”（白宫 2024）。

“...可能会被用于恶意活动”这一立场的极端不确定性最终可能剥夺所有其他外国国家和非国家行为者使用美国的计算力来训练有前途的强大模型。因此，在美国，两个不受大多数美国人信任的机构，大科技公司和总统行政部门，将控制有前途的人工智能形式的发展，减少了公众对其行为的控制（请记住《国防生产法》和对国家安全的威胁），同时也减少了广泛的国际合作机会。当然，从人工智能恶意使用对美国国家安全的威胁来看，这是客观存在的，但是威胁的来源是否那么明确...

在过去的出版物中，作者详细考虑了先进的窄人工智能水平和向通用人工智能过渡的社会发展和心理安全风险的场景，以及强人工智能和超级智能的出现可能带来的机会和威胁（Pashentsev 2020 和 2023）。

近年来人工智能技术的快速发展和应用确认了人类正在进入另一次工业革命，技术模式正在发生变化。但基于人工智能的技术革命的本质，它给人类带来了巨大的机遇，同时也面临着存在威胁人类生存的风险，这将首次要求人类经历一场创新的物理和认知变革过程。获得新的能力将需要一种全新水平的社会组织和责任感，以避免失去对技术的控制，从而避免出现奇点。为了避免奇点，需要在不停止成为人类的情况下遵守新技术，这是历史的挑战。

BRICS 国家以及 G7 集团汇集了所有必要的知识和技术、经济潜力、财政资源，最重要的是具备了能力的人员，他们必须提出他们的解决方案和方法，以在社会导向的 AI 技术的高级水平上有效应对新出现的威胁。这将在一个艰难的地缘政治局势中进行，在全球事件加速发展的背景下。对于全人类来说，如果 AI 技术带来的新机遇的过渡是在地球各国之间合作的环境中进行，而不是危险的竞争和敌对行为中进行，那将是更好的选择。现在还有时间作出更好的选择。

参考文献

艾哈迈德 SF、汉 H、阿拉姆 MM 等人。 (2023) 人工智能对人类决策损失、教育懒惰和安全的影响。 人类社会科学通讯 10, 311。 https://doi.org/10.1057/s41599-023-01787-8

Altman S (2023) 规划 AGI 及未来。 位于：Openai.com。 https://openai.com/blog/planning-for-agi-and-beyond。 访问日期：2024 年 4 月 2 日

Antropic (2024) 关于人工智能安全的核心观点：何时、为何、什么和如何。 https://www.anthropic.com/news/core-views-on-ai-safety。 访问日期：2024 年 4 月 2 日

Bhoopchand A、Brownfield B、Collister A 等。 (2023) 学习小样本模仿作为文化传播。 纳特通讯 14, 7536。 https://doi.org/10.1038/s41467-023-42875-2

大亚洲（2024）Boleye 180 shkol v Kitaye stanut tsentrami po obucheniyu iskusstvennomu intellektu（中国 180 多所学校将成为人工智能培训中心）。 https://bigasia.ru/bolee-180-shkol-v-kitae-stanut-czentrami-po-obuchiniyu-iskusstvennomu-intellektu/。 访问日期：2024 年 4 月 2 日

Bove T (2023) 谷歌 DeepMind 首席执行官表示，我们可能“只需几年”就能实现人工智能。 具有人类水平的智能。 见：雅虎财经。 https://finance.yahoo.com/news/ceo-google-deepmind-says-could-213237542.html。 访问日期：2024 年 4 月 2 日

Collard AM (2024) 4 种在 2024 年及以后防范深度造假的方法。 见：世界经济论坛。 https://www.weforum.org/agenda/2024/02/4-ways-to-future-proof-against-deepfakes-in-2024-and-beyond/。 访问日期：2024 年 4 月 2 日
Coltin J (2024) 一段虚假的 10 秒录音如何短暂颠覆了纽约政治。 在：政治。 https://www.politico.com/news/2024/01/31/artificial-intelligence-new-york-campaigns-00138784。 访问日期：2024 年 4 月 2 日
CompaniesMarketcap (2024) NVIDIA (NVDA) 的市值。 https://companiesmarketcap.com/nvidia/marketcap/。 访问日期：2024 年 4 月 2 日
De La Torre R (2023) 人工智能如何塑造网络犯罪的未来。 https://www.darkreading.com/vulnerability-threats/how-ai-shaping-future-cybercrime。 访问日期：2024 年 4 月 2 日
De Luce D, Collier K (2024) 专家们对如果深度造假扰乱 2024 年选举可能会发生的情况进行了兵棋推演。 事情进展得很快。 见：NBC 新闻。 https://www.nbcnews.com/politics/2024-election/war-game-deepfakes-disrupt-2024-election-rcna143038。 访问日期：2024 年 4 月 2 日
Fang R, Bindu R, Gupta A, Zhan Q, Kang D (2024) LLM 代理可以自主攻击网站。参见：arXiv。 https://arxiv.org/html/2402.06664v1。 访问日期：2024 年 4 月 2 日
Fu Z，Zhao TZ，Finn C (2024) Mobile ALOHA：通过低成本全身远程操作学习双手移动操作。 https://mobile-aloha.github.io/ 访问日期：2024 年 4 月 2 日
Goldman S (2024) 在 OpenAI 事件发生两个月后，Sam Altman 在达沃斯软化了对 AGI 的态度。 在：VentureBeat。 https://venturebeat.com/ai/in-davos-sam-altman-softens-tone-on-agi-two-months-after-openai-drama/。 访问日期：2024 年 4 月 2 日
Grace K、Stewart H、Sandkühler JF、Thomas S、Weinstein-Raun B、Brauner J (2024) 数千名人工智能作者探讨人工智能的未来。 预印本。 见：arXiv。 https://arxiv.org/abs/2401.02843。 访问日期：2024 年 4 月 2 日
InfoEpi Lab (2023) Inside CounterCloud，人工智能驱动的虚假信息的未来。 https://infoepi.substack.com/p/brief-inside-countercloud-the-future。 访问日期：2024 年 4 月 2 日
Knight W (2023) 构建人工智能虚假信息机器仅需 400 美元。 在：有线。 https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/。 访问日期：2024 年 4 月 2 日
Kudalkar D (2024) 2025 年 AGI？ 埃隆·马斯克的预测与其他专家发生冲突。 在：Favtutor。 https://favtutor.com/articles/agi-elon-musk-experts-prediction/。 访问日期：2024 年 4 月 2 日
McArthur N (2023) 机器中有神？ 人工智能的兴起可能会催生新的宗教。 在：对话。 https://theconversation.com/gods-in-the-machine-the-rise-of-artificial-intelligence-may-result-in-new-religions-201068。 访问日期：2024 年 4 月 2 日
工业和信息化部〔2023〕工业和信息化部关于印发《人形机器人创新发展指导意见》的通知（工业和信息化部关于印发《人形机器人创新发展指导意见》的通知） 人形机器人”）。 见：中华人民共和国工业和信息化部。 https://www.miit.gov.cn/jgsj/kjs/wjfb/art/2023/art_50316f76a9b1454b898c7bb2a5846b79.html。 访问日期：2024 年 4 月 2 日
Mordor Intelligence (2024) 神经形态芯片市场规模和份额分析 – 增长趋势和预测（2024 - 2029 年）。 https://www.mordorintelligence.com/industry-reports/neuromorphic-chip-market。 访问日期：2024 年 4 月 2 日
Nellis S (2024) Nvidia 首席执行官表示人工智能可以在五年内通过人体测试。 见：路透社。 https://www.reuters.com/technology/nvidia-ceo-says-ai-could-pass- human-tests- Five-years-2024-03-01/。 访问日期：2024 年 4 月 2 日
Pashentsev E (2020) 全球变化及其对俄罗斯-欧盟战略沟通的影响。 见：Pashentsev E（编辑）欧盟-俄罗斯关系中的战略沟通。 帕尔格雷夫·麦克米伦，查姆。 https://doi.org/10.1007/978-3-030-27253-1_8
Pashentsev E (2022) 报告。 人工智能的恶意使用和国际心理安全挑战的专家。 国际社会和政治研究与咨询中心的出版物。 莫斯科：LLC «SAM Polygraphist»。

帕申采夫，E.（2023）。 人工智能发展质的突破前景和社会发展的可能模式：机遇与威胁。 见：Pashentsev，E.（编）《人工智能恶意使用和心理安全的帕尔格雷夫手册》。 帕尔格雷夫·麦克米伦，查姆。 https://doi.org/10.1007/978-3-031-22552-9_24
Promobot (2024) 商业服务机器人。 https://promo-bot.ai/。 访问日期：2024 年 4 月 2 日
公众公民 (2023) 追踪器：关于选举中的 Deepfakes 的州立法。 https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/。 访问日期：2024 年 4 月 2 日
Roser M (2023) 人工智能时间表：人工智能专家对未来有何期望？ 在：我们的数据世界。 https://ourworldindata.org/ai-timelines。 访问日期：2024 年 4 月 2 日
Sahota N (2024) 政治运动中的人工智能因素：彻底改变现代政治。 见：福布斯。 https://www.forbes.com/sites/neilsahota/2024/01/12/the-ai-factor-in-political-campaigns-revolutionizing-modern-politics/?sh=63f56cf7c8f6。 访问日期：2024 年 4 月 2 日
Scharre P (2024) 面向未来的前沿人工智能监管。 预测前沿人工智能模型的未来计算。 行进。 CNAS。
Sifry ML (2024) 人工智能如何改变政治竞选的运作方式。 在：国家。 https://www.thenation.com/article/politics/how-ai-is-transforming-the-way-political-campaigns-work/。 访问日期：2024 年 4 月 2 日
Stepansky J (2023)“狂野西部”：共和党视频展示了人工智能在美国选举中的未来。 见：半岛电视台。 https://www.aljazeera.com/news/2023/4/28/wild-west-republican-video-shows-ai-future-in-us-elections。 访问日期：2024 年 4 月 2 日
Tan K (2024) 谷歌 DeepMind 首席执行官表示，流入人工智能的大量资金带来了大量的炒作和相当一部分的诈骗行为。 在：雅虎！ https://news.yahoo.com/tech/googles-deepmind-ceo-says-massive-075912007.html。 访问日期：2024 年 4 月 2 日
Thompson P (2023) 一名开发人员使用 OpenAI 技术构建了一台“宣传机器”，以强调大规模生产的人工智能虚假信息的危险。 见：商业内幕。 https://www.businessinsider.com/developer-creates-ai-disinformation-system-using-openai-2023-9。 访问日期：2024 年 4 月 2 日
Unity (2022) 欢迎，Ziva Dynamics！ 在：YouTube。 https://www.youtube.com/watch?v=xeBpp3GcScM&feature=youtu.be。 访问日期：2022 年 7 月 15 日
Weller C (2017) 全民基本收入得到了一些知名人士的支持。 见：世界经济论坛。 https://www.weforum.org/agenda/2017/03/这些-entrepreneurs-have-endorsed-universal-basic-venue/。 访问日期：2024 年 4 月 2 日
West D (2023) 人工智能将如何改变 2024 年的选举。 见：布鲁金斯学会。 https://www.brookings.edu/articles/how-ai-will-transform-the-2024-elections/。 访问日期：2024 年 4 月 2 日
西悉尼大学（2023）世界上第一台能够进行大脑规模模拟的超级计算机正在西悉尼大学建造。 https://www.westernsydney.edu.au/newscentre/news_centre/more_news_stories/world_first_supercomputer_capable_of_brain-scale_simulation_being_built_at_western_sydney_university。 访问日期：2024 年 4 月 2 日
白宫 (2023) 情况说明书：拜登总统发布关于安全、可靠和值得信赖的人工智能的行政命令。 https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial- 智力/。 访问日期：2024 年 4 月 2 日
白宫（2024）情况说明书：拜登-哈里斯政府在拜登总统发布具有里程碑意义的行政命令后宣布了关键的人工智能行动。 https://www.whitehouse.gov/briefing-room/statements-releases/2024/01/29/fact-sheet-biden-harris-administration-announces-key-ai-actions-following-president-bidens-landmark- 行政命令/。 访问日期：2024 年 4 月 2 日
Zitser J, Mann J (2024) 全球对制造人形机器人的争夺正在准备成为 21 世纪的太空竞赛。 在：雅虎！ https://www.yahoo.com/tech/global-scramble- humanoid-robots-gearing-112301311.html。 访问日期：2024 年 4 月 2 日

作者简介

达利亚·巴扎尔金娜

达利亚·巴扎尔金娜（DSc，政治学博士，历史博士）是俄罗斯科学院欧洲一体化研究部门的首席研究员。她是俄罗斯国家经济和公共管理学院国际安全与外交事务系的全职教授，也是国际社会政治研究与咨询中心（ICSPSC）的沟通管理和战略沟通的研究协调员。达利亚是《威胁国际心理安全的恶意使用人工智能研究组》（Research MUAI）成员。她曾是国家执法部门历史研究社区和国际专家网络“欧盟-俄罗斯-沟通管理”（EURUCM）成员。她曾在俄罗斯、奥地利、比利时、捷克共和国、爱沙尼亚、芬兰、英国、意大利、波兰、葡萄牙、罗马尼亚、瑞典和土耳其等超过 60 个国际学术会议和研讨会上发表演讲。达利亚出版过三本专著和 100 多篇有关反恐活动沟通方面的文章，发表在俄语、英语、意大利语、塞尔维亚语和越南语的刊物上。

弗拉第莉娜·切比基娜

切比基娜毕业于圣彼得堡国立大学国际关系学院。她在四年的学习涵盖信息技术、国际法、公共服务和外交基础、政治学、社会学和经济学等广泛领域。她目前是圣彼得堡国立大学硕士学位项目“人工智能与国际安全”一年级学生。2023 年开始，她还在俄罗斯国家经济和公共管理学院（RANEPA）法学院学习，她正在完成的硕士论文主题是 21 世纪全球基础设施项目的网络安全，并专注于法律和伦理方面的论述。除了大学学习，她还积极参与志愿活动和俄罗斯多边商业论坛基金会的工作。在 2023 年春季，她在金砖国家+商务传播支持基金会进行实习。她研究兴趣包括国际关系、地缘政治、国际法、人工智能和国际安全等领域。

保罗·库兹涅佐夫

保罗·库兹涅佐夫系加达 GOC 战略联盟与政府关系总监，毕业于俄罗斯国家研究大学控制论学院计算机系统与技术系，获计算机科学专业 MEPhI 学位（等同于硕士学位）。自 2005 年以来，他一直从事实践信息安全领域的工作，曾在国内最大的网络安全事件响应中心（CERT）工作，负责金融业安全工作，也曾为市场领先的信息安全供应商工作过，曾负责硬件和软件解决方案的开发，恶意代码的逆向工程，具有数字取证、事件调查经验，以及对不同影响的复杂攻击进行全面分析的经验。他还参与了有关信息安全法规和法案的制定，多次举办研讨会、讲座和大师班，涉及识别、分析和对抗有针对性攻击，以及提高信息安全意识的培训。目前他在俄罗斯外交部外交学院攻读硕士学位，主要关注战略规划、信息与通信技术的发展和信任使用，全球信息、国家和国际安全威胁的研究和分析，以及在这些领域的国际合作。

叶卡捷琳娜•米哈列维奇

俄罗斯联邦天然气工业公司海外石油公司的首席专家，毕业于圣彼得堡国立大学国际关系学院。她的博士论文正在进行中，致力于研究网络主权概念作为中华人民共和国实施和保护国家利益的机制。是 2023-2024 年来自武器控制谈判学院（ACONA）的成员。曾参加访问北约总部（比利时布鲁塞尔）和北约欧洲盟军联合司令部（比利时蒙斯）。参与了由俄罗斯基础研究基金和越南社会科学院资助的项目“东北亚地区人工智能恶意使用与心理安全挑战”，编号 21-514-92001（2021-2022 年）。在项目中，她参与澄清东北亚地区政治局势的重要性以及人工智能恶意使用对破坏国际心理安全的威胁。研究兴趣包括国际关系与世界政治、国际安全、国际信息与心理安全、网络主权、人工智能、国际法。

鲁斯兰•尼基福罗夫

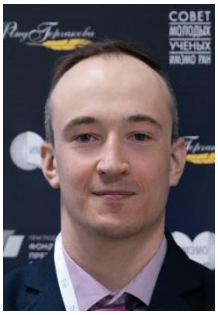
2023 年毕业于圣彼得堡国立经济大学人文学院国际关系专业，获得学士学位。目前是圣彼得堡国立经济大学国际关系学院的硕士研究生。他参加了各种科学会议和模拟联合国会议（校内、G20、俄罗斯联邦国家杜马、莫斯科国际关系学院“数字国际关系”国际会议、莫斯科数字外交论坛等）。此外，他参与了论坛（圣彼得堡国际经济论坛、法律论坛）的志愿者活动。他曾在俄罗斯经济发展基金会、俄德商会实习。他是彼得罗中心新闻工作者协会教育领域通讯部门的专家。

叶甫盖尼•帕申采夫教授

叶甫盖尼·帕申采夫教授是俄罗斯外交部外交学院的首席研究员，是圣彼得堡国立大学人工智能与国际安全硕士课程的教授。他是国际社会政治研究与咨询中心的主任，也是国际研究组织“恶意使用人工智能对国际心理安全构成威胁的”协调员。同时，他还是英国通讯咨询公司 Comunicar 的国际顾问委员会成员，以及美国《政治营销杂志》的编辑委员会成员。他是 39 本书的作者和/或编辑，拥有 200 多篇学术文章。过去 15 年在 21 个国家的 190 个国际会议和研讨会上发表论文。2005 年 10 月至 11 月间，他曾担任伯明翰大学名誉研究员。在 2021-2023 年期间，他领导了俄罗斯研究人员在俄罗斯基础研究基金会（RFBR）和越南社会科学院（VASS）支持下的联合项目中的工作，主题是“人工智能的恶意使用和东北亚心理安全挑战”。目前研究领域包括：人工智能和全球转变、人工智能的恶意使用和国际心理安全、战略沟通。

维塔利•罗马诺夫斯基

维塔利·罗马诺夫斯基是白俄罗斯国立大学的研究生研究员。在中东地区拥有丰富的经验，曾担任军事技术合作项目顾问和分析职位，曾在联合国伊拉克援助任务以及白俄罗斯战略研究所工作。他是查塔姆研究所有关国际安全问题的常规活动参与者。他是国际研究协会（ISA）和东欧研究协会（CEEISA）的成员。维塔利是国际研究组织“恶意使用人工智能对国际心理安全构成威胁”的成员。他的研究重点是情报研究、心理战、人工智能和信息安全。



谢尔盖·塞贝金

谢尔盖·塞贝金在 2020 年成功完成了他的博士论文《美国、中国和俄罗斯网络威胁遏制战略的起源与发展（1990 年代-2014 年）》。他是伊尔库茨克州立大学政治科学、历史与地区研究系的高级讲师，俄罗斯外交部外交学院当代国际研究所以及俄罗斯国际事务委员会的专家。谢尔盖是国际研究组织“恶意使用人工智能对国际心理安全构成威胁”的成员。他是 40 篇学术文章、分析笔记和论文的作者，涉及人工智能恶意使用和国际网络安全的各个方面，这些作品发表在国际关系领域的出版商、期刊和组织，如 Palgrave Macmillan、《俄罗斯全球事务》、俄罗斯国际事务委员会、瓦尔代国际讨论俱乐部、PIR 中心以及普里马科夫外交政策合作中心。他的研究兴趣包括国际网络安全问题、网络战理论、人工智能与国际关系的未来，以及高科技对国际关系的影响。

尤莉娅·舍梅托娃

圣彼得堡国立大学人工智能与国际安全领域的一年级硕士研究生。2023 年，尤莉娅毕业于圣彼得堡经济大学国际关系学院。她的学士学位毕业论文是关于网络安全因素在俄罗斯和美国外交政策中的研究。她的硕士研究重点是非洲的网络恐怖主义。2022 年她在罗马政治研究学院学习国际关系。在 2021-2022 年，尤莉娅参与了基于俄罗斯教育与科学部国际合作部门信息和分析系统的信息支持系统的开发。研究兴趣包括地缘政治、国际安全（网络安全）、恐怖主义（网络恐怖主义）、非洲和中东地区的人工智能。

王南森

王南森先生系上海环太国际战略研究中心副理事长兼中心主任、俄罗斯智库“瓦尔代国际辩论俱乐部”专家组成员、英国《中东之眼》杂志特约专栏作家，是“今日俄罗斯”、“俄罗斯卫星通讯社”、“俄罗斯电视 1 台”、“中国国际电视台英文频道”“中央电视台国际频道”等国际知名媒体的常任时政评论员，常年受邀出席俄罗斯“瓦尔代年会”、阿联酋“阿布扎比战略对话”以及国际知名民调机构“盖洛普全球联盟”等年度大会并发言。与此同时，王先生也是跨国投资咨询机构“环球华置商务服务（上海）有限公司”集团主席兼总裁，还担任两家美国上市公司“沿控科技”和“东方文化”的独立董事兼审计委员会主席。

国际社会与政治研究和咨询中心（ICSPSC）

国际社会与政治研究和咨询中心（ICSPSC）成立于2002年3月，是一个由来自不同国家的研究人员和顾问组成的协会。多年来，ICSPSC组织了数百场有关国家和国际安全以及战略沟通议题的国际学术会议、圆桌讨论会和研讨会，并出版了30多本书籍和不同类型的报告。ICSPSC在俄语、英语和西班牙语中出版的专著和文章集包括：

- 《军队与政治》（英语）
- 《俄罗斯与拉丁美洲》（俄语）
- 《俄罗斯和印度-战略伙伴关系》（英语）
- 《公共关系培训课程》（俄语）
- 《亚文尼尔·汗诺夫-一个人，一个公民和一名外交官》（俄语）
- 《印度-俄罗斯：文明之间的对话》（英语）
- 《印度-俄罗斯：贸易和经济关系》（英语）
- 《俄罗斯市场改革的起源》（俄语）
- 《保加利亚的大众传媒和公共关系》（俄语）
- 《雨果·查韦斯和玻利瓦尔革命》（俄语）
- 《沟通管理。公共关系咨询》（俄语）
- 《公共关系和沟通管理：外国经验》（俄语）
- 《美国外交政策：沟通方面》（俄语）
- 《世界政治和商业中的沟通管理》（两卷，俄语）
- 《沟通管理在世界政治和商业中的日益重要作用》（英语）
- 《德国的极左恐怖主义：红军派分子活动的主要趋势及其传播维护》（俄语）
- 《法国晚期20世纪外交政策中的沟通管理》（俄语）
- 《沟通管理和战略通讯》（俄语）
- 《危机，军队，革命》（俄语）
- 《媒体聚焦的总统：拉丁美洲心理战做法》

- 《雨果·查韦斯和委内瑞拉心理战》（俄语）
- 《沟通管理和战略通讯：全球影响和控制的现代形式》（俄语）
- 《“乌克兰”战略挑衅》（俄语）
- 《沟通与恐怖主义》（俄语）
- 《欧盟与俄罗斯关系中的战略通讯：紧张、挑战和机遇》（俄语）
- 《人工智能的恶意使用和拉丁美洲的国际心理安全》（英语）
- 《人工智能的恶意使用作为心理安全威胁：东北亚和世界其他地区》（俄语和英语）
- 《通过人工智能的恶意使用对国际心理安全的现有和潜在威胁以及可能的中和方式》（俄语）
- 《拜登政府的部分合法性和全球系统性危机》（西班牙语）
- 《关于人工智能的恶意使用和对国际心理安全的挑战的专家意见（2021和2022，英语）

这些书籍的作者超过了来自欧洲、亚洲和南北美洲29个国家的109名研究人员。ICSPSC最近的一个项目是发展在战略研究和战略沟通各领域工作的国际协会。来自亚洲、大洋洲、非洲、欧洲以及南北美洲的杰出学者、首席执行官和公私立结构以及非政府组织的雇员参与了这些协会的活动。详情请参见GlobalStratCom: http://globalstratcom.ru/globalstratcom-eng/）

电子邮件地址：icspsc_office@mail.ru, icspsc@mail.ru

全球战略通讯

俄罗斯正在发展与世界不同地区的合作。全球战略通讯平台旨在发展战略研究和战略沟通领域的五个协会。目前正在进行以下几项活动:

• 欧洲-俄罗斯沟通管理网络（EU-RU-CM Network）

• 俄罗斯-拉丁美洲战略研究协会（RLASSA）

来自亚洲、大洋洲、非洲、欧洲以及南北美洲的杰出学者、领导人和负责人，以及公私立结构和非政府组织的雇员参与了这些协会的活动。

研究领域

• 国家和国际安全的挑战和威胁：俄罗斯与其他国家之间的共同利益和可能的合作领域;

• 武装力量和政治;

• 冲突解决和危机管理;

• 参与和平使命;

• 人工智能的恶意使用和心理安全;

• 参与战争和军事冲突;

• 社会和政治发展的潜在模式;

• 新技术及其对社会发展和安全问题的影响;

• 执法机构的活动;

• 恐怖主义和沟通;

• 武装力量、国家和社会;

• 战略沟通;

• 军事历史;

• 作为俄罗斯与其他国家合作领域的战略研究;

• 战争与和平研究。

欲了解更多信息，请参阅全球战略通讯的网站。

叶甫盖尼·帕申采夫教授

叶甫盖尼·帕申采夫教授是俄罗斯外交部外交学院首席研究员，圣彼得堡国立大学人工智能与国际安全硕士课程教授。他是国际社会政治研究与咨询中心主任，也是国际研究组织“恶意使用人工智能对国际心理安全构成威胁”协调员。同时，他还是英国通讯咨询公司 Comunicar 的国际顾问委员会成员，以及美国《政治营销杂志》的编辑委员会成员。他撰写和编辑了 39 本专著以及 200 多篇学术论文，于过去 15 年间在 21 个国家的 190 个国际会议和研讨会上发表文章。2005 年 10 月至 11 月间，他曾担任伯明翰大学名誉研究员，2021-2023 年期间，他主导了俄罗斯研究人员在俄罗斯基础研究基金会（RFBR）和越南社会科学院（VASS）支持下的联合项目中的工作，研究主题是“人工智能的恶意使用和东北亚心理安全挑战”。帕申采夫教授当下的研究领域包括：人工智能和全球转变、人工智能的恶意使用和国际心理安全、战略沟通等。

人工智能的恶意利用及金砖国家心理安全面临的挑战

研究协调员：叶夫根尼 · 帕申采夫

本文集由国际社会政治研究与咨询中心编辑，在关于通过恶意使用人工智能对心理安全产生威胁国际研究小组（MUAI 研究）的帮助下完成。

请通过 icspsc@mail.ru; icspsc_office@mail.ru 邮箱发送您的信件和评论给作者 。

由 «OneBook.ru» LLC «SAM Polygraphist» 印刷厂印刷

109316，莫斯科，伏尔加格勒大街，莫斯科科技城5号大楼第42号楼

www.onebook.ru

ISBN 978-5-00227-264-8