人工智能伦理问题的回顾与前瞻

XXX

（2019XXXXXX，清华大学 计算机科学与技术系，北京 100084）

摘 要：自1956年达特茅斯会议提出人工智能的概念以来，从符号主义到联结主义再到如今的行为主义，人工智能获得了多领域跨学科的跨越式发展。当前，借助深度学习、神经网络等技术，人工智能已经在模式识别、数据分析、智能生活等方面有了众多的创新应用。本文从回顾人工智能伦理问题的起源和发展出发，对当前人工智能伦理出现的问题作出简要的回顾、分析与探究，同时对未来人工智能算法构建中的道德框架部分作出展望。

关键词：人工智能；深度学习；机器伦理；非中立性道德；算法治理

Ethical Issues of Artificial Intelligence in the Past and Future

Wang Zihan

(2019011342, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract. The concept of “artificial intelligence” since 1956 have been proposed in the Dartmouth Summer Research Project, afterwards from Symbolicism to Connectionism to Actionism, the technology obtained significant development multidisciplinarily and interdisciplinarily. Nowadays, through deep learning, neural networks and other technologies, artificial intelligence has many innovative applications in pattern recognition, data analysis, and smart life. This article starts with a review of the origin and development of artificial intelligence ethics issues, makes a brief review, analysis, and inquiry into the current problems of artificial intelligence ethics, and at the same time makes a prospect on the ethical framework part of future artificial intelligence algorithm construction.

Key words artificial intelligence; deep learning; machine ethics; non-neutral morality; algorithm governance

1 人工智能概述

1.1 第一代人工智能

1956年,由美国达特茅斯学院的约翰·麦卡锡约请了10位相关领域的专家在达特矛斯夏日人工智能探究计划(Dartmouth Summer Research Project on Artificial Intelligence)进行了一个月的思想碰撞后,第一次正式使用了“人工智能”的术语,标志着人工智能这一新兴学科的正式诞生。[1]之后的40年时间，符号主义的方法一直是研究的主流。而第一代人工智能的发展以1994年爱德华·费根鲍姆与拉吉·瑞迪因大规模人工智能系统而获得图灵奖宣告结束。自此之后，人工智能陷入了一段低潮时期。

1.2 第二代人工智能

第二代以概率统计学习、深度学习为基础的数据驱动方法在第一代人工智能的发展后期逐渐兴起。2012年，在ImageNet大规模图像模式识别比赛上，辛顿教授和他的两个研究生亚历克斯·克里热夫斯基、伊利亚·萨茨克维尔使用基于ANN的卷积神经框架的深度学习网络，获得巨大的成功。他们的模型是一个全数八层的卷积神经网络，预测的头五个类别的错误率只有15.3%。这意味着图像领域的神经网络识别方法获得了大幅度提升，已经远远胜过其他的图像识别算法，同时也证明了卷积网络的叠加对深度神经网络构建的巨大意义，这一事件成为第二代人工智能技术突破的转折。此后，第二代人工智能在语音识别、图像处理和围棋等方面逐渐超越人类。

然而，这些第二代人工智能都满足以下五个条件：（1）具备充足、多样化、完整的数据库（2）算法实现具有确定性（3）获取完整的信息，比如围棋是完全信息博弈（4）问题静态可预测（5）针对有限的领域。[2]这意味着对于复杂路况下的自动驾驶、牌类的不完全博弈、增强人工智能“社交性”的尝试，都难以在第二代人工智能的框架下实现。并且由于第二代人工智能基于概率统计而非真正的理解，我们无法通过增加神经网络层数或复杂性达到真正的智能。

1.3 第三代人工智能

在2019年召开的第五届中国人工智能大会上(CCAI2019)，张钹院士谈到，第三代AI将会改变第二代AI会被欺骗、鲁棒性差、不可解释等缺点，“建立可解释、鲁棒的人工智能理论，发展安全、可信、可用的人工智能技术，促进人工智能的创新应用。”[3]

2 人工智能伦理问题的起源与发展

2.1 人工智能伦理问题的初始探究

“人工智能”这一概念诞生伊始，关于人工智能伦理问题的思考就存出不穷。我们不妨引入 Pre-future （过去预测的未来）这一概念，本节就将着重讨论人工智能伦理问题的 Pre-future 讨论。

科幻作者艾萨克·阿西莫夫(Isaac Asimov)在其1942年出版的短篇小说"Runaround"中初次提出了机器人三定律，但这一套“义务伦理学”（deontology）框架始终未得到计算机界、工业制造界和机器伦理学界的认可。

阿兰·麦席森·图灵(Alan Mathison Turing, The father of AI)在1950年发表的论文《计算机器与智能》(Computing Machinery and Intelligence)中提出问题：机器能思考吗(Can machines think)?并在论文中提到人工智能迟早会对人类的生存造成威胁。[4]

值得注意的是，这些 Pre-future 的讨论内容都早于1951年的第一台神经网络机SNARC。在这之后，尤其是第一代人工智能20世纪70-80年代的低谷期，关于人工智能的伦理讨论更多的转向理论层面。

2.2 人工智能伦理问题的理论发展

德国学者汉斯·论克在《责任伦理：技术文明时期的伦理学探索》中第一次将“责任伦理学“单独作为一个重要的伦理概念进行论述，这种科技伦理学的独特视角对之后的机器伦理甚至是整伦理学和神学界产生了巨大的影响。“责任伦理学”重点讨论了机器人出现“意外”故障该由谁负责的问题。

而后，美国达特茅斯大学哲学学院教授詹姆斯H.摩尔(James H. Moor)提出了道德行为体的划分概念。他将所有有着内在伦理化运行规则的智能体按照伦理判断水平与行为能力分为三种：隐性道德执行者(Implicit Ethical Agent)、显性道德执行者(Explicit Ehitcal Agent)和完全道德执行者(Full Ethical Agent)。而在詹姆斯H.摩尔的体系下，构建显性道德执行体作为一种明确、可解释、可变通的智能体解决方案在避免人为因素或自然因素造成的道德选择场景中可能是作为一种最佳的智能体机器伦理而存在。而且机器相比人类有着计算与决策速度上的巨大优势，因此是否部分伦理道德决策移交智能体，从某种方面来说即是否相对信任智能体自主决策成为此后很长一段时间的机器伦理学讨论重点。当前，在人工智能无人驾驶达到L4，向L5和L6发展的趋势下，这一讨论就显得愈发重要，我们将在第三节重点阐述。

然而，伦理问题表述通常具有的模糊性、不确定性和多态性常常与机器语言的准确、可解释性相矛盾。有趣的是在第二代人工智能发展至今天，尤其是连接主义派别的ANN和SNN网络，通常都是不可解释的“黑箱”。于是，机器伦理学的发展路径有了些许分叉，一是尽量界定在人类看来都难以解释的各种伦理规则，比如电车难题或者战场中的伤亡价值计算，称为“机械伦理”(Mechanical ethics)。

在“机械伦理”的探索过程中，前有美国冯·赖德(von Wright)在1950年代创立机器伦理重要工具——道义逻辑；后有美国伦斯勒理工学院(Rensselaer Polytechnic Institute)的康斯坦丁·阿库达斯(Konstantine Arkoudas)与赛尔玛·布林斯加德罗(Selmer Bringsjord)在2005年的论文中构建了一套可行性高的机器伦理判断算法，证明在机器上部署道义逻辑的可行性。并且，他们认为道义逻辑给伦理道德提供了一条具体化的道路，并可以通过这种方式确保智能体的行为符合人类的普遍道德期望。

二是构建类人的智能体，只提供一个公认的具有普世道德意义的伦理决策大数据样本，训练出一个具有普世意义下认可的智能体伦理决策系统，一切决策都由多层神经网络 的“黑箱”决定，而不直接由人类干预，称为“智能伦理”(Intelligent ethics)。“智能伦理”的许多基础技术仍不成熟，对这种伦理解决方案的讨论始终在对未来技术的想象基础之上展开，拥有太多的不确定性，因此始终难以得到认可。第三节中会做相应的展望。

当然，究竟选择“机械伦理”还是“智能伦理”解决方案最终还是取决于人工智能技术的未来发展和人类伦理道德的动态变化以及对机器伦理决策普遍性接受程度。人工智能伦理问题始终一个动态变化、不断发展且发展速度不断变快的螺旋式上升过程。

3 人工智能伦理问题前瞻

本节将讨论的问题集中在近未来，而对于远未来的人工智能的人权问题不做讨论。为了更好地讨论，本节会引入一些未来学的理论与观点。

3.1 技术的非中立性道德

彼得·保罗·韦贝克（Peter-Paul Verbeek）在其著作《道德化的技术：理解与设计事物的道德》中希望扩大伦理学的范围，以便更好地适应这个技术时代，并由此揭示了人类与技术不可分割的本质。在韦贝克思想中，技术可以被认为是“道德调解者”，它塑造了我们对世界的感知方式，从而揭示和指导了我们可能的行为。由于每种技术都会影响我们感知方式，甚至影响我们的思维方式，因此没有任何一种技术在道德上是中立的——任何技术都会改变我们的生活，而这种改变好坏并存——人工智能也不例外。

面对人和技术不可分割的关系，对人类目的和行动道德性的质疑无疑会扩展到技术的道德性。技术可以用来改变我们周围的世界，个人的目标，无论是私人的还是社会的，都可以借助技术来实现。考虑到人类在创建技术时所追求的目标与技术本身被赋予的属性具有相对的一致性，我们可以得出结论，技术具有内在的道德特性。

3.2 人工智能技术的不可预测性

如果我们将“驯化”属性拓展到人类与其创造的技术，比如“造纸术”是人类“驯化”的产物，我们可以完全按照我们的自主意愿使用它。那么，当前的第二代人工智能仍然是一个“未驯化”的技术，基于对大数据处理得到的模型难以解释不可预测，因此我们往往难以预见大部分可能产生的黑天鹅类风险——我们更多的只能风险发生后在补救和完善。

2015年，Google相册从用户数据中学习自动标记照片。然而结果具有歧视性，例如将有色人种标记为大猩猩。

2016年，Microsoft在推特上发布具备深度学习能力的名叫Tay的AI程序，它基于与其他人的在线交流来塑造世界观，并基于这些人产生真实的表达。然而，不到一天，Tay就开始模仿其他用户的攻击性发言，包括种族主义、性别歧视和反犹太主义出版物，这使得微软不得不在不到24小时的时候停用Tay。

我们可以看出，这些不可预测的负面后果是由于人工智能程序将亚文化内容扩散到广阔的公共平台，换句话说，它们不仅可以被人为附加“不道德性”，如 deepfake 技术，也会自主的从人类身上学习“不道德”性。这种“不道德”性的不可预测还体现在程序漏洞方面，比如微软辩称Tay的故障是用户利用程序漏洞的发起的攻击，以及2019年出现的图像识别领域的对抗样本。

更加广泛的不可预测性来自于使得人工智能自我迭代的尝试，例如 AlphaZero 与AlphaStar 的自我训练。人工智能不仅仅只是充当在其所使用的领域中执行预先建立的功能的工具，而且还开始发展更加智能的行为方式——值得注意的是使算法能够自我调整调整以产生新算法和完成新任务的新方法[7]——如果使用这种算法，即使对于创建该算法的程序员来说，实现结果的方式也将难以解释，这使得它们呈现出更少可预测的行为，从而带来新的风险。

更加重要与危险的不可预测性体现在无人驾驶与军事机器人领域。这些方面的讨论十分普遍，在此不再赘述

3.3 人工智能可能的道德框架

建立人工智能必要的道德框架是“算法治理”的必要部分，它将是未来创造新算法时减少算法被用于控制、审查、社会歧视、侵犯隐私权和财产权等行为的软性准则。

欧盟委员会于2019年4月发布了文件《可信赖人工智能(Trustworthy AI)道德准则》以及关于人工智能伦理的准则。根据该准则，可信赖的人工智能应该具有以下特点：（1）合法——遵守所有适用法律和法规；（2）道德——尊重道德原则和价值观；（3）稳定——从技术角度出发。

准则还提出了AI系统应当被视为值得信赖的七个关键要求：

（1）人工代理和监督：人工智能系统应赋予人类权力，使他们能够做出明智的决定并促进其基本权利；

（2）技术的鲁棒性和安全性：安全、准确、可靠和可重复的AI系统是确保最大程度地减少或防止意外伤害的唯一方法；

（3）隐私和数据治理：除了确保充分尊重隐私和数据保护外，还必须确保充分的数据治理机制，同时考虑到数据的质量和完整性，并确保对数据的合法访问；

（4）透明度：数据，系统和AI业务模型应透明、可追溯性，并且人们需要意识到他们正在与AI系统进行交互且被告知AI系统的功能和局限性；

（5）多样性，非歧视性和公平性：必须从算法层面避免歧视；

（6）社会和环境福祉：人工智能系统应使全人类包括后代受益，因此必须确保它们是可持续的和环境友好的；

（7）问责机制：应建立问责机制以明切AI系统的责任归属，此外还应确保有适当的补救措施。[9]

与这份倡议类似，许多国家、公司和专业团体都在发布具有类似价值和原则的AI指南，以减少人工智能在开发阶段可能产生的风险。

4 结 语

本文围绕人工智能的伦理问题，从伦理学、逻辑学、计算机科学、未来学等多角度出发，对人工智能伦理问题的起源、当下和将来作出相应的总结、分析和预测。

如本文所介绍的，人工智能伦理问题的讨论仍然在模糊的环境中进行，我们能够采取的准确路径仍然具有极大的不确定性。用意大利哲学家卢西亚诺·弗洛里迪（Luciano Floridi）的话来说：“新的挑战不是技术创新，而是数字化治理”。[10]

参 考. 文. 献

1. McCarthy J, Minsky M L, Rochester N, et al. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955[J]. AI magazine, 2006, 27(4): 12-12.
2. 《信息科学技术概论》系列讲座. 张钹：人工智能发展现状与未来
3. 张钹院士：人工智能的魅力就是它永远在路上 | CCAI 2019[OL]. [2019-09-21]. https://www.chainnews.com/articles/433609698360.htm
4. Machinery C. Computing machinery and intelligence**-AM Turing[**J]. Mind, 1950, 59(236): 433.
5. Moor J H. The nature, importance, and difficulty of machine ethics[J]. IEEE intelligent systems, 2006, 21(4): 18-21.
6. Arkoudas K, Bringsjord S, Bello P. Toward ethical robots via mechanized deontic logic[C]//AAAI fall symposium on machine ethics. 2005: 17-23.
7. Domingos P. The master algorithm: How the quest for the ultimate learning machine will remake our world[M]. Basic Books, 2015.
8. Doneda D, Almeida V A F. What is algorithm governance?[J]. IEEE Internet Computing, 2016, 20(4): 60-63.
9. High-Level Expert Group on AI (HLEG AI). (2019). Ethics guidelines for trustworthy AI[OL] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
10. Floridi L. Soft Ethics and the Governance of the Digital[J]. Philosophy & Technology, 2018, 31(1): 1-8.