人工智能行业：可信人工智能白皮书

可信人工智能发展建议

打造可信任的人工智能系统已经成为各界关注的焦点和努力方向。通过实践可信人工智能方法论，有助于提升人工智能的可信水平，让其更好的被社会大众接受。可信人工智能并非一成不变，而是伴随着人工智能技术、伦理、法律的发展，将会不断演进以适应新的发展需要。这也将对所涉及到的各类主体提出新的要求。

（一）政府层面加快推动我国人工智能监管及立法进程

构造体系化的人工智能法律监管框架。一是完善现行法律法规以适应发展需要，在《网络安全法》、《数据安全法》，以及未来将发布的《个人信息保护法》等基础上，梳理人工智能系统监管过程面临的适用问题，不断完善法律法规。二是推进新立法工作主动应对新风险，深入研究人工智能引发的新问题和新态势，及时梳理形成立法建议。

三是创新手段推进法律的落地执行，探索采用试点、沙箱等监管方式，研发智能化监管工具，不断提高监管的效率和灵活性。此外，要坚持统筹推进人工智能领域国内法治和涉外法治，积极参与多双边区域合作机制，推动国际间人工智能治理规则制定，寻求共识、弥合分歧。

（二）技术研究层面需全面做好体系化前瞻性布局可信人工智能一体化研究将是未来重要趋势。当前针对可信人工智能的研究，多是从安全、隐私、公平等单一维度展开。已有研究工作表明，安全性、公平性、可解

释性等不同要求之间存在相互协同或相互制约的关系，若仅考虑某一个方面的要求则可能会造成其他要求的冲突。因此需要针对可信人工智能构建一体化研究框架，以保持不同特征要素之间的最优动态平衡。

面向可信通用人工智能（AGI）的研究需要提前布局。目前无论是人工智能治理还是可信人工智能的工作，大多是面向弱人工智能技术及应用来进行的，通用人工智能甚至是超级智能尚未引起足够关注，而这些一旦出现将是关乎人类命运的重大事件，需要具有前瞻性的布局，如通过发展超级深度学习、量子机器学习等前沿技术探寻通用人工智能的发展路径。同时，我们也需要在探索强人工智能时开展可信相关的研究工作。

（三）企业实践层面需匹配业务发展实现敏捷可信企业拓展人工智能技术应用过程中应注重可信人工智能敏捷迭代。随着人工智能技术与不同行业的广泛融合，其应用深度与日俱增，企业所面临的可信特质要求将不断扩充，这就对企业应具备的可信实践能力提出了更高的要求。一方面应研发可信人工智能检测和监测工具，以匹配业务发展需要，并针对行业应用的独特性进行升级和迭代。

另一方面应积极与监管部门对接，主动配合参与数字沙盒、安全港、试点应用、标准合规等监管措施，构建内部和外部相协调的敏捷可信机制。

（四）行业组织层面需搭建交流合作平台打造可信生态

鼓励行业组织围绕可信人工智能领域搭建专门交流平台，号召产业各

方共同打造可信人工智能生态。可信人工智能是一项复杂的系统化工程，需要多方共同参与，应充分发挥行业组织优势，广泛吸纳各方优秀实践经验，编制可信人工智能操作指引；围绕人工智能研发管理、技术保障、产品应用等方面，建立完善可信人工智能标准体系；加快研发人工智能测评和监测能力，运用评估测试、跟踪监测等多种手段，持续推动可信人工智能在产业界落地。

关键词：人工智能 网络安全