人工智能安全发展上海倡议

为了应对人工智能安全挑战，守护智能时代人类未来，在上海举办2018世界人工智能大会之际，我们，世界人工智能大会安全高端对话专家组，在充分吸纳国内外人工智能安全共识基础上，联合国内外人工智能安全领域的专家学者和产业界同仁，共同发起“人工智能安全发展上海倡议”，内容如下：

面向未来：人工智能发展需要实现创新与安全的协同，以安全保障创新，以创新驱动安全。在确保人工智能自身安全的同时，积极将人工智能技术应用于解决人类社会的安全难题。

以人为本：国际社会应一道科学规划人工智能的发展路径，确保人工智能按照人类预期并服务人类福祉方向发展，机器自主进化和自我复制等关键进程须进行风险评估和安全监管。

责任明晰：人工智能发展应当建立完备的安全责任框架，需要针对人工智能的应用场景，不断创新法律法规和伦理规范，明确人工智能安全责任认定和分担机制。

隐私保护：人工智能发展需要保障用户的数据安全，人工智能发展不得以牺牲用户隐私为代价，需要加强数据保护立法，丰富人工智能的技术路线，不断强化人工智能应用中的用户隐私保护。

算法公正：人工智能发展应避免由于算法设计对公众产生的危害，须明确算法动机和可解释性，克服算法设计和数据收集引发的不公正影响。

透明监管：人工智能发展应当避免技术黑箱导致的安全风险，需要通过建立可审查、可回溯、可推演的监管机制，确保目标功能和技术实现的统一。

和平利用：人工智能技术应当审慎运用在军事领域，自主化武器的研发和使用须遵循严格的风险评估，避免由于人工智能技术在军事领域滥用对全球和平稳定的威胁。

开放合作：人工智能发展需要各国、各方的协同共进，应当积极在国际范围建立人工智能安全发展的规范和标准，避免技术和政策不兼容导致的安全风险。

2018世界人工智能大会安全高端对话专家组

2018年9月

Shanghai Initiative for Safe and Secure AI Development

On the occasion of the 2018 World Artificial Intelligence Conference (WAIC) held in Shanghai, the WAIC high-level security dialogue expert group draws from the international consensus on AI security and proposes the Shanghai Initiative for Safe and Secure AI Development, together with Chinese and international experts, academics and industry practitioners in the field of AI security. The initiative is as follows:

Oriented to the future. AI development must be achieved with a proper balance between innovation and security. Security provides guarantee for innovation, and innovation enhances security. While AI security is ensured, AI technologies should be applied to address security challenges confronting mankind.

People-centered. Countries need to plan, in a scientific way, the paths for AI development to ensure AI will progress as expected by mankind and deliver benefits to mankind. There must be risk assessment and security oversight on critical processes such as self-improvement and self-replication of machines.

Clearly-defined responsibilities. AI must be developed with a well-established security accountability framework. A mechanism for ascertaining and sharing AI security responsibilities should be established for different scenarios of AI application and in accordance with laws and ethical norms.

Privacy protection. AI development must not undermine user privacy and data security or come at the expense of user privacy. Laws and technology roadmaps should be improved to strengthen the protection of user privacy in AI applications.

Algorithm fairness. The harm of algorithm design to the public must be avoided in AI development. There must be clear motives and interpretability of algorithms to address the unfair biasesproduced and magnified by algorithms and datasets.

Transparent regulation. Security risks caused by black box technology must be avoided in AI development. There must be a regulation mechanism that is accountable, trackable and deducible to ensure unity between intended functions and technical realization.

Peaceful use. Prudence must be exercised when AI technologies are to be applied in military fields. Strict risk assessment must be conducted for automatic R&D and use of weapons. Efforts must be made to prevent the threat to global peace and stability caused by the abuse of AI technologies in military fields.

Open cooperation. AI development calls for concerted efforts of all countries and all sectors. Proactive efforts must be made to establish norms and standards for safe and secure AI development worldwide and prevent security risks caused by technology and policy incompatibility.