

2019 年第 3 期（总第 3 期）

国际战略与安全研究报告

INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

链接人工智能准则



清华大学战略与安全研究中心

CENTER FOR
INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

链接人工智能准则^①

曾毅 鲁恩萌 皇甫存青

【内容提要】人工智能准则决定了人工智能未来发展的社会因素与伦理道德考量,这些准则由研究机构、政府部门及业内人士共同制定。既存的人工智能准则的各个版本所考虑的内容、出发点与侧重点都各不相同,也没有一个完整的版本能够涵盖所有准则的内容。因此,本文提出链接人工智能准则,意在构建可以联结、分析各类人工智能准则的平台。笔者将明确指出不同机构提出的人工智能准则之间的共性与联系,同时也会对各个准则的特性进行探索。经分析,笔者认为人工智能未来的发展不应单单采取某一种准则,而是有必要综合不同人工智能准则,得出更为全面的框架,并且应注重各个准则之间的牵动与完善。

【关键词】人工智能,人工智能伦理道德,人工智能治理,国际合作

【作者简介】曾毅,中国科学院自动化研究所研究员、类脑智能研究中心副主任,中英人工智能伦理与治理研究中心主任,国家新一代人工智能治理专业委员会委员,北京智源人工智能研究院人工智能伦理与安全研究中心主任。

电子邮箱: yi.zeng@ia.ac.cn

^①本报告基于同名论文对相关数据及分析进行重要更新,该论文发表于2019年国际人工智能学会 AAAI 年会“人工智能安全”研讨会论文集《美国人工智能协会—安全的人工智能2019》。本报告在原论文的基础上新增了对全球22份最新人工智能准则提议(共50份)的研究与分析,新增提议均为原论文发表后提出。原文详见参考文献[Zeng 2019]

一、人工智能准则：百家争鸣

人工智能的伦理道德与社会影响引起了广泛关注，各组织机构也纷纷提出政策框架。本研究着眼于与人工智能总体治理相关的各类人工智能准则（包括指导方针、规则及倡议）。本文还将这些准则逐条列出，清晰明了，目的在于传达准则提出者对于人工智能理解、开发与利用的价值观与态度。本研究不包括技术标准等技术细节的讨论，也不包括传统的机器人原则。

基于以上考虑，笔者收集了迄今为止提出的 50 份人工智能准则，并从每一份中节选出最能表现提出者看法的文字内容（大多数情况下为准则的标题），其中还包含了原始文本中对节选部分进行解释的重要评论。所有准则均按照提出的背景分类：

学术界、非营利组织及非政府组织提出的准则：（1）《Partnership on AI：准则》（PAI，2016）。（2）《合乎伦理的设计：一般准则（第二版）》（电气电子工程师协会，2017）。（3）《阿西洛马人工智能准则》（生命未来研究所，2017）。（4）《算法透明与问责准则》（美国公共政策委员会计算机协会，2017）。（5）《日本人工智能学会伦理准则》（日本人工智能学会，2017）。（6）《人工智能系统的三条准则》（Etzioni，2017）（7）《人工智能治理准则》（未来学会，2017）（8）《人工智能政策准则》（信息技术产业协会，2017）（9）《人工智能负责任发展蒙特利尔宣言》（蒙特利尔大学，2017）（10）《人工智能伦理十大准则》（全球工会联盟，2017）（11）《IBE 在商业中使用人工智能的基本价值和准则的交互式框架》（商业道德研究所，2018）（12）《面向工作场所人工智能的 G20 框架》（加拿大国际治理创新中心，2018）（13）《和谐人工智能准则》（HAIP）（HAIP，2018）（14）《斯坦福以人为中心的人工智能倡议》（斯坦福大学，2018）（15）《人工智能通用准则》（The Public Voice，2018）（16）《人

工智能负责任发展蒙特利尔宣言》(蒙特利尔大学, 2018)(17)傅莹提出的六条人工智能准则(清华大学战略与安全研究中心, 2019)(18)《经合组织人工智能准则》(经合组织, 2019)(19)《人工智能北京共识》(北京, 2019)(20)《自主和智能系统在伦理方面的考量》(电气电子工程师协会, 2019)。

政府提出的准则:(21)《人工智能研发准则》(日本总务省, 2017)(22)《伦理准则和民主先决条件》(欧洲科学与新技术伦理小组, 2018)(23)《人工智能伦理准则》(英国上议院, 2018)(24)《人工智能利用准则草案》(日本总务省, 2018)(25)《“可信赖的人工智能”伦理准则草案》(欧盟委员会人工智能高级专家组, 2018)(26)《以人类为中心的人工智能社会准则(草案)》(日本内阁, 2018)(27)《迪拜人工智能准则》(智慧迪拜, 2019)(28)《人工智能治理框架模型: 指导准则》(新加坡个人资料保护委员会, 2019)(29)《“可信赖的人工智能”的关键要求》(欧盟委员会, 2019)(30)《G20人工智能准则》(G20, 2019)(31)《新一代人工智能治理准则——发展负责任的人工智能》(中华人民共和国科学技术部新一代人工智能发展研究中心, 2019)。

业界人士提出的准则:(32)《微软 CEO 的人工智能十大准则》(Nadella, 2016)(33)《认知时代的准则》(IBM, 2017)(34)《伦理道德规范: 人工智能商用开发的五大核心准则》(Sage, 2017)(35)《DeepMind 伦理与社会准则》(DeepMind, 2017)(36)《人工智能公共政策准则》(英特尔, 2017)(37)《微软人工智能准则》(微软, 2018)(38)《OpenAI 纲领》(OpenAI, 2018)(39)《德国电信的人工智能指导方针》(德国电信, 2018)(40)《人工智能伦理道德的四大准则》(百度, 2018)(41)《信任与透明准则》(IBM, 2018a)(42)《谷歌人工智能: 我们的准则》(谷歌, 2018b)(43)《人工智能的日常伦理: 五个重点关注的领域》(IBM, 2018)(44)《SAP

的人工智能指导准则》（SAP，2018）（45）《人工智能的可用、可靠、可知、可控》（腾讯，2018）（46）《索尼集团人工智能伦理道德指导方针》（索尼，2018）（47）《GE 医疗的人工智能准则》（GE 医疗，2018）（48）《Telefónica 的人工智能准则》（Telefonica，2018）（49）《Unity 的人工智能伦理指导准则》（Unity，2018 年）（50）《人工智能行业自律公约（征求意见稿）》（中国人工智能机器人产业联盟，2019）。

二、人工智能准则的语义关联

本文致力于将各类准则的共性链接起来。由于遣词造句可能不同，因此应考虑语义对等或类似的用语。

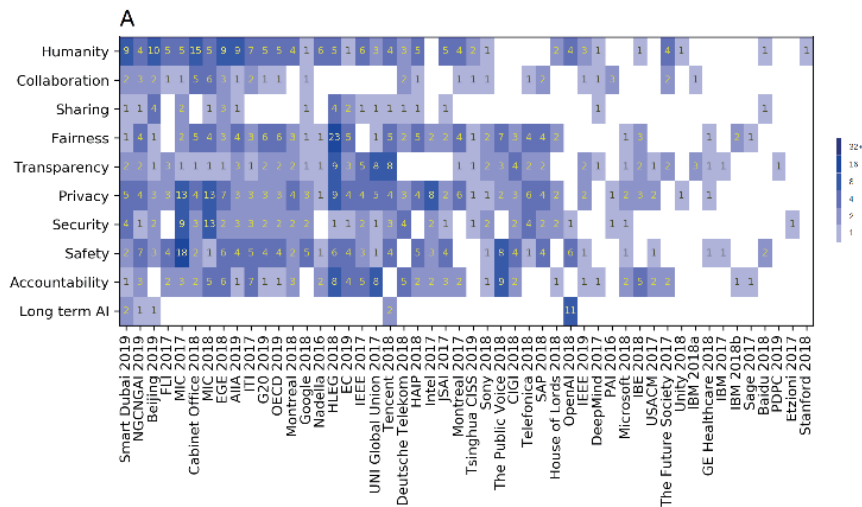
首先，本文通过人工选择确定了一系列关键词作为核心词汇，共分为十类。然后采用谷歌词向量（word2vec）模型来指代词汇，从而找到拥有相似含义的词，该模型的数据库来源于新闻。通过计算原关键词词向量与其他词汇词向量之间的余弦相似度，可以得出原关键词与其他词汇的相似度，随之按照相似度排名产生备选关键词扩展列表。列表上第一个与原关键词产生明显语义偏离的词为临界点，在该词相似度以下的一切词汇都弃之不用。一些具有相似含义的短语也会添加至关键词扩展列表中。例如，就“合作”（collaboration）一词而言，扩展列表中也会包括合作的复数形式（collaborations）、合作的、合作地以及“合作”的各种动词形态。而在处理“公平”（fairness）一词时，扩展列表上包括的词汇有公平的、更公平、不公平的与不公平。

表 1. 人工智能准则主题与人工选择关键词

主题	关键词
人类	人类、有益、福祉、尊严、自由、教育

协作	协作、伙伴关系、合作、对话
共享	共享、平等
公平	公平、公正、偏见（bias）、歧视、偏见（prejudice）
透明	透明、可解释的、可预测的、清晰的、审核、追踪
隐私	隐私、知情、掌控数据
（外部）安全	（外部）安全、网络安全、保密性
（内部）安全	（内部）安全、确认、验证、测试、可控性、人类控制
追责	追责、责任
远期人工智能	远期人工智能、高阶人工智能、通用人工智能（AGI）、超级智能（ASI）

本文所提议的准则主题涵盖率指的是提议中提及主题的百分比。若任何词汇或扩展关键词在某项提议中出现，则笔者将标记该词，表明该提议涵盖了该词的相关主题。表1列明了人工智能准则的10个主题与相关词汇，该列表基于语义相似度而进行词汇扩展，从而涵盖更多词汇。



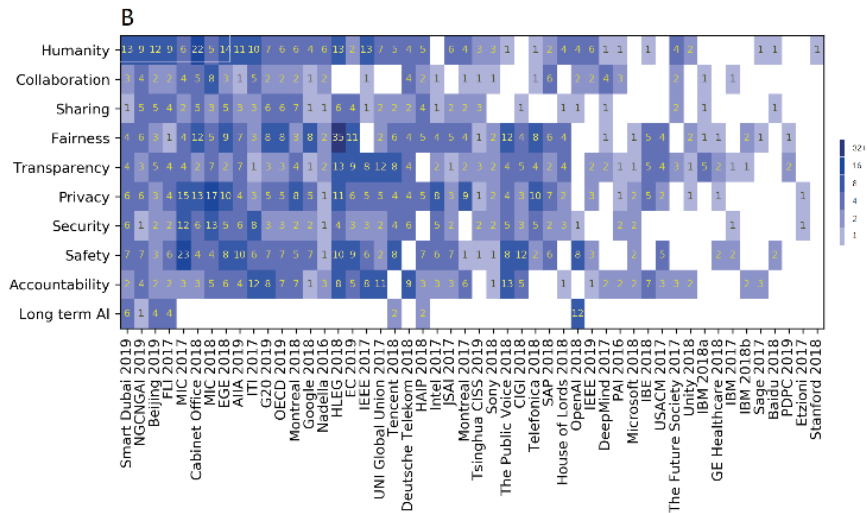


图 1 基于人工选择关键词得出的准则标题与解释性文字中的主题涵盖率（A）与根据语义相似度分类的扩展关键词（B）

图 1 展示了 10 个主题在不同准则中的涵盖率。图中的颜色与关键词在准则中出现的次数相关。由图可见，根据语义相似度扩展关键词之后，准则中的主题出现次数大幅提高，相似词语与表达使用的语义分析也更加准确稳定。

三、各组织、各人工智能准则之间的互补性

如图 2 所示，对不同的准则提议在主题与关键词的涵盖率上进行了比较。在所有准则提议中，有 4 个涵盖了所有主题。在关键词涵盖率的前 10 名中，有 7 个也在主题涵盖率排名中位列前 10。欧盟委员会人工智能高级专家组（2018）在关键词涵盖率中排名更加靠前（第二名），却在主题涵盖率排名中相对靠后（并列第 17 名），这是因为在其提案中，大量涉及关于公平、透明、隐私与追责的主题，却可能忽视了合作与人工智能长远发展的主题。生命未来研究所（2017 年）涵盖了全部 10 个主题（并列第一），但没有进行细致探讨，因此关

关键词排名较为靠后（第 12 名）。但要注意的是，提议涵盖率的排名靠后并不一定表示这些提议没有考虑到某些主题，仅说明这些提议所关注的重点或许各不相同。从另一角度来看，不同的考量也可以互为补充。

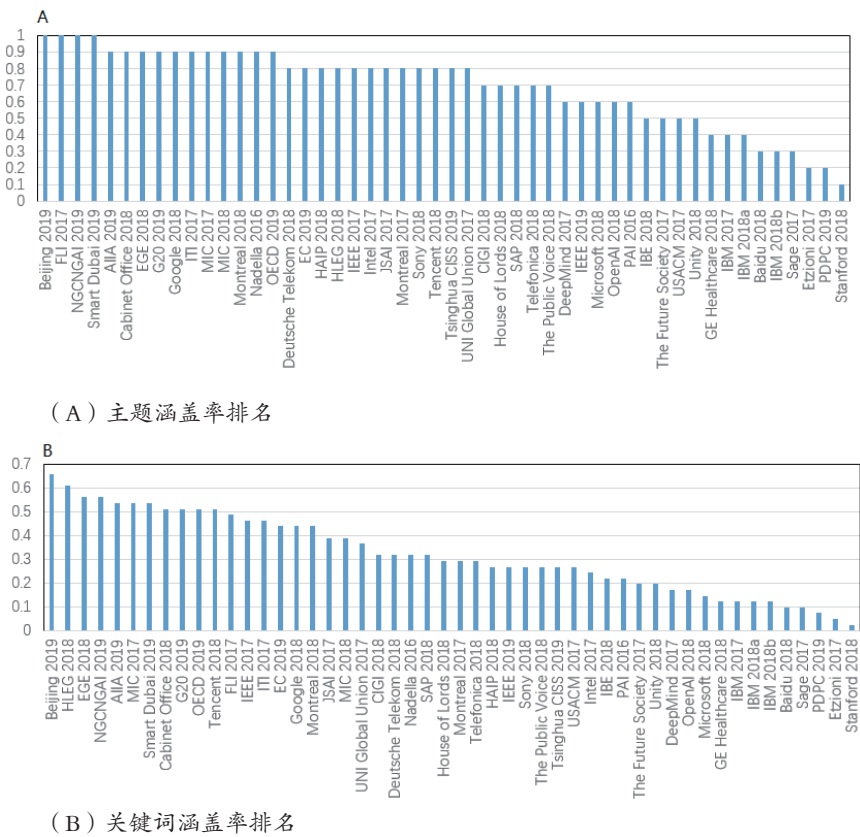


图 2 各人工智能准则的主题涵盖率 (A) 与关键词涵盖率 (B) 排名

虽然在大部分情况下，由不同组织制定的准则通常使用的词汇具有一致性，但是文本分析还是比较模糊。这种模糊性可能源自词汇与语境的多义现象。例如，“race” 这个词汇会在“军备竞赛”（arms race）以及“避免竞赛（race avoiding）（生命未来研究所，2017）的语境中出现，指的是研究人员与国家之间的竞争（即“合作”主

题），但该词也会在“性别、种族、性取向”（gender, race, sexual orientation）（全球工会联盟，2017）的语境中出现，指的是人工智能系统可能存在的偏见（即“公平”主题）。再例如，高级人工智能系统的“self-improvement”（自我进化）是一种我们应当警惕的特性（生命未来研究所，2017），但人工智能研究人员的“self-improvement”（自我提高）却值得鼓励（日本人工智能学会，2017）。

此类模糊性也会在同一主题中出现。比如，我们会出于公平考量，要求人工智能系统的决策过程透明化（transparency），也会要求人工智能系统保持透明（transparency），从而使人工智能更加安全、可追踪且可控。《阿西洛马人工智能准则》在文中已经明确对透明化进行了区分（见“司法透明”与“错误透明”（生命未来研究所，2017）），而其他准则通常仅使用一种语义或是混淆两种语义。在这些情况下，模糊性来源于某个概念的高度概括，也表明各个话题之间有着内在联系。

人工智能准则的主题有不少共性，但很多准则也反映出不同组织的特性。例如，《蒙特利尔宣言》提倡保障“一切有知觉的生物”的福祉，根据该宣言的定义，“一切有知觉的生物”包括“任何能够感受喜悦、痛苦、感情的生物，即有感觉的生物”（蒙特利尔大学，2017）。《日本人工智能学会伦理准则》指出，人工智能必须同日本人工智能学会会员一样遵守这些准则，才能成为社会成员或准成员（日本人工智能学会，2017）。

电气电子工程师协会报告中的一般准则提出建议，“在不久的将来，人工智能系统不应拥有与人类同等的权利与特殊待遇：人工智能系统应当永远受人类评判、受人类控制”（电气电子工程师协会，2017）。IBM认为，“认知系统不会真正获得意识或独立行动力”，因此强调要通过推广人工智能与认知系统来“强化人类智慧”（IBM，2017）。不同提议涉及到的不同方面反映出了人工智能的多样性，也

意味着我们有必要去辨别并整合这些多样考量，从而构建起更加全面的框架。

据本文分析，笔者对未来人工智能准则的研究与提案提出以下建议：

提高学术界、非营利组织与非政府组织的公平意识。人工智能的裨益应当由全人类共享，数据、人才与技术一经垄断，便会阻碍人工智能的发展进程并加重不平等。

通用人工智能和超级智能的长远战略规划。本文研究的大多人工智能准则都没有考虑到通用人工智能和超级智能。从当前的人工智能水平来看，通用人工智能和超级智能尚且遥不可及，但若未雨绸缪，定会带来战略性的未来规划，也能够防患于未然。

从以人为本到和谐共处。当前的人工智能准则提议主要着眼于以人为本、为人谋益，缺乏对人类社会不断变革的考量。因此，在考虑人与未来人工智能更加和谐共处的同时，也应将认知生命系统纳入考虑范围。

四、结论

不同的人工智能准则对当下与未来人工智能战略的考虑角度与包含内容也各不相同。笔者不建议仅仅采取一种人工智能准则，而是应当综合不同人工智能准则，得出更为全面的框架，并且注重各个准则之间的相互牵动与完善。点击 <http://www.linking-ai-principles.org> 即可进入链接人工智能准则（LAIP）平台。该平台支持关键词检索与段落检索语义相近的准则供读者参考。

致 谢

本研究受中国科学技术信息研究所“人工智能社会伦理核心问题”项目及科技部新一代人工智能发展研究中心支持。

References

DeepMind. 2017. DeepMind Ethics & Society Principles. <https://deepmind.com/applied/deepmind-ethics-society/principles/>.

Etzioni, O. 2017. How to Regulate Artificial Intelligence. <https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html>.

European Group on Ethics in Science and New Technologies (EGE). 2018. Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems. http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

Future of Life Institute (FLI). 2017. Asilomar AI Principles. <https://futureoflife.org/ai-principles/>.

Google. 2018. AI at Google: Our Principles. <https://ai.google/principles>.

HAIP Initiative. 2018. Harmonious Artificial Intelligence Principles (HAIP). <http://bii.ia.ac.cn/hai/index.php>.

House of Lords, UK. 2018. AI in the UK: ready, willing and able? <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>.

IBM. 2017. Principles for the Cognitive Era. <https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/>.

IBM. 2018. Principles for Trust and Transparency. <https://www.ibm.com/blogs/policy/trust-principles/>.

Information Technology Industry Council (ITI). 2017. AI Policy Principles. <https://www.iti.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf>.

Microsoft. 2018. Microsoft AI Principles. <https://www.microsoft.com/en-us/ai/our-approach-to-ai>.

Ministry of Internal Affairs and Communications (MIC), the Government of Japan. 2017. AI R&D Principles. http://www.soumu.go.jp/main_content/000507517.pdf.

Ministry of Internal Affairs and Communications (MIC), the Government of Japan. 2018. Draft AI Utilization Principles. http://www.soumu.go.jp/main_content/000581310.pdf.

Nadella, S. 2016. The Partnership of the Future: Microsoft's CEO explores how humans and A.I. can work together to solve society's greatest challenges. <https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html>.

OpenAI. 2018. OpenAI Charter. <https://blog.openai.com/openai-charter/>.

PAI 2016. Partnership on AI. (September 28, 2016). Partnership on AI: Tenets. Retrieved from <https://www.partnershiponai.org/tenets/>.

IEEE 2017. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (December 12, 2017). Ethically Aligned Design (v2): General Principles. Retrieved from <https://standards.ieee.org/industry-connections/ec/autonomous-systems.html>.

FLI 2017. Beneficial AI 2017. (Jan 8, 2017). Asilomar AI Principles. Retrieved from <https://futureoflife.org/ai-principles/>.

USACM 2017. ACM US Public Policy Council (USACM). (January 12, 2017). Principles for Algorithmic Transparency and Accountability. Retrieved from https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

JSAI 2017. The Japanese Society for Artificial Intelligence (JSAI). (February 28, 2017). The Japanese Society for Artificial Intelligence Ethical Guidelines. Retrieved from <http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf>.

Etzioni 2017. Oren Etzioni. (September 1, 2017). Three Rules for Artificial Intelligence Systems. Retrieved from <https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html>.

The Future Society 2017. The Future Society, Science, Law and Society (SLS) Initiative. (October 3, 2017). Principles for the Governance of AI. Retrieved from <http://www.thefuturesociety.org/science-law-society-sls-initiative/#1516790384127-3ea0ef44-2aae>.

ITI 2017. Information Technology Industry Council (ITI). (October 24,

2017). AI Policy Principles. Retrieved from <https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf>.

Montreal 2017. University of Montreal. (November 3, 2017). The Montreal Declaration for a Responsible Development of Artificial Intelligence. Retrieved from <https://www.montrealdeclaration-responsibleai.com/the-declaration>.

UNI Global Union 2017. UNI Global Union. (December 11, 2017). Top 10 Principles For Ethical Artificial Intelligence. Retrieved from http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.

IBE 2018. Institute of Business Ethics (IBE). (January 11, 2018). IBE interactive framework of fundamental values and principles for the use of Artificial Intelligence (AI) in business. Retrieved from https://www.ibe.org.uk/userassets/briefings/ibe_briefing_58_business_ethics_and_artificial_intelligence.pdf.

CIGI 2018. Centre for International Governance Innovation (CIGI), Canada. (July 19, 2018). Toward a G20 Framework for Artificial Intelligence in the Workplace. Retrieved from <https://www.cigionline.org/publications/toward-g20-framework-artificial-intelligence-workplace>.

HAIP 2018. HAIP Initiative. (September 16, 2018). Harmonious Artificial Intelligence Principles (HAIP). Retrieved from <http://www.harmonious-ai.org/>.

Stanford 2018. Stanford University. (October 19, 2018). The Stanford Human-Centered AI Initiative (HAI). Retrieved from http://hai.stanford.edu/news/introducing_stanfords_human_centered_ai_initiative/.

The Public Voice 2018. The Public Voice coalition. (October 23, 2018). Universal Guidelines for Artificial Intelligence. Retrieved from <https://thepublicvoice.org/ai-universal-guidelines/>.

Montreal 2018. University of Montreal. (December 4, 2018). The Montreal Declaration for a Responsible Development of Artificial Intelligence. Retrieved from <https://www.montrealdeclaration-responsibleai.com/the-declaration>.

Tsinghua-CISS 2019. Center for International Security and Strategy,

Tsinghua University (Tsinghua-CISS). (January 23, 2019). Six AI Principles proposed by Mme Fu Ying. Retrieved from <http://www.imir.tsinghua.edu.cn/publish/iis/7238/20190123180702719613866/1554792786844.pdf>.

OECD 2019. The Organisation for Economic Co-operation and Development (OECD). (May 22, 2019). OECD Principles on Artificial Intelligence. Retrieved from <http://www.oecd.org/going-digital/ai/principles/>.

Beijing 2019. Beijing Academy of Artificial Intelligence (BAAI) et al. (May 25, 2019). Beijing AI Principles. Retrieved from <https://baip.baai.ac.cn/en>.

IEEE 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (June 24, 2019). Ethical Aspects of Autonomous and Intelligent Systems. Retrieved from <http://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE19002.pdf>.

MIC 2017. Ministry of Internal Affairs and Communications (MIC), Japan. (July 28, 2017). AI R&D Principles. Retrieved from http://www.soumu.go.jp/main_content/000507517.pdf.

EGE 2018. European Group on Ethics in Science and New Technologies. (March 9, 2018). Ethical principles and democratic prerequisites. Retrieved from http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

House of Lords 2018. House of Lords, UK. (April 16, 2018). AI Code. Retrieved from <https://social.shorthand.com/LordsAICom/32KXpihQLj/ai-in-the-uk>.

MIC 2018. Ministry of Internal Affairs and Communications (MIC), Japan. (July 17, 2018). Draft AI Utilization Principles. Retrieved from http://www.soumu.go.jp/main_content/000581310.pdf.

HLEG 2018. The European Commission's High-Level Expert Group on Artificial Intelligence. (December 18, 2018). Draft Ethics Guidelines for Trustworthy AI. Retrieved from <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>.

Cabinet Office 2018. Cabinet Office, Japan. (December 27, 2018). Social Principles of Human-centric AI (Draft). Retrieved from <https://www8.cao.go.jp/cstp/stmain/aisocialprinciples.pdf>.

Smart Dubai 2019. Smart Dubai. (January 8, 2019). Dubai's AI Principles.

Retrieved from <https://smartdubai.ae/initiatives/ai-principles>.

PDPC 2019. Personal Data Protection Commission (PDPC), Singapore. (January 23, 2019). A Proposed Model AI Governance Framework: Guiding Principles. Retrieved from <https://www.pdpc.gov.sg/resources/model-ai-gov>.

EC 2019. European Commission. (April 8, 2019). Key requirements for trustworthy AI. Retrieved from <http://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAIN-PART-1.PDF>.

G20 2019. G20 Ministerial Meeting on Trade and Digital Economy. (June 9, 2019). G20 AI Principles. Retrieved from <https://www.mofa.go.jp/mofaj/files/000486596.pdf>.

NGCNGAI 2019. National Governance Committee for the New Generation Artificial Intelligence, China. (June 17, 2019). Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence. Retrieved from <https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html?from=singlemessage&isappinstall=0>.

Nadella 2016. Satya Nadella. (June 28, 2016). Microsoft CEO's 10 AI rules. Retrieved from <https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html>.

IBM 2017. IBM. (January 17, 2017). Principles for the Cognitive Era. Retrieved from <https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/>.

Sage 2017. Sage. (June 27, 2017). The Ethics of Code: Developing AI for Business with Five Core Principles. Retrieved from <https://www.sage.com/ca/our-news/press-releases/2017/06/designing-AI-for-business>.

DeepMind 2017. DeepMind. (October 3, 2017). DeepMind Ethics & Society Principles. Retrieved from <https://deepmind.com/applied/deepmind-ethics-society/principles/>.

Intel 2017. Intel. (October 18, 2017). AI public policy principles. Retrieved from <https://blogs.intel.com/policy/2017/10/18/naveen-rao-announces-intel-ai-public-policy/>.

Microsoft 2018. Microsoft. (January 17, 2018). Microsoft AI Principles. Retrieved from <https://www.microsoft.com/en-us/ai/our-approach-to-ai>.

OpenAI 2018. OpenAI. (April 9, 2018). OpenAI Charter. Retrieved from <https://blog.openai.com/openai-charter/>

Deutsche Telekom 2018. Deutsche Telekom. (May 11, 2018). Deutsche Telekom's guidelines for artificial intelligence. Retrieved from <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366>.

Baidu 2018. Baidu. (May 26, 2018). Four principles of AI ethics. Retrieved from <http://www.fonow.com/view/208592.html>.

IBM 2018. IBM. (May 30, 2018)., Principles for Trust and Transparency. Retrieved from <https://www.ibm.com/blogs/policy/trust-principles/>.

Google 2018. Google. (June 7, 2018). Artificial Intelligence at Google: Our Principles. Retrieved from <https://ai.google/principles>.

IBM 2018. IBM. (September 6, 2018)., Everyday Ethics for Artificial Intelligence: Five Areas of Ethical Focus. Retrieved from <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.

SAP 2018. SAP. (September 18, 2018). SAP's Guiding Principles for Artificial Intelligence. Retrieved from <https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/>.

Tencent 2018. Tencent. (September 18, 2018). The ARCC (Available, Reliable, Comprehensible, and Controllable) for AI. Retrieved from https://mp.weixin.qq.com/s/_CbBsrjrTbRkKjUNdmhuqQ.

Sony 2018. Sony. (September 25, 2018). Sony Group AI Ethics Guidelines. Retrieved from https://www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf.

GE Healthcare 2018. GE Healthcare. (October 1, 2018). GE Healthcare AI principles. Retrieved from <http://newsroom.gehealthcare.com/ethics-healthcare-arent-new-application-important/>.

Telefonica 2018. Telefónica. (October 30, 2018). AI Principles of Telefónica. Retrieved from <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>.

Unity 2018. Unity Technologies. (November 28, 2018). Unity's Guiding Principles for Ethical AI. Retrieved from <https://blogs.unity3d.com/2018/11/28/introducing-unitys-guiding-principles-for-ethical-ai/>.

AIIA 2019. Artificial Intelligence Industry Alliance (AIIA), China. (May 31, 2019). Joint Pledge on Artificial Intelligence Industry Self-Discipline (Draft for Comment). Retrieved from <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/?from=timeline&isappinstalled=0>.

Zeng 2019. Yi Zeng, Enmeng Lu, Cunqing Huangfu. Linking Artificial Intelligence Principles. The AAAI 2019 Safe Artificial Intelligence (AAAI SafeAI-2019) Workshop, January 27, 2019.

清华大学战略与安全研究中心

办公地点：清华大学明斋 217

联系电话：010-62771388

电子邮箱：ciss_thu@163.com