**成绩**



**学术道德与科研诚信**

**结课论文**

**日 期： 2022.10.24**

**题　　目：从工程伦理角度分析人工智能**

**学 院： 自动化学院**

**专业名称： 控制科学与工程**

**学 号： 1120181540**

**姓 名： 李澍凡**

**任课教师： 范春萍**

**评 阅 人： 范春萍**

**从工程伦理角度分析人工智能**

**大纲：**

人工智能在社会生活中的发展和应用十分迅速并且已经无处不在。这种颠覆性技术正在彻底改变整个行业，并带来了巨大的机遇。然而，在备受瞩目的损害用户 个人权利案例（例如，Facebook的Cambridge Analytica丑闻）的引发下，人们对人工智能的担忧正在上升。这促使利益相关者（政府、非政府组织、学术界、工业界）通过以道德为重点的研究、原则以及框架向社会做出回应。以道德为重点的人工智能文章得到了蓬勃的发展，通常被称为“值得信赖的AI”。本文从更高的层次介绍了人工智能伦理，人工智能伦理具有很强的跨学科性，因此本文还参考了哲学、法律以及计算机科学相关的知识。

**摘要：**

人工智能（AI）伦理学是一个领域，作为对人们日益关注人工智能所造成的影响的回应而产生的一个新兴领域。人工智能伦理是更加广泛的数字伦理领域的一个子集，也是在《学术道德与科研诚信》课程中所提及的科技伦理的一个分支学科。对人工智能进行科技伦理角度的分析，可以解决开发和应用新型数字技术所引发的担忧，例如人工智能、大数据分析和区块链技术等。本文主要目的是通过介绍人工智能伦理的基本概念、草图方法以及中心主题，从更高的层次对该领域进行讨论。本文第一部分将介绍人工智能伦理的基本概念。第二部分探讨了人工智能伦理学的基础，包括工程伦理学、技术哲学和科学技术研究等。第三部分讨论当前人工智能伦理的三种方法，即原则，过程和道德意识。第四部分讨论将人工智能伦理应用至工程实践当中的中心主题。最后，本文总结并着重强调了人工智能伦理学的未来方向。

**基本概念介绍：**

在介绍人工智能伦理之前，需要对数字伦理进行简单的了解。数字伦理是为了应对近些年来数字技术快速发展过程中产生的一些令人们担心的问题而发展的一门新兴学科，其中涉及到的数字技术包括区块链、物联网、大数据分析以及机器学习等。数字伦理所涉及的范围相较人工智能伦理更加广泛且更具有一般性，对于数字伦理的讨论主要集中在新型数字技术对心理、社会环境以及政治的影响上，对心理的影响主要体现在个体上，对社会的影响主要集中在集体以及环境上，对政治的影响主要体现在社会的组织结构上。

人工智能伦理是数字伦理的一个重要分支，对人工智能所造成的影响进行的研究与分析主要也是从心理、社会以及政治这三个角度进行的。人工智能伦理的大部分研究与分析角度同数字伦理相似，但也有人工智能所具有的独特之处。其中关于心理层面还涉及到心理自治，防止人工智能不当使用以及知晓是否在于人工智能进行交互的权益；社会层面涉及到公平以及正义的问题；政治层面涉及到人工智能对民主进程以及经济的影响。总而言之，人工智能伦理学是一门综合性很强的学科，并且随着科技的发展以及思想的进步仍在发生快速的进化。本文将对人工智能伦理中的基本概念进行简要介绍。

1. 何为人工智能

人工智能是近几十年来新兴的数字技术之一，其快速发展驱动着第四次工业革命的进程。人工智能的基础就是算法，用最容易理解的语言来描述即是通过一套完整的规则或者流程去解决问题或者任务，在数字领域中可以用编程语言在计算机中表达。关于人工智能的算法主要分为两类，分别是静态算法以及动态算法。静态算法是仅执行固定动作的传统程序，主要为基于知识的系统。动态算法则可以在与环境的交互中学习以及迭代，常见的有机器学习算法。机器学习算法是近年来最为成功的人工智能算法，其在不同环境中可以实现自适应的能力极大推动了人工智能的快速发展。机器学习算法主要分为监督学习、无监督学习以及强化学习。机器学习算法主要通过提供数据集以及标签来训练，从而实现算法的准确度。面对非结构性的环境，还可以使用无监督学习以及强化学习的方法，从环境中获得反馈以及奖励，对算法实现调整并最终达到最佳的效果。在机器学习的基础上，人工智能领域又涌现出更多的算法包括深度学习、对抗性学习以及元学习等，这些算法也在推动人工智能的进一步发展，并且也使得人工智能在社会生活中的应用更加普遍，不仅是在学术界，在日常生活当中人工智能也开始发挥其重要的作用，例如最近流行的AI绘画、AI通话等，当然许多问题也伴随着其兴起而产生，这就涉及到本文的核心人工智能伦理，了解人工智能的基本概念对后文的分析有很大的帮助。

1. 何为伦理学

伦理学是一门相当广泛的学科，涉及的范围极大并且对其有着多元化的解读。伦理道德体系起源于西欧以及北美，主要是从英语系的国家中产生。在人工智能伦理道德体系产生之后，一些非西方的伦理道德体系例如日本以及非洲的伦理道德体系也可以融入人工智能伦理道德体系中并且开始被人们着重强调。考虑到当今伦理道德框架面临的挑战越来越大，我们所进行的研究当中人工智能伦理学中的“伦理”成为了这个学科的关键概念。我们广泛地将伦理学理解为理性系统地研究什么是正确的标准；而道德则是英语当中区别好与坏的相关概念的常用术语。

除了伦理以及道德，法律也属于本文所研究的广义的人工智能伦理学中的一个重要部分。法律和道德在人工智能伦理学中是高度相互依存的两个重要部分。管理、问责制以及透明度的关键主要依赖于法理，我们可以通过法律在特定的范围内建立成文的规则和准则。另外法律是可执行，并且存在强制执行法律的机构即政府部门。从哲学上来解释，伦理道德是对人工智能伦理的高标准，法律则是对人工智能伦理的最低要求。

**人工智能伦理的发展历程：**

人工智能伦理学是一个新兴领域，仍处于萌芽阶段并还在不断地发展。然而在人工智能伦理学还未产生之时，其相关的研究已经达到了较为成熟的阶段。从另一个角度看，人工智能伦理学是这些相关研究的交叉以及延续，其相关研究主要包括工程伦理学、技术哲学以及科学技术研究。

1. 关于对相关研究分析的选择

人工智能伦理学的相关研究并不仅仅是包括这三个领域，实际上与人工智能伦理学最为相关的研究是机器人伦理学。之所以选择这三个领域对人工智能的相关研究进行介绍，主要是因为这三个领域跨越了交叉学科的极端。一方面，技术哲学属于理论研究，属于“抽象”的非经验极端。另一方面，工程伦理则属于实践研究，是对伦理道德体系的具体应用。科学与技术研究作为另一个极串联起了技术哲学以及工程伦理。除了这些先前的学科之外，“负责任的创新”作为一种新的创新方法，在当今的科研界成为一股新的浪潮。其具体的优势体现在社会和道德方面，明确考虑到了经济、社会文化和环境方面的平衡，值得在人工智能伦理学的研究过程中参考借鉴。

1. 工程伦理

工程伦理可以被视为与工程实践相关的价值观以及伦理道德体系。工程用专业术语描述具体指的是构建、设计以及建造。工程大概是广义科学中本质上最具有实践性的学科，其通常被称为应用科学。工程的相关概念早在人类进入文明社会之初就已经产生，不断推动人类社会的发展，但是同时也曾造成过许多负面影响。因此，在学术界已经存在大量讨论工程对社会以及环境影响的观点。尽管工程学包括许多分支学科（包括土木、机械以及计算机等），不同学科根据其学科特点会对社会造成特定的影响。但是由于该领域已经发展至极为成熟的阶段，因此已经归纳出具有普适性的工程规范，也可以称为工程伦理。

英国皇家工程师学会就曾发布过“工程伦理”指南。该指南分为两个部分：第一部分是“伦理道德原则声明”，即要求工程中需要做到：1）诚实和正直2）尊重生命、法律、环境以及公众3）准确性和严谨性4）领导力和积极沟通；第二部分是“实践中的工程伦理学”，主要通过一些具体案例对伦理道德原则进行补充。

但是需要注意的是，工程伦理主要是由一些私人聚集的社区或者组织所驱动的，他们以过程为导向发展出了一套自己的标准，有时候这些标准会超出法律范围，甚至与法律相冲突。

1. 技术哲学

技术哲学主要是研究技术的本质以及其是如何影响个人、社会以及政治的。人工智能伦理乃至数字伦理中有许多的哲学分支，其中包括伦理哲学、政治哲学以及社会哲学。但是在这之中与人工智能伦理最相关的是被称为其前身的“技术哲学”。

技术哲学出现于上世纪20年代左右，它有别于科学哲学，科学哲学在哲学史上有着更长的历史遗产，涉及方法和知识。而技术哲学则是技术创新的结果，它涉及科学发现的应用和用途。主要重点是评估技术如何影响人类生存环境以及技术是否存在价值。从历史角度来看，技术哲学可以看作是对一直存在着核威胁和上世纪60年代兴起的以社会意识为代表的反主流文化的后二战世界的回应。实际上，技术哲学对科学发现和科技进步之间存在必要联系的观点提出了挑战。相关文献通常都是较为负面的，研究人员认为新技术的发展会对社会秩序以及人类的生活带来危险，会对自然造成破坏，还产生了科学技术异化观。当然，会中也存在着积极乐观的技术哲学观，认为技术的进步对人类的生活带来了很多的便利，同时进一步解放了生产力，推动社会的进步。

1. 科学与技术研究

科学和技术研究是对科学研究和技术创新与文化、社会和政治之间的相互影响进行的研究。上文所述的技术哲学属于典型的伦理道德恐慌，即是对新技术以及新事物的出现而产生的下意识的抗拒，过于放大了新技术的缺点，对其中的好处则视而不见。事实上，技术哲学已经成为干预政治的工具，而不是对技术创新的性质和影响进行全面的调查后产生的思想。过去的技术哲学仅仅是社会科学的发展过程中对新技术的非经验性的指责，所以当今越来越多的研究人员开始通过自然科学的方法，通过调查、检验假设以及数据分析对新兴科学和技术进行经验性的研究，从而提出了科学和技术研究这一概念。科学和技术研究通过指出以不同方式设计和使用技术，可能会导致完全不同的社会结果。通过实证调查，研究可以对技术应用的价值进行评估。研究可以根据经验调查技术由谁构建、由谁使用以及技术对社会的影响。

**人工智能伦理的基本原则：**

关于人工智能伦理的基本原则有着许多的说法。在2019年对人工智能伦理准则的全面回顾中总结出了11项“伦理准则”：（1）透明度（2）公平、公正、平等（3）无恶意（4）责任和问责制（5）隐私（6）福利（7）自由和自治（8）信任（9）尊严（10）可持续性（11）相互支持。本文将伦理原则与不断发展并变化的工程专业知识相结合，提出了更加符合当今社会环境的，可以落实到工程实践中的伦理原则：人类幸福、安全、隐私、透明度、公平以及责任。本文所提出的原则均是在“值得信赖的AI”的范畴之内的。接下来将从这些方面对人工智能进行分析。

1. 人类幸福
2. 安全
3. 隐私
4. 透明度
5. 公平
6. 责任