分类号：　　密级：

U D C ：　　学号：405000315025

南 昌 大 学 硕 士 研 究 生

学 位 论 文

人工智能技术风险研究

Research of Artificial intelligence technology risk

钱玲

培养单位（院、系）：　人文学院　　哲学系

指导教师姓名、职称：　黄承烈　　副教授

申请学位的学科门类：哲学

学科专业名称：科学技术哲学

论文答辩日期：2018 年 6 月 3 日

答辩委员会主席：_____

评阅人：_____

_____

2018 年 6 月 3 日

一、学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得南昌大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名（手写）：钱玲　　签字日期：2018 年 6 月 3 日

二、学位论文版权使用授权书

本学位论文作者完全了解南昌大学有关保留、使用学位论文的规定，同意学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权南昌大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文。同时授权北京万方数据股份有限公司和中国学术期刊（光盘版）电子杂志社将本学位论文收录到《中国学位论文全文数据库》和《中国优秀博硕士学位论文全文数据库》中全文发表，并通过网络向社会公众提供信息服务，同意按“章程”规定享受相关权益。

学位论文作者签名（手写）：钱玲　　导师签名（手写）：黄承义

签字日期：2018 年 6 月 2 日　　签字日期：2018 年 6 月 2 日

论文题目	人工智能技术风险研究				
姓　名	钱玲	学号	405000315015	论文级别	博士□ 硕士☑
院/系/所	人文学院		专业	科学技术哲学	
E_mail					
备注：					

☑公开　□保密（向校学位办申请获批准为“保密”，_____年___月后公开

摘　要

人类正处于智能时代，智能搜索引擎、机器翻译、指纹识别、人脸识别、自动驾驶等技术的应用，使得人类社会生活、工作变得更加便利。但随着人工智能技术的普及，越来越多的学者开始对人工智能技术的发展表示担忧，人工智能技术在给社会带来便利的同时，也会给社会公共安全带来了巨大隐患，因此，对人工智能技术风险进行哲学分析，具有一定的现实意义与理论意义。本文基于风险社会视角对人工智能技术风险研究，当今社会是风险社会，技术风险是风险社会中最主要的原因，人工智能技术作为一种新兴技术，由于它的特殊性必将给社会带来巨大的风险。首先，通过智能、人工智能等概念的区分，明确人工智能的基本内涵及特征。其次，从本体论、认识论和社会层面对人工智能技术风险产生的原因进行了初步研究，明确了人工智能技术风险是它的内在属性，也是人类在技术发明和使用过程中不可避免的风险。最后通过加强技术与人文的沟通，提高公众认知水平和科学家道德责任意识，完善相关法律法规，构建“亚政治”风险管理机制等具体措施，对人工智能技术进行风险规避。本文的目的是通过对人工智能技术全方面的分析，并提出有利于人类社会发展的具体措施建议，以期发展对人类有益的人工智能技术。

关键词：风险社会，技术风险，人工智能技术风险，人工智能技术风险规避

ABSTRACT

Human beings are in the era of intelligence. The application of intelligent search engines, machine translation, fingerprint recognition, face recognition, and automatic driving have made human social life and work more convenient. However, with the popularization of artificial intelligence technology, more and more scholars have begun to express concern about the development of artificial intelligence technology. Although artificial intelligence technology brings convenience to the society, it also brings great hidden dangers to social public safety. The philosophical analysis of the risk of artificial intelligence technology has certain practical and theoretical significance. This article is based on the risk social perspective on the research of artificial intelligence technology risk. Our society is a risk society now. Technology risk is the most important reason in risk society. Artificial intelligence technology is regarded as a new technology, and its particularity will inevitably bring huge risks to the society. First of all, by distinguishing the concepts of intelligence and artificial intelligence, the basic connotation and characteristics of artificial intelligence are clarified. Secondly, from the ontological, epistemological and social perspectives, a preliminary study was conducted on the causes of artificial intelligence technology risks. It was clarified that the risk of artificial intelligence technology is its intrinsic property, and it is also an inevitable risk for human beings in the process of technological invention and use. Finally, by strengthening the communication between technology and humanities, raising public awareness and the awareness of scientists' ethical responsibilities, improving relevant laws and regulations, and establishing specific measures such as “sub-political” risk management mechanisms, we will gain the risk-avoidance of artificial intelligence technology. The purpose of this paper is to analyze all aspects of artificial intelligence technology and to propose concrete measures that are beneficial to the development of human society, in order to develop artificial intelligence technologies that are beneficial to humans.

Key Words： Social Risk, Technology Risk, Artificial Intelligence Technology, Risk Aversion Artificial Intelligence Technology

目录

第 1 章 绪论

1.1 选题背景和选题意义

1.1.1 选题背景

随着人工智能、大数据、互联网、虚拟现实等现代技术的发展，社会环境发生了翻天覆地的变化，特别是人工智能的蓬勃发展，正在颠覆人类社会现有的组织、生产和生活形态。2016 年阿尔法狗（Alpha Go）在围棋领域首次战胜了人类的世界冠军， 2017 年 10 月 19 日，谷歌的 Deep Mind 团队发布了强化版的阿法元（Alpha Go Zero），该版本的 Alpha Go 进一步实现了在 AI 发展中非常有意义的一步——“无师自通”，验证了即使在像围棋这样最具挑战性的领域，也可以通过纯强化学习的方法自我完善达到目的。

近年来，又掀起人工智能研究热潮，人工智能技术被应用到越来越多的行业，无人驾驶给交通行业带来出行上的改变，利用算法识别帮助警察抓获嫌犯，智能机器人有望解决医疗行业的资源分配问题等，人工智能技术深入生活中各个方面，将会给未来社会带来巨大变革。不仅如此，美国、日本、英国和中国等世界科技强国都纷纷加入人工智能研究浪潮之中，出台相关的战略规划，将人工智能上升到了国家战略高度。人工智能目前处于高速发展状态，科学家们希望研究出一种可以独立思维的智能体，以实现对人类智能的模拟，并希望通过智能机器的研究达到对人类智能进一步的研究。但是人工智能在给我们生活带来种种益处的同时，也带来许多风险，比如，人工智能机器人 Sophia 的资本骗局造成公众对智能机器的恐慌，美国自动驾驶汽车撞死人事件，等等，人工智能技术带来的风险也一一突显，因此人工智能技术风险研究具有现实意义和理论意义。

1.1.2 选题目的与意义

人工智能技术在日常生活中应用方面众多，给人类生活带来了巨大的变化，也会社会环境带来未知的风险。目前对于人工智能技术带来的风险问题，还没有形成完善风险防范机制。希望通过本文关于人工智能技术的风险分析，能为人工智能技术在风险防范上提供有益的建议。本文从风险社会视角出发，分析人工智能技术风险相关问题，具有重要的现实意义。对人工智能技术发展现状的风险研究，有利于我们加强社会风险防范，构建新的风险防范机制，让人工智能技术向人类有利的方向发展。同时人工智能技术风险研究也可以加强风险社会理论实践，增加风险社会的理论意义。对人工智能技术风险问题研究，可以揭示现代技术在风险社会中的地位，有助于厘清当下对智能、人工智能的概念，将目前人工智能技术风险研究提升至哲学层面。

1.2 研究现状综述

1.2.1 风险社会理论研究综述

乌尔里希·贝克认为现代性是社会风险的根源，并且认为当今社会最大的风险是技术风险。[①]吉登斯进一步提出风险社会中最主要的风险不是自然风险而是人为制造的风险,[②]现代化和人为制造的风险是风险社会最为根本的特征。风险社会理论的构建，为现代风险问题提供了一个新的理论视角。

玛丽·道格拉斯和拉什认为贝克等人的风险社会理论存在缺陷，他们试图从文化角度出发认为风险可以作为一种文化符号，不同的文化群体，对待风险和接受风险是有差别的，因此构建了社会文化风险理论。[③]斯洛维奇与伦内，他们主张心理学风险社会理论。他们提出所有风险都是个体对风险的“主观建构”，是人们主观认识上的风险，并且提出了“心理测量学”对人类关于风险感知的差别，所有人关于风险的感知都会受自己的主观意识影响，公众与科学家都一样。[④]另外一个主流视角，卢曼从复杂系统角度去认识风险社会，他认为当代社

[①] 乌尔里希·贝克(Ulrich Beck)著.风险社会[M].何博闻译,南京:译林出版社,2004.
[②] 安东尼·吉登斯.失控的世界[M].周红云译.江西:江西人民出版社，2001.
[③] 斯科特·拉什.风险社会与风险文化[C].李惠斌编,全球化与公民社会,桂林:广西师范大学出版社,2003
[④] 保罗·斯洛维奇，风险的感知[M]，赵岩东译北京：北京出版社，2007,20.

会是偶然性的、假设的和自相矛盾的社会。我们生活在一个没有选择的社会中，不得不面对风险。同时他强调避免风险要增强反省，并采取克制的态度通过各职能部门的调整来适应社会系统的分化和自治。[①]

国内对风险社会的研究起步于本世纪初，随着贝克和吉登斯等人关于风险社会著作的出版，学术界开始对风险社会理论进行多维度的研究。贝克在二十一世纪初到中国的交流会议，也促进了国内对于风险理论的相关研究。近年来，对于风险社会的研究可以分为以下几种。

第一，探究风险社会理论，杨雪冬教授通过系统分析贝克、吉登斯相关理论，为我们认识当下所处的历史阶段和面临的问题提供很好的理论基础。[②]同时杨雪冬还出版了《风险社会与秩序重建》，从风险社会理论角度对当下的风险与秩序问题进行详细分析，书中详细地介绍了风险社会理论，并通过对风险社会与秩序的理论建构，最后运用于解决我国制度建设中存在的问题。[③]

第二、从其他哲学理论视角分析风险社会，庄友刚教授从马克思主义视角探讨风险社会，发表了一系列文章如《风险社会与反思现代性：马克思主义的批判审视（2004 年）》[④]，《从马克思主义视野对风险社会的二重审视（2004 年）》[⑤]，及相关书籍《跨越风险社会》（2008 年）等，在书中作者站在历史唯物主义立场，对当下风险社会理论进行考察，试图运用历史辩证法去揭示风险社会的历史本质和历史规律，构建一个系统的历史唯物主义风险社会理论。[⑥]杨海博士同样从马克思主义哲学出发，通过对国外风险社会根本原因的系统分析，试图提出合理的风险治理建议，为中国面临的风险问题提出合理的应对策略[⑦]。

第三，基于风险社会理论背景下，研究社会各个方面问题，如风险社会中的法律问题。劳东燕从风险社会和刑法体系中的安全问题，探讨了刑法相关理论的变动[⑧]。孙粤文关对当下国内公共安全治理面临的众多问题，从社会角度进行分析，试图通过大数据来为公共安全治理提供新的技术方法。[⑨]毛明芳则是通

① 尼克拉斯·卢曼著. 信任:一个社会复杂性的简化机制[M].瞿铁鹏,李强译,上海:上海人民出版社,2005.
② 杨雪冬，风险社会理论述评[J]，国家行政学院学报，2005(01):87-90.
③ 杨雪冬，风险社会与秩序重建[M],社会科学文献出版社，2006.
④ 庄友刚，风险社会与反思现代性:马克思主义的批判审视[J],江海学刊,2004(06):38-42.
⑤ 庄友刚，从马克思主义视野对风险社会的二重审视[J],探索,2004(03),131-134.
⑥ 庄友刚，跨越风险社会:风险社会的历史唯物主义研究[M]，北京：人民出版社，2008.
⑦ 杨海，风险社会的哲学研究[D]，中共中央党校，2014.
⑧ 劳东燕，风险社会与变动中的刑法理论，中外法学，2014(01):70-102.
⑨ 孙粤文，大数据：风险社会公共安全治理的新思维与新技术，求实，2016(12):69-77.

过风险社会理论对现代技术风险问题系统分析并提供合理的风险解决和规避措施。[①]

本文是基于风险社会视角下进行的研究，对人工智能技术风险进行分析，风险社会相关理论对人工智能技术风险成因及规避措施提供了理论基础。

1.2.2 技术风险研究综述

技术风险是风险社会的主要风险，也是现代社会中最主要的一种风险。现代社会人们通过技术改变世界，同时也给世界增加了风险。西方国家对于技术风险研究关注比较早，18 世纪卢梭在《论科学与艺术》中人性论角度批判了科学与工艺，他认为人的这种美德“被人造物和文明的习俗掩盖了”，以至于损害了人性的自然状态。马尔库塞《单向度的人》一书中从社会批判与文化批判出发，认为现代技术社会的单向度化，无所不在的技术控制会导致“单向度”的病态社会。[②]1981 年召开的“新技术风险”为主题的全球会议，提出技术风险这个概念，随后出版了一系列关于技术风险的书籍与文章。1984 年，查尔斯·佩罗他在《高风险技术与“正常”事故》[③]一书中，从技术的复杂性和紧密配合性的角度分析了技术系统事故的成因。美国学者 H·W·刘易斯《技术与风险》[④]，从认识论角度谈到了风险的测量、认识、评估和管理等问题。

国内对技术风险的研究从上个世纪 90 年代开始。随着科技的发展，对于技术风险的研究日益增多，试企图创立“技术风险学”。近期国内学者对技术风险的研究大致可以分为以下几类：其一，对技术风险的本质的考察，比如王世进在 2012 年发表的论文《多维视野下技术风险的哲学探究》，他从技术风险的本体论、技术史、唯物史观、内外在根源和社会建构功能等多维视角下对技术风险展开探讨，增加了多视角分析技术风险的哲学研究。[⑤]其二，探讨技术风险的产生根源及对策方面的研究，毛明芳博士关于《现代技术风险的生成与规避研究》，她从主观建构和客观形成两个角度分析了现代技术风险形成的原因，并通过提出有效的建议措施以防范技术风险。[⑥]其三，关于新兴技术风险的研究，如

[①] 毛明芳，现代技术风险的生成与规避研究[D]，中共中央党校，2010.
[②] 马尔库塞,单向度的人[M],李继译,上海:上海译文出版社,1989.
[③] [美]查尔斯·佩罗.高风险技术与"正常"事故[M].寒窗译，北京：科学技术文献出版社，1988.
[④] [美] H.W·刘易斯，技术与风险[M]，中国对外翻译出版公司，1990.
[⑤] 王世进，多维视野下技术风险的哲学探究[D]，复旦大学，2012
[⑥] 毛明芳，现代技术风险的生成与规避研究[D]，中共中央党校，2010

刘中梅博士关于《公众参与纳米技术风险沟通的影响因素研究》，通过风险认知、科学素养、风险放大等理论，对公众关于纳米风险沟通的影响因素进行分析，试图提出有利于公众与纳米技术风险沟通的建议。[①]虽然技术风险研究时间很短，但还是有很多值得借鉴的地方，可以为本文在探究人工智能技术风险提供研究思路和理论依据。

1.2.3 人工智能技术风险研究综述

人工智能技术作为新兴技术之一，应用领域众多，产生的风险范围更为广泛。对于人工智能技术风险研究具有重大意义。关于人工智能技术风险问题的反思，最为著名的就是美国阿西莫夫提出的“机器人三定律”[②]，这个定律后来成为学术界一致公认的研发原则。之后，维纳在 1950 年写作的《人有人的用处：控制论与社会》[③]一书中，提出“坤之轮思想”，他认为自动化技术或机器人技术很有可能会造成“人脑的贬值”。维纳提供一个全新的社会视角及方法去解决社会问题（自动化机器方面），而且还对未来社会发出了警示，预示人类可能要面临的问题并提出应对的措施。1967 年，美国哲学家芒福德在其著作《机器的神话》[④]中以人文主义者的立场，表达了对自动机器应用看法。他认为机器不仅会剥夺了人类工作的权利，而且还会消灭个体的差异性，使人类面临丧失个性的威胁，从而使整个社会机制变得机械化。

随着人工智能技术的发展越来越多的研究人员加入其中，国内许多学者关于人工智能技术风险的看法也是不同的。杜严勇教授从伦理角度讨论人工智能技术相关问题，他从自反性伦理治理概念的基础上，分析了人工智能伦理及其治理问题。他认为要实现自反性治理需要提高理性和认知，培养行动者的治理能力，并且注重治理过程中的开放性与多样性。[⑤]王东浩博士通过梳理机器人在应用中伦理问题，从研究路径角度对机器人问题进行深入探讨，并且试图从文化和宗教的角度阐释应用伦理学对于机器人伦理的影响。[⑥]徐英瑾教授从人工智

[①] 刘中梅，公众参与纳米技术风险沟通的影响因素研究[D], 大连理工大学,2016
[②] 机器人三定律：第一、机器人不得伤害人类，或看到人类受到伤害而袖手旁观；第二、在不违反第一定律的前提下，机器人必须绝对服从人类给与的任何命令；第三、在不违反第一定律和第二定律的前提下，机器人必须尽力保护自己。
[③] 维纳.人有人的用处:控制论与社会[M].陈步译,北京大学出版社,2010.
[④] 刘易斯•芒福德.机器的神话[M].宋俊岭等译.中国建筑工业出版社出版,2009.
[⑤] 杜严勇，论人工智能的自反性伦理治理[[J]，新疆师范大学学报（哲学社会科学版），2018(02):111-119
[⑥] 王东浩,机器人伦理问题研究[D],南开大学，2014

能军事化用途角度对人工智能在军事领域的五个危害性论证，分析了人工智能军事化可能会造成的伦理问题，最后得出要使军用人工智能满足人类现有价值体系，重点在于开发具有人类意义上的伦理推理能力的军用人工智能。[①]陈晋的《人工智能技术发展的伦理困境研究》论文分析人工智能技术产生到应用所带的伦理问题，对人工智能技术产生伦理问题，最后利用马克思主义理论视角提出了相应的解决方案。[②]吴汉东从法律层面考虑人工智能带来的风险问题，对于人工智能技术引发的现代性的负面影响提出了法律上的建议和措施。认为应该以社会法理体系对人工智能发展带的风险进行控制，制定以法律为主导的国家层面的发展战略。司晓等人通过对具体人工智能技术在自动驾驶汽车以及智能机器人上的应用所带来的威胁，讨论了人工智能的民事责任的归属问题。

通过对国内的研究的分析，我们发现人工智能技术风险的研究比较片面，大部分只是考虑人工智能技术的一部分或者几部分风险问题。因此对人工智能技术风险进行全面的研究分析很有必要。

1.3 研究思路与研究方法

1.3.1 研究思路

本文的研究思路，本文以风险社会理论和技术风险理论为基础，对人工智能技术风险进行分析。通过研究人工智能技术风险问题提出了一些可行性规避方案。全文可分为以下几个方面：

第一部分，说明选题的背景及选题的目的与意义，归纳分析了风险社会、技术风险以及人工智能技术风险相关文献的研究现状，表明了自己的研究思路和研究方法。

第二部分，先对风险社会、技术风险的相关概念进行阐述，通过前面的理论铺垫分析了人工智能技术风险相关内涵，并阐述了人工智能技术风险的特征。

第三部分，通过对人工智能技术风险所造成的不同层次的后果的详细分析，进一步解释人工智能技术风险的内因以及外因，从本体论、认识论和社会层面等角度对人工智能成因分析，说明了人工智能技术风险产生是它的内在属性。

[①] 徐英谨. 技术与正义:未来战争中的人工智能[J]，人民论坛·学术前沿，2016(07):34-53.
[②] 陈晋，人工智能技术发展的伦理困境研究[D],吉林大学，2016.

第四部分，根据成因分析提出具体的风险防范措施，帮助解决人工智能技术风险问题，发展有益人类的人工智能技术。

1.3.2 研究方法

为了对人工智能技术风险进行研究，本文主要采用了以下几种研究方法：第一，文献分析法：通过查阅并收集人工智能技术风险相关文献，充分归纳分析人工智能技术风险相关文献，并总结为论据进行论证。第二，概念分析法：通过对人工智能技术风险相关概念进行梳理分析，厘清人工智能相关概念，有助于规避概念模糊而导致的错误。第三，采用举例论证和应用论证等方法，在进行理论阐述时，运用具体的案例对人工智能技术风险的具体问题进行说明，更加清晰的表达出要论证的观点。

第 2 章 风险、技术风险和人工智能技术风险

在对人工智能技术风险相关问题进行分析之前，需要厘清风险社会、技术风险以及人工智能技术风险的概念，为后文阐述风险成因以及提出建议措施提供理论依据。

2.1 风险社会的内涵

2.1.1 风险的概念

什么是风险？从字面意义来看，风险意味着不确定性，是一种可能的、潜在的危险。另一方面，风险是相对人而言，离开人的风险都谈不上是风险。自然风险和人为制造的风险都是相对于人而言，在哲学层面上讲，风险是人的一种存在状态。过去我们对风险的认识都是可以直观感受的，通过感官判断这些风险是否会对人有危害，现在的风险都是很难感知的，增加了我们对风险感知的不确定性。随着技术的发展，风险的威胁程度也增加了，无法预知的风险成为了当代社会人们关注的焦点。某些情况下这些风险不会对当下的人产生作用，而是威胁到他们的后代，即使在这种情况下，人们都迫切需要识别并使之变成可见和可解释的危险。

同时要注意风险与危险的区别，不能将风险等同于危险。吉登斯说，“风险与冒险或者危险是不同的。风险指的是在与将来可能性关系中被评价的危险程度。”[①]而危险是现在可知的，具有威胁性的存在；风险意味着不确定性，潜在性和可能性。

[①] 安东尼·吉登斯，失控的世界[M].周红云译.江西：江西人民出版社，2001，18-19

2.1.2 风险社会的概念

贝克最早提出“风险社会”用以区别于过去的社会。他认为当今社会具有一种现代化的特征，现代化这一特征标志着现代社会进入了风险社会。现代风险与以往的工业社会风险相比具有一种特殊性，解释这种特殊性关键在于区分外部风险和被制造出来的风险。吉登斯对这两种风险作了说明，外部风险就是来自外部的传统的不变性和固定性，而被制造出来的风险就是指我们在没有历史经验的情况下利用现有的知识对这个世界所做出的改变之后产生的风险。[①]风险社会更多是被制造的风险占据主导地位，区分是否进入风险社会关键在于，我们目前担心的风险是趋向于外部风险还是被制造的风险，当被制造的风险占据主导地位时就标志着我们进入了风险社会。

随着科技的发展，越来越多新的风险出现，技术成为推动当今社会进入风险社会的最大动因，因此风险社会的主要特征表现为技术风险。新科技的出现不仅给社会带来了便利，同时也增加了风险。因此对新兴技术风险研究，有利于对社会风险的把握，同时也有利于将现代社会发展成对人类有利的方向。

2.2 技术风险的内涵

2.2.1 技术的概念

什么是技术？历来不少哲学家对技术都下过定义，第一种是工具论，“认为技术是一种手段和人类行为，可以被叫做工具的和人类学的技术规定”[②]。第二种是目的论，海德格尔是从技术的目的性对技术进行定义，他认为“技术是人类有意识地改变或控制客观环境以满足自身需要的手段或活动。”[③]第三种是从人性出发认识技术，芒福德是从人的角度出发考虑技术的本质。[④]因此技术是作为人一种存在方式。广义上讲，技术是指人类有意识改造自然、改造社会和改造人自身的全部活动过程中所应用的一切手段、方法、知识等活动方式的总和。

[①] [英]安东尼·吉登斯,失控的世界[M].周红云译.江西:江西人民出版社,2001,22
[②] 吴国盛，技术哲学经典读本，上海：上海交通大学出版社，2008.301
[③] 殷正坤,试析技术的本质[J],天津社会科学,2001(4):41-44
[④] [美]刘易斯•芒福德.机器的神话[M]l.宋俊岭等译.中国建筑工业出版社出版，2009.

2.2.2 技术风险的内涵

（1）技术风险概念

技术风险是由技术带来的风险，是技术在应用过程中产生的对人的危害以及对其他生命体和环境的危害。现代技术相对于传统技术来说，带来的风险特征也是不一样的。首先，现代技术风险更多的是整体的，会影响到整个人类社会根本发展，比如核技术风险、基因技术风险等。其次，现代技术风险是全球性的，由于信息技术的发展，使得技术风险会造成全球范围的危害，比如电脑病毒等。

（2）技术风险的分类

我们以技术本身是否会有风险为标准进行分类，现有的技术风险类型大致分为三类。

第一类，认为技术本身是无风险的，这一类是技术乐观主义有的看法，认为技术本身不会存在任何风险，即使技术有风险也可以通过自己的发展而消除。拉普拉斯的妖理论就是典型代表，他认为世界的未来是可预知的并且完全确定性的。所以在拉普拉斯看来，世界是完全可以控制的，是不会产生任何风险的。

第二类，认为技术是中立的，本身是无风险的，风险产生于技术的应用，并且通过有效规避就可以避免。技术不带有任何善恶之分，技术仅仅为了达到人类目的的一种手段或工具体系，这种工具会产生什么样的作用，是对人类有益或者有害，都源自于人类对这种工具的运用。

第三类，认为技术风险包含在技术本质属性中，我们认为技术风险是包含在技术本质属性之中的。风险不仅是当代社会的重要特征，也是现代技术的重要属性。贝克提出风险社会，是因为现代化所带的风险，而技术是实现现代化的手段，因此现代技术在社会发展中不仅表现为积极的力量，推动社会发展，而且也表现为消极的力量，蕴藏着巨大的风险。吉登斯指出，“技术进步表现为积极力量，但它并不总是如此。科学技术的发展和风险问题紧密相关。”[①]所以由此可知，风险并不是外在于技术的社会运行特征，而是包含在技术的内在属性之中。

[①] 安东尼·吉登斯，第三条道路及其批评[M]，北京：中共中央党校出版社，2002，139.

2.3 人工智能技术风险

人工智能技术作为技术之一，同样也会产生风险，而人工智能技术因为它应用层面广，影响程度深，所以当下对于人工智能技术风险研究很有必要。它的发明与发展给现代生活带来了质的飞跃。

在对人工智能技术风险展开讨论之前，还需要厘清几个基本概念。其一：智能，人类智能与人工智能的概念？其二：人工智能技术及其应用，其三：人工智能技术风险，以及它的特殊性。

2.3.1 人工智能的内涵

（一）智能

什么是智能？平克的心智计算理论将心智定义为：“心智是一套由计算器官组成的系统，它经自然选择的设计来解决我们祖先在茹毛饮血的生活中所面对的那类问题，具体包括：理解和操纵物体、动物、植物以及他人。”[①]他认为心智不会仅仅是我们神经元简单逻辑计算后的结果，它还要与复杂的外在环境交互，它虽然也针对特定问题，是问题求解一种方式，但我们解决特定问题或达成特定目标往往不仅仅求助于心智。这种心智或心智计算理论是为人类所独有的突出表现是我们的语言，这种表现是我们与其他动物差别的最为明显的特征。

智能不是简单的语言就定义的，丹尼尔·丹尼特在《心灵种种：对意识的探讨》中写道：“我们是怎么知道你有心灵的呢？因为任何能够理解我的话的人就自动为我所用的代词‘你’所称谓，而唯有具备心灵之物能够理解这些话。”[②]我们通常都是不假思索地通过语言来理解他人的心灵，语言可以使我们进行合理对话参与到政治生活中来，同时我们也可以通过语言进行一些欺骗性活动，但语言并不构成心灵的必要条件。

日常我们所说的“智能”是智慧与能力，从感知、记忆再到思维的过程，可以称之为“智慧”，智慧的结果产生了语言与行为，再通过语言和行为表达出来的过程，就称为“能力”，将智慧与能力结合使用就是可以称之为“智能”。

（二）人类智能与人工智能的区别

[①] 平克. 心智探奇:人类心智的起源与进化[M]. 郝耀伟,译. 杭州:浙江人民出版社, 2016:22.
[②] 丹尼尔·丹尼特. 心灵种种：对意识的探索[M]. 上海:上海科学技术出版社, 2012:8.

最早明确提出人类智能与人工智能区分的是笛卡儿，他在《谈谈方法》这本书中，将人称作是“神造的机器”，并且认为人类是不可能制造出与人类智能比肩的机器，即使人造机器设计的非常巧妙，做出的动作跟人类非常相似，也都不能跟人类相比。为了说明机器与人类不会一样，他还用了两条标准去解释机器不是真正的人。第一条标准是：它们不能像人类一样去使用语言，并且通过语言表达自己的想法；第二条标准是：机器人不能像人类一样去做所有的事，即使能在许多事都做的比人类好。[①]其次，莱布尼茨也通过“莱布尼茨磨坊”思想实验去证明机器不能具备人才有的思想和知觉，他把机器内部运作比作磨坊，通过观察磨坊里的动作，去试图找出一些有关思想或者知觉的影子。他通过“单子”和“自然机器”两个核心概念去表达机器不具备知觉是因为无法组合成只有人类才能组成灵魂的高级单子。

人工智能从一开始就是为了模拟人类智能并希望能实现智能，人类智能是否能模拟本文不作阐述，但是从人工智能发展的现状来看，人工智能和人类智能还是存在很大差异的。从智能本质来看，人工智能并不具有人类智能。比如金观涛在“反思人工智能革命”[②]，通过对智能的本质思考给出了一个人工智能如何才能实现的几个条件，首先，主体是自由的，其次，能用语言传递知识，组织社会，产生社会行动，第三，主体能意识到自己是有自由意识的，并能通过意志创造出应然世界，最后，在创造应然世界时还能进一步放大主体自由。人工智能只有在满足以上所有条件才能到达人类智能。所以目前人工智能与人类智能还存在本质的区别，人工智能只是实现了人类智能的部分延伸与模拟。

（三）人工智能定义及程度区分

在上文我们已经指出人类智能与人工智能存在本质不同，也为我们区分人工智能定义做了一个很好铺垫。由于目前还没有建立统一的人工智能标准，加之人工智能所涉及的学科众多，所以对人工智能的定义也是各有不同。1956 年约翰·麦卡锡第一次提出了“人工智能”，学界从此兴起了人工智能的相关研究，并对人工智能进行定义。斯图尔特·拉塞尔（Stuart Russell）在《人工智能：一种现代方法》中，通过 4 种途径“像人一样行动”，“像人一样思考”，“合理地思维”和“合理地行动”[③]来定义人工智能。并通过 Agent（智能体）介绍了各

[①] 笛卡儿.谈谈方法[M].王太庆译,北京：商务印书馆,2000，43-44.
[②] 金观涛.反思人工智能革命[J],文化纵横,2017(8):20-29.
[③] 斯图尔特·拉塞尔等著,人工智能：一种现代方法[M].北京：清华大学，2013，3-4

种关于人工智能方法。日本东京大学大学院工学系研究副教授松尾丰，在他的《人工智能狂潮：机器人会超越人类吗？》一书中，谈到日本许多知名人工智能专家关于人工智能的定义。比如公立函馆未来大学校长中岛秀之认为，人工智能是“采用人工方法制造的、具有智能的实体，或者以制造智能为目的的、对智能本身进行研究的领域。”[①]北陆尖端科学技术大学院大学教授沟口理一郎认为，人工智能是“采用人工方法制造的、能做出智能行为的东西（或系统）”。[②]而松尾丰教授认为人工智能是：“采用人工方法制造的类人智能，以及其制造技术。”[③]关于人工智能的定义众说纷纭，从广义上讲，“人工智能是关于人造物的智能行为，而那个行为包括知觉、推理、学习、交流和在复杂环境中的行为。人工智能的一个长期目标发明出可以像人类一样或能更好完成以上行为的机器；另外一个是理解这种智能行为是否存在于机器、人类或其他动物中”[④]本文采用这个定义。

目前学术界对 “强—弱人工智能”的界限也是模糊不清。但厘清智能程度，对后文的人工智能技术所带来的风险程度，以及区分人工智能技术真正的风险及后果都有重要意义。关于强弱人工智能的界限划分对本文意义重大，我们采用塞尔的定义。强人工智能和弱人工智能最早是由约翰·塞尔提出来的，他在《心灵、大脑与程序》这本书中，定义了强—弱人工智能。塞尔认为，弱人工智能（弱 AI）可以表达为：“计算机在心灵研究中的主要价值是为我们提供了一个强有力的工具。”[⑤]弱人工智能只能是为我们在研究心灵时的一个工具并不具备其他认知。而强人工智能（强 AI）可以表示为：“计算机不只是研究心灵的工具，更确切地说，带有正确程序的计算机确实可被认为具有理解和其他认知状态，在这个意义上，恰当编程的计算机其实就是一个心灵。”[⑥]在很大程度上，生活中我们所说的强人工智能其实都只是弱人工智能，也就是目前所说的通用人工智能（AGI）。

我们以塞尔关于“强—弱人工智能” 划分的智能程度，对目前人工智能技术所实现的智能进一步做出以下三种说明：其一，目前的人工智能离塞尔定义的强 AI 还有很远的距离，而强人工智能可能带的风险我们只能先做一个前期的

[①][日]松尾丰著,人工智能狂潮:机器人会超越人类吗?[M],赵函宏,高华彬译.北京:机械工业出版社,2015,25
[②][日]松尾丰著,人工智能狂潮:机器人会超越人类吗?[M],赵函宏,高华彬译.北京:机械工业出版社,2015,25
[③] [日]松尾丰著,人工智能狂潮:机器人会超越人类吗?[M],赵函宏,高华彬译.北京:机械工业出版社,2015,26
[④] [美]Nils J·Nilsson 著，人工智能，郑扣根等译，北京：机械工业出版社，2000,1.
[⑤] [英]玛格利特·博登,人工智能哲学,刘西瑞,王汉琦译,上海:上海译文出版社,2001,92.
[⑥] [英]玛格利特·博登,人工智能哲学,刘西瑞,王汉琦译,上海:上海译文出版社,2001,92.

风险分析；其二，塞尔定义的弱人工智能[①]，实际上是我们说的通用人工智能（AGI）程度，当下的人工智能程度也还尚未达到，但是人工智能学界都希望可以实现 AGI；其三，目前人工智能技术所实现的智能只是某个方面的人类心智能力，虽然目前人工智能技术只是模拟人类智能的某些方面，但是由于人工智能技术的特殊性，加之现代其他新兴技术的兴起并与之结合，所以目前人工智能技术所造成的风险也是不容忽视的。

2.3.2 人工智能技术

（一）人工智能技术概念

维基百科将人工智能技术定义为“普通计算机程序的手段实现的类人智能技术”[②]。换句话说就是人工智能技术模拟人类智能的技术。人工智能在实际领域的技术，比如专家系统、人工生命、模式识别、定理证明、机器人学、机器学习、自动程序设计、自然语言处理、问题求解、人工神经网络、智能决策系统等。

（二）人工智能技术应用的领域

人工智能技术是一门综合性很强的学科，加上心理学、生理学、数学、哲学等学科的发展与贡献，其研究和应用领域也变得越加广泛，并且随着时间的推移，各种新思想、新理论、新技术层出不穷。人工智能技术发展到目前，其主要应用领域有人工智能技术具体在生活应用领域很广，在工业，服务业，自动驾驶，医疗及行业，金融行业，教育行业，法律行业，军事方面等等都有涉及应用。

2.3.3 人工智能技术风险

（一）人工智能技术风险

人工智能技术作为一种高新技术，人类通过人工智能技术来改造社会同时，也给人类社会带来了风险。吉登斯说过，技术的进步表现出来的力量并不总是积极的，科学技术的发展和风险问题都是紧密相关的。比如 Uber 自动驾驶汽车

[①] 注释：为了防止后文说的“弱人工智能”概念与塞尔定义的“弱人工智能”概念搞混，本文将塞尔定义的“弱人工智能”称为“通用人工智能”，后文说的弱人工智能只是模拟人类心智能力一部分的表达。
[②] 注释：该词解释来源于维基百科

致死案例[①]，自动驾驶车辆会大范围的改变人们日后的出行方式，设计的初衷是减少驾驶的车祸风险，并可以有助于社会出行方便。但是由于技术设计还不够完善导致不幸，让人类开始思考人工智能技术在给社会带来便利的同时，它可能带的风险问题。正如斯坦诺维奇在《机器人叛乱：在达尔文时代找到意义》中说的“等你真正拿到那把剑的时候，你才发现，那把剑不那么好使，它甚至会伤着你。”[②]虽然目前的人工智能技术并未像该书的作者所说的机器人那样会思考去叛乱，改写主人的命运，调用更多的认知资源为自己所用，就连塞尔所定义的弱人工智能还未实现，但是人工智能技术给社会带来的风险是的的确确存在的。人工智能技术在发明及应用时就可能给社会带来巨大的风险。

（二）人工智能技术风险的特殊性

人工智能技术风险之所以会给人类社会带来巨大的风险，在于它的技术本身的一个特殊性。第一，人工智能技术应用领域众多。它可以在人类社会的方方面面都有应用。在工业，服务业，自动驾驶，医疗行业，金融行业，教育行业，法律行业，军事方面等等都有及应用。人工智能技术应用层面广会导致风险程度的扩大。第二，人工智能技术的智能性。智能性是其他技术所没有的，即使像核技术这样可以对人类造成巨大危害的技术，智能性是人工智能所特有的。人工智能技术是模拟人类智能的技术，所以它会取代人类部分岗位的，由此会导致社会结构的变化，社会结构变化可能会导致人类脑力劳动逐步被智能机器所取代。第三，人工智能技术的不可预测性。由于目前人工智能技术发展还处于初级阶段，是否可以实现通用人工智能，甚至是强人工智能，还不可预见，所以未来可能带来的风险程度也是不可以预见的。

综上所述，人工智能技术的应用层面广、智能性和不可预测性足以使人们将人工智能技术风险置于特殊位置，这也是为什么业界人士对人类智能风险一直争论不休的根本原因。

[①] 新浪新闻中心，关于无人驾驶汽车撞死人系首个自动驾驶车撞死人案例，[EB/OL]. http://news.sina.com.cn/w/2018-03-20/doc-ifysmpev5202955.shtml

[②][加]斯坦诺维奇[著]，机器人叛乱：在达尔文时代找到意义，吴宝沛译.北京:机械工业出版社，2015.5

第 3 章 人工智能技术风险的后果及成因

风险社会理论表明，人工智能技术的发展，必将产生一系列风险。本章可以分为两个部分。第一，对人工智能技术风险的后果分析，如对社会秩序、社会伦理、未来人类生存发展的威胁。其二，对人工智能技术成因分析，可分为内在因素和外在因素，从本体论、认识论和社会层面对人工智能技术风险内因分析；再从技术理性与社会断裂以及政治因素、制度缺少等社会因素分析人工智能技术风险的外因。

3.1 人工智能技术风险的后果

3.1.1 对社会秩序的威胁

（一）失业问题

从第一次工业革命开始，机器取代手工，将农业社会推向工业社会，开启了人类用技术解放双手的历程，从而导致失业问题。之前两次工业革命都是用机器取代工人的体力劳动，把人类从体力劳动中解放出来，但现在人工智能技术应用是取代人类脑力劳动，这是与前两次产生的最大不同。与工业革命相比，人工智能带来的革命程度更深，人工智能技术在日常生活中应用广泛，在各个领域帮助人类进行分析、判断和决策。脑力劳动的大量解放，必然会导致人类大面积的失业。因此大量的机器应用必将对我们社会关系产生影响，增加社会的不安定因素。

智能机器的普及究竟导致什么后果，以制造业为例，大量廉价工人推动了制造业的发展，但是随着智能机器的使用，大量的工人将面临失业风险。特斯拉汽车公司作为硅谷东部弗利芒特市最大的汽车配件厂，很少看到工人装配汽车，都是使用机器人取代人类在流水线作业，这不但不会影响汽车个性化设计。特斯拉的案例说明，在未来的制造业中，大量机器将取代人类在制造业中的岗位，还大大降低了成本。但是由智能机器取代人类之后，大量的劳动力又该如

何安排？智能机器越来越多地剥夺人的工作机会，将造成很多社会问题。比如失业人员难以生存，社会财富更加集中，财富分配更加不均，从而加剧社会不平等。未来社会的无用之人越来越多，人生存的意义就如何定义，或许新兴的岗位可以满足部分人，但是大范围的机器取代人类工作，是否可以产生新兴行业并提供足够多的就业机会以满足失业人员的需要，将是一个极不确定的问题。

（二）公共安全威胁程度放大

人工智能技术滥用将成为未来公共安全最大的威胁，并且还是最难以控制的威胁。人工智能技术被一些恐怖分子利用，对于恐怖袭击更加方便，只需要利用一些人工智能新产品以及自主功能的机器就可以实现。早前伯克利开发的新技术，利用人工智能技术攻击文本—语音转换系统，只需通过该人工智能技术，就可以改变任何想改变的语音，利用这个方法任何音频都可以变成攻击者想要输出的文本。如今世界到处都是智能音响和语音助手，新的人工智能技术无疑给社会公共安全带来巨大的隐患。另外在 2017 年，浙江警方破获了一起利用人工智能技术侵犯大众的个人信息案。黑客们利用深度学习技术对机器进行训练，让机器可以实现自主操作，批量识别验证码。民警当时说这些黑客利用人工智能在很短时间就可以识别上千上万个验证码。[①]由此可见，目前利用人工智能技术犯罪的门槛太低，给人类社会带来一记预警，提防人工智能技术滥用导致社会公共安全威胁很有必要。

（三）军事化威胁

自动武器应用是众多威胁中最值得注意的方面，由于军事需要，人工智能将用于生产更高效的杀人武器，比如通过人工智能技术来操作无人机进行攻击，以前是由人来做出决定攻击目标的，现在是由人工智能技术操控无人机，去指出攻击目标就可以自动锁定目标并执行攻击，这是将杀人的权力让渡给一个机器。其中，令我们不安的一点是，普通人或者团体都可以控制自动武器进行攻击，这样会极大增加预防难度，并会造成更大伤害。然而，发明生产的商家不会去承担这个责任，研究这项技术的科学家也会表示对此不负责，这无疑是给社会带来一个难题——责任归属的问题，由于问责的过程不明确，导致技术的滥用得不到根本上的控制。再者，加上国与国之间的竞争，谁都不愿意落后于其他国家，因此，自动武器的开发不但不会止步，而且还会加快。

[①] 人民网，关于当人工智能成为“矛”,“盾在哪里”[EB/OL]. http://finance.people.com.cn/n1/2018/0226/c1004-29834825.html

3.1.2 对社会伦理的威胁

（一）生命不平等问题

在现有的技术发展过程中，还要堤防一种威胁，就是对人类生命不平等的威胁。前面所说的风险都是威胁社会秩序安全，但是随着人工智能技术的发展，还会在生命平等上造成威胁。我们都知道现代医疗水平的提高，使得现代人类生命大幅度的延长。越来越多的技术比如器官移植、再生医学、基因工程以及纳米机器人等新技术，将大幅度的改变人类生命年限。最近不少专家认为人与机器的结合的“赛博格”可能才是最为可怕的威胁，人可以通过人工智能与生物技术和仿生技术制造出人工器官，并替代人类衰竭的器官，由人与机器的结合产生的新的生命体“赛博格”，但其高昂的费用，可能只能被少数极端富裕的人享用，并由此也会导致政治生活上的不平等。

3.1.3 对未来人类生存发展的威胁

随着人类对人工智能技术研究，人工智能或许有望实现通用人工智能或者强人工智能。腾讯研究院在 2017 年开展了一次关于人工智能相关问题的问卷调查，参与问卷的 2968 名人士是来自人工智能直接或间接相关的研发人员、技术人员、产品人员、法律政策与人文社科研究者。据研究调查，不论是从事人工智能技术研究人员、产品设计者还是人文社科的学者或者其他行业的人员，大部分都对强人工智能能够到来表示肯定，只不过在实现的时间上有差异，其中只有少数人员认为不会实现强人工智能[①]。虽然大部分人员认为人工智能不能控制人类，但是其中 33.5%的人认为人工智能是有可能控制人类的[②]。只要存在不确定性，就会产生风险。所以对强人工智能的风险问题，我们还是需要提前预防。

（一）人类主体地位的威胁

当人工智能技术达到通用人工智能程度时，我们是否应该考虑赋予人工智能人权。在未来强人工智能出现的时候，其实我们面临着三种情况。第一，当我们发明的机器具有与人类同等智能，就算机器不会产生与人一样的意识，但是很难保证在未来不会出现，这时人类的主体地位是否遭受威胁？如果人工智

[①]腾讯研究院等著，人工智能[M],北京：中国人民大学出版社，2017,16
[②]腾讯研究院等著，人工智能[M],北京：中国人民大学出版社，2017,15

能出现智能，并且这种智能与我们智能不一样，那是否会与人类产生冲突，对人类造成伤害呢？如果这种智能与人类智能一致，那是否会甘心被人类所统治？

为什么说当人工智能出现智能时，人类的主体地位就会受到威胁？以大数据下隐私权为例，大数据可能造成的是个人隐私的暴露，但更深层次反映了人的主体性的丧失。首先人作为主体，他具有主体性，而主体性又是人作为主体的根据和条件。这种主体性表现为自主、能动、自由、有目的地活动。当人工智能技术与大数据结合，广泛的个人数据暴露，导致个人隐私的曝光。“可以预见，有利海量隐私数据为学习材料，人工智能可以迅速获得‘解剖’我们的能力，进而以机器的方式和思维来规制我们。”[①]这样看来，人工智能就不只是我们人类能力的延伸了，更将会导致我们的主体地位的丧失。为了避免这种可能情况的发生，有必要开始对人工智能技术进行风险原因分析，尽量让它向着有益的方向发展。

3.2 人工智能技术风险的内在成因

3.2.1 人工智能本体论的不确定性

从本体论上说，人工智能是人造物的智能行为，是人类将科学知识与技术相结合试图模拟人类智能的一种技术行为。科学本身就具有不确定性，由于对科学知识的应用，造成技术也带有不确定性，这种不确定性是技术本身的内在属性，不会随着对技术认识深化而减少，也不会随着时间和空间的改变而变化。

牛顿作为经典力学的建立者，同时他也是确定性科学的代表人物。牛顿认为只要我们掌握物体的一个初始运动的状态，就可以精确的推断出世界将来会发生的任意状态。我们可以在经验范围内证实这种数学建构，我们获得科学知识的真理性。拉普拉斯的决定论被看作是典型的因果决定论，关于他的“拉普拉斯妖”理论，在他看来世界没有什么事物是不确定的，不确定是人类认识的不足，确定性是事物的客观属性。

[①] 罗岗等,基本收入·隐私权·主体性——人工智能与后人类时代（上）[J]，读书，2017（10）

随着量子力学与测不准关系的兴起，确定性观念已经被打破，越来越多的人认为科学是不确定性的。海森堡的“不确定性原理”（或者测不准原理）是量子力学的一个基本原理，该原理表明观察某个微观粒子时，不可能同时确定它位置和动量，其中的一个量值越是确定，另外一个量的不确定性程度就越大。普里戈金认为，在人类处于一种新的理论的转折的时候，科学就不再是确定的了，而是对现实世界复杂性和不确定性的反映。[①]

科学的这种不确定性，会导致技术的不确定性，人工智能技术也具有不确定性，表现在以下几个方面。第一，人工智能技术对常识的判断就具有不确定性。不确定性推理是人工智能的一个重要研究领域，贝叶斯网络是一种基于概率的不确定性推理网络。它是用来处理人工智能中的不确定性信息，虽然它对分析不确定性和概率性的事件、知识或信息中做出推理。但是并不可能完全降低这些事物本身的不确定性。第二，神经网络算法是人工智能技术最为重要的一种技术，由于算法的不确定性，增加人工智能技术不确定性的程度。神经网络算法不确定性是由于“算法黑箱”导致的，在我们给定一个输入输出值时，算法如何提出特征过程我们不得而知。第三，神经网络算法主要由数据推动，具有不确定性。第四，从数据背后的推动力量来看，主要是企业和资本的推动其向前发展，对于公众安全性问题，存在着不透明和黑箱操作的空间，增加了它的技术的不确定性。第五，还有一个不确定来源于数据本身，即使是同一个数据不同的人操作也会导致结果的不一样，所以人工智能技术本身的不确定性是构成人工智能技术风险一个重要原因。由于技术本身的不确定性导致结果的不确定性，结果的不确定性就导致风险的不确定性。所以人工智能技术在本体论上是不确定的，并且由这种不确定性就会导致风险的不确定性。

3.2.2 人工智能认识论的不可知性

由于目前人工智能技术存在很多不确定性和不可解释性，人类对人工智能在认知上还是存在不可知性。人类对事物发生、发展的过程及其规律的认识，是人类认识的本质。但是由于人工智能技术在本体论上是不确定的，而这种不确定性会增加我们认识和解释事物过程的难度。人工智能技术目前的所有功能主要通过算法来实现，依据算法实现功能会面临几个问题，首先，算法本身的

[①] [比]伊利亚·普里戈金，确定性的终结[M]，湛敏译，上海：上海科技教育出版社，2009.5

不确定性，目前的算法主要是机器学习和深度学习算法，而在输入输出的过程中，存在着一个算法隐层。在这个机器学习模型中间的隐层，我们并不能知道机器提取了什么特征，对于这个过程我们无法解释。所以在算法层面上存在了不可解释性。其次，在机器学习算法中还有一个重要分分支是神经网络算法，而神经网络算法是通过穷尽局部的采样特性和不明确的非线性来实现算法功能的。[①]第三，它的算法是由数据推动的，而深度学习又是穷尽经验采样分别的方法，增加了人工智能技术不可解释和不确定性。所以算法的“黑箱操作”的不可解释性是产生人工智能技术风险的内在成因之一。由于人工智能技术不确定性的本质，导致认识的不可知性，即使处于这样的一种情况，但是人工智能研究人员并不会因为不可知而放弃研究，反而，人类会增加人工智能技术研究，必将导致风险不可预测性加强。

3.3 人工智能技术风险的外在因素

3.3.1 技术理性与社会理性的断裂

理性是人性的一种展现，是人类社会的一种文化表现。理性分为技术理性与社会理性，技术理性是技术按着自己的逻辑发展，在条件允许的情况下，它会具有强大的技术力量，可以导致人的生活按照技术逻辑的方式发展。技术理性的无限发展情况下，会严重损害人的自由、道德、情感，使之成为单向度的人。社会理性则是人在社会中情感、道德、自由等人性的满足。

现代技术通过技术给人类社会带来无限便利，使人们忽视了来自技术的奴役。马尔库塞指出，“因为这种不自由既不表现为不合理的，又不表现为政治性的，而是表现为对扩大舒适生活、提高劳动生产率的技术装置的屈从”[②]技术成为了一种控制性的力量，这种力量存在于政治、经济、文化、思维方式等各个方面。然而由于技术给人带来一种满足的假象，它们为人们提供了舒适的生活，消解了人们对技术的反抗意识。人们根本感受不到技术异化带来的痛苦，反而很享受技术带来的生活上的便利。人工智能技术作为高新技术，它给人类社会带来的便利随处可见，比如手机上的人工智能技术，智能搜索引擎、语音识别、

[①] 熊红凯，人工智能技术下对真理和生命的可解释性[J],探索与争鸣，2017(10):14-15.
[②] 马尔库塞.单向度的人 [M].李继译,上海:上海译文出版社,1989,126.

图像识别等，让我们生活交流更便捷。手机给人们带来的依赖程度越高，技术异化的表现程度就越明显。技术理性使得人工智能技术人员尽可能想发展出更智能的技术，而忽视人工智能技术可能会导致人的主体地位的丧失。

现代社会的技术异化，源自于技术理性和社会理性的断裂。这种断裂不仅表现为人工智能技术对人的人格的损害，还表现在人工智能内部概念的模糊不清。人工智能研究是探究人类智能的一种方法，业界对于“智能”的界定模糊不清，恰恰反映出科学被技术异化和人文精神的丧失。[①]因为缺乏人文方面的思考，导致很多概念说不清，道不明。贝克曾说过“没有社会理性的科学理性是空洞的，但没有科学理性的社会理性是盲目的。”[②]

不少学者认为当下人工智能所带的社会问题都是因为人文反思跟不上科学技术的发展，由于科技发展速度太快，导致我们忽略了人文关于事物发展本质。比如关于哈佛一名计算机科学家讲述利用人工智能技术识别犯罪团伙，被问到在帮助警察尽快破案的同时，是否考虑过因为技术强大导致的后果时，他说“我只是一名工程师”。这个事件的背后我们可以看到更深层次的问题就是技术异化与人文精神的断裂，人工智能技术人员为了技术实现越来越强大的功能，而忽视了人文思考，从而使技术理性和社会理性的断裂加剧人工智能技术风险。

3.3.2 社会因素的影响

（一）社会群体对风险认知的影响

社会不同群体对风险感知能力是不一样的，对风险感知也是带有主观意识的，并且公众与科学家之间虽然在看待风险上会受到科学素养和公共教育程度不同的影响而产生偏差，但公众与科学家都会受到自身世界观、意识形态和价值观的影响。即使是最为权威的专家也会受制于自己的主观意识来看待风险。

1. 公众对风险认知的影响

公众对风险认知依赖于自己的直觉判断，往往对高科技的风险感到恐慌，即使该技术对人类产生的伤害的例子少之又少，但是人类对该科技认知和判断都很少，大多数的消息来自媒体或者自己的主观意识，通过这些心理、社会和

[①] 金观涛，反思“人工智能革命”[J]，文化纵横，2017(8)：20-29

[②] [德] 乌尔里希•贝克(Ulrich Beck)著,风险社会[M],何博闻译，南京：译林出版社，2004，30

文化等因素的作用将会造成风险的扩大。比如当人工智能驾驶汽车发生危险的时候，人类会对这种新兴的技术风险感到恐慌，对新技术发展表示担忧，由此在风险的社会影响上造成恐惧的心理影响。公众对人工智能技术风险的感知会间接地影响到风险判断和风险预防政策的制定。

2.科学家对风险认知的影响

一项关于人工智能是否有可能控制人类的调查报告显示，有 59%的人认为人工智能不可能控制人类，但是还是存在 33.5%的人认为有可能。在认为人工智能能够控制人类的人群中不太了解人工智能的人占 38.47%，有些了解和非常了解人工智能的人占 36.76%和 27.8%。[①]

首先要肯定的是，科学家对风险的感知也是会产生偏差的。科学家一样会受到来自自己的科学专业限制以及自己的主观意识、情感、价值取向等等影响。所以说科学家对风险的判断会直接影响到技术风险的判断与预防。错误的判断风险，会导致风险无法控制，危害人类。还有一种情况就是认知偏差可能会导致灾难的发生，技术人员知道该技术存在风险，但是他认为这个风险是可以控制的，其实并不能控制，由于认知的偏差就会导致灾难的发生。[②]

其次，关于风险管理会导致一种权力的追逐，当你有权定义一项风险时，你就可以通过对风险的定义进而提出相对的解决方案。“当你按一种方式定义风险时，某一种方式就会成为最有效的、最安全的或最好的方案。而当你按另一种方式（比如包括了风险定性特征和其他背景因素）来定义时，你很可能会对你的解决方案进行重新排序。”[③]当你对技术风险定义有一定权力时，你就可以将它定义成你愿意看的那种结果，比如做产品设计的技术人员，一般他们不愿意承认这项技术是具有风险的。因此这种在风险认知上的主观决定也会加剧技术风险的产生。

（二）政治的影响

[①]腾讯研究院等著，人工智能[M],北京：中国人民大学出版社，2017,

[②] Yudkowsky, Eliezer. Artificial Intelligence as a Positive and Negative Factor in Global Risk[J], Global Catastrophic Risks, New York: Oxford University Press. 2008. 308–345

[③]保罗·斯洛维奇，风险的感知[M]，赵岩东译北京：北京出版社，2007,20

政治是上层建筑领域中各种权力主体维护自身利益的特定行为以及由此结成的特定关系，是人类历史发展到一定时期产生的一种重要社会现象。因此政治也是一种权力的表现，而这种权力的表现在现代社会更多体现为技术的权力。技术发展不仅推动了经济的发展，更加体现了一个国家的实力，所以现代社会更多的表现为一种技术权力。由于这种国与国的政治影响，必然会导致技术的不透明和技术的垄断。当下是民族国家和全球资本主义推动的世界，政府的支持，资本的推动，人工智能已经不可能停止研究了。而这种权力的竞争也会影响技术资源配置，开发和研发以及技术应用上，掌握这种权力的政府和机构可以决定是否开发某一项技术，是让社会承受技术所带来的风险，还是避免。更有甚者可以利用人工智能的这种技术的特殊性，作为满足自己政治权力手段。

比如 Facebook 数据泄露事件，一家名叫 Cambridge Analytica 的公司违规窃取数据并利用大量选民数据，通过数据模型分析用户行为，帮助制定有利于特朗普竞选的口号，并帮助其赢得大选。这件事件反映出了几个问题，一方面反映了政治权力的使用不当，会使人工智能技术存在黑箱操作的可能，并使风险承受方公众承受更大的风险。另一方面，反映了制度的缺失，使得企业存在不良操作还得不到有效制止。

（三）制度的缺失

在新奥尔良举办的人工智能、道德与社会（AIES）会议上，哈佛大学的 Crowd Innovation Lab 发表一项研究关于“使用算法自动识别一起犯罪是否属于团伙犯罪”，其主要研究内容是关于利用一种新的算法，帮助警察识别在犯事的人是否是属于某个犯罪团伙。谷歌的软件工程师 Blake Lemoine 在会上提问说，是否有考虑过算法歧视，怎样让算法不产生偏见？如果某人被误以为是犯罪团伙怎么办？是否考虑过这么强大的技术会被用在哪里以及如果罪犯也利用这种技术来躲避警察的追捕怎么解决？但是这位哈佛大学的计算机科学家 Hau Chan 被问这些问题时，他只是回答说，他不确定新工具将被怎样使用，他只是一名工程师。由此引发了我们对技术这种不负责任态度的思考。这种不负责的态度是由于技术与人文的断裂造成责任意识缺失的，除此之外还有另外一个原因，就是制度。技术研究人员在研究一项可能带来巨大影响的技术时，对于技术后果的考虑显得很有必要，但是这种不负责任的想法是由制度上的缺失造成的，贝克提到过一个概念“组织化不负责”，由于现代技术作用的广泛性和复杂性，人们往往难以预估其作用范围和作用期限；同时，由于现代管理机构的日益庞大，各管理

机构之间的层级关系日益复杂以及各种管理部门之间的职能交叉与重复，往往难以确定技术风险的责任主体。所以由于这种制度的缺失，未来可能会出现人工智能技术的滥用，带来社会公共安全隐患。

（四）资本的推动的影响

索菲亚“机器公民身份”骗局就是背后资本推动的结果。现在我们处于经济全球化，经济是推动人工智能发展的最有力推手。政策的支持，使越来越多的企业在国家扶持下，大量融资，发展人工智能产业，使人工智能产品融入生活的方方面面。经济利益是企业发展人工智能最终归宿，所以作为人工智能技术开发主体的企业会以利益作为选择产品的重要因素。如今人工智能技术风险尚未建成完善的管理体制，企业在追经济效益的过程中，可能会有意淡化或忽视自己的社会责任和生态责任，不会去关注技术的风险与负面影响，在他们追求利益最大化的同时，尽可能降低自身的成本。

以 Facebook 信息泄露事件为例,分析资本对风险推动的作用。Facebook 内部曝光了一份备忘录，在明知风险的情况下，Facebook 还是不择手段，用以增加用户数量。曝光的那份备忘录中记录了以下几句话：“不管是那些可能让人质疑的导入联系人的做法，还是使用特殊词让用户能被朋友搜索到。我们所做的一切都是为了增加我们的用户量。”“如果玩儿砸了，那结果会非常糟糕。有可能会导致一些人的生活被彻底暴露给某些不法分子。”[①]企业为了自身的利益，完全不顾公众面临的风险，即使公众有可能受到恐怖袭击。在追求企业利益最大化时，忽略企业应该承当的社会责任。人工智能技术对数据依赖程度非常大，大量数据在企业手中，使得个人隐私得不到有效保护。企业在开发产品时，通过非法利用人工智能操纵用户判断并获得经济利益或政治利益，都是会给公众带来巨大的风险。

[①] AI 前线，关于 FB 隐私门再爆内部猛料：哪怕被利用搞恐袭，只要用户增长就行！[EB/OL]. https://mp.weixin.qq.com/s/CHuZEmaanLzFoA6hpnW9mg,2018-3-30[2018-4-1].

第 4 章 人工智能技术风险规避现状及建议

4.1 加强技术与文化的沟通

由于科技与文化的断裂，致使现代技术的发展总是缺乏人文反思。因此在发展技术的同时，也需要考虑人文因素，将人作为发展技术的底线。

加强人文与技术的沟通可以通过以下几种方式进行，首先，在制定一项科技政策时，要结合多方领域人员的建议，比如人文工作者。技术人员可能更多期望以技术给社会带来改变或者通过技术获得某种权力，而忽视了技术可能给社会带来的风险。人文工作者可以在技术研发之前，提前对技术进行分析，比如哲学家对人工智能技术风险的哲学层面的探讨，对人工智能技术的研究与应用提供一个可参考的意见。其次，可以通过加强技术人员的人文熏陶，从事人工智能技术的人员通过对他们传授人文知识，增加对技术应用研发上的人文思考，有助于减少技术人员开发对人类有巨大风险的产品。虽然通过这些建议对于改变这一现实可能力量有限，但是人文对人工智能技术的作用，会为人工智能技术发展指引方向。

4.2 提高公众风险认知水平

人工智能技术风险也由于公众的认知水平的限制，导致对风险评估的不准确甚至夸大。所以提高公众的风险认知水平很有必要，可以通过以下几种方式。第一，加强对公众的科学知识教育，对科学技术知识的了解，会降低事物的风险的恐惧。第二，通过引导，让公众对未知风险感知不要产生恐惧。

科学素养会影响公众对风险的认知。贝克曾说过，公众如果认识的风险越少，就会对风险感知的越多。当我们对一项技术了解的少，就会对该技术的超强能力感觉到恐怖。就像当下的人工智能技术一样，因为我们对人工智能技术了解都是通过媒体或者科幻电影，对于这种模拟人类智能的机器，我们不了解它是否会产生超越人的智能时，我们就会对它感到害怕，就算它只是一个懂得

下棋的博弈机器人，在其他方面它可能连三岁小孩都不如。例如之前机器人索菲亚在各档节目中表露出“自我意识”、“甚至是我会毁灭人类”的话语，从而引起了不少人的担忧，加上科幻电影前期的伏笔，人们开始相信并担心机器人会引发叛乱统治人类。这是因为对人工智能的不了解，缺乏专业的判断。所以提高公众科学素养，对于人们去分析风险问题有很大的帮助。

此外，还应因此应加强公众与专家、政府的风险沟通。首先我们可以通过加强专家与公众的沟通，让专家向公众解释某一项具体技术是如何发展的，并且该技术会产生什么的风险。其次，现代媒体是公众了解新技术和技术风险来源的主要途径，所以对媒体建立有效的管理机制，不要让一些不实新闻误导公众。最后还要通过日常的科普，向人们传输一种正确的科学观，并且能在心理感知风险时提供可靠心理支持。

4.3 科学家的道德责任

科学家是否应该为自己研究负责？答案是肯定的。得让科学家们明白，他们为什么要承担责任。布里奇斯托克曾论证过科学家为什么在研究成果的应用上，是具有道德责任的，但是这种情况是特定的场合下，科学家对其行为是有责任的。布里奇斯托克用了几个命题来推理论述得出，科学家在何时需要对自己的研究应用负上道德责任。他最后给出了一个命题，“当且仅当 P 对 X 承担因果责任，且 P 能够知道 X 将要发生，又 P 的所为是在自由、理性的状态下做出的，那么，P 对 X 承担道德责任。”[①]（这里的 P 指代科学家，X 指发生的具体后果）。首先我们得明白科学家承担因果责任是什么，就是科学家在发明一项技术时，该技术导致危害发生，这个发生关系就是因果关系。如果科学家没有发明那项技术，就不会导致这样的后果。其次，科学家是在自愿并自由的情况下，去开发该项技术的。综上所述，科学家在这种情况下要为他的研究负上道德责任。布里奇斯托克认为，不仅仅要考虑到主体的道德责任问题，还要考虑行为的实施对象——客体，他用了另外一词“道德考虑”对道德责任进行划分，对于被实施对象会受到什么威胁也是作为科学家应该考虑的范围。所以说科学家在从事某项科学研究时，都必须对社会的影响承担一定的责任。

[①] 布里奇斯托克（Bridgstock,M.）等著，科学技术与社会导论[M]，刘立等译，北京：清华大学出版社，2005，72

人工智能技术是一项可以对社会产生出巨大影响的技术，从事人工智能技术研究的科学家应当对自己所做的研究，负上自己应承担的道德责任。之前在人工智能、道德与社会（AIES）会议上，由一项研究成果引发了一场关于科学家道德责任的争论。引发争论的是哈佛大学关于“使用算法自动识别一起犯罪是否属于团伙犯罪”这项研究，当时在发布这项技术时，被问及这项技术被用于实际，是否考虑过该技术可能带来的负面影响时，报告人只说“我只是一名工程师”，不知道具体会应用在何处，由此引发了争议。[①]当下科学家对自己的责任意识还很薄弱，正如前面所说的，科研人员对自己的研究应用应该考虑对社会的影响，然而这种不负责的话语更加突出了科技与人文断裂，当科学家在创造一个强大的事物时，应该考虑下它可能被如何使用，以及它可能造成的风险程度。如今使用人工智能作恶是一件很容易的事，Open AI 发布了一份关于恶意使用人工智能的报告。该报告从技术角度出发，指出了未来 5 年内人工智能技术被滥用的风险，因此加强科学家的道德责任教育显得更有必要。

加强科学家的道德教育，我们可以通过制定道德规范。目前业界已经有出台了关于人工智能技术伦理方面的准则，试图通过道德教育来约束从事人工智能技术相关人员。比如在 ASILOMAR 会议上制定的“阿西洛马人工智能原则”，就是通过一系列的伦理准则对科学研究及从事相关人工智能领域的人员，提供一个可以参考的道德规范准则。2017 年 IEEE 电气电子工程师学会也提出了一个关于自主与智能系统伦理全球倡议项目（《人工智能设计的伦理准则》）。希望通过教育、培训和授权，确保从事自主与智能系统设计开发的利益相关方优先考虑伦理问题，只有这样，技术进步才能增进人类的福祉。仅仅通过制定的伦理准则告知科学家们对人工智能技术的研究和发明应承担的具体责任是不够的，还需要通过日常教育，加强科学家的道德教育，让他们了解自己应负的社会责任。但是这个层面上的规避可能达到的效果并不尽如人意，下面可以通过完善的法律法规，约束从事人工智能研究的科学家、企业、政府。

[①]大数据文摘，关于算法识别团伙犯罪引发巨大争议，[EB/OL] 2018-03-07[2018-03-28]. http://mp.weixin.qq.com/s/PoLvqyjIze-28NpV9NMnOw

4.4 完善相关法律法规

立法是地方和国家层面历史悠久的风险管理方法，立法和行政机构制定相关的法律法规，规定许可的制度范围，提供人工智能技术风险的管理方法，并希望可以通过完善的法律约束人工智能相关行业，按照对人类有益的方向发展。目前已经有一些国家出台了一些相关的法律法规，也有些组织制定的行业准则希望能有助于法律法规的制定。

目前已出台的人工智能相关的法律法规，在技术层面为了解决算法歧视问题，2017 年美国计算机协会发布关于算法歧视的七项原则，试图去解释人工智能技术的算法歧视问题。同年又出台了“算法问责法案”，并成立了专门工作组和联邦人工智能发展应用咨询委员会，监督市政机构使用的自动决策算法的公平性、问责性和透明度。在自动驾驶技术上，美国在 2017 年推动了自动驾驶汽车立法，通过了两部法案，SELF DRIVE ACT 和 AV START ACT，预计在在 2018 年正式通过。在机器人应用上，2017 年欧盟会议通过了关于制定机器人民事法律规则的决议，并制定了《机器人民事法律规则》，提出诸多制度。决议希望能成立统一的机器人与人工智能监管机构；引入电子人格的登记、保险和管理；提出伦理原则和《机器人宪章》来保障负责任的创新等一系列规则。韩国也提出《机器人基本法案》，并成立了专门的国家机器人伦理政策委员会，旨在确定机器人相关伦理和责任的原则，保障机器人的发展对人类有益。爱沙尼亚提出了机器人法案，并赋予人工智能代理人法律地位，虽然还在讨论阶段，但为全球范围考虑人工智能代理法律地位提供了样本。

这些立法都是通过某个方面对人工智能建立法律制度，通过立法规范人工智能开发与应用可能会造成风险。当下的法律法规的制定跟不上人工智能的发展速度，并且还没有完善的国际法律法规约束，很多科学家、企业可以因为利益开发一些对社会不利的产品。国与国之间由于技术的垄断，使致很多算法不透明，导致弱势一方可能会承担更大程度的风险等等。所以迫切需要国际上制定完善的法律法规，对当下人工智能的开发利用出台完善的法律法规进行约束，提前做好人工智能技术防范工作。建立完善的法律体制是当下对人工智能技术风险的防范最为有力的方式。

4.5 建立“亚政治”风险管理机制

人工智能技术开发及应用广泛，对于它的风险防范，只从现代政治下构建风险规避机制，已经远远不能防止大规模的风险问题。希望提出一种新的政治形式，从政治制度和政治运行机制上面来预防技术风险。这种政治形式就是贝克的“亚政治”，也就是“工业社会中的政治格局正在变成非政治性的，而工业主义中曾是非政治性的东西正变成政治的。”[①]“亚政治”最主要的特征表现为，非政治体系的群体可以参与到政策的制定之中，对风险社会管理有相应的政治权利。对人工智能技术建立“亚政治”风险管理机制，就意味着适当地限制政府管理部分的权利，防止政府部门管理太多，同时还要考虑技术专家因为利益纠缠导致的风险问题。“亚政治”风险管理机制就是让社会群体和公众共同加入技术政策决策当中，与政府部门、技术专家共同协商制定风险规避政策。

目前国外成立有不少风险研究机构，比如研究可能会导致人类灭绝的风险研究组织 Centre For The Study of Existential Risk (CSER 存在风险研究中心)。为了确保未来强大的科技对人类有益无弊的机构 Future of life Institute（FLI 生命未来研究所），该机构着重于如何保持人工智能的研发对人类有益，正在探索如何减低核武器和生物技术所带来的风险。为了保证人工智能会产生积极影响的 Machine Intelligence Research Institute（MIRI 机器智能研究所）；还有牛津大学创立的 Future of Humanity Institute（FHI 未来人类研究所），该组织提出“Governance of AI Program”是为了追踪人工智能在正义，经济，网络安全和军事领域的当代应用，并认真对待它们对透明度，公平性，问责性和安全性成的紧迫问题。

政府在制定人工智能政策时，可以加入上面这些非政治性的组织，除此之外还需要咨询公众的建议，让公众有参与技术决策和享有政治权利，制定一个符合人类总体的发展的政策。通过这种“亚政治”风险管理机制，可以有效防范政府滥权，并尽可能让人工智能技术发展为对人类有益的技术。

[①] [德] 乌尔里希•贝克等著,自反性现代性[M]，赵文书译，北京：商务印书馆，2004,24

结论

人工智能技术作为人类社会的一种重要技术，它对我们生活产生了巨大影响。因此在探讨人工智能技术所带的风险是很有必要的。本文通过对人工智能技术风险本质的探讨，分析了人工智能技术风险的内因和外响，提出了风险预防的策略和方案。从技术理性与社会理性断裂角度，对人工智能技术提出人文和技术相互融合的观点，其次在公众认知上，强调了提高公众风险认知水平的重要性。对人工智能风险在认识上提高风险意识。第三，通过科学家的道德教育，防范人工智能技术可能被滥用，通过立法，明确人工智能研究者的法律道德责任。加强防范对人工智能所造成的滥用或者开发不合理的技术的规避。最后还要通过法律法规的建设，完善人工智能技术风险的归责作用。

但是本文对人工智能技术风险的研究程度不够，对当下人工智能技术的了解还存在不足，并且就当下对人工智能技术风险规避的措施来看，未必有效降低人工智能技术所带的风险。因此我们希望在今后的学习和研究中，能进一步推进本研究，使之更加完善。

致谢

研究生三年的学习阶段即将告一段落，毕业前夕真的感触良多。在校学习到很多的专业理论知识，也学到了很多工作生活中的一些经验，研究生期间的这段学习过程对我未来的生活工作中有很多帮助。这次论文写作过程真的记忆深刻，我要感谢的人真的太多了。

首先非常感谢三年以来，我的导师黄承烈教授对我的栽培，让我对科技哲学专业有了最初的兴趣。不仅在学业上教授我专业知识和学习方法，还在工作和生活上给予我帮助与鼓励。在我学业和工作遇到困惑时，总是悉心的帮我解惑并且还给予我精神上的鼓励。没有导师的耐心解答和帮忙，我是不可能完成毕业论文的，所以由衷的感谢导师孜孜不倦的教导。

其次，在詹世友老师、胡小安老师、刘剑凌老师、潘博老师以及叶庆华老师的教学和指导下，我对科技哲学专业加深了认识和拓展了学习视野，对于学习方法和写作方面有了更加深层次的认识和提高。在此次论文写作上，各位老师都指出了我在写作的内容和方法上需要加强和改进的地方。由此，我对老师们的教学和指点表示深深的感谢。

最后，我要感谢父母兄弟对我无限包容、疼爱，在学习和生活上让我无后顾之忧的学习。还要感谢同窗刘君、陈赫、童平以及学妹袁婧对我帮助和忍让，在我遇到困难的时候及时帮我解决问题，并且时常在生活上给予我照顾，若没有她们的同进退，我是不可能这么顺利就写好我的论文。当然还要感谢一同奋斗的好友范静、程思燕及方璐，我们一同在图书馆相互鼓励，督促对方学习。从她们身上学到不同的学习方法，给我在写作毕业论文提供不同的视角。

在大家的帮助之下，我终于写完了我的毕业论文。这次论文写作真的给我一次很好的学习机会，让我认识到自己的不足，以及自己在写作时的问题。在今后的学习中，我也要去帮助别人，并时常怀着感恩的心，将自己的援助之手传递给需要帮助的人，并且要坚持不懈学习，争取做一名优秀的有志青年。

参考文献

[1] 乌尔里希•贝克(Ulrich Beck)著.风险社会[M].何博闻译,南京:译林出版社,2004.
[2] 安东尼•吉登斯.失控的世界[M].周红云译,江西:江西人民出版社,2001.
[3] 斯科特•拉什.风险社会与风险文化[C].李惠斌编.全球化与公民社会,桂林:广西师范大学出版社,2003.
[4] 保罗·斯洛维奇.风险的感知[M].赵岩东译,北京:北京出版社,2007,20
[5] 尼克拉斯·卢曼著.信任:一个社会复杂性的简化机制[M].瞿铁鹏,李强译,上海:上海人民出版社,2005.
[6] 杨雪冬.风险社会理论述评[J],国家行政学院学报,2005(01):87-90.
[7] 杨雪冬.风险社会与秩序重建[M],社会科学文献出版社，2006.
[8] 庄友刚.风险社会与反思现代性:马克思主义的批判审视[J],江海学刊,2004(06):38-42.
[9] 庄友刚.从马克思主义视野对风险社会的二重审视[J],探索,2004(03),131-134.
[10] 庄友刚.跨越风险社会:风险社会的历史唯物主义研究[M],北京:人民出版社，2008
[11] 杨海.风险社会的哲学研究[D],中共中央党校,2014
[12] 劳东燕.风险社会与变动中的刑法理论,中外法学,2014(01):70-102.
[13] 孙粤文.大数据:风险社会公共安全治理的新思维与新技术,求实,2016(12):69-77.
[14] 毛明芳.现代技术风险的生成与规避研究[D],中共中央党校,2010.
[15] 查尔斯•佩罗.高风险技术与“正常”事故[M].寒窗译,北京:科学技术文献出版社,1988.
[16] H.W•刘易斯.技术与风险[M],中国对外翻译出版公司，1990.
[17] 王世进.多维视野下技术风险的哲学探究[D],复旦大学，2012.
[18] 刘中梅,公众参与纳米技术风险沟通的影响因素研究[D],大连理工大学,2016
[19] 杜严勇.论人工智能的自反性伦理治理[[J],新疆师范大学学报（哲学社会科学版）,2018(02):111-119.
[20] 王东浩.机器人伦理问题研究[D],南开大学，2014
[21] 维纳.人有人的用处:控制论与社会[M].陈步译,北京大学出版社,2010.
[22] 徐英谨.技术与正义:未来战争中的人工智能[J],人民论坛·学术前沿,2016(07):34-53.
[23] 陈晋.人工智能技术发展的伦理困境研究[D],吉林大学,2016.
[24] 吴国盛.技术哲学经典读本[M],上海:上海交通大学出版社,2008.301.
[25] 殷正坤.试析技术的本质[J],天津社会科学,2001(4):41-44.
[26] 刘易斯•芒福德.机器的神话[M]I.宋俊岭等译.中国建筑工业出版社出版,2009.
[27] 安东尼•吉登斯.第三条道路及其批评[M],北京:中共中央党校出版社,2002
[28] 平克.心智探奇:人类心智的起源与进化[M].郝耀伟译,杭州:浙江人民出版社,2016:22.
[29] 丹尼尔•丹尼特.心灵种种:对意识的探索[M].上海:上海科学技术出版社,2012:8.
[30] 斯图尔特·拉塞尔等著.人工智能:一种现代方法[M].北京:清华大学，2013.
[31] 笛卡儿.谈谈方法[M].王太庆译,北京:商务印书馆,2000,43-44.

[32] 松尾丰著.人工智能狂潮:机器人会超越人类吗?[M].赵函宏,高华彬译,北京:机械工业出版社,2015,26
[33] Nils J•Nilsson 著.人工智能[M].郑扣根等译,北京:机械工业出版社,2000.
[34] 玛格利特•博登.人工智能哲学[M].刘西瑞,王汉琦译,上海:上海译文出版社,2001.
[35] 新浪新闻中心.关于无人驾驶汽车撞死人系首个自动驾驶车撞死人案例,[EB/OL]. http://news.sina.com.cn/w/2018-03-20/doc-ifysmpev5202955.shtml
[36] 斯坦诺维奇[著].机器人叛乱:在达尔文时代找到意义[M].吴宝沛译,北京:机械工业出版社，2015.
[37] 人民网.当人工智能成为"矛","盾在哪里" [EB/OL].(2018-3-30)[2018-4-1]. http://finance.people.com.cn/n1/2018/0226/c1004-29834825.html
[38] 腾讯研究院等著.人工智能[M],北京:中国人民大学出版社,2017.
[39] Max Tegmark. Life 3.0: Being Human in the Age of Artificial Intelligence[M],New York:Knopf,2017.
[40] 罗岗等.基本收入・隐私权・主体性——人工智能与后人类时代（上）[J],读书,2017(10).
[41] 伊利亚・普里戈金.确定性的终结[M].湛敏译,上海:上海科技教育出版社,2009.5.
[42] 熊红凯.人工智能技术下对真理和生命的可解释性[J],探索与争鸣,2017(10):14-15.
[43] 马尔库塞.单向度的人[M].李继译,上海:上海译文出版社,1989,126.
[44] 金观涛.反思人工智能革命[J],文化纵横,2017(8):20-29.
[45] Yudkowsky, Eliezer. Artificial Intelligence as a Positive and Negative Factor in Global Risk[J], Global Catastrophic Risks, New York: Oxford University Press. 2008. 308–345.
[46] 保罗・斯洛维奇.风险的感知[M].赵岩东译,北京:北京出版社,2007,20.
[47] AI 前线, FB 隐私门再爆内部猛料:哪怕被利用搞恐袭,只要用户增长就行![EB/OL]. (2018-3-30)[2018-4-1]. https://mp.weixin.qq.com/s/CHuZEmaanLzFoA6hpnW9mg.
[48] 布里奇斯托克（Bridgstock,M.）等著.科学技术与社会导论[M].刘立等译,北京:清华大学出版社,2005,72.
[49] 大数据文摘,关于算法识别团伙犯罪引发巨大争议 [EB/OL].(2018-03-07)[2018-03-28].http://mp.weixin.qq.com/s/PoLvqyjIze-28NpV9NMnOw.
[50] 乌尔里希•贝克等著.自反性现代性[M].赵文书译,北京:商务印书馆,2004,24.
[51] 玛格丽特・A・博登编著.人工智能哲学[M].刘西瑞,王汉琦译,上海译文出版社,2005.
[52] 罗素(Russell .S.J),诺维格(Norvig .P.)著,殷建平等译.人工智能:一种现代的方法[M],北京:清华大学出版社,2013.
[53] 冯•诺伊曼著.计算机与人脑[M].北京:北京大学出版社,2010.06.
[54] 波斯特洛姆著.张体伟,张玉清译.超级智能[M].北京:中信出版社,2015.2.
[55] 尤瓦尔•赫拉利著.未来简史 从智人到神人[M].林俊宏,译.北京:中信出版社,2017.
[56] 乌尔里希•贝克(Ulrich Beck)原著,世界风险社会[M].南京大学出版社,2004.
[57] 谢尔顿•克里姆斯基(Sheldon Krimsky),(英)多米尼克•戈尔丁(Dominic Colding)编著,风险的社会理论学说[M]. 北京出版社,2005.
[58] 芭芭拉•亚当(Barbara Adam),(英)乌尔里希•贝克(Ulrich Beck),(英)约斯特•房•龙(Joost Van Loon)编著,风险社会及其超越[M]. 北京：北京出版社, 2005.

[59] 皮金,风险的社会放大[M].中国劳动社会保障出版社,2010.
[60] 费多益著,全球化与风险社会[M]. 社会科学文献出版社,薛晓源,周战超主编, 2005.
[61] 拜纳姆,罗杰森.计算机伦理与专业责任[M],李伦等译,北京大学出版社,2010.