迈向可信赖的人工智能：可验证声明的支持机制[1]

执行择要

人工智能技术的最新发展使其在商业、科学与其他创新领域得到了广泛的应用。随着此波应用浪潮的涌现，人们越来越意识到人工智能系统所带来的风险，并认识到现有法律与业界、学界规范仍不足以保证人工智能的可靠研发[1] [2] [3]。

机器学习领域的研发人员和科技公司已经采取了一些措施来弥补这些规范不足，其举措包括广泛采用人工智能行业认可的道德准则。然而，道德准则缺乏法律约束力，也往往难以转化为实际行动。而且，外界人员很难评估人工智能开发者到底有是否表里如一，也没有办法让他们在违反道德原则的时候承担责任。这就导致了很多人谴责人工智能开发者在谈论道德问题时口惠而实不至[4]。人工智能开发者要想赢得系统用户、客户、政府、社会和其他利益相关方的信任，就不应该只谈原则，而要集中精力建立合理的机制来保证行为的负责性[5]。作出可检验、能追责的承诺是朝这个方向迈出的重要一步。

如果能提供精准的声明和充足的证据，人工智能开发人员就能更好地向监管机构、公众和其他开发者证明其行为的负责。如果有关人工智能开发的声明更容易被验证，就能实现更有效的政府监管并且减少开发者为获得竞争优势而偷工减料的压力[1]。相反地，如果没有能力验证开发人员的声明，用户或其他利益相关方就更有可能因模棱两可、有误导性或虚假的说法而利益受损。

本报告提出了诸多建议，意在使不同利益相关方能够更容易地检验人工智能开发者的对外声明，特别是有关其安全性、安保性、公平性和隐私性的声明。保证可信任的人工智能发展是一个多方面的问题，我们认为，执行这些机制有助于促进该目标的达成。[2]本报告中提出的机制可以帮助我们处理不同利益相关方可能面对的问题：

- 作为用户，我能否在使用新的人工智能系统机器翻译敏感文件时，检验其对隐私保护级别声明的真实性？
- 作为监管者，我能否追踪无人驾驶汽车导致事故的过程，并且知道用哪种标准来评判汽车公司的安全声明？
- 作为学者，我能否在缺乏业界计算资源的条件下，对大型人工智能系统所带来的影响进行客观的研究？
- 作为人工智能研发者，我能否确信我在某一领域的竞争对手遵循最佳实践，而不是偷工减料以获得竞争优势？

即使人工智能开发者有意愿或者需求使自己的产品声明具体而可验证，他们也可能缺乏达成这一目标的相关机制。人工智能开发社群需要一系列有效的机制，为检验人工智能系统和开发过程的声明提供支持。

---

[1] 本文中文翻译的通讯作者是谢旻希(Brian Tse)。在翻译过程中，我们得到了肖文泉(Jenny W. Xiao)的宝贵帮助。

[2] 当然，仅仅是拥有验证开发人员声明的能力还不足以确保负责任的人工智能研发。这是因为并非所有的重要声明都能被验证。而且，保证负责任的人工智能研发还需要政府和标准制定组织之类的监管机构来确保开发人员的激励机制和公众利益保持一致。

从这个角度出发，本报告的作者于2019年4月举行了一次研讨会，旨在构思促进研发者提出声明、验证声明的机制。[3]本报告以该研讨会上的讨论成果为基础，提出的机制主要致力于达成以下两个目标：

- 增加沟通渠道，便利人工智能开发者对外验证有关其系统属性的声明。

- 加强应对能力，使利益相关方（如用户、政府决策者和更广大的社会)能够对人工智能开发者提出特殊而多样的要求。

针对妨碍人工智能声明有效评估的具体问题，本报告提出了一一对应的一系列机制和建议。其中部分机制已经存在，但仍需完善，而另一部分则是前所未有的。本报告旨在为进一步增强人工智能研发声明的可验证性作出贡献。

该报告提出的机制作用于*制度*、*软件*和*硬件*三个层面。制度、软件和硬件也是人工智能系统和开发过程中相互重叠、相互影响的三大关键要素。

- **体制机制**：这些机制改变或阐明开发者面临的激励机制，并且增强其行为的能见度，以保证其研发的系统具有安全性、可靠性、公平性和隐私保护。体制机制是有效验证人工智能研发声明的基础，因为人类和人类行为将最终决定人工智能的发展方向。本报告在讨论中提出，可以利用**第三方审核**来替代自我评估声明；利用**红队测试练习（red teaming exercises)**以增强开发人员的防范意识，减少系统被误用或袭击的可能性；利用**误差和安全隐患侦查激励制度 (bias and safety bounties)**以建立激励机制，促进对人工智能系统缺陷的及时发现、及时报告；以及加强**人工智能安全事故信息共享**，以增进社会对人工智能系统的认识，理解到人工智能可能带来意外或非理想的后果。

- **软件机制**：这些机制让人工智能系统的属性更易于理解和监督。具体措施包括**审计跟踪 (audit trails)**，通过收集有关开发和部署过程的关键信息来强化高利害人工智能系统的问责制；保证**可解释性**以增进对人工智能系统特征的理解和审查；以及**隐私保护的机器学习 (privacy-preserving machine learning)**，使开发人员对隐私保护的承诺更有鲁棒性。

- **硬件机制**：与计算硬件有关的机制可以在多方面发挥关键作用，包括证实有关隐私和安全性的声明、提高组织如何使用资源的透明度、以及影响谁具有验证不同声明所必需的资源。探讨的机制包括**机器学习的硬件安全设施**以提高隐私和安全性声明的可验证性；**高精度计算资源的测量**，以提高关于计算能力使用的声明的价值和可比性；以及**为学术界提供计算资源支持**，以提高业界以外人士评估有关大型人工智能系统的声明的能力。

每种机制都提供额外的途径来检验开发者的承诺，有潜力为建立可信赖的人工智能生态作出贡献。下一页和报告末尾详细地列举了不同机制的相关建议，并且包含完整的列表。

[3] 见附录一 “Workshop and Report Writing Process”。

建议

制度机制和建议

1. 一个利益相关方的联盟应组建工作小组，研究如何建立**第三方人工智能审计机制**并为该机制提供资源。

2. 人工智能研发机构应该参与**红队测试 (red-teaming)**的练习，从而发现系统潜在的风险，并分享相关的最佳实践和应对问题的工具。

3. 人工智能开发者应试行**误差和安全隐患侦查激励制度 (bias and safety bounties)**，以建立广泛监督人工智能系统的激励机制和标准流程。

4. 人工智能开发者应该通过不同的合作渠道，分享更多**人工智能事故的信息**。

软件机制和建议

5. 标准制定机构应该和学界、业界合作，要求对安全攸关的人工智能系统实行**审计跟踪 (audit trails)**。

6. 人工智能研发和资助机构应该支持人工智能系统的**可解释性**研究，并将重点放在风险评估和监察上。

7. 人工智能开发者应开发、共享并使用**隐私保护的机器学习 (privacy-preserving machine learning)**的工具与指南，并且其中必须包括衡量性能的标准。

硬件机制和建议

8. 业界和学界应共同努力为人工智能加速器开发**硬件安全功能**，或者确立在机器学习环境中使用安全区（包括商品硬件上的“安全飞地”）的最佳实践。

9. 一个或多个人工智能实验室应该对单个项目进行**高精度计算资源的测量**，并报告其实践能否被广泛采用。

10. 政府资助机构应大幅增加对学界研究人员的**计算能力资源的资助**，以增强学术研究人员验证商业人工智能声明的能力。