生成式人工智能模型治理框架

新加坡资讯通信媒体发展局 (IMDA) / AI Verify Foundation

2024年6月

中文翻译版 · 仅供参考，以英文原文为准

原文来源：https://aiverifyfoundation.sg/wp-content/uploads/2024/06/Model-AI-Governance-Framework-for-Generative-AI-19-June-2024.pdf

---

目录

执行摘要

生成式AI（Generative AI）已引发全球的关注与想象。虽然它具有巨大的变革潜力，但也伴随着风险。因此，构建一个可信的生态系统至关重要——它有助于人们放心地拥抱AI，为创新提供最大空间，并作为利用AI造福公众的核心基础。

AI作为整体技术已经发展多年。此前的开发与部署有时被称为传统AI（Traditional AI）。为了奠定促进传统AI负责任使用的基础，新加坡于2019年发布了首版《AI模型治理框架》，并于2020年进行了更新。生成式AI的最新出现强化了一些相同的AI风险（如偏差、滥用、缺乏可解释性），并引入了新的风险（如幻觉、版权侵权、价值对齐）。这些关注在我们此前于2023年6月发布的《生成式AI：对信任与治理的影响》讨论文件中已有强调。相关讨论和反馈具有重要参考价值。

现有的治理框架需要审视，以培育更广泛的可信生态系统。需要在保护用户和推动创新之间取得谨慎的平衡。国际上也围绕问责制、版权和虚假信息等相关且重要的议题展开了各种讨论。这些问题相互关联，需要以务实和整体化的方式来审视。没有任何单一干预措施能够成为万能良方。

因此，**本《生成式AI模型治理框架》旨在提出系统且平衡的方法，在继续促进创新的同时应对生成式AI的关注**。 它要求所有关键利益相关者——包括政策制定者、行业、研究界和广大公众——共同尽其职责。该框架提出了九个维度，需要整体审视以培育可信的生态系统。

a) 问责制（Accountability） ——问责制是激励AI开发链上各方对终端用户负责的关键考量。在此过程中，我们认识到生成式AI如同大多数软件开发一样，涉及技术栈中的多个层级，因此责任分配可能并不立即明确。虽然生成式AI开发有其独特特征，但仍可以与当今的云计算和软件开发栈进行有益的类比，并采取初步的实际步骤。

b) 数据（Data） ——数据是模型开发的核心要素。它对模型输出质量有重大影响。因此，输入模型的内容很重要，需要确保数据质量，例如使用可信的数据来源。在模型训练数据的使用可能存在争议的情况下（如个人数据和版权材料），给予商业明确性、确保公平对待并以务实的方式处理也很重要。

c) 可信的开发与部署（Trusted Development and Deployment） ——模型开发及其上层的应用部署是AI驱动创新的核心。尽管终端用户可能看不到这些，但围绕基线安全和卫生措施的有意义的透明度至关重要。这涉及行业采用开发、评估方面的最佳实践，以及随后的"食品标签"式透明度和披露。这可以逐步提升更广泛的意识和安全性。

d) 事件报告（Incident Reporting） ——即使拥有最健全的开发流程和保障措施，我们今天使用的软件也不是万无一失的。AI同样如此。事件报告是一种成熟的做法，允许及时通知和补救。因此，建立事件监控和报告的结构和流程至关重要。这也支持AI系统的持续改进。

e) 测试与保障（Testing and Assurance） ——对于可信的生态系统，第三方测试与保障发挥着互补作用。我们今天在许多领域（如金融和医疗保健）都这样做，以实现独立验证。虽然AI测试是一个新兴领域，但企业采用第三方测试与保障来向终端用户展示信任是有价值的。制定AI测试的通用标准以确保质量和一致性也很重要。

f) 安全（Security） ——生成式AI引入了针对模型本身的新型威胁向量的可能性。这超越了任何软件栈固有的安全风险。虽然这是一个新兴领域，但现有的信息安全框架需要适应调整，并开发新的测试工具来应对这些风险。

g) 内容溯源（Content Provenance） ——AI生成的内容由于其创建的便利性，可能加剧虚假信息。关于内容生成的位置和方式的透明度使终端用户能够以知情的方式决定如何消费在线内容。各国政府正在关注数字水印和密码学溯源等技术解决方案。这些技术需要在正确的上下文中使用。

h) 安全与对齐研发（Safety and Alignment R&D） ——当今模型安全的科学水平尚未完全覆盖所有风险。需要加速投资研发以改善模型与人类意图和价值观的对齐。全球AI安全研发机构之间的合作将是关键，以优化有限资源实现最大影响，并跟上模型能力的商业化增长步伐。

i) AI造福公众（AI for Public Good） ——负责任的AI不仅仅是风险缓解。它也关乎提升和赋能我们的民众和企业在AI驱动的未来中蓬勃发展。民主化AI访问、改善公共部门AI采用、提升劳动者技能以及可持续地开发AI系统将支持将AI引向公众利益的努力。

本框架建立在我们的《生成式AI讨论文件》中强调的政策理念之上，并借鉴了与主要司法管辖区、国际组织、研究界和领先AI组织的洞察和讨论。该框架将随着技术和政策讨论的发展而演进。

培育可信的AI生态系统

1. 问责制（Accountability）

问责制是培育可信生态系统的关键考量。AI开发链上的各方需要对终端用户负责，结构性激励应与这一需求保持一致。这些参与者包括模型开发者、应用部署者和云服务提供商（通常提供AI应用托管的平台）。生成式AI如同大多数软件开发一样，涉及技术栈中的多个层级。虽然责任分配可能不立即明确，但可以与当今的云计算和软件开发进行有益的类比，并采取实际步骤。

> 注：虽然本框架侧重于分配AI开发的责任，但终端用户在AI使用方面也有独立的责任（如遵守使用条款）。
>
> 注：我们认识到生成式AI开发链是复杂的。应用开发者（开发利用AI技术的解决方案或应用程序的一方）和应用部署者（向终端用户提供AI解决方案或应用程序的一方）有时可能是不同的主体。为简便起见，本文使用"应用部署者"一词来指代应用开发者和部署者。

设计

要全面做到这一点，应考虑如何在开发过程中预先（事前，ex-ante）分配责任作为最佳实践，并在事后发现问题时如何获得补救（事后，ex-post）提供指导。

事前——预先分配

责任可以根据生成式AI开发链中每个利益相关者拥有的控制程度来分配，以确保有能力的一方采取必要行动来保护终端用户。作为参考，虽然开发链中可能涉及各种利益相关者，但云计算行业已经随时间建立并编纂了全面的共享责任模型。其目标是确保云环境的整体安全。这些模型通过说明云服务提供商（提供基础设施层的一方）和其客户（在上层托管应用程序的一方）各自承担的控制和措施来分配责任。

将这种方法扩展到AI开发是有价值的。云服务提供商最近已将其云共享责任模型的某些元素扩展到涵盖AI，最初侧重于安全控制。这是一个好的开始，类似的方法可以用于解决其他安全问题。AI共享责任方法可能还需要考虑不同的模型类型（如闭源、开源或开放权重），因为应用部署者对每种模型类型具有不同程度的控制权。在这种情况下，例如在使用开源或开放权重模型时，责任应要求应用部署者从信誉良好的平台下载模型，以最小化篡改模型的风险。作为对自身模型及

其部署方式最了解的一方，模型开发者适合以协调的方式引领这一发展。这将为利益相关者提供更大的前置确定性，并培育更安全的生态系统。

事后——安全网

共享责任模型作为问责制的重要基础——当问题发生时提供补救的明确性。然而，它们可能无法涵盖所有可能的场景。在出现新的或未预期的问题时分配责任也可能在实践中具有挑战性。值得考虑额外措施——包括关于赔偿和保险的概念——以更好地保障终端用户。

这在今天以有限的形式存在。在较为明确的需要补救的领域，行业已做出相应举动。一些模型开发者（如Adobe、Anthropic、Google、Microsoft和OpenAI）已开始为某些风险提供担保，例如因使用其AI产品和服务引发的第三方版权索赔。这样做时，开发者隐含地承认了他们对模型训练数据及其模型使用方式的责任。

不可避免地会有其他不太明确且未得到充分覆盖的领域。这可能包括对社会整体产生不成比例影响的风险，以及可能仅在AI使用过程中才会出现的风险。因此，考虑更新法律框架使其更加灵活，并允许新兴风险能够被轻松且公平地解决是有用的。这类似于今天实物产品的终端用户享有的安全保护。此类努力的一个例子是欧盟拟议的《AI责任指令》和即将批准的《修订产品责任指令》。这些指令旨在使终端用户更容易证明AI产品和服务造成的损害。这确保没有任何一方在赔偿过程中受到不公平对待。

最后，总会有一些遗留问题无法被覆盖。这是一个非常新兴的讨论，替代解决方案如**无过错保险（No-fault Insurance）** 可以作为安全网加以考虑。

---

2. 数据（Data）

数据是模型和应用开发的核心要素。训练稳健可靠的AI模型需要大量数据语料库。鉴于其重要性，企业需要在模型开发中如何使用数据方面获得明确性和确定性。这包括潜在争议领域，如公开可用的个人数据和版权材料，这些通常包含在网络爬取的数据集中。在此类情况下，认识到相互竞争的关切、确保公平对待并以务实的方式处理是很重要的。此外，良好地开发模型需要高质量的数据，在某些情况下还需要有代表性的数据。确保可用数据集的完整性也很重要。

> 注：数据投毒（Data Poisoning）通过引入、修改或删除特定数据点来攻击训练数据集。例如，了解模型开发者从Wikipedia等来源收集内容的确切时间后，不良行为者

> 可以用虚假内容"投毒"Wikipedia网页，这些内容将被爬取并用于训练生成式AI模型。

设计

个人数据的可信使用

由于个人数据在现有法律制度下运作，政策制定者阐明现有个人数据法律如何适用于生成式AI是一个有用的起点。这将促进在仍然保护个人权利的方式下使用个人数据。例如，政策制定者和监管机构可以澄清同意要求或适用的例外情况，并为AI中数据使用的良好商业实践提供指导。

一组新兴技术——统称为**隐私增强技术（Privacy Enhancing Technologies，PETs）**——有潜力允许数据在保护数据机密性和隐私的同时用于AI模型的开发。一些PETs如匿名化技术并不新鲜，而其他技术仍在发展中。理解PETs如何应用于AI将是一个重要的推进领域。

在版权与数据可访问性之间取得平衡

从模型开发的角度看，在训练数据集中使用版权材料以及版权所有者的同意问题开始引发关注，特别是在报酬和许可方面以促进此类使用。模型也越来越多地被用于生成创意输出——其中一些模仿现有创作者的风格，引发了这是否构成合理使用（Fair Use）的考量。

鉴于AI训练涉及的大量数据，制定方法来以明确和高效的方式解决这些困难问题是有价值的。今天，法律框架尚未围绕此类方法达成共识。一些版权所有者已在美国和英国法院对生成式AI公司提起诉讼。各国也在探索非立法解决方案，如版权指南和面向开发者和终端用户的行为准则。

鉴于各方利益攸关，政策制定者应促进所有相关利益相关者之间的公开对话，以了解快速演变的生成式AI技术的影响，并确保潜在解决方案是平衡的且符合市场现实。

促进高质量数据的可获取性

作为组织层面的整体卫生措施，AI开发者采取数据质量控制措施和数据治理方面的一般最佳实践是一种良好的自律，包括一致且准确地标注训练数据集，以及使用数据分析工具促进数据清洗（如去偏差和去除不当内容）。

在全球层面，值得考虑协调努力来扩大可用的可信数据集池。参考数据集在AI模型开发（如用于微调）以及基准测试和评估中都是重要工具。政府也可以考虑与当地社区合作，为其特定上下文（如低资源语言）策划一个有代表性的训练数据集库。这有助于提高反映一国文化和社会多样性的优质数据集的可用性，从而支持开发更安全、更具文化代表性的模型。

3. 可信的开发与部署（Trusted Development and Deployment）

模型开发及其上层的应用部署是AI驱动创新的核心。然而，目前对于确保可信模型所采取的方法缺乏信息。即使在"开源"模型的情况下，一些重要信息如方法论和数据集可能也不会公开。

未来，行业围绕开发和安全评估的最佳实践达成共识是很重要的。此后，围绕基线安全和卫生措施的有意义的透明度也将是关键。这将使AI生态系统中的所有利益相关者能够更安全地使用模型。此类透明度需要与合法考量（如保护商业和专有信息以及不让不良行为者钻空子）之间取得平衡。

设计

安全最佳实践需要由模型开发者和应用部署者在AI开发生命周期中围绕开发、披露和评估来实施。2020年版《AI模型治理框架》已为此奠定基础，该框架为传统AI解决方案的开发和部署制定了最佳实践。其中阐述的原则仍然具有相关性，并在此针对生成式AI进行扩展。

开发——基线安全实践

安全措施正在快速发展，模型开发者和应用部署者最适合决定使用什么。即便如此，行业实践开始围绕一些共同的安全实践达成共识。

例如，在预训练之后，**人类反馈强化学习（Reinforcement Learning from Human Feedback, RLHF）** 等微调技术可以引导模型生成与人类偏好和价值观更加一致的、更安全的输出。安全方面的关键步骤还包括考虑用例的上下文并进行风险评估。例如，进一步微调或使用用户交互技术（如输入和输出过滤器）可以帮助减少有害输出。**检索增强生成（Retrieval-Augmented Generation, RAG）** 和少样本学习（Few-shot Learning）等技术也常被用于减少幻觉并提高准确性。

披露——"食品标签"

围绕这些构成AI模型核心构成的安全措施的透明度随后至关重要。这类似于"食品或成分标签"。通过向下游用户提供相关信息，他们可以做出更明智的决策。虽然领先的模型开发者已经披露了一些信息，但标准化披露将促进模型之间的可比性并推动更安全的模型使用。相关的披露领域可能包括：

a) **使用的数据**：训练数据来源类型及训练前数据处理方式的概述。

b) **训练基础设施**：使用的训练基础设施概述，以及在可能的情况下估计的环境影响。

c) **评估结果**：已完成的评估及关键结果的概述。

d) **缓解和安全措施**：已实施的安全措施（如偏差纠正技术和防止敏感数据泄露的保障措施）。

e) **风险和局限性**：模型的已知风险及解决这些风险的举措。

f) **预期用途**：明确说明模型预期用途范围的声明。

g) **用户数据保护**：关于用户数据将如何使用和保护的概述。

此类披露为所有模型提供了标准基线。定制化或高级模型的开发者可以考虑披露额外信息。

披露的详细程度可以根据透明需求与保护专有信息之间的平衡进行调整。向前迈进的一步是行业就作为一般披露的一部分向所有方提供的基线透明度达成一致。这涉及模型开发者和应用部署者双方。或者，此基线的制定可以由政府和第三方推动。

对于可能构成高风险的模型——如具有国家安全或社会影响的高级模型——也需要向政府提供更高的透明度。因此，政策制定者有空间定义模型风险阈值，超过该阈值将适用额外的监督措施。

评估

当今评估生成式AI主要有两种方法：（i）**基准测试（Benchmarking）**——通过问答数据集测试模型以评估性能和安全性；（ii）**红队测试（Red Teaming）**——红队作为对抗性用户"攻破"模型，诱导安全、安全及其他违规行为。

虽然基准测试和红队测试目前被广泛采用，但在提供模型性能和安全性的稳健评估方面仍远远不够（参见"安全与对齐研发"维度）。

即使在基准测试和红队测试框架内，今天的大多数评估集中在生成式AI的前端性能上，而较少关注其后端安全性。评估工具也存在不足（如针对多模态模型的工具），以及对危险能力的测试缺乏。另一个问题是一致性——许多测试和评估需要根据特定模型定制，有时可比性是一个挑战。

因此，有必要朝着更全面和系统化的安全评估方法努力。这将产生更有用和可比的洞察。为提供额外的保障，标准化方法还可以包括定义一套基线必需的安全测试和开发共享资源，并与政策制定者协商。

> **标准化安全评估的起点**
>
> AI Verify Foundation和IMDA推荐了一套初始的LLM标准化模型安全评估，涵盖鲁棒性、事实性、偏差倾向、毒性生成和数据治理。详见2023年10月发布的《LLM评估汇编》（Cataloguing LLM Evaluations）论文。该论文提供了全景扫描和关于可

> 考虑哪些安全评估的实用指导。鉴于生成式AI领域的快速进展，这些建议需要持续改进。

行业和领域可能有需要额外评估的独特需求（如对高风险用例如医疗诊断规定严格的准确性阈值）。此外，应用部署者更可能专注于针对其用例的领域特定评估。在某些情况下，如具有非常小众能力的模型，可能需要定制化评估。行业和行业政策制定者因此需要共同改进评估基准和工具，同时保持基线和行业特定要求之间的一致性。

---

4. 事件报告（Incident Reporting）

即使拥有最健全的开发流程和保障措施，我们今天使用的软件也不是万无一失的。AI同样如此。事件报告是一种成熟的做法，包括在电信、金融和网络安全等关键领域。它允许及时通知和补救。因此，建立事件报告的结构和流程至关重要。这反过来通过洞察、补救和修补支持AI系统的持续改进。

设计

漏洞报告——预防性行动的激励

在事件发生之前，软件产品所有者将漏洞报告作为整体主动安全方法的一部分。他们邀请和支持白帽黑客或独立研究人员发现其软件中的漏洞，有时通过策划的漏洞赏金计划。一旦发现，漏洞被报告，产品所有者随后有时间（通常为90天，基于行业惯例）来修补其软件、发布漏洞（如通过提交CVE）并感谢白帽黑客或独立研究人员。这允许软件产品所有者和用户采取主动措施来增强整体安全性。

> **通用漏洞和暴露（Common Vulnerabilities and Exposures, CVE）计划**
>
> CVE计划由MITRE公司管理，编纂了一份公开已知的安全漏洞和暴露列表。全球网络安全团队广泛参考此列表以查找可能影响其组织的新漏洞。软件产品所有者可将漏洞提交为CVE。发现零日CVE的能力在白帽社区中也被视为一项成就。

AI开发者可以应用类似的概念，为其AI系统中发现的安全漏洞提供报告渠道。他们可以应用与漏洞报告相同的最佳实践，包括评估事件、修补和发布的时间窗口。这也应辅以持续的监控努力，在终端用户注意到之前检测故障。

事件报告

在事件发生后，组织需要内部流程来报告事件以实现及时通知和补救。根据事件的影响和AI参与的程度，这可能包括通知公众和政府。因此，定义"严重AI事件"或设定正式报告的重要性阈值至关重要。AI事件也可能涉及广泛领域。因此原则需要与现有报告制度的原则相协调。借鉴网络安全，AI事件可以报告给相当于"信息共享与分析中心"（Information Sharing and Analysis Centres）的机构，这些是促进信息共享和良好实践的受信任实体，以及根据法律要求向相关当局报告。

报告应是适度的，这意味着在全面报告和实用性之间取得平衡。这需要根据具体的当地情况进行调整。在此方面，欧盟《AI法案》为法律报告要求提供了一个参考点。

> **欧盟《AI法案》下的事件报告**
>
> 高风险AI系统的提供者被要求在AI系统提供者得知事件后15天内，向事件发生所在成员国的市场监督当局报告严重事件。"严重事件"被定义为任何直接或间接导致人员死亡、严重损害人员健康、严重且不可逆地破坏关键基础设施、违反欧盟法律下基本权利，或严重损害财产或环境的AI系统事件或故障。

---

5. 测试与保障（Testing and Assurance）

在可信的生态系统中，第三方测试和保障通常发挥互补作用。我们今天在许多领域（如金融和医疗保健）都这样做，以实现独立验证。虽然公司通常进行审计以证明合规，但越来越多的公司开始将外部审计视为提供透明度并与终端用户建立更大可信度和信任的有用机制。

虽然这是一个新兴领域，但我们可以借鉴成熟的审计实践来发展AI第三方测试生态系统。第三方测试还将受益于围绕AI评估的全面且一致的标准（在"可信的开发与部署"维度中已有讨论）。

设计

培育第三方测试生态系统的发展涉及两个关键方面：

a) 如何测试（How to Test）——标准化

在短期内，第三方测试将包括开发者自身使用的同一套基准和评估。最终，这需要以标准化的方式进行，第三方测试才能有效，并促进模型之间的有意义可比性。

因此，应更加重视设定共同基准和方法论。这可以通过提供共同工具来减少跨不同模型或应用测试所需的摩擦来推动。此后，对于更成熟的领域，AI测试可以通过ISO/IEC和IEEE等标准组织进行编纂，以支持更协调和稳健的第三方测试。

随着测试生态系统的发展，也有空间标准化第三方测试的范围。

b) 谁来测试（Who to Test）——可信的认证

独立性是确保测试结果客观性和完整性的关键。建立一批合格的第三方测试人员至关重要。行业机构和政府的协调努力对于发展该领域的能力将是有益的。最终，可以开发认证机制以确保独立性和能力。这在许多领域（如金融）是常见做法。许多审计和专业服务公司理所当然地越来越有兴趣发展初步的AI审计能力和服务。

---

6. 安全（Security）

生成式AI使人们重新关注AI本身的安全。许多问题是熟悉的，例如AI/ML中间件中的供应链风险。在解决AI安全问题时，将通过当前方法解决的传统软件安全问题与针对AI模型本身的新型威胁向量区分开来是有用的。后者是一个新兴领域。尽管如此，类似的安全概念可能仍然适用。

设计

适应"安全设计"（Security-by-Design）

安全设计是一个基本的安全概念。它旨在通过将安全融入系统开发生命周期（SDLC）的每个阶段来最小化系统漏洞并减少攻击面。关键的SDLC阶段包括开发、评估、运维和维护。

然而，鉴于生成式AI的独特特征，可能需要进行改进。例如，将自然语言作为输入注入的能力可能对设计适当的安全控制构成挑战。此外，生成式AI的概率性质挑战了在SDLC中用于系统改进和风险缓解的传统评估技术。因此，需要为生成式AI开发或调整新概念。

开发新的安全保障措施

需要开发新工具，可能包括：

a) **输入过滤器（Input Filters）**：输入审核工具检测不安全的提示（如阻止恶意代码）。这些工具需要针对领域特定风险进行定制。

b) **生成式AI数字取证工具（Digital Forensics Tools for Generative AI）**：数字取证工具用于调查和分析数字数据（如文件内容）以重建网络安全事件。应探索新的取证工具，以帮助增强识别和提取可能隐藏在生成式AI模型中的恶意代码的能力。

除了这些工具外，MITRE的《AI系统对抗性威胁格局》（Adversarial Threat Landscape for AI Systems）等数据库提供了关于ML系统（包括生成式AI）的对手战术、技术和案例研究的信息。AI开发者可以使用这些来支持风险评估和威胁建模，并识别有用的工具或流程。

---

7. 内容溯源（Content Provenance）

生成式AI的兴起——能够大规模快速创建逼真的合成内容——使消费者更难区分AI生成的内容和原始内容。此类关注的常见表现形式是深度伪造（Deepfakes）。这加剧了虚假信息等危害，甚至可能威胁选举的公正性等社会安全。

各国政府、行业和社会普遍认识到需要技术解决方案——如数字水印和密码学溯源——来跟上AI生成内容的速度和规模。数字水印和密码学溯源都旨在标记并提供额外信息，用于标识由AI创建或修改的内容。

> **数字水印（Digital Watermarking）** 技术将信息嵌入内容中，可用于识别AI生成的内容。今天有多种数字水印解决方案来标记AI生成的内容（如Google DeepMind的SynthID和Meta的Stable Signature）。然而，目前只能通过编码水印的同一公司来解码水印，这是因为缺乏可互操作的标准。
>
> **密码学溯源（Cryptographic Provenance）** 解决方案跟踪和验证数字内容的来源及所做的任何编辑，记录受到密码学保护。内容溯源和真实性联盟（Coalition for Content Provenance and Authenticity, C2PA）正在推动开发一个开放标准，以实现内容溯源的跟踪。

技术解决方案本身可能不够充分，很可能需要执法机制来补充。

设计

政策需要精心设计，以在正确的上下文中实现实际使用。实际上，短期内可能无法让所有内容创建、编辑或显示工具都包含这些技术。溯源信息也可能被删除。此外，消费者对这些工具的了解程度较低。恶意行为者也会找到规避这些工具的方法，更糟的是，利用它们来制造虚假的真实感。

因此，有必要与内容生命周期中的关键方合作，例如与出版商合作支持数字水印和溯源详情的嵌入和展示。由于大多数数字内容通过社交媒体平台、浏览器或媒体机构消费，出版商的支持对于向终端用户提供跨多渠道验证内容真实性的能力至关重要。还需要确保正确和安全的实施，以防止不良行为者以任何方式利用它。

不同类型的编辑（例如图像是完全由AI生成的还是仅有一小部分是AI生成的）将影响终端用户对内容的感知。为改善终端用户体验并使消费者能够区分非AI内容和AI生成内容，标准化需要标记的编辑类型将是有帮助的。

终端用户需要更深入地了解内容生命周期中的内容溯源，并学习利用工具来验证真实性。关键利益相关者（如内容创作者、出版商、解决方案提供商）可以与政策制定者合作提升认识。展示的溯源详情也应尽可能简化，以便终端用户理解。

---

8. 安全与对齐研发（Safety and Alignment R&D）

当今的安全技术和评估工具尚不能完全解决所有潜在风险。例如，即使是目前价值对齐的主要方法RLHF也有其局限性。现有的大型模型也缺乏可解释性，且可能无法始终如一地复现。鉴于模型进步的速度，有必要确保人类对齐和控制生成式AI的能力能够跟上潜在风险的步伐，包括当前风险（如偏差、幻觉）和未来灾难性风险。

设计

虽然呼吁加大研发投入是一个不会后悔的举措，但可能存在一些实际步骤来加速新研发洞察的转化和使用。例如，需要理解和系统地映射安全与对齐领域中出现的多样化研究方向和方法——并以协调的方式应用它们。

a) 一个广泛的研究领域涉及开发更加对齐的模型（也被一些人称为"正向对齐"，Forward Alignment），例如通过**AI反馈强化学习（Reinforcement Learning from AI Feedback,**

RLAIF）。RLAIF旨在通过增强反馈效率和质量以及实现对高级模型的可扩展监督来改进RLHF。然而，它也有其自身的缺陷。

b) 另一个研究领域是在模型训练后对其进行评估，以验证其对齐（也被一些人称为"反向对齐"，Backward Alignment）。这包括测试涌现能力（Emergent Capabilities），以便及早发现潜在的危险能力，如自主复制和长期规划。**机械可解释性（Mechanistic Interpretability）**——旨在理解模型的神经网络以找到问题行为的来源——也作为一个研究领域越来越受到关注。

为了跟上模型能力的进步，模型安全和对齐的研发需要加速。目前，大多数对齐研究由AI公司进行。因此，英国、美国、日本和新加坡等地设立AI安全研发机构或同等机构是一个积极的发展，表明承诺利用现有的研发生态系统并投入额外资源（可能包括算力和模型访问权限）来推动面向全球福祉的研究。

然而，全球合作将是关键，以优化有限的人才和资源实现最大影响。可以根据格局映射共同确定和优先排序有影响力的研究领域。目标是使更有影响力的研发工作能够提前开发安全和评估机制。

---

9. AI造福公众（AI for Public Good）

生成式AI的变革潜力是巨大的。如果我们的方法正确，全球社区将收获指数级的效益。当务之急是为发达国家和发展中国家加速增长和生产力，同时以AI潜在的民主化力量赋能全球的个人和企业。在此方面，各国必须走到一起相互支持，特别是通过国际和区域组织。这对于发展中国家和小型国家尤其重要，可通过联合国数字小国论坛（Digital Forum of Small States, Digital FOSS）和东南亚国家联盟（ASEAN）等关键平台来实现。目标是建立一个全球数字共同体（Digital Commons）——一个拥有共同规则和所有公民平等机会的场所，无论其地理位置如何。

设计

AI可以在四个具体领域产生有益和长远的效果。

民主化技术访问（Democratising Access to Technology）

所有社会成员都应以可信的方式获得生成式AI的访问。生成式AI由于其自然语言的特性本质上是直观的，但确保整体产品（生成式AI只是其中一个组件）以人为中心的方式设计仍然很重要。世界上大多数公民可能不了解其所使用应用程序背后的技术和"黑箱"。因此，设计应用程序以引出预期的社会和人类成果是关键。

为更广泛地支持这一点，政府可以与公司和社区合作开展数字素养计划，鼓励安全和负责任的AI使用。主题可以包括教育终端用户如何安全使用聊天机器人、提高他们对AI"拟人化"的警觉性，以及识别深度伪造。

生成式AI的采用也可能具有挑战性，特别是对于中小企业（SMEs）。政府和行业伙伴可以提高认知并提供支持，推动中小企业的创新和AI使用。一个例子是新加坡的生成式AI沙箱，为中小企业提供生成式AI企业解决方案的工具和培训。

公共服务交付（Public Service Delivery）

AI应以有影响力的方式服务公众。今天，AI为许多公共服务提供动力，如学校中的自适应学习系统和医院中的健康管理系统。这释放了新的价值主张，创造了效率并改善了用户体验。

政府协调资源以支持公共部门AI采用是可取的。这包括促进不同政府机构之间的负责任数据共享、高性能计算访问和其他相关政策。AI开发者通过帮助政府识别用例和提供解决公民痛点的AI解决方案发挥贡献作用。

劳动力（Workforce）

要释放AI的生产价值，劳动力的协调技能提升至关重要。这是对抗技术替代劳动力可能产生的负面结果的关键。除了使用AI工具的特定技能集外，创造力、批判性思维和复杂问题解决等核心技能对于帮助人们有效利用AI也很重要。

行业、政府和教育机构可以共同合作重新设计工作并为劳动者提供技能提升机会。随着组织采用企业生成式AI解决方案，它们也可以为员工开发专门的培训计划。这将使员工能够应对工作中的转变并享受由工作变革带来的效益。

可持续发展（Sustainability）

可持续增长至关重要。生成式AI的资源需求（如能源和水）并非微不足道，很可能影响可持续发展目标。因此，生成式AI生态系统中的利益相关者需要共同合作，开发适当的技术（如节能计算）来支持我们的气候责任。

为了指导此类计划，生成式AI的碳足迹（如模型训练和推理）也需要被追踪和衡量。AI开发者和设备制造商更适合进行绿色计算技术的研发并采用节能硬件。此外，AI工作负载可以托管在推行一流节能实践、使用绿色能源或能源转换路径的数据中心。

结语

随着生成式AI继续发展和演进，全球在政策方法上的合作是必要的。本框架中的九个维度为全球对话提供了基础，以在最大化创新空间的同时应对生成式AI的关注。所提出的理念也进一步推进了问责制、透明性、公平性、鲁棒性和安全性的核心原则。它们重申了政策制定者与行业、研究人员和志同道合的司法管辖区合作的必要性。我们希望这能作为迈向发展可信AI生态系统的下一步，在这个生态系统中，AI被用于公众利益，人们安全、自信地拥抱AI。

致谢

我们对以下个人和组织为《生成式AI模型治理框架》提供的宝贵反馈表示衷心感谢（按字母顺序排列）：

行业

Aimodels.org、Adobe、Alteryx、Amazon Web Services、Apollo Research、BIGO Technology、Braithwate、Cisco、Data Protection Schemes Limited、Ernst & Young、Google、HM、IBM、Kaspersky、KPMG、Lazada、LexisNexis、LLMware.ai、Mastercard、Mediacorp、Meta、Microsoft、MNT3、NovaSync Labs、NCS、OpenAI、Resaro、Salesforce、SAP、Singapore Airlines、SymphonyAI、Telenor Group、Temasek、TT-Logic、Visa、Workday、Zühlke

政府、研究机构和协会

A*STAR、Academy of Medicine Singapore、AI Professionals Association、AI 2030、APAC Gates、Asia Internet Coalition、Asia Securities Industry and Financial Markets Association、Association of Chartered Certified Accountants、BSA | The Software Alliance、Centre for AI and Digital Policy、Chartered Software Developer Association、Computer & Communications Industry Association、Copyright Licensing and Administration Society、Department of Commerce (United States)、Digital Prosperity for Asia Coalition、Digital Trust Centre、Future of Privacy Forum、International Federation of the Phonographic Industry、Ministry of Health of Singapore、Motion Picture Association、Pragmagility、Recording Industry Association Singapore、Rhythmisis Institute、SHE、The American Chamber of Commerce in

Singapore、The App Association、The Dialogue、US-ASEAN Business Council、Vibrations

其他贡献者

Aidan O'Gara、Alex Toh、Jamie Bernardi、Lim Zheng Xiong、Merlin Stein、Nicholas Ni、Oliver Guest、Oscar Delaney、Pankaj Jasal、Raymond Chan、Saad Siddiqui、Shaun Ee、Simon Chesterman、Soh Teck Foo、Srikanth Mahankali、Steven David Brown、Will Hodgkins、Xavier Tan、Zaheed Kara

---

后续发展

《生成式AI模型治理框架》是迈向培育生成式AI可信生态系统的第一步。在收到反馈的基础上，通过实施指南和资源提供更大的确定性方面仍有进一步工作要做。参照框架的九个维度，我们将继续与关键利益相关者合作开发这些指南和资源，以实现一种系统且平衡的方法，在为生成式AI创新提供最大空间的同时建立护栏。

© 版权所有 IMDA 和 AI Verify Foundation 2024。保留所有权利。

关于AI Verify Foundation

认识到合作和汇集专业知识的重要性，新加坡成立了AI Verify Foundation，以利用全球开源社区的集体力量和贡献来构建AI治理测试工具。AI Verify Foundation的使命是培育和协调开发者社区，为AI测试框架、代码库、标准和最佳实践的发展做出贡献。它将建立一个中立的空间，用于思想交流和开放合作，并培养一个多元化的AI测试倡导者网络，通过教育和推广推动广泛采用。愿景是建立一个为人类更大福祉做出贡献的社区，通过实现AI的可信开发来实现。

关于IMDA

在IMDA，我们视自己为新加坡数字未来的建筑师。我们从端到端覆盖数字空间，作为政府机构独特地同时戴着三顶帽子——作为经济发展者（从企业数字化到资助研发）、作为构建可信生态系

统的监管者（从数据/AI到数字基础设施），以及作为社会平衡者（推动数字包容并确保没有人被落下）。因此，我们不是孤立地看待AI治理，而是在经济和更广泛社会的交汇点上审视。通过将三顶帽子结合起来，我们希望不仅在新加坡，而且在亚洲乃至更远的地方推动边界，在实现这一新兴和动态技术的安全和可信使用方面有所作为。