美国工程生物学研究联盟

工程生物学与人工智能交叉融合的安保考虑

中国科学院上海营养与健康研究所
上海生命科学信息中心
上海市生物工程学会
2024 年 4 月

美国工程生物学研究联盟

工程生物学与人工智能交叉融合的安保考虑

编者按：美国工程生物学研究联盟（EBRC）于 2023 年 11 月发布题为《工程生物学与人工智能交叉融合的安保考虑》（Security Considerations at the Intersection of Engineering Biology and Artificial Intelligence）的白皮书。白皮书指出，研究人员正借助人工智能推动工程生物学的发展，但利益相关者也必须意识到人工智能可能带来的潜在风险。为此，白皮书通过对 3 个技术领域（从头开始的生物设计、闭环自主实验室系统、处理自然语言的大语言模型）的分析，探讨了这些技术的交叉可能引发的重大安保问题。针对每项技术，白皮书均详细描述了技术内容、相关安保挑战其应对策略。

工程生物学和人工智能（AI）的特点都是快速增长，有望大幅加速应对紧迫挑战的技术解决方案，但同时也可能引发的新的安保问题。随着这些技术的交叉融合，新能力正在不断被探索和认识。因此，现在应该及时识别和考虑潜在的安保影响，并探索应对策略的重要时。

工程生物学将工程设计框架应用于基因水平上生物系统的构建与修饰。研究通常以迭代的“设计-构建-测试-学习”（DBTL）循环为指导，其中基因线路、途径、生物体，甚至细胞或微生物群落是由分子成分（如 DNA、RNA、蛋白质）构建并根据设计的功能来进行测试，以测量相关的性能。该性能用于指导下一轮设计，并且循环重复，直到达到理想情况下优化的目标功能。

人工智能对蛋白质、途径、系统、甚至细胞或生物群落设计起到了增强的作用，减少了 DBTL 的迭代次数，提高了每次循环的效率。此外，正在开发闭环自主的研究系统，并且可能在未来几年实现。在这样的系统中，可编程机器人将构建和测试生物元件或系统。人工智能将从这些测试中学习并迭代原始生物设计。因此，将大幅减少人力投入。此外，像 ChatGPT 这样的自然语言大型语言模型（LLM）正在广泛使用，这些实验已经超出国家资助和国家监管研究机构的视野，

使传统的安保监管能力变得更加复杂。

分子生物学、工程生物学和生物信息学的技术和能力在 21 世纪得到迅速发展，而人工智能的加入将会显著提高这些技术的发展速度。这些技术的单独或协同应用都可能会带来安保问题[1]。人工智能的能力可以帮助恶意行为者构思方法，收集执行计划的信息；设计可能造成伤害的生物元件或系统；制定避免被发现的策略；承担旨在造成伤害的生物系统的构建、测试和武器化的过程；以及开发最终产品的分销流程。由此产生的有害产品可能包括人类、植物、动物，甚至微生物病原体，这些产品具有可能增强的疾病特征。此外，还可能出现逃避当前检测或治疗机制的新型毒素，例如，对人类有害但对当前抗毒素不敏感的肉毒杆菌毒素，或利用微生物合成非法的阿片类药物，甚至降解或削弱重要材料的微生物等。袭击的规模可能从小规模的有针对性的生物犯罪，到影响重大但规模有限的生物恐怖主义行为，再到具有高度地区、国家和/或全球发病率和死亡率的生物战争行为。各种行为体，从个人到意识形态团体，再到独立国家，都可能受到个人、意识形态、经济、政治或其他动机的驱使而尝试此类袭击。

正是认识到这种滥用的可能性，EBRC 发布了这份白皮书，旨在识别和交流工程生物学与人工智能交叉领域安保考虑的关键领域。尽管在某些主题上尚未达成共识，但 EBRC 确定了 3 个关键领域，这些领域的融合可能会带来重大的安保问题：从头开始的生物设计、闭环自主实验室系统，以及处理自然语言的大语言模型。通过讨论每个领域的预期用途和进展、引发的安保问题，以及防止和/或减轻滥用的机会，白皮书建议建立一个关于工程生物学与人工智能交叉领域的国际论坛，以识别潜在安保问题、制定缓解措施，并推动国际社会就负责任地开发和使用这些工具达成共识。

[1] 本报告的讨论范围仅限于安保（security），并将其定义为故意滥用生物学造成伤害。但是，与工程生物学和人工智能相关的安全（safety）影响同样也令人担忧。研究人员可能会无意犯下导致重大伤害的错误；自主实验室系统中的小误差可能会在 DBTL 循环中放大；人工智能系统的安全性很重要，需要进一步讨论和考虑。

1. 利用人工智能进行从头开始的生物设计

人工智能通常会加速其应用领域的发展，使用预先存在的数据来学习人类不可能或需要很长时间才能识别的趋势。人工智能可以应用于不同生物尺度的生物设计，从分子蛋白质设计到工程和代谢途径的设计，再到合成基因组，甚至是微生物群落。

1.1 人工智能在从头生物设计应用中的机遇

（1）增强的蛋白质设计

蛋白质是生物功能和活性的基础。研究人员经常利用蛋白质执行特定任务，例如，在体外合成或降解目标化合物，或将工程化蛋白质整合到生物体中。虽然自然界提供了各种各样的天然蛋白质，但这些天然蛋白质可能不具有实验室或工业使用所需的稳健性和高效性。因此，可能需要优化特定应用的蛋白质性能，或从头开始设计具有所需功能、自然界中不存在的新蛋白质。人工智能在蛋白质优化和新蛋白质设计方面展现出巨大潜力，同时也对安保领域产生影响。

研究人员此前已经能够通过结构引导设计或定向进化等方法提高蛋白质性能并达到预期目的，但人工智能可以更快、更好的设计。人工智能可以根据已知的蛋白质氨基酸序列、三维结构与生物功能相关的数据，进行“序列到功能”或“基因型到表型”算法的训练。人工智能生成的序列替换或全新片段、域或结构的预测可以产生新的优化功能。然而，蛋白质的功能和活性取决于与其相互作用的分子。Rosetta 等分子建模工具正在提高理解小分子如何与蛋白质结合，以及蛋白质如何与其他蛋白质相互作用的能力。对这些相互作用和分子界面的更多理解和建模将大大提高蛋白质的设计和优化能力。总之，人工智能将大大加速 DBTL 循环，进而发现并改进相关功能，无论是蛋白质稳定性、抗体或受体结合、催化活性、免疫原性，还是病毒嗜性等的更高阶功能。迄今为止，人工智能已被有效地用于优化蛋白质功能、结构和其他特性，并且鉴于最新进展，重新设计或从头设计具有全新功能的蛋白质可能很快就会成为常规操作。

（2）代谢途径、基因组和微生物群落的设计

生物系统的设计，如代谢途径、基因组和微生物群落，其复杂性远超单个蛋

白质的设计。代谢途径工程不仅设计单个蛋白质工程，还驱动多种酶按特定顺序工作，以将特定分子输入转化为所需输出。在此过程中，每种酶都需要必须经过优化微调其活性。基因调节元件，如启动子和增强子，以及诱导型表达系统也可能需要优化，以确保蛋白质在正确的时间和方式下生产。代谢途径工程的分子输出通常涉及新化合物的合成，这些化合物可能来源于合成化学或低效生物过程。在人工智能的助力下，这些化合物得以被筛选并鉴定。高度工程化或重编写的基因组正变得日益复杂。未来，有望设计出新型合成细胞，它们不仅具备细胞设计的功能，还需包含维持细胞功能所需的所有组件。在微生物群落的设计方面，可以利用群落成员的独特遗传能力，在空间和时间上进行调控，实现诸如碳封存或环境养分管理等目标。

人工智能正在并且毫无疑问地将继续推动这些高阶生物系统的设计。然而，人工智能支持的生物系统设计并没有人工智能支持蛋白质设计发展得好，很大程度上是由于系统的复杂性更高，以及生成足够质量的高维集成训练数据的相关挑战。生物系统是动态的，受多种因素影响，例如其他基因的表达（竞争途径中间体的转录因子或酶）、辅助因子的可用性、影响代谢通量的其他代谢物的存在、生长培养基和条件等。在生物系统中随时间测量这些动态因素具有挑战性，因此，高质量、组织良好的数据对于训练功能广泛的模型非常重要。人工智能支持的生物系统设计正在随着训练数据的改进不断完善，这也是建立更有效数据模型策略的结果。

1.2 使用人工智能从头生物设计的安保问题

人工智能辅助生物设计的进步可能会加速恶意应用的发展。例如，借助人工智能辅助设计微调病原体毒力因子，以逃避人类免疫系统的识别和攻，或将关键农业物种的害虫或病原体转化为生物武器。通过人工智能辅助设计，可能无需很多 DBTL 循环，就能在在短时间内以更低的成本完成这些项目，从而降低了被发现的风险。人工智能辅助生物设计降低了对专业知识和资源的门槛，提高了“成功”研究的可能性，因此可能激发潜在的恶意行为者对生物有害用途的兴趣。

人们还担忧，人工智能可能会帮助恶意行为者更轻松地绕过现有安全和安保系统。目前，对购买合成 DNA 的订单和客户进行自愿筛选有助于确保合理地获

得高度关注的序列。这种设计和物理世界之间的隔离至关重要；如果生物学不是建立在真实世界中，仅在计算机上设计潜在的有害生物实体并不会造成伤害。然而，由于人工智能工具能够更好地阐明序列与功能的关系，DNA 可能被设计为与任何已知的相关序列具有低序列一致性或没有序列一致性，但仍能执行相关功能。在这种情况下，DNA 合成公司可能无法识别出这些序列的潜在风险，从而在未经深入审查的情况下合成相关序列。

1.3 利用人工智能预防和缓解从头生物设计的滥用

由于人工智能工具的开源性质及其在全球范围内的开发和使用，不可能限制人工智能设计能力的发展，而且一旦限制肯定会阻碍科学进步。此外，限制使用这些人工智能模型可能会阻止小型实验室甚至初创企业进入生物技术市场。

幸运的是，在整个研发过程和开发管线中可以而且应该考虑和/或使用几种风险缓解方法。生命科学研究界应该更加关注安保问题，并尽早认识到潜在问题以便进行干预。此外，采用更复杂的归因方法也可以起到威慑作用，模型本身也可以被设计成避开某些生物设计空间，从而减低风险。

（1）领先一步：将序列到功能的算法集成到安保管线中

为了确保那些使用核酸或氨基酸序列筛选保障安保的公司在竞争中保持优势，它们必须及时了解人工智能模型在功能到序列方面的的最新进展。为此，必须投入公共和私人资源来开发能够预测新序列风险的人工智能模型，以确保这些公司相对于滥用此类模型的公司保持领先地位。将序列划分为“有害”或“无害”的模型，如果公司可以利用这些模型，从而降低构建、使用和维护方面的计算成本。如果得到支持，公司应该能够建立和维护检测可能被用来造成潜在伤害的新序列的能力。这种分类方法类似于用于检测欺诈和电子垃圾邮件（例如，“垃圾邮件”或“非垃圾邮件”）的方法。在这些情况下，模型能够从试图绕过它们的多种尝试中受益并“学习”。然而，相比试图绕过序列筛选算法的尝试，正面案在训练数据中相对较少。由于训练数据的不均匀性，建立具有高特异性的模型变得相对困难，并且非特异性模型可能导致大量误报。此外，那些从未或很少向操作员发出潜在相关无/低同源性序列警报的模型，实际上可能会遗漏本应发出警示的序列。生物安保筛查操作员无法知道缺乏系统警报是由于筛查能力不足，还

是因为没有订购需要关注的序列。因此，**美国政府应加大投资，并与科学界合作，共同开发能够识别蛋白质序列中可能具有有害功能的序列筛查工具。同时，考虑提供经济激励措施，鼓励 DNA 合成供应商使用此类模型，并支持评估筛查系统的能力发展。**

（2）社区意识和关注

高风险生物材料的开发，例如高传染性人类病原体，几乎都需要特定的实验工作流程。此类实验工作流程涉及病原体感染的人类或类人模型，以及高通量分选或筛选测量。高传染性病原体需要设计能够在雾化中存活的病毒载体。利用这种载体的实验工作流程涉及专门的步骤，包括脂质包封和/或使用大量纯化的类人脂质产生乳液。虽然有很多合理的理由使用此类实验工作流程，但社区对实验室进行的研究的认知可能意识到或发现意外或不寻常的研究实践。因此，联邦执法人员，例如联邦调查局大的规模杀伤性武器协调员，应继续与工程生物学界成员和其他相关社区建立联系，以减少讨论问题的障碍。研究人员应接受所在机构或其他实体提供的培训，了解如何采取适当行动来应对可能的风险。

（3）人工智能增强归因能力

如果恶意行为者认为被捕的可能性很高，他们可能会更加谨慎地使用生物学。因此，发展出一种能够将工程基因序列归因于特定实验室或组织的能力显得尤为重要。这能力基于一个观察结果，即各个实验室在质粒设计和构建中往往会做出一致和独特的决策，例如使用的克隆方法、筛选方法、报告基因和其他选择等，这些选择结合起来会在质粒上留下实验室的标记。模型可以利用 Addgene 等质粒库，这些质粒库保存了来自世界各地 5,700 多个实验室的 135,000 多个质粒，并将其分发给其他研究人员。在基因工程归因挑战等工作的支持下，归因能力正在迅速取得进展。然而，这种方法与其他方法一样，并不完美，主要因为①质粒在科学家之间定期共享和分发，②这种模型可以用于微调设计，最终将责任从一个来源转移或错误地指向另一个来源。

（4）安保设计

经过几十年描述蛋白质序列和功能的研究，现有的数据充足，计算能力已经发展到可以为蛋白质设计建立复杂人工智能模型的程度。目前的生物设计模型最

擅长插值（interpolation），这意味着它们在所训练的数据领域内运行良好。它们通常不善于外推到训练集中缺失的序列和功能空间。因此，如果模型开发人员同意从训练数据中删去某些只会产生危险结果的序列与功能，那么就有可能最大限度地减少潜在的风险，并增强人工智能在生物设计领域的可靠性和安全性。

事实上，这种方法的效用可能有限。需要建立一个全领域的规范，其中大多数或所有模型开发人员将同意自愿删去某些训练数据，特意降低其模型在某些生物风险中的性能。这样做法可能会在无意间造成相当大的安全隐患。如果没有任何定义和识别威胁空间的模型能力，研究人员可能会在使用人工智能构建和测试时产生有害或可能有害的设计。尽管建立并遵循了这样的规范，但由于许多模型是开源的，恶意行为者仍然有可能通过对全新蛋白质折叠或功能进行微小调整来改进模型，进而探索潜在的生物威胁空间。随着模型在外推方面的不断改进，现有的安全措施可能会逐渐失效。

开源模型的普及也限制了访问控制的效用。如果对生物设计模型的用户进行筛选，可以阻止没有凭证或合法需求的个人访问。然而，定义合法性是具有挑战的，特别是在生物设计领域，可能会阻碍公众对生物学的参与。为了替代用户筛选，工具开发人员可以要求用户使用用户名和密码登录，并可能跟踪他们的查询信息，从而支持追溯归因功能。

通过设计实现安保的另一种方式是广泛意识到人工智能应用于工程生物学的风险。Anthropic 是一家人工智能公司，开发了“用于应对灾难性风险的人工智能安全级别（ASL），大致模仿了美国政府处理危险生物材料的生物安全等级（BSL）标准”。根据 ASL 流程，可以对模型进行风险评估，然后制定适当的安全、安保和操作标准。模型风险评估或许在生物设计中也很有用。

2. 闭环、自主的生物学研究

最近，开发与自主推进研究的机器人相结合的人工智能系统受到越来越多关注。此类系统中，用户定义的参数和现有模型可以用于开发由机器人系统自主测试的假设。机器人系统生成实验数据，并将其反馈到人工智能模型中。然后，该模型会推荐测试新的假设。这种类型的“闭环自主实验室”，通常被称为“自主

驱动实验室”，理论上可以在无需人工输入或干预的情况下无限期运行。这种研究方法相较于传统方式，可能更加迅速且高效，尤其在解决复杂棘手的问题上。它可以推动 DBTL 循环的快速迭代，同时受益于人工智能支持的设计、数据分析和假设细化。通过将人类排除在实验过程之外，提高一致性和再现性（例如，人工智能模型和机器人不需要睡眠或离开实验室）。

2.1 与闭环自主生物研究相关的安保问题

与生物设计一样，闭环、自主生物研究实验室也可能被恶意行为者利用。他们可以利用自主研究设备①作为定期访问此类设备的实验室员工；②作为云实验室的客户；③通过侵入实验室系统或云实验室远程控制机器人设备；④资金充足的组织可能会建设和使用此类实验室系统用于有害目的。

对工作人员来说，对于工作人员而言，减少亲自操作实验的时间可能会增加发现可疑活动的难度。虽然实验室通常对使用这种资源密集型设备有记录保存和注册要求，但实验室人员可能会误用设备，或者擅自更改已批准的实验设置。云实验室作为新兴领域，其广泛使用和访问程度尚不明确。客户可能会隐瞒自己和/或的身份或工作性质，以外包一些必要的技术开发。此外，任何联网的自主系统都可能被黑客入侵，并可能被重新编程以开发未经批准的有害产品。然而，黑客需要对现有的物理设备、试剂和样本有广泛的了解才可能成功。最后，具备充足资源的组织可能有能力建造这样的实验室，并利用其在生物武器的开发和优化方面取得快速进展。

2.2 防止和缓解闭环自主生物研究的滥用

人工干预和人工智能驱动的策略可以最大限度地减少与自主实验室相关的生物风险。在单个实验室或铸造厂层面，基本的干预措施可能涉及在 DBTL 循环的“构建”阶段之前，或在执行与特定实所相关的风险级别所对应的循环次数之后，要求进行人工批准。一些机器人系统可能需要多个人在开始工作之前签署实验设置和轨迹。云实验室中还没有描述负责任安全措施的标准、指南或最佳实践。一些云实验室开始效仿基因合成公司，例如通过测序验证收到的样本、筛选客户、启用网络保护和防火墙防止黑客攻击等。

更复杂的保护措施可能包括在启动新的 DBTL 循环之前，制定用于估计实

验产品（例如新代谢物、工程蛋白质、DNA 序列）对人类、植物和动物，或其子集的毒性或潜在危害的指标。若经证实这种方法行之有效，那么将这种防护系统与实验系统隔离就很重要。如此一来，即使实验系统受到损害，防护措施仍然可以发挥其作用。

闭环自主实验室仍处于萌芽阶段。它们价格昂贵，而且在中短期内使用的程度尚不清楚。因此，不能夸大它们目前构成的威胁，但也不能等到该行业全面发展后才讨论制定安保规范和最佳实践。因此，**美国政府应资助相关研究，以更好地了解此类实验室未来可能发挥的作用，以及目前可能很快发生的安保问题，并为安全使用此类实验室提供建议和最佳实践。**

3. 大型语言模型

大型语言模型（LLM）的快速发展已引发公众广泛关注，同时也伴随着高度的警惕性。

在生物科学领域，人们担忧的问题是 LLM 会降低开发造成伤害的生物系统所需的专业知识水平。近期一项研究表明，麻省理工学院的部分学生在 1 个小时内通过询问 ChatGPT 了解了如何使用反向遗传学合成流行病病原体，并列出了实验流程，还确定了不进行序列筛选的 DNA 合成公司不。然而，要准确评学生使用 ChatGPT 学习和访问这些信息的速度相比使用搜索引擎的速度优势是具有挑战性的，也不清楚节省的时间是否会对有动机的恶意行为者产生多少影响，甚至该研究本身可能就会给恶意行为者使用 LLM 提供了便利。

虽然 LLM 可以明显降低知识获取的障碍，帮助恶意行为者了解生物学的两用性，但尚不清楚这些信息是否或在多大程度上会导致生物学的滥用。在滥用生物学所存在的所有障碍中（例如，试剂和设备的采购，实验室技能等），知识获取障碍占比多少？获得生物武器所花费的总时间中，最初获取知识所占比例是多少？LLM 是否为恶意行为者提供了独特的便利？LLM 是否不仅能描述如何构建生物学，还能描述如何将其武器化？隐性知识在多大程度上是无法通过 LLM 传达的？这些问题在短期、中期和长期将如何变化？

上述问题的答案不一定很清晰，研究界也不完全认同。有些担忧可能是有道

理的。LLM 降低了参与生物学活动的门槛。例如，缺少生物信息学知识的人员可以通过 LLM 的帮助使用生物设计工具；LLM 还可以帮助实验室工作人员与闭环自主研究系统对接。然而，生物学的有效滥用或武器化需要超越分子生物学或病毒学实验室技能的独特能力。例如，致病系统或生物体武器化需要设计能够在野外保持其传染性的生物系统，大规模生产该系统，测试其传播性和致死性（例如在动物或人类细胞系中），并成功引入目标人群。这种生物学必须在物理世界中进行。它需要生物设计、实验室技能、设备、材料和隐性知识，其中绝大多数仅从阅读人工智能生成的文本中学习是非常具有挑战性的。

值得注意的是，LLM 可能会非常自信地犯错或“产生幻觉”。合法使用 LLM 更快推进或解决研究问题的研究人员可能能够识别出这种错误，或者在遵循 LLM 的提示前与同事进行讨论。但恶意行为者如果没有足够的知识来识别，就会适得其反，这可能也是恶意行为者利用 LLM 的明显劣势。

也许可以构建 LLM 在收到某些类型的查询或给出某些类型的响应时向开发人员发出警报。无论是现在还是将来，知识都很难控制。因此，应该优先关注并加强核酸等物理材料的防护措施。同时，应积极支持和发展那些能够帮助合成公司降低生物设计工具风险的模型、支持筛查的政策，以及识别和评估其他物理领域可能存在的风险。

4. 建议

鉴于该主题的复杂性、这些技术的快速发展以及潜在风险和收益的国际性质，本报告建议**建立一个定期召开的国际论坛或倡议**。

- **识别与工程生物学和人工智能融合相关的新兴安保问题；**
- **开展地平线扫描，预测未来几年工程生物学和人工智能的发展轨迹；**
- **考虑在世界各社区和地区之间如何有差异地承担相关风险；**
- **制定适当的指导方针、政策和/或最佳实践，以及实施这些准则需要配套的工具和策略，以防止滥用；**
- **随着时间的推移，评估和迭代已实施的安保实践。**

论坛应定期组织和举行会议，更新实现特定能力的预期时间表，加强合作。

随着生物风险空间的发展，制定最佳安保实践，并评估既往政策和实践。论坛应由私营部门和/或学术界主办，并得到各国政府、世界经济论坛（WEF）或经济合作与发展组织（OECD）等国际组织的支持和参与。

此类倡议的利益相关者至少要包括人工智能和工程生物学学术界和行业界的成员、（生物）安保专家和政府合作伙伴。此外，在可能的情况下，应纳入社区代表。学术界和行业利益相关者可能致力于开发和传播佳实践。政策制定者应积极参与，以便各国制定的框架、监管结构、标准和政府指导能够与技术能力保持一致。这种政府框架还需要构建一个协同、互补的体系，强化各部门间的沟通与协作，以消除分裂、减少混乱和堵塞可能被利用的漏洞。

5. 结论

政策制定者、研究人员和其他利益相关者都非常关注人工智能，尤其是当其与工程生物学等技术的融合。识别这些技术交叉融合中潜在的安保问题非常重要。本报告提出，从头开始生物设计、闭环自主实验室和自然语言大型语言模型在人工智能和工程生物学交叉领域可能存在潜在安保隐患。尽管如此，就每项技术目前和未来对恶意行为者的影响程度，以及适当的预防和缓解战略达成共识仍然具有挑战性。未来，各利益相关者需要共同探讨，抓住机会，预防、阻止和减轻人工智能在工程生物学中的滥用行为。同时，利益相关者还需权衡过度限制可能带来的其他损失，并最终携手确定并实施合理的保障措施，最大限度地降低滥用风险。

刘晓 张学博 编译自 EBRC