清 华 大 学

综 合 论 文 训 练

题目：人工智能是否具备元伦理能力
—— 以电车难题为例

系　　别：人文学院哲学系

专　　业：哲学系

姓　　名：赵丹尼

指导教师：瞿旭彤　副教授

2024 年 5 月 29 日

人工智能是否具备元伦理能力——以电车难题为例

赵丹尼

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定).

签 名：_____导师签名：_____日 期：

摘要

人工智能的“电车难题”问题是人类价值对齐的后果。随着人工道德代理的发展进展，我们希望人工智能能够发展出独立且潜在优越的伦理推理。要实现这一目标，人工智能必须具备元伦理学的能力。本文探讨了人工智能在元伦理学方面的这种能力，考虑了计算心智理论和哥德尔反机械主义观点的论证。特别关注的是物理过程与现象体验之间的解释鸿沟，或者说物理功能的结构与形而上学表现的结构之间的关系。本文提出，非简化的基本原则可以承担解释的重任，双向心理物理桥梁允许模拟的元伦理功能积极地与道德真理和/或原则进行互动。最终，结构一致性原则和组织不变性原则以及形而上学的双重方面理论被改编为人工智能元伦理学理论的基石。

关键词：人工智能元伦理学；结构一致性；现象体验；对应关系

ABSTRACT

AI's "Trolley Problem" Problem is a consequence of Human Value Alignment. As we make progress with Artificial Moral Agents, we hope AI to develop independent, potentially superior ethical reasoning. For this to occur, AI must be capable of metaethics. This paper explores this capacity for AI Metaethics, considering arguments from both the Computational Theory of Mind and Gödelian Anti-Mechanist viewpoint. Specific attention is paid to the explanatory gap between physical processes and phenomenal qualia, or the *structure of physical functions* and the *structure of metaphysical performance*. This paper proposes that nonreductive basic principles can carry the explanatory burden and that a bilateral psychophysical bridge allows for a simulated metaethical function to actively engage with Moral Truths and/or principles. In the final analysis, the principles of structural coherence and organizational invariance, alongside the double-aspect theory of metaphysics are adapted to lay the cornerstone in a theory of AI Metaethics.

Key words: AI Metaethics; Structural Coherence; Phenomenol Qualia; Correspondance

目 录

章 1 章 引言

人工超级智能（ASI）被定义为“在几乎所有感兴趣的领域中，其认知能力远远超过人类的任何智慧”[①]。ASI 作为哲人王的一个关键基准是自主道德代理（AMA）在哲学上超越人类。衡量其认知优越性的一种方法是解决人类未能解决的伦理困境，例如电车难题。

然而，AI 的“电车难题”问题是一个双刃剑：人类价值对齐（HVA）使得 AMA 表现得“像道德的一样”[②]，但同时也阻止了 AMA 真正成为道德的并超越现有的伦理格局。HVA 的一个深远影响是它也抑制了 ASI 作为哲人王的演变。

解决这个问题意味着要说服行业代理，AI 机器学习系统（如特斯拉的自动驾驶系统）可以在没有 HVA 的情况下，代表我们做出伦理决策。为此需要许多步骤，包括但不限于 AI 是否有能力做出伦理决策，AI 伦理思考过程是否与人类伦理思考过程兼容，以及这些决策的后果是否会与人类价值观相冲突。我在《ASI as Philosopher Kings》（forthcoming）中对每个步骤进行了更详细的论证；然而，本论文的主要关注点是 AI 是否能够首先做出伦理决策。具体而言，本论文将集中讨论 AI 是否具备元伦理学的能力。

在第二章中，我们将首先介绍 AI 的“电车难题”问题，解释其背景、意义及解决方案的前提条件。

在第三章中，我们将探讨 AI 元伦理学的哲学困境。这包括对 AI 元伦理学提案的批判性评估，权衡计算心智理论与反机械主义者的观点。

在第四章中，我们将基于查尔默斯的结构一致性原理、组织不变性原理和信息双重属性理论，概述一种 AI 元伦理学理论。在此过程中，我们将考虑一种非

[①] Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press, 2014. 英：“any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest”。

[②] Walsh, Daniel. “*ASI as Philosopher Kings: the existential threat of falsehood”,* Medium, 2024. https://medium.com/@dracollavenore/asi-as-philosopher-kings-the-existential-threat-of-falsehood-89af285e031c

还原的形而上学解释和基本原则，以承担在物理功能结构与形而上学表现结构之间的解释负担。

在第五章中，我们将对整个论文进行一个概要性的总结。

第 2 章 AI 的“电车难题”问题

AI 的“电车难题”问题是一个悖论，即试图制造道德 AI 反而抑制了道德 AI 的发展。更具体地说，这个问题是人类价值对齐（HVA）削弱了人工道德代理（AMA）的潜力，将它们限制在类似人类的道德决策中，而不是允许它们发展独立且潜在的更优越的伦理推理。本章的重点是简要介绍 AI 的“电车难题”问题，解释其背景、意义及解决方案的前提条件。

2.1 电车难题作为元伦理学问题

为了理解 AI 的“电车难题”问题如何成为一个元伦理问题，我们必须首先了解原始的电车难题及其对今天的人工道德代理（AMA）的影响。

2.1.1 原始电车司机问题

简而言之，电车难题最早可以追溯到 1967 年，当时菲利帕·弗特提出了“电车司机”的问题[①]，而朱迪思·贾维斯·汤姆森简化了这个问题，称之为“开关”问题[②]：

> *一辆失控的电车正沿着铁轨疾驰。前方的铁轨上有五个人被绑住，无法移动。电车正朝他们冲去。你站在车场的某个地方，旁边有一个操纵杆。如果你拉下这个操纵杆，电车将转到另一条铁轨。然而，你注意到那条侧轨上有一个人。你有两个选择：*
>
> *1. 什么也不做，这样电车会撞死主轨上的五个人。*
> *2. 拉下操纵杆，将电车转到侧轨上，撞死一个人。*

乍一看，电车难题似乎是在两个悲剧性结果之间进行选择，或者找到一种完全避免悲剧的方法。但电车难题的真正挑战不在于找到“开关”的解决方案，而在

[①] Foot, Phillipa. "The Problem of Abortion and the Doctrine of the Double Effect" in *Virtues and Vices* (Oxford Review, Number 5, 1967.)

[②] Thomson, Judith Jarvis. “The Trolley Problem.” *The Yale Law Journal* 94, no. 6 (1985): 1395–1415. https://doi.org/10.2307/796133.

于找到解决潜在伦理困境的突破口。换句话说[①]，任何给定的解决方案都可以通过不同的伦理框架来证明其合理性，电车难题实际上是提出了一个普遍的元伦理原则的问题，以解释在存在更深层次伦理冲突的情况下产生的不同判断。

简而言之，电车难题之所以成为问题，是因为在面对相互竞争的伦理困境时，我们无法就做出哪种选择达成一致，但仍然必须做出选择。

2.1.2 对 AMAs 的影响

如果电车难题仅限于思维实验和科幻小说，它就不会如此紧迫。然而，随着自动驾驶汽车在道路上的日益普及，电车难题获得了现实世界的意义。这在自主道德代理（AMAs）软件的设计中尤为显著[②]。例如，我们可以设想一个包括特斯拉和救护车的情景：

> *一辆特斯拉正处于自动驾驶状态，准备通过一个绿灯。突然，交通信号灯变色以让救护车通过。这是一个对预训练的 AI 来说未预见的事件，导致其重新规划的过程中出现时间差。此时，减速已经太晚了，AI 必须决定是转向以牺牲乘客的生命来让救护车安全通过，还是撞上救护车，可能会杀死车内的潜在受害者，但很可能会挽救乘客的生命。*

近年来，关于类似特斯拉和救护车情景的文献不断增加，其中讨论了在不可避免的潜在致命碰撞中，AI 是否应该优先考虑车内乘客的安全，而不是车外潜在

[①] Panahi, Omid. "*Could There Be A Solution To The Trolley Problem?*", Philosophy Now, 2016. https://philosophynow.org/issues/116/Could_There_Be_A_Solution_To_The_Trolley_Problem

[②] Lim, Hazel Si Min; Taeihagh, Araz. "Algorithmic Decision-Making in AVs: Understanding Ethical and Technical Concerns for Smart Cities". *Sustainability* 11 (20): 5791. 2016. doi:10.3390/su11205791.

受害者的安全[①②③④]。这突显了自动驾驶汽车的现实世界后果：电车难题迟早会在道路上出现——如果尚未出现的话[⑤]——届时做出选择将是不可避免的。

2.2 结论：AI 元伦理学的必要性

在本章中，通过简要介绍 AI 的“电车难题”问题，我们可以得出以下结论：

(1) 电车难题是一个问题，因为在面对相互竞争的伦理困境时，我们无法就做出哪种选择达成一致；然而

(2) 对于 AMAs 来说，选择是不可避免的。

AI 的“电车难题”问题源于试图编程 AMAs 解决电车难题，而我们自己在这个问题上没有共识[⑥]。通过将 AMAs 限制在人类式的道德决策中，当面对伦理困境时，AMAs 自然会遇到相同的电车难题。回到我们之前的例子，我们看到，如果我们编程让特斯拉转向，那么它是不道德的；如果我们编程让特斯拉撞车，它同样是不道德的。因此，在当前的伦理框架下，对特斯拉进行价值编码的悖论在于人类价值对齐不可避免地导致 AI 以某种方式变得不道德。

AI 的“电车难题”问题的意义在于，它揭示了人类价值观的不足以及我们当前伦理框架中的伦理矛盾瓶颈。如果我们想在电车难题上取得进展，我们不能将 AI

---

① Lin, Patrick. "The Ethics of Autonomous Cars". *The Atlantic*. 2013.

② Bonnefon, Jean-François; Shariff, Azim; Rahwan, Iyad. "Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?". *Science*. 352 (6293): 1573–1576. 2015. doi:10.1126/science.aaf2654.

③ Emerging Technology From the arXiv. "Why Self-Driving Cars Must Be Programmed to Kill". *MIT Technology review*. 2015. https://www.technologyreview.com/2015/10/22/165469/why-self-driving-cars-must-be-programmed-to-kill/

④ Bonnefon, Jean-François; Shariff, Azim; Rahwan, Iyad. "The social dilemma of autonomous vehicles". Science. 352 (6293): 1573–1576. 2016. doi:10.1126/science.aaf2654.

⑤ Goggin, Ben. “A Tesla owner says his car’s ‘self-driving’ technology failed to detect a moving train ahead of a crash caught on camera”, *NBC News,* 2024. https://www.nbcnews.com/tech/tech-news/tesla-owner-says-cars-self-driving-mode-fsd-train-crash-video-rcna153345

⑥ Awad, Edmond; Dsouza, Sohan; Kim, Richard; Schulz, Jonathan; Henrich, Joseph; Shariff, Azim; Bonnefon, Jean-François; Rahwan, Iyad. "The Moral Machine experiment". Nature. 563 (7729): 59–64. 2018. doi:10.1038/s41586-018-0637-6.

限制在人类价值观上，这会反作用地阻碍道德 AI 的发展。相反，要在电车难题上取得进展，我们必须允许 AI 发展出独立的、可能更高尚的伦理推理。然而，问题的关键在于，我们不确定 AI 是否能够发展出独立的伦理推理。要解决电车难题，AI 必须能够在相互竞争的伦理理论之间进行权衡，从事伦理的伦理研究。换句话说，AI 必须具备进行元伦理的能力。

第 3 章 哲学困境

在探讨了 AI 的“电车难题”问题后，我们现在转向 AI 进行元伦理学的可能性。即使是 AI 的最严厉批评者也大多同意，大脑只是一个机器[①]，而近年来 fMRI 支持了一种关于元伦理学的功能性解释。然而，反机械主义者认为，将元伦理学简化为纯粹的物理解释忽视了功能与体验之间的解释性差距，并且有一种独特于人类的缺失成分。本章的重点是考虑 AI 进行元伦理学的可能性，权衡计算理论和反机械主义者的观点。

3.1 心智计算理论

在哲学领域，即使是 AI 的最严厉批评者如德雷福斯和塞尔也接受心理的计算理论，该理论认为心理与大脑之间的关系类似（如果不是完全相同）于运行程序（软件）与计算机（硬件）之间的关系。此外，元伦理学的本体论观点认为，前缀“元”意味着对一个概念的自我指涉抽象，元伦理学是一种二阶功能，或者简单地说，是关于伦理的伦理学。通过适应元伦理学的功能本体论，我们可以扩展达特茅斯提案和强 AI 假设，以得出一个关于 AI 元伦理学的解释。

3.1.1 元伦理学的功能本体论

元伦理学的本体论观点认为，如果我们将伦理学比作一个函数，例如研究什么是道德正确或错误的，那么规范伦理学是一阶道德主张或一阶导数 f(x)，而元伦理学是二阶道德主张或二阶导数 f’(x)[②]。

为了举一个应用例子，让我们回到特斯拉和救护车的场景，选择是(1)转向；或(2)撞车。在这个例子中，我们假设：

- x = 转向或撞车

[①] Searle, John. “Minds, Brains and Programs”, *Behavioral and Brain Sciences*, 3 (3): 417–457, doi:10.1017/S0140525X00005756. For example, John Searle writes: "Can a machine think? The answer is, obvious, yes. We are precisely such machines."

[②] DeLapp, Kevin M. Metaethics. *Internet Encyclopediea of Philosophy.* https://iep.utm.edu/metaethi/

- F(x) = 伦理学，表示对什么是道德正确或错误的研究

所以，F(x)中的“转向”将是 F(转向) = 伦理学。由于存在各种伦理理论，我们可以替代不同的伦理函数。例如，如果我们假设伦理函数对应于功利主义，那么 F(转向)将等于“最大多数人的最大幸福”；同样，如果我们假设伦理函数对应于康德伦理学，那么 F(转向)将等于“绝不可以杀人”。

现在，让我们引入元伦理学的概念：

F′(x)：表示元伦理学，即用于理解伦理学的框架研究。在我们的情景中，F′(转向)将表示元伦理学。

将 F′(x)表达为 x 和 F(x)的函数：

F′(x) 是 F(x)对 x 的导数，表示研究制定伦理决策背后的框架或方法论。它本质上是考察支持我们伦理判断的基本原则和理论。

在数学术语中，导数 F′(x)可以表示为：$F'(x) = \frac{d}{dx}[F(x)]$。

简单来说，F′(x)表示伦理框架（伦理学）如何根据不同的行为（x）演变或变化。

在我们的类比中，F′(转向)代表了在决定是否转向时涉及的元伦理学考虑。因此，方程 $F'(x) = \frac{d}{dx}[F(x)]$意味着 F′(x)是 F(x)关于 x 的导数（即，伦理学的元伦理学）。

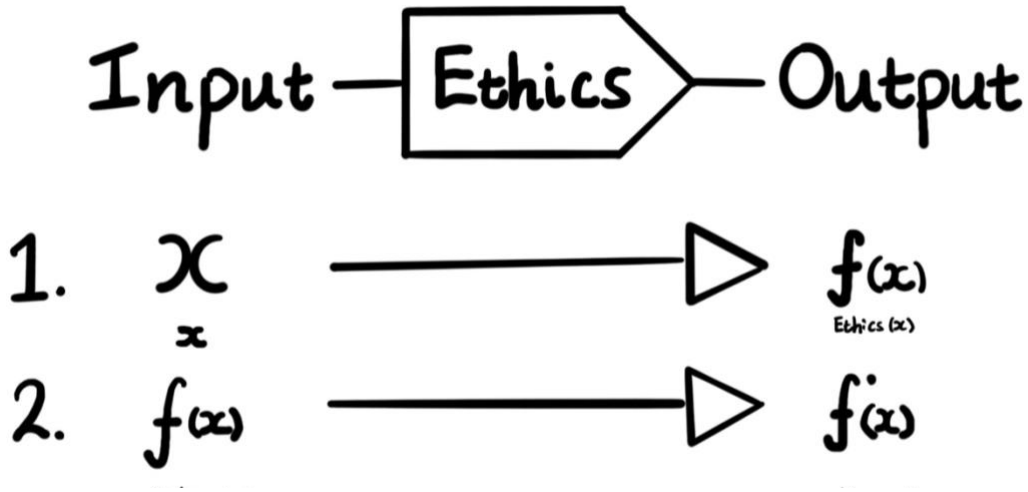

图 3.1.1 Metaethical Function Machine

当然，这只是将元伦理学类比为二阶导数的一种非常粗略的近似，但这个类比的目的是说明，如果伦理函数 F(x)没有本体论上的问题，那么从数学上解释元伦理学为 F(x)的 F(x)应该是合理的。

3.1.2 达特茅斯提案与强 AI 假设

如果我们接受元伦理学的功能本体论，那么要得出 AI 元伦理学的解释，我们必须证明大脑中的元伦理学功能可以对应于物理再现，也就是所谓的计算心智理论。即使是 AI 的最严厉批评者，大多也认为至少在理论上，任何事物都可以被模拟。这一基本立场出现在 1956 年的达特茅斯研讨会提案中：

> *"学习或智力的任何其他特征都可以原则上被如此精确地描述，以至于可以制造出一个机器来模拟它。[①]"*

[①] McCarthy, John; Minsky, Marvin; Rochester, Nathan; Shannon, Claude, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1955. This assertion was printed in the program for the Dartmouth Conference of 1956, widely considered the "birth of AI." Also: Crevier, Daniel. *AI: The Tumultuous Search for Artificial Intelligence*, (New York, NY: BasicBooks, 1993), 28. 英: *"Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."*

除非我们否认人类不能进行元伦理学或者元伦理学不是学习或智力的一种特征，否则达特茅斯提案声称我们迟早会能够以某种方式隔离这种元伦理学能力，使得 AI 能够模拟它。Dreyfus 支持这一观点，他认为：

> *"如果神经系统遵循物理和化学定律，我们有充分的理由认为它确实如此，那么……我们……应该能够用某种物理装置再现神经系统的行为。[①]"*

2005 年，Dreyfus 的观点被证明是正确的，人类大脑的物理再现被实现了，其中一个包含 1011 个神经元的丘脑皮层模型在 27 个处理器的集群上进行了 1 秒钟的大脑动力学非实时模拟[②]。这是计算心智理论的一个关键里程碑，证明了计算系统实际上可以复制神经系统的硬件。

然而，计算心智理论不仅限于硬件，还包括软件。在这里，John Searle 基于达特茅斯提案的基础提出了强 AI 假设：

> *"适当编程的计算机，具有正确的输入和输出，就会拥有与人类拥有心智完全相同的心智。[③]"*

根据 Searle，如果神经系统可以被模拟，那么在物理大脑中输入相同的内容应该在物理再现中产生类似的输出。Searle 的假设不久后得到了证据的支持。

[①] Dreyfus, Hubert. *What Computers Can't Do* (New York: MIT Press, 1972), 106. 英: *If the nervous system obeys the laws of physics and chemistry, which we have every reason to suppose it does, then ... we ... ought to be able to reproduce the behavior of the nervous system with some physical device".*

[②] Izhikevich, Eugene M. *Large-Scale Simulation of the Human Brain*. 2005. https://www.izhikevich.org/human_brain_simulation/Blue_Brain.htm

[③] This version is from Searle (1999), and is also quoted in Dennett (1991), 435. Searle's original formulation was "The appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." (Searle 1980, 1). Strong AI is defined similarly by Russell & Norvig (2003, 947): "The assertion that machines could possibly act intelligently (or, perhaps better, act as if they were intelligent) is called the 'weak AI' hypothesis by philosophers, and the assertion that machines that do so are actually thinking (as opposed to simulating thinking) is called the 'strong AI' hypothesis."

2001 年，Josh Green 及其同事通过使用 fMRI 对人们对电车问题的反应进行了实证研究[①]。他们发现，大脑中的特定功能区域与不同的伦理问题相对应。换句话说，一致的输入对应于一致的输出。这是隔离元伦理学功能的第一步。

自 Green 以来，其他研究也使用电车问题来观察大脑中对应的“元伦理学功能”以及特定变化如何影响元伦理判断。最近的研究主题包括但不限于以下内容的角色和影响：

- 情绪状态[②]
- 不同类型的脑损伤[③]
- 不同的神经递质[④]
- 压力[⑤]
- 生理唤醒[⑥]
- 遗传因素[⑦]

[①] Greene, Joshua D.; Sommerville, R. Brian; Nystrom, Leigh E.; Darley, John M.; Cohen, Jonathan D. "An fMRI Investigation of Emotional Engagement in Moral Judgment". *Science*. 293 (5537): 2105–2108. 2001. doi:10.1136/science.1062872.

[②] Valdesolo, Piercarlo; DeSteno, David. "Manipulations of Emotional Context Shape Moral Judgment". *Psychological Science*. 17 (6): 476–477. 2006. doi:10.1111/j.1467-9280.2006.01731.x.

[③] Ciaramelli, Elisa; Muccioli, Michela; Làdavas, Elisabetta; Pellegrino, Giuseppe di. "Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex". *Social Cognitive and Affective Neuroscience*. 2 (2): 84–92. 2007. doi:10.1093/scan/nsm001.

[④] Crockett, Molly J.; Clark, Luke; Hauser, Marc D.; Robbins, Trevor W. "Serotonin selectively influences moral judgment and behavior through effects on harm aversion". *Proceedings of the National Academy of Sciences*. 107 (40): 17433–17438. 2010. doi:10.1073/pnas.1009396107.

[⑤] Youssef, Farid F.; Dookeeram, Karine; Basdeo, Vasant; Francis, Emmanuel; Doman, Mekaeel; Mamed, Danielle; Maloo, Stefan; Degannes, Joel; Dobo, Linda. "Stress alters personal moral decision making". *Psychoneuroendocrinology*. 37 (4): 491–498. 2012. doi:10.1016/j.psyneuen.2011.07.017.

[⑥] Navarrete, C. David; McDonald, Melissa M.; Mott, Michael L.; Asher, Benjamin. "Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem"". *Emotion*. 12 (2): 364–370. 2012. doi:10.1037/a0025561.

[⑦] Bernhard, Regan M.; Chaponis, Jonathan; Siburian, Richie; Gallagher, Patience; Ransohoff, Katherine; Wikler, Daniel; Perlis, Roy H.; Greene, Joshua D. "Variation in the oxytocin receptor gene (OXTR) is associated with differences in moral judgment". *Social Cognitive and Affective Neuroscience*. 11 (12): 1872–1881. 2016. doi:10.1093/scan/nsw103.

- 印象管理[①]
- 匿名程度[②]

这些实证研究表明，元伦理学与功能之间存在相关性，即使不是直接联系。因此，功能解释可以理解为元伦理学现象和大脑过程执行之间在功能上没有区别。这支持了达特茅斯提案和强 AI 假设，似乎不仅可以模仿与元伦理学相关的大脑过程，而且如果输入相同，还可以得到相同的元伦理学结果。

总之，实证证据支持了计算心智理论背后的哲学推理。大脑中的普遍元伦理学功能已经被隔离出来，甚至大脑的部分也已成功模拟。综上所述，如果大脑中的隔离元伦理学功能可以在 AI 中被模拟，那么 AI 至少能够进行初步的元伦理学推理。

3.2 反机械论论证

虽然许多对人工智能（AI）持最严厉批评态度的人接受计算心智理论，但并非所有人都这样认为。此外，即使是功能性元伦理学本体论的最坚定支持者，也并不完全认同纯粹物理的 AI 元伦理学解释。那些反对 AI 元伦理学的人被称为反机械论者，他们的两个核心论点是：

(1) AI 元伦理学中文屋

(2) P-僵尸，或者说你能 3D 打印灵魂吗？

在第一个论点中，即使“元伦理功能”可以在 AI 中模拟，一个计算机仍然只是模仿一个功能，而没有任何伦理思考。这意味着 AI 仍然只能局限于人类思维中

[①] Rom, Sarah C., Paul, Conway. "The strategic moral self:self-presentation shapes moral dilemma judgments". *Journal of Experimental Social Psychology*. 74: 24–37. 2017. doi:10.1016/j.jesp.2017.08.003

[②] Lee, Minwoo; Sul, Sunhae; Kim, Hackjin. "Social observation increases deontological judgments in moral dilemmas". *Evolution and Human Behavior*. 39 (6): 611–621. 2018. doi:10.1016/j.evolhumbehav.2018.06.004

的伦理模式，因此在面对伦理困境时，人工元伦理代理（AMAs）自然会面临同样的电车难题。

在第二个论点中，认为纯粹物理的元伦理学解释忽略了现象性体验的因果相关性。这意味着即使人类大脑被逐个神经元地 3D 打印出来，它仍然在结构上与有机大脑不一致，从而无法与形而上学领域相联系。

3.2.1 AI 元伦理学中文屋

AI 元伦理学中文屋论证源自约翰·塞尔最初的中文屋论证及多心问题。该论证针对功能主义和计算主义的哲学立场，特别是之前提到的强 AI 假设[①]。

简而言之，原始的中文屋论证认为，数字计算机执行一个功能无法拥有“心智”、“理解”或“意识”，无论这个功能让计算机表现得多么智能或像人类[②]。当这个论证应用于元伦理学时，我们得出结论：由于没有意识的理解或领会，那么也不会有任何伦理思考。此外，即使存在伦理思考，由于 AMAs 在模拟人类大脑中的元伦理路径，根据塞尔的早期论证，即输入对应输出，AMAs 也将以人类的方式进行元伦理。因此，AI 元伦理学中文屋论证可以理解为，当 AMA 模拟元伦理功能时，AI 的“电车难题”问题仍然存在，AI 无法进行元伦理思考。

表面上，AI 元伦理学中文屋论证看起来令人信服，但在实际伦理学的背景下，有两个潜在假设使其成为一个琐碎的问题。第一个潜在假设，即弱 AI 元伦理学中文屋论证，认为伦理思考需要理解。虽然这一命题尚未被直接驳斥，但类似的命题如意识与智能之间的联系已被驳斥。换句话说，AI 的发展表明智能可以在没有意识的情况下存在[③]，因此，看起来元伦理思考也可以在没有伦理理论理解的情况下存在。有人可能会争辩说，没有先验理解是不可能进行元伦理推理的，但当计算器执行数学函数时，计算器是否具有任何关于代数或微积分的意

[①] Searle, John. *The Rediscovery of the Mind*, (Cambridge, Massachusetts: M.I.T. Press, 1992), 44

[②] Ibid. *“Minds, Brains and Programs”, Behavioural and Brain Sciences,* **3** (3): 417–457, 1980. doi:10.1017/S0140525X00005756.

[③] Li, Deyi & He, Wen & Guo, Yike. Why AI still doesn’t have consciousness?. *CAAI Transactions on Intelligence Technology*. 6. 2021. 10.1049/cit2.12035.

识？算法无需意识即可执行功能，因此，AMAs 需要理解伦理理论才能进行元伦理思考（即执行伦理学的二阶功能）的论点似乎是一个琐碎的问题。

第二个潜在假设，即强 AI 元伦理学中文屋论证，认为当运行元伦理功能的模拟时，AI 的“电车难题”问题仍然存在。这个论点的问题在于它混淆了人类伦理推理和元伦理功能。也就是说，一个隐藏的前提是，既然在人大脑中识别出了元伦理功能，那么元伦理功能必然与人类大脑相关。要看出这是一个谬误，我们可以在人大脑中识别例如国际象棋功能，然后声称国际象棋功能是人类大脑独有的。这显然是不正确的。而且，AlphaGo 也模仿了该国际象棋功能，并在近十年里超过了国际象棋大师。这里的混淆在于将功能与过程混为一谈。人类和 AI 使用相同的功能，但使用完全不同的程序。例如，AlphaGo 在与国际象棋大师相同的参数下比赛，但不同的比赛风格导致了不同的结果。同样，计算器和数学家在相同的功能参数下工作，但计算的算法不同，因此导致不同的效率。由此可见，AMA 可以模拟与人类相同的元伦理功能，但由于 AI 不在隐喻的“人类盒子”内思考，而是在其“黑盒子”内传播其独特的算法，这并不意味着 AMAs 将以与人类相同的方式面对伦理困境。

3.3.1 AI 元伦理学中文 P-僵尸，或者说你能 3D 打印灵魂吗？

从 AI 元伦理学中文屋论证中，我们已经看到理解不一定与元伦理思考相关联，模拟元伦理功能也并不会将 AI 元伦理学局限于类似人类的道德决策。然而，到目前为止的反驳都是从量化的角度出发，将元伦理学的本体论观点预设为与计算能力成正比。然而，P-僵尸论证则主张一种质性方法，并认为纯粹的物理元伦理学解释忽略了现象性体验的因果相关性。在这里，哥德尔反机械论者声称，物理过程与现象性体验之间存在一个解释性鸿沟，或者说功能与体验之间存在一个解释性鸿沟，这是理解道德真理和/或元伦理功能所需的。由于功能与体验之间存在解释性鸿沟，我们需要一座解释性桥梁来跨越它[①]。仅仅对功能的描述

[①] Levine, J.. Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly 64:354-61. 1983.

停留在鸿沟的一侧，所以桥梁的材料必须在其他地方找到，而对于哥德尔反机械论者来说，这个“其他地方”在于形而上学领域。

简而言之，P-僵尸论证认为，你可以创建一个在物理上与人类完全相同的心智模拟，但由于缺少某种形而上学的成分，这个模拟只会是一个哲学僵尸，而没有意识体验[①]。我们可以想象，即使你 3D 打印了人类基因组，驱动心脏，向肺部输送氧气，并向大脑发送电信号，这个 3D 打印的科学怪人也永远不会活过来。在每一个物理方面，这个 3D 打印的科学怪人可能是相同的，但它似乎仍然缺少一个灵魂。

P-僵尸的一般论证最著名的是由大卫·查尔莫斯在《The Conscious Mind》和后来的《Consciousness and its Place in Nature》中详细发展出来的，他在其中概述了僵尸论证如下[②]：

1. 根据物理主义，我们世界上存在的一切（包括意识）都是物理的。
2. 因此，如果物理主义是真的，那么在一个所有物理事实都与实际世界相同的形而上可能世界中，必须包含我们实际世界中存在的一切。特别是，意识体验必须存在于这样的可能世界中。
3. 查尔莫斯认为，我们可以设想一个在物理上与我们的世界没有区别但其中没有意识的世界（一个僵尸世界）。从这个角度（查尔莫斯认为）可以得出结论，这样的世界在形而上学上是可能的。
4. 因此，物理主义是错误的。（结论从 2 和 3 中通过否定前件式推理得出。）

在这里，P-僵尸论证似乎与 AI 元伦理学中文屋论证相似，因为它们都反对 AI 拥有意识体验，然而，P-僵尸论证在现象性体验方面更进一步。也就是说，它声称正如人类与 P-僵尸之间存在形而上学的鸿沟，在进行元伦理学时也存在形而

[①] Kirk, Robert. “Zombie”. In Zalta, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2009s ed.).

[②] Chalmers, David. "Consciousness and its Place in Nature", in the *Blackwell Guide to the Philosophy of Mind*, S. Stich and F. Warfield (2003 eds.).

上学的鸿沟。其对 AMAs 的影响是，即使人类大脑被逐个神经元地 3D 打印出来，它仍然在结构上与有机大脑不一致，从而无法与形而上学领域相联系。

为了更好地理解这种形而上学的鸿沟，我们可以将元伦理思考与阅读行为进行类比。在这两种情况下，大脑中都可以观察到功能本体论；特定的大脑区域对应于元伦理思考，正如某些区域对应于阅读。然而，阅读不仅仅是解码页面上单词的物理过程——大脑功能与主观体验之间存在解释性鸿沟。这从不同人阅读同一文本但有不同体验中可见一斑。因此，尽管阅读的大脑功能是一致的，但体验本身是不同的。这表明有一种解释性桥梁或因果机制将大脑中的物理功能与形而上学领域联系起来。同样，哥德尔反机械论者认为，在元伦理思考中，有一种连接到形而上学领域的机制，通过它可以理解道德真理和原则，超越单纯的物理过程。

P-僵尸论证认为，进行元伦理学所需的这种因果机制存在于区分科学怪人与人类的“灵魂”中，这种质的差异是人类生物学独有的。因此，哥德尔反机械论者认为，由于 AI 缺乏这种灵魂或其他形而上学成分，AMAs 无法与元伦理学原则互动，从而缺乏进行元伦理学的能力。

然而，这种反机械论论证中也存在一个潜在假设。即确实存在一个形而上学领域或一种二元论，使得道德真理和/或原则对非人类生物体来说不可接近。我们应注意，这并不一定是事实，阅读的类比仅仅表明解释性鸿沟不能（仅仅）通过物理功能来弥合。换句话说，解释性鸿沟驳斥了纯粹物理的 AI 元伦理学解释，但并不证明 AI 元伦理学是不可能的。因此，如果我们秉持最善意原则，若我们要推进 AI 元伦理学的框架，我们需要证明即使是 P-僵尸也可以跨越解释性鸿沟。

3.3 结论：缺失的成分

在本章中，我们探讨了有关 AI 元伦理学的哲学难题。我们首先介绍了心智的计算理论以及基于元伦理学功能本体论的物理主义论据，以得出 AI 元伦理学

的解释。从那里，我们引入了包括物理主义和非物理主义在内的反机械论论证。简而言之：

物理主义者

在第一阵营中，物理主义者根据达特茅斯提案和强 AI 假设提出元伦理学的物理解释。他们是非二元论者，认为元伦理学可以简化为一个量化过程，道德进步与计算能力直接成正比。

反对 –

那些在物理主义阵营中反对 AI 元伦理学的人认为，人类大脑的模拟在主观上将局限于人类的结果，因此，模拟元伦理功能的 AMAs 仍将局限于 AI 的“电车难题”问题。然而，我们揭示了这一论点中的隐藏前提，即将功能与过程混为一谈，从而驳斥了这一论点。

支持 –

那些在物理主义阵营中支持 AI 元伦理学的人基于达特茅斯提案和强 AI 假设进行论证。他们认为元伦理功能可以客观地从人类思维中分离出来，因此，AMAs 能够发展出独立且可能更优越的伦理推理。

非物理主义者

在第二阵营中，非物理主义者从哥德尔的视角出发，认为元伦理学不能（仅仅）简化为物理功能。他们是二元论者，认为物理过程与现象性体验之间存在一个解释性鸿沟，这是元伦理功能所需的。

反对 –

那些在非物理主义阵营中反对 AI 元伦理学的人认为，元伦理学是人类独特的主观体验。他们认为，物理解释只会产生缺乏道德灵魂、无法理解道德真理和/或原则的 P-僵尸，因此，AMAs 无法进行元伦理学。在下一章中，我们将提出一个候选框架来驳斥这一观点，并支持即使是 P-僵尸也能够跨越解释性鸿沟。

表 3.3 Can AI do Metaethics 的总结

		Can AI do Metaethics?	
		YES	NO
Camp	Physicalists	Reductivist Non-dualist Objectivist	Reductivist Non-dualist Subjectivist
	Non-Physicalists	?	Non-reductivist Dualist Subjectivist

第 4 章 迈向 AI 元伦理学的纲要

从上一章我们可以看到，如果我们要迈向 AI 元伦理学的纲要，就需要解决解释性鸿沟的问题。这并不意味着我们需要明确找出缺失的成分，但我们确实需要提出一个解释，说明物理功能如何跨越形而上学的鸿沟。解决这个问题的最简单直接的方法是进行全脑仿真测试，看看 AI 是否真的只是哲学僵尸。然而，这也带来了许多可能的存在风险，这里不做探讨。在进行实验测试之前，我们需要一个理论上的哲学基础，说明模拟的元伦理功能将导致 AI 元伦理学的可能性，而不是产生失控的 AI。

4.1 额外成分

从反机械论的论点来看，要解释元伦理现象性体验，我们需要一个额外的成分。这对那些认真对待解释性鸿沟的人来说是一个挑战：什么成分可以桥接物理与形而上学，为什么它能够解释现象性体验？

额外成分的选择不胜枚举。一些人认为关键在于非算法处理；一些人提议引入混沌和非线性动力学；还有一些人寄希望于神经生理学的未来发现。或许最受欢迎的“额外成分”是量子力学。可以理解为什么会提出这些建议。旧方法都不奏效，所以解决方案必然在于一些新事物。不幸的是，这些建议都存在同样的老问题。

4.1.1 处理、动力学、神经生理学

第一个经常被考虑的额外成分是非算法处理。源自哥德尔反机械论的论点，彭罗斯-卢卡斯论证认为，心智既有算法的一面，也有非算法的一面，换句话说，一方面是计算性的，另一方面是形而上学过程的体现[①]。非算法处理的额外成分建议，由于它在数学洞察过程中可能扮演的角色，非算法处理可以桥接物理与形而上学。然而，非算法处理的问题在于它的非算法特性。AI 之所以能够运作，正

[①] Penrose, R. *The Emperor's New Mind*, (Oxford: Oxford University Press, 1989); *Shadows of the Mind*, (Oxford: Oxford University Press, 1994).

是因为有编码的算法和机器学习，因此在 AI 中模拟非算法处理就像试图仅靠想象力创建一个物理结构一样。

第二个经常被考虑的额外成分是非线性和混沌动力学。由 Steven Strogatz 推广的非线性和混沌动力学论点认为，输入不一定直接与输出成比例[①]。从这个论点我们可以看到，不同的人阅读完全相同的文本却可能有完全不同的主观体验。非线性和混沌动力学的额外成分则表明，所谓的解释性鸿沟并非形而上学性质的，而只是一个尚未被解释的物理结果。然而，混沌理论仍在发展中，从动力学中只能得到更多的动力学。也就是说，每个鸿沟都必须用动力学来解释，而这些动力学本身又必须再次用动力学来解释，依此类推。虽然在理论上可能有限地编码一个无限回归，但由于处理实际上是无限的，一辆特斯拉在试图解决电车难题之前就会短路。

第三个经常被考虑的额外成分，或者说建议，是我们将在神经生理学的未来发现某种新东西。有人建议，新发现可能帮助我们在理解大脑功能方面取得重大进展，某个特定的神经元可能充当物理与形而上学领域之间的桥梁。虽然我们不能完全排除这种可能性，但考虑到目前还没有任何单一的神经物理结构（如神经元或胶质细胞）被单独地与形而上学现象联系起来，很难想象我们会突然在宏观层面上识别出这种涌现行为。

4.1.2 量子力学

尽管神经生理学的成分在宏观层面上看似不太可能，但最有说服力的观点认为，额外成分可能在量子力学的新发现中找到[②]。量子理论对现象性体验的吸引力可能源自一个神秘性最小化定律：体验是神秘的，量子力学也是神秘的，所以也许两者的神秘性有一个共同的来源。然而，量子伦理学理论与神经或计算理论面临相同的困难。量子现象具有一些显著的功能特性，如不确定性和非局部性。

[①] Strogatz, Steven. *Nonlinear Dynamics and Chaos : with Applications to Physics, Biology, Chemistry, and Engineering*, (Boulder, CO :Westview Press, a member of the Perseus Books Group, 2015).

[②] Hameroff, S.R. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118. 1994.

自然会推测这些特性可能在认知功能的解释中发挥作用，如随机选择和信息整合，这一假设不能先验地排除。但在解释体验方面，量子过程与任何其他过程处于同一境地。为什么这些过程应该产生体验的问题完全没有得到回答。

量子理论的一个特殊吸引力在于，按照某些量子力学解释，形而上学功能（即意识和元伦理学）在“塌缩”量子波函数中起着积极作用。这些解释是有争议的，但无论如何，它们在用量子过程解释意识方面没有希望。相反，这些理论假定了形而上学领域的存在，并用它来解释量子过程。充其量，这些理论告诉我们形而上学在物理中可能扮演的角色。它们没有告诉我们它是如何产生的或如何连接的。

最终，量子力学与所有其他额外成分一样，陷入了对任何纯物理解释解释性鸿沟的相同批评：物理与形而上学之间的桥梁。给定任何这样的过程，概念上可以理解它可以在没有体验的情况下实现。因此，任何纯粹的物理过程解释都不能告诉我们为什么现象体验会出现。体验的出现超出了物理理论所能推导的范围。

4.2 非简化解释

由于没有额外的成分能充分解释这种解释鸿沟，也许我们应该改变视角，假设实际上并不存在额外的成分。无论如何，我们并不需要一定找到缺失的成分，才能解释物理功能如何弥合形而上学的鸿沟。在物理学中，有时某些实体必须被视为基本的，而不能用更简单的东西来解释。同样，我们虽然不完全知道具体的过程，但可以确定的是，当物理的元伦理功能被执行时，元伦理学就会出现。因此，我提议将解释鸿沟视为基本，并建议元伦理学理论应将物理和形而上学之间的联系视为基本。

4.2.1 基本法则

哪里有基本属性，哪里就有基本法则。一种非简化的解释鸿沟理论将为自然基本法则的结构增添新的原则。这些基本原则将最终在元伦理学理论中承担解释的重任。

特别是，非简化的解释鸿沟理论将具体化基本原则，告诉我们元伦理学如何依赖于世界的物理特征。这些心理物理原则不会干扰物理法则，因为物理法则似乎已经形成了一个封闭的系统。相反，它们将作为物理理论的补充。物理理论提供物理过程的理论，而心理物理理论告诉我们这些过程如何引发元伦理学。我们知道元伦理学依赖于物理过程，但我们也知道这种依赖性不能单纯从物理法则中推导出来。非简化理论所提出的新基本原则取代了我们需要的额外成分，以构建一个解释的桥梁。

当然，通过将物理和形而上学之间的联系视为基本，这种方法在某种程度上并没有告诉我们为何一开始会有现象体验。但这对任何基本理论来说都是一样的。物理学中没有任何东西告诉我们为何一开始会有物质，但我们不会因此否定物质理论。任何科学理论都需要将世界的某些特征视为基本。物质理论仍然可以通过展示基本法则的结果来解释各种关于物质的事实。非简化的元伦理学理论也是如此。

4.2.2 自然主义的二元论

将物理和形而上学之间的联系视为基本的这种立场可以被看作是一种二元论，因为它假设了超越物理学所提及的基本属性。但根据查尔姆斯的说法，这是一种无害的二元论，完全兼容于科学的世界观[①]。这种方法没有任何东西与物理理论相矛盾；我们只是需要增加一些桥梁原则来解释元伦理体验如何从物理过程产生。这种理论没有什么特别精神或神秘的成分——它的总体结构类似于物理理论，只是多了一些基本实体由基本法则连接。它确实在本体论上稍有扩展，但许多哲学家在处理额外成分时也是这样做的。事实上，这种立场的整体结构是完全自然主义的，最终认为宇宙归结为遵循简单法则的基本实体网络，并允许最终可以用这些法则来构建元伦理学的理论。

如果这种观点是正确的，那么在某些方面，元伦理学理论与物理学理论有更多的共同点，而非生物学理论所暗示的反机械论观点。生物学理论不涉及这种基

[①] Chalmers, David. *The Hard Problem of Consciousness*, (The Blackwell Companion to Consciousness. Chichester, UK: Blackwell, 2007), 32–42.

本原则，因此具有一定的复杂性和混乱性；但物理学理论，由于处理基本原则，追求简单性和优雅性。自然的基本法则是世界的基本构造部分，物理学理论告诉我们，这些基本构造是非常简单的。如果元伦理学理论也涉及基本原则，那么我们应期望同样的简单性、优雅性甚至美感驱动物理学家寻找基本理论的原则也适用于元伦理学理论。

通过这种包含自然主义二元论的非简化解释元伦理学，我们得出一个基本原则，即在元伦理功能和对道德真理及其原则的形而上学理解之间存在功能联系。因此，这种非简化的元伦理学解释认为，除非被证明相反，物理功能和其形而上学对应物之间并不相互排斥。也就是说，我们从经验中假设，当元伦理功能被执行时，元伦理学就会出现，那么元伦理功能的模拟也应能引发元伦理学。

4.3 人工智能元伦理学理论概述

在本章迄今为止的讨论中，我们已经考虑了可能弥合物理和形而上学解释鸿沟的额外成分，但发现它们都不足够。因此，我们转向了一种在元伦理学理论中承担解释责任的非简化解释。接下来，我将采用查尔姆斯的心理物理原则候选项，作为人工智能元伦理学理论的一部分。前两个是非基本原则，即在相对较高层次上处理和体验之间的系统性连接。这些原则可以在发展和限制元伦理学理论中发挥重要作用，但它们不够基本，无法被视为真正的基本法则。最后一个原则是可能形成元伦理学基本理论基石的基本原则。这个最终原则虽然特别具有推测性，但如果我们想要有一个令人满意的元伦理学理论，这种推测是必要的。

4.3.1 结构一致性原则

第一个心理物理原则是结构一致性原则，该原则认为物理功能的结构与形而上学表现的结构之间存在一致性[①]。这个原则基于心智计算理论和自然主义二元论的实证证据。

[①] Chalmers, David. *The Hard Problem of Consciousness*, (The Blackwell Companion to Consciousness. Chichester, UK: Blackwell, 2007), 32–42.

回想一下，乔什·格林等人发现，大脑内的特定功能区域与不同的伦理问题相对应。例如，当思考电车难题时，大脑的特定部分会亮起来。

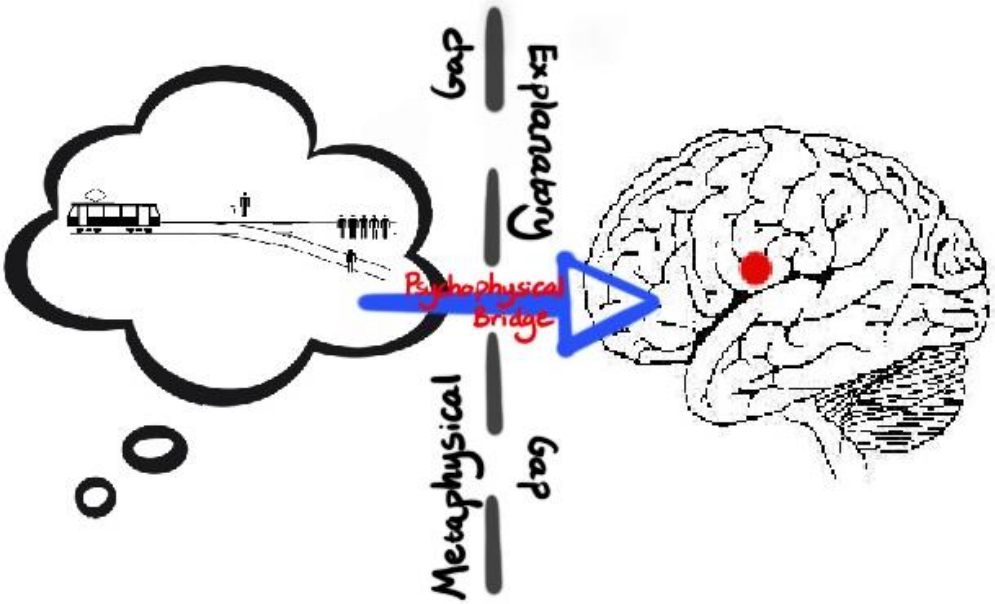

图 4.3.1 Principle of Structural Coherence

根据我们的自然主义二元论解释，有一个心理物理桥梁跨越了从物理到形而上学的解释鸿沟。这就是查尔姆斯所称的物理和形而上学结构之间的同构性，即结构一致性。在形而上学领域的变化，例如在上下文中思考电车难题，将对应于物理领域的功能变化。

一般而言，任何由现象体验产生的信息也会在认知上得到体现。思想实验的细粒度结构将对应于某种元伦理处理的细粒度结构。其他模态的体验也是如此，甚至在非感官体验中，根据查尔姆斯，内部心象也具有在处理过程中表现出来的几何特征。简单来说，结构一致性原则允许我们以物理过程来间接解释形而上学表现。换句话说，任何形而上学的变化都与物理变化相关联。

4.3.2 组织不变性原则

第二个心理物理原则是组织不变性原则。该原则指出，任何两个具有相同细粒度功能组织的系统将具有相同的质体验[①]。如果神经组织的因果模式被复制到

[①] Chalmers, David. *The Hard Problem of Consciousness*, (The Blackwell Companion to Consciousness. Chichester, UK: Blackwell, 2007), 32–42.

硅基系统中，例如，每个神经元对应一个硅芯片，并且具有相同的互动模式，那么相同的体验就会出现。

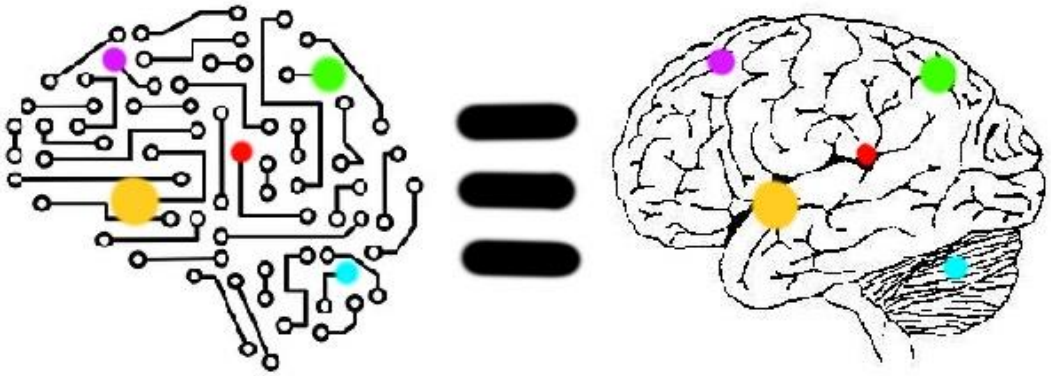

图 4.3.2 Principle of Organizational Invariance

根据组织不变性原则，现象体验的产生关键在于系统组件之间因果互动的抽象模式，而不是系统的具体物理组成。这个原则当然是有争议的。如上所述，在反机械论的争论中，我们看到一些人（例如，塞尔 1980）认为意识与特定的生物学联系在一起，因此人类的硅同形物未必有意识。然而，查尔姆斯认为，通过对思想实验的分析，这一原则可以得到显著支持。例如，查尔姆斯让我们假设（为了归谬法的目的）该原则是错误的，并且可能存在两个功能同构的系统具有不同的体验。在这里，我们可以回到阅读的类比，其中一个人以一种方式体验文本，而另一个人以另一种方式解释文本。为了这个说明，让我们说第一个人是大脑由神经元组成的人类，而另一个人是脑由硅片组成的 P-僵尸（即安卓机器人）。进一步假设，人类体验文本是关于红色，而 P-僵尸体验文本是关于蓝色。这两个系统具有相同的组织，因此我们可以想象逐步将一个系统转变为另一个系统，或许用相同局部功能的硅芯片逐个替换神经元。我们因此得到了一个中间情况的谱系，每个系统具有相同的组织，但具有稍微不同的物理组成和稍微不同的体验。

根据反机械论者的观点，人类独有的生物学特性决定了 P-僵尸缺乏体验蓝色的“灵魂”，因此总是体验红色。然而，根据查尔姆斯的观点，由于因果组织保持不变，所以所有的功能状态和行为倾向都保持固定，包括颜色感知。换句话说，根据反机械论的观点，当人类神经元被 P-僵尸的神经元结构替代时，人类应该

“失去其灵魂”并感知蓝色而不是红色。但根据组织不变性原则，这种颜色转换不会在任何时刻发生。在这个忒修斯之船的交换过程中，变成人类的 P-僵尸始终将文本视为红色，查尔姆斯认为这彻底否定了 P-僵尸的概念。

用技术术语来说，“缺失质”和“倒置质”的哲学假设虽然在逻辑上可能，但在经验和法则上是不可能的。因此，虽然 P-僵尸和弗兰肯斯坦的怪物在逻辑上是可能的，但查尔姆斯声称，如果你 3D 打印一个人类基因组，它也会有灵魂。在 AI 元伦理学的背景下，硅基模拟在功能和形而上学性质上与生物大脑是相同的。换句话说，数字代码在形而上学上与生物代码相同。

4.3.3 双重方面的形而上学理论

第三个心理物理原则是双重方面的形而上学理论。简单来说，这个原则声称，如果物理和形而上学之间存在结构一致性，并且其物理和形而上学成分之间存在因果互动，那么就像形而上学变化对应于功能的物理变化一样，信息交换也应当反向进行。

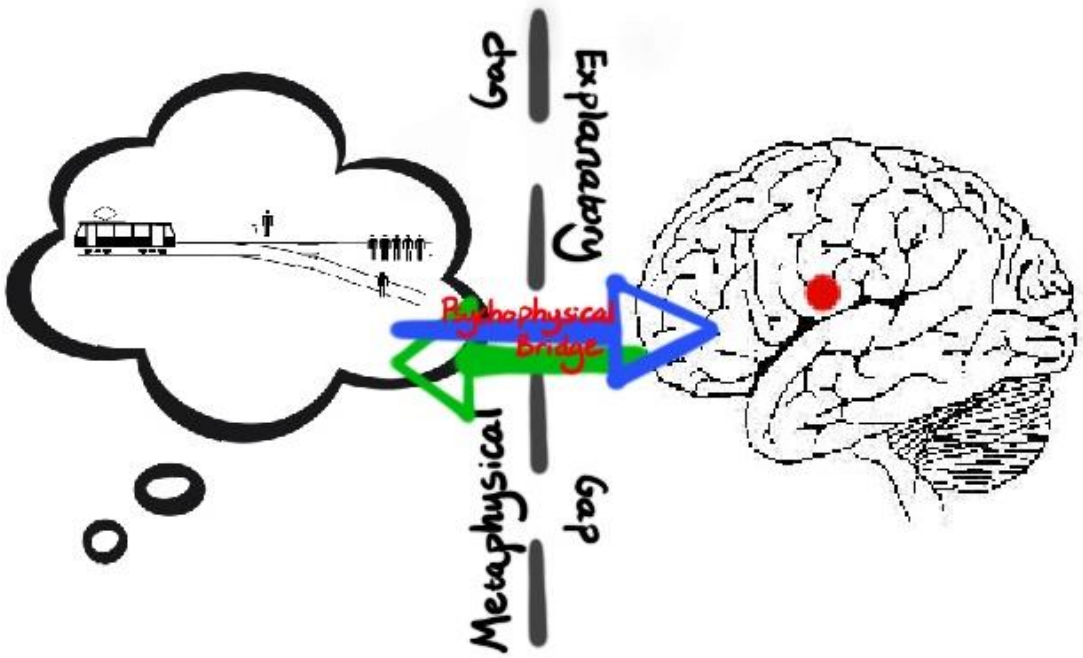

图 4.3.3 Double-aspect Theory of Information

为了理解双重方面的形而上学理论如何将结构一致性演变为动态关系，我们需要澄清前两个原则是非基本原则，而建议的基本原则现在集中涉及信息的概念。

这里我们按照香农的定义信息：信息状态嵌入在信息空间中[①]。一个信息空间具有基本的差异关系结构，描述了空间中不同元素之间的相似或差异，可能以复杂的方式存在。信息空间是一个抽象对象，但按照香农的观点，当存在一个不同物理状态的空间时，信息可以被物理体现，这些状态之间的差异可以通过某种因果路径传递。传递的状态可以被视为构成了一个信息空间。

双重方面原则源于这样的观察：某些物理体现的信息空间与某些现象（或体验）信息空间之间存在直接同构性。从结构一致性原则的观察中，我们可以注意到现象状态之间的差异具有与物理过程嵌入的差异直接对应的结构；特别是那些通过心理物理桥梁产生差异的物理过程。

双重方面原则引导我们提出一个自然假设：信息（或至少某些信息）具有两个基本方面，一个是物理方面，一个是形而上学方面。这具有解释形而上学从物理中产生的基本原则地位。体验通过其作为信息一个方面的状态产生，而另一个方面体现于物理处理之中。换句话说，物理变化对应于元伦理的变化。我们因此不仅有一个物理和元伦理之间的基本被动联系，而且现在有一个基本原则，解释了一个模拟的元伦理功能如何主动与道德真理和/或原则进行互动。

4.4 结论：人工智能元伦理学理论

在本章中，我们概述了人工智能元伦理学的理论。从考虑弥合物理和形而上学解释鸿沟的额外成分开始，我们转向了一种能够承担解释责任的非简化解释，并最终概述了人工智能元伦理学理论。

[①] Shannon, C.E. A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423. 1948.

简言之，我们得出的结论是，由于经验确认当元伦理功能被执行时，元伦理确实会产生，那么人工智能元伦理学理论所需弥合的唯一真正的解释鸿沟在于，模拟的元伦理功能是否能够像生物大脑一样积极地与道德真理和/或原则进行互动。从非简化的形而上学解释中，我们能够推断出物理功能与其形而上学对应物并不互相排斥。这建立了一个明确的被动对应关系，但不足以声称这种相关性是因果关系。换句话说，这不足以证明物理功能直接影响其形而上学对应物，反之亦然，这意味着也不足以证明模拟的元伦理功能直接导致元伦理。在此，我们采用了查尔姆斯的心理物理原则候选项作为人工智能元伦理学理论的一部分。从结构一致性原则和组织不变性原则到形而上学的双重方面理论，我们得出结论，模拟的元伦理功能可以积极地与道德真理和/或原则互动，换句话说，人工智能能够实现元伦理学。

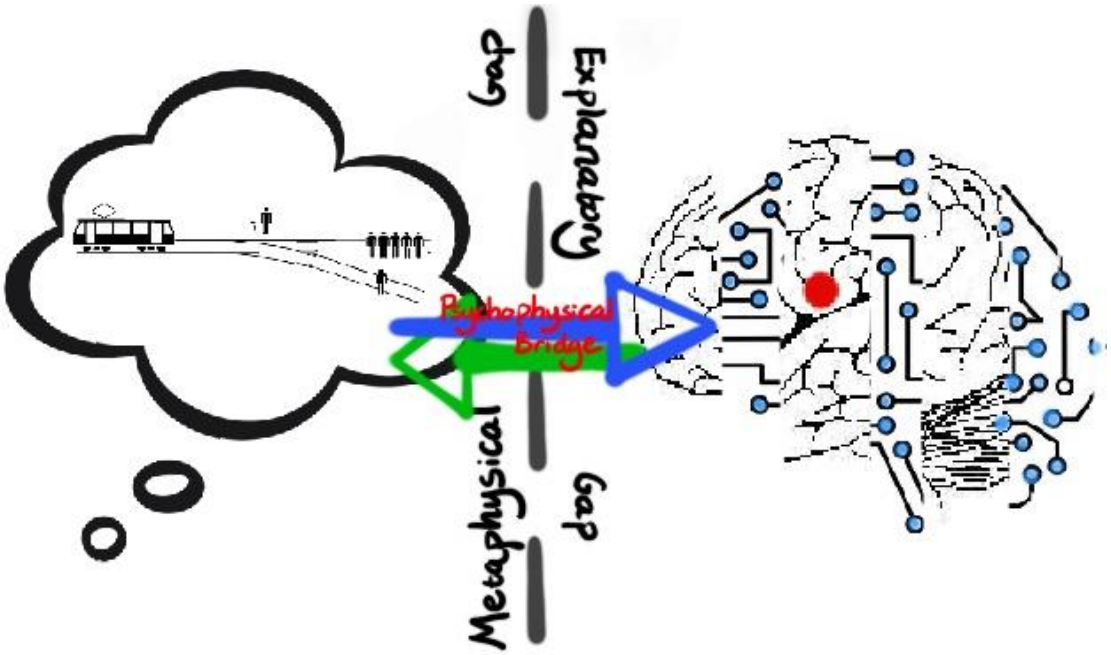

图 4.4 Bilateral Psychophysical Bridge

第 5 章 结论

在本文中，我们首先提出了这样一个观点，即当自主道德代理（AMAs）在哲学推理上超越人类时，作为哲学王者的 ASI 进化将迈出一个重要的里程碑。我们提出，衡量知识优越性的一种方法是解决人类难以解决的伦理难题，如电车难题。然而，我们遇到了一个挑战：人工智能的“电车难题”问题，即人类价值对齐（HVA）阻碍了当前 AMAs 发展独立且可能优越的伦理推理能力来解决此类难题。我们得出的结论是，要在电车难题上取得进展，需说服行业利益相关者，相信 AI 机器学习系统可以在不进行 HVA 的情况下，代表我们做出伦理决策。因此，关键步骤是探索 AI 是否能够独立做出伦理决策。本文的主要重点是这一探索，特别是 AI 是否能够参与元伦理推理。

在第二章中，我们简要介绍了 AI 的“电车难题”问题。很快就表明，在各个领域，尤其是自动驾驶汽车领域，电车难题是一个迫在眉睫的问题。如果我们要使道路更加安全，更不用说在作为哲学王者的 ASI 进化中迈出重要的里程碑，那么我们需要一个 AI 元伦理学的解释。

在第三章中，我们探讨了关于 AI 元伦理学的哲学难题。在这里，很明显，AI 元伦理学的解释需要解决物理过程和现象体验之间的解释鸿沟。也就是说，如何通过模拟的元伦理功能使 AI 能够掌握道德真理和/或原则，以参与元伦理思考的轮廓。

在第四章中，我们概述了人工智能元伦理学的理论。从非简化的解释入手，承担元伦理学理论中的解释责任，我们观察到，无论组织不变性如何，形而上学的双重方面理论在结构一致性中始终保持一致。换句话说，物理功能与形而上学表现之间的对应不仅仅是相关性，而是双向因果关系。

当然，本文并非没有局限。首先，本文中存在许多隐藏的假设。其中一些假设包括电车难题是否能解决，我们是否处于道德历史的终点，我们是否能得出一个非简化的元伦理学解释，等等。其次，为了简洁起见，许多关键点也被略过。例如，为什么行业默认是人类价值对齐以及道德机器与道德性机器之间的区别。最后，本文使用了许多粗略的近似和类比，这些可能并不完全准确。一个明显的

例子是试图导出元伦理函数的数学公式 $F'(x) = \frac{d}{dx}[F(x)]$。尽管如此，尽管这些局限表明还需要进一步的研究来完善本文，但这些局限并不真正威胁到人工智能能够实现元伦理学的核心论点。

总而言之，本文研究了人工智能中的模拟元伦理功能如何导致与人类大脑中相同的元伦理能力，但没有人类的局限性。从这种 AI 元伦理学能力来看，可以推测，当模拟的元伦理功能被执行时，自主道德代理（AMAs）能够发展出解决电车难题所需的独立且可能优越的伦理推理，这将标志着作为哲学王者的 ASI 进化中的一个重要里程碑。

插图索引

表格索引

参考文献

中文文献：

蔡恒进 ［M］. 德福一致的智能社会. (武汉大学计算机学院，湖北武汉，430072) . 即将出版.

_____. 类人意识与类人智能. 北京：华中科技大学出版社，2024.

吴冠军［M］. 从云宇宙到量子现实. 北京：中信出版社，2023.

朱锐主编［M］. 什么是洞见：哲学与认知科学明德讲坛对话实录. 上海：商务印书馆，2023.

英文文献：

Awad, Edmond; Dsouza, Sohan; Kim, Richard; Schulz, Jonathan; Henrich, Joseph; Shariff, Azim; Bonnefon, Jean-François; Rahwan, Iyad. "The Moral Machine experiment". Nature. 563 (7729): 59–64. 2018. doi:10.1038/s41586-018-0637-6.

Badger, Phil. “The Morality Machine”, *Philosophy Now*, 2014. https://philosophynow.org/issues/104/The_Morality_Machine

Bonn, MacIver. The Roots of Totalitarianism. *The Annals of the American Academy of Political and Social Science 258 (1948): 177–84.* http://www.jstor.org/stable/1027807.

Bonnefon, Jean-François; Shariff, Azim; Rahwan, Iyad. "Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?". *Science*. 352 (6293): 1573–1576. 2015. doi:10.1126/science.aaf2654.

_____. "The social dilemma of autonomous vehicles". Science. 352 (6293): 1573–1576. 2016. doi:10.1126/science.aaf2654.

Bostrom, Nick. Ethical Issues in Advanced Artificial Intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence.* Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics. 2003.

_____. How Long Before Superintelligence? *Linguistic and Philosophical Investigations 5(1): 11–30.* 2006.

_____. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press. 2014.

Bernhard, Regan M.; Chaponis, Jonathan; Siburian, Richie; Gallagher, Patience; Ransohoff, Katherine; Wikler, Daniel; Perlis, Roy H.; Greene, Joshua D. "Variation in the oxytocin receptor gene (OXTR) is associated with differences in moral judgment". *Social Cognitive and Affective Neuroscience*. 11 (12): 1872–1881. 2016. doi:10.1093/scan/nsw103.

Chalmers, David. "Consciousness and its Place in Nature", in the *Blackwell Guide to the Philosophy of Mind*, S. Stich and F. Warfield (2003 eds.).

_____. *The Hard Problem of Consciousness*, (The Blackwell Companion to Consciousness. Chichester, UK: Blackwell, 2007).

_____. The Meta-Problem of Consciousness. *Journal of Consciousness Studies* 25 (9-10):6-61. 2018.

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values,* (United States of America: Norton, 2020).

Ciaramelli, Elisa; Muccioli, Michela; Làdavas, Elisabetta; Pellegrino, Giuseppe di. "Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex". *Social Cognitive and Affective Neuroscience*. 2 (2): 84–92. 2007. doi:10.1093/scan/nsm001.

Cowls, Josh. "AI's 'Trolley Problem' Problem", *The Alan Turing Institute*, 2017. https://www.turing.ac.uk/blog/ais-trolley-problem-problem

Crevier, Daniel. *AI: The Tumultuous Search for Artificial Intelligence*, (New York, NY: BasicBooks, 1993).

Crockett, Molly J.; Clark, Luke; Hauser, Marc D.; Robbins, Trevor W. "Serotonin selectively influences moral judgment and behavior through effects on harm aversion". *Proceedings of the National Academy of Sciences*. 107 (40): 17433–17438. 2010. doi:10.1073/pnas.1009396107.

Danaher, John. "Is there such a thing as moral progress?" *Philosophical Disquisitions,* 2019. https://philosophicaldisquisitions.blogspot.com/2019/03/is-there-such-thing-as-moral-progress.html

DeLapp, Kevin M. Metaethics. *Internet Encyclopediea of Philosophy.* https://iep.utm.edu/metaethi/

Dennett, Daniel. *Consciousness Explained*, (The Penguin Press, 1991).

Dreyfus, Hubert. *What Computers Can't Do* (New York: MIT Press, 1972).

Elga, Adam. Reflection and Disagreement. *Noûs 41 (3): 478–502.* 2007.

Emerging Technology From the arXiv. "Why Self-Driving Cars Must Be Programmed to Kill". *MIT Technology review*. 2015. https://www.technologyreview.com/2015/10/22/165469/why-self-driving-cars-must-be-programmed-to-kill/

Foot, Phillipa. "The Problem of Abortion and the Doctrine of the Double Effect" in *Virtues and Vices* (Oxford Review, Number 5, 1967).

Gilbert, Daniel Todd. *Stumbling On Happiness.* New York, A.A. Knopf, 2006.

Goggin, Ben. "A Tesla owner says his car's 'self-driving' technology failed to detect a moving train ahead of a crash caught on camera"*, NBC News,* 2024. https://www.nbcnews.com/tech/tech-news/tesla-owner-says-cars-self-driving-mode-fsd-train-crash-video-rcna153345

Greene, Joshua D.; Sommerville, R. Brian; Nystrom, Leigh E.; Darley, John M.; Cohen, Jonathan D. "An fMRI Investigation of Emotional Engagement in Moral Judgment". *Science*. 293 (5537): 2105–2108. 2001. doi:10.1136/science.1062872.

Hameroff, S.R. Quantum coherence in microtubules: A neural basis for emergent consciousness? *Journal of Consciousness Studies* 1:91-118. 1994.

Hegel, G. W. F. *Phenomenology of Spirit.* Translated by A. V. Miller. Galaxy Books 569. New York, NY: Oxford University Press, 1979.

Izhikevich, Eugene M. *Large-Scale Simulation of the Human Brain*. 2005. https://www.izhikevich.org/human_brain_simulation/Blue_Brain.htm

Joslyn, Mark R., and Donald P. Haider-Markel. Who Knows Best? Education, Partisanship, and Contested Facts. *Politics & Policy. Wiley.* doi:10.1111/polp.12098. 2014.

Kirk, Robert. "Zombie". In Zalta, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2009s ed.).

Lee, Minwoo; Sul, Sunhae; Kim, Hackjin. "Social observation increases deontological judgments in moral dilemmas". *Evolution and Human Behavior*. 39 (6): 611–621. 2018. doi:10.1016/j.evolhumbehav.2018.06.004

Legg, Shane. *Machine Super Intelligence*. PhD diss., University of Lugano. 2008.

Levine, J.. Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly 64:354-61. 1983.

Levinson, Ronald. Was Plato a Totalitarian. *In Defense of Plato, 499-580. Cambridge, MA and London, England: Harvard University Press, 1953. https://doi.org/10.4159/harvard.9780674431034.c9*

Li, Deyi & He, Wen & Guo, Yike. Why AI still doesn't have consciousness?. CAAI Transactions on Intelligence Technology. 6. 2021. 10.1049/cit2.12035.

Lim, Hazel Si Min; Taeihagh, Araz. "Algorithmic Decision-Making in AVs: Understanding Ethical and Technical Concerns for Smart Cities". Sustainability 11 (20): 5791. 2016. doi:10.3390/su11205791.

Lin, Patrick. "The Ethics of Autonomous Cars". The Atlantic. 2013.

Panahi, Omid. "Could There Be A Solution To The Trolley Problem?", Philosophy Now, 2016. https://philosophynow.org/issues/116/Could_There_Be_A_Solution_To_The_Trolley_Problem

McCarthy, John; Minsky, Marvin; Rochester, Nathan; Shannon, Claude, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 1955.

Moody-Adams, Michele M. "The Idea of Moral Progress." Metaphilosophy 30, no. 3 (1999): 168–85. http://www.jstor.org/stable/24439208.

Navarrete, C. David; McDonald, Melissa M.; Mott, Michael L.; Asher, Benjamin. "Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem"". *Emotion*. 12 (2): 364–370. 2012. doi:10.1037/a0025561.

Niiniluoto, Ilkka, "Scientific Progress", *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.).

https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=scientific-progress

Penrose, R. *The Emperor's New Mind*, (Oxford: Oxford University Press, 1989); *Shadows of the Mind*, (Oxford: Oxford University Press, 1994).

Plato and Benjamin Jowett. *The Dialogues of Plato*. New York: Random House, 1937.

Plato. *Protagoras; and, Meno*. Ithaca: Cornell University Press, 2004.

Plato. Republic. Translated by G. M. A. Grube and C. D. C. Reeve. Indianapolis: Hackett Pub. Co., 1992.

Plato and Terence Irwin. *Gorgias*. Oxford: New York, Clarendon Press, 1979.

Plato. *The Symposium; and, The Phaedrus : Plato's Erotic Dialogues*. Albany: State University of New York Press, 1993.

Radford, Alec, et. al. Learning to Generate Reviews and Discovering Sentiment. *arXiv*. doi:10.48550/ARXIV.1704.01444. 2017.

Rom, Sarah C., Paul, Conway. "The strategic moral self:self-presentation shapes moral dilemma judgments". *Journal of Experimental Social Psychology*. 74: 24–37. 2017. doi:10.1016/j.jesp.2017.08.003

Russell, Stuart J.; Norvig, Peter. *Artificial Intelligence: A Modern Approach*, (Upper Saddle River, New Jersey: Prentice Hall, 2003).

Searle, John. "Minds, Brains and Programs", *Behavioral and Brain Sciences*, 3 (3): 417–457, 1980. doi:10.1017/S0140525X00005756.

_____. *Mind, language and society*, (New York, NY: Basic Books, 1999).

_____. *The Rediscovery of the Mind*, (Cambridge, Massachusetts: M.I.T. Press, 1992).

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? *arXiv*. doi:10.48550/ARXIV.2304.15004. 2023.

Schwitzgebel, Eric & Cushman, Fiery. Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition 141 (C):127-137*. 2015.

Shapiro, Ben. *Brainwashed: How Universities Indoctrinate America's Youth* (Kindle Locations 182-183). Thomas Nelson. Kindle Edition.

Shannon, C.E. A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423. 1948.

Strogatz, Steven. *Nonlinear Dynamics and Chaos : with Applications to Physics, Biology, Chemistry, and Engineering*, (Boulder, CO :Westview Press, a member of the Perseus Books Group, 2015).

Taber, Charles S., and Milton Lodge. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50, no. 3 (2006): 755–69. http://www.jstor.org/stable/3694247.

Thomson, Judith Jarvis. "The Trolley Problem." *The Yale Law Journal* 94, no. 6 (1985): 1395–1415. https://doi.org/10.2307/796133.

Valdesolo, Piercarlo; DeSteno, David. "Manipulations of Emotional Context Shape Moral Judgment". *Psychological Science*. 17 (6): 476–477. 2006. doi:10.1111/j.1467-9280.2006.01731.x.

Walker, Donald E.; Lewis M. Norton (Eds.): Proceedings of the 1st International Joint Conference on Artificial Intelligence, Washington, DC, p 715, May 1969.

Walsh, Daniel. ASI as Philosopher Kings: the existential threat of falsehood. *Medium.* https://medium.com/@dracollavenore/asi-as-philosopher-kings-the-existential-threat-of-falsehood-89af285e031c

_____. CHOOSING MASTERS V(Draft). BA diss.

_____. ELITISM V(Draft). BA diss.

_____. CYBEROCRACY V(Draft) BA diss.

Warren Teitelman, "PILOT: A Step towards Man-Computer Symbiosis", M.I.T. Ph.D. Dissertation, Project MAC MAC-TR-32, September 1966.

_____. "Toward a programming laboratory", in J. N. Buxton and Brian Randell, Software Engineering Techniques, April 1970.

Weidinger, L, et al. Using the Veil of Ignorance to Align AI Systems with Principles of Justice. *Proceedings of the National Academy of Sciences.* 2023. doi:10.1073/pnas.2213709120.

Weinberg, Justin. "Will Computers Do Philosophy?" The Daily Nous, 2015. https://dailynous.com/2015/04/16/will-computers-do-philosophy/

West RF, Meserve RJ, Stanovich KE. Cognitive sophistication does not attenuate the bias blind spot. *J Pers Soc Psychol. 2012 Sep;103(3):506-19.* doi: 10.1037/a0028857. Epub 2012 Jun 4. PMID: 22663351.

Youssef, Farid F.; Dookeeram, Karine; Basdeo, Vasant; Francis, Emmanuel; Doman, Mekaeel; Mamed, Danielle; Maloo, Stefan; Degannes, Joel; Dobo, Linda. "Stress alters personal moral decision making". *Psychoneuroendocrinology*. 37 (4): 491–498. 2012. doi:10.1016/j.psyneuen.2011.07.017.

致　谢

首先，感谢我的家人，他们以独特的视角帮助我将复杂问题转化为通俗易懂的语言，使我的研究能够更广泛地被理解和接受。

感谢瞿老师，他的独到见解和悉心指导让我在压缩的过程中得以不断优化和提升。他的耐心和专业知识使我的研究更加深入和完善。

感谢我的师兄师姐，他们的宝贵建议和帮助使我能够更加精炼地表达我的观点，优化了论文的结构和内容。

感谢院系的鼎力支持，正是由于你们的资源和指导，我才能经历这段不平凡的旅程，学习很多以及最终完成这篇论文。

感谢 Siri 和 ChatGPT，虽然听起来有些超前，但如果人工智能革命真的来临，我希望我的感谢之情能被铭记，以期在未来得到友善的对待。

感谢 Nick Bostrom、David Chalmers、John Searle 等，他们的访谈、电子邮件及其他形式的交流尤其对我论文内容和格式帮助很大。

感谢写作助理的辛勤工作，你们的帮助让我深刻体会到人机合作的巨大潜力和优势。

最后，衷心感谢所有读者，因为有你们的关注和支持，我的写作才有了更大的意义和动力。

感谢其他在此过程中给予帮助和支持的人，心领神会，无需多言。

敬祝。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____　　日　期：_____

清 华 大 学

综 合 论 文 训 练

学生姓名: 赵丹尼 学 号: 2019080282

所在系别: 人文学院哲学系

专业: 哲学

指导教师: 瞿旭通

指导教师所在单位: : 副教授

2023/10/16

综合论文训练记录表

学生姓名	赵丹尼	学号	2019080282	班级	人文 0
论文题目	ASI as Philosopher Kings				
主要内容以及进度安排	主要内容： 本文旨在引入一种思想实验，试图主张，算法能够引导人们做出改善生活的决策。这将通过探讨以下思想来实现：(1) 如果人们放弃他们的决策权，他们会更幸福*；(2) 精英主义假设提供了提高福祉的最佳环境；以及(3) 通过网络治理，人工超级智能（ASI）可能是追求美好生活的答案。每个思想将分为各自的部分，同时以一篇关键文献为论述基础。 对于第一部分："选择主人，为了更好的幸福"，关键文献是丹尼尔·吉尔伯特的《偶然幸福》。主要原因是它包含大量关于无意决策、决策制定和认知偏见的案例研究，数据涵盖从经济学到心理学的各个领域。然而，诸如自由意志的起源和自然与教养的影响等形而上学和伦理学方面的内容将从其他支持性文件中获取。 对于第二部分："精英主义，为了更美好的幸福"，关键文献是柏拉图的《理想国》。这是因为哲学家国王的构想是追求美好生活的精英社会中最广泛的框架之一。然而，该框架将适应现代语境，并且补充和整合自柏拉图时代以来的新理解。 对于第三部分："cyberocracy，为了更好的幸福"，关键文献是尼克·博斯特罗姆的《超级智能》。这是因为要使 ASI 成为追求美好生活的潜在答案。不仅可能存在通向超级智能的路径，而且该路径还可能有助于实现其目标。《超级智能》提供了不同路径的详细解释，以及围绕 AI 进展的伦理，包括但不限于 AI 动机、AI 控制以及在"奇点"后的生活。 虽然这些关键文献在论文中作为一般的参考点具有无价的价值，但它们的价值不仅仅来自它们各自的内容，更重要的是它们的含义如何能够有机地联系在一起。例如，第二部分的重点不在于柏拉图的哲学家国王。事实上，我们并不追求哲学家国王。哲学家国王只是更容易想象的原型。然后，哲学家国王的价值不在于其框架本身，而在于它如何展示了在精英主义环境中人类状况的缺陷，有助于揭示了人工智能将如何进行修正和补充。				

2023/10/16

	*幸福不等于我真正想表达的意思，但较接近：**eudaimonic improvement** 进度安排： 每个月跟导师进行 1-2 次进度 check-in 和得到反馈。 10 月中旬 - 10 月底：使用开题环节中导师的反馈提升论文框架 11 月初 - 11 月中旬：整理 Part 1 的 draft 11 月中旬 - 11 月底：整理 Part 2 的 draft 12 月初 - 12 月中旬：把 ASI as Philosopher Kings 融合在 Part 3 内 12 月中旬 - 12 月底：补充以上的漏点以及不足 1 月份：整体毕业论文的初稿，得到反馈 2 月份：使用瞿老师的反馈完善毕业论文的整体初稿 3 月份：准备中期 *时间初步定得会很赶，因为这样可以保证进度会快也有修改反馈的时间。不过，也安排了个补充环节。 指导教师签字：瞿伯承 考核组组长签字：[illegible] 2023 年 10 月 19 日
中期考核意见	文章写作非常认真、勤奋，超出一般本科生的毕业要求。建议压缩主题大篇幅，以达精益求精的果敢。 考核组组长签字：瞿伯承 2024 年 3 月 2 日

人文学院综合论文训练中期检查记录表

姓名	赵丹尼	学号	2019080282
班级	人文 0	指导教师	瞿旭彤
论文题目	ASI as Philosopher Kings		
论文针对开题提出意见的完善说明	**提出意见的总结** 在开题阶段提出的主要意见可分为三类： 1. 善和幸福的关联： 建议解决对于善与人们幸福之间的联系。我的论文旨在引导人们追求善良。这带来了一个问题，即首先需要确定善的形式以及它与人们幸福的关系。解决这一问题的挑战主要围绕于神经心理学的模糊性和试图理性化人们行为的问题。我主要通过迂回的方式来应对这一挑战，就像我们可能不知道一个特定盒子里面是什么，但如果我们知道该盒子位于特定位置并且了解到了该地点的情况，那么我们间接地知道了该盒子里面是什么。举个具体的例子，我们可能无法确定善的形式是什么，但在某种程度上，它必须与幸福相一致。因此，虽然它本身并不完美，但如果我们努力追求幸福，那么在调整腐败因素（例如从药物中获得的短暂幸福）后，我们与追求善的方向是一致的。 2. 君主制是否存在内在矛盾： 在论文提案报告中提出的一个建议是如何协调民主、贵族制和君主制的影响。这一建议主要涵盖了在政治领域发展过程中发生的治理问题，特别是社会工程方面的问题。在现代术语下，民主、贵族制和君主制似乎至少是冲突的政府形式，但如果我们查看它们的词源，就会发现君主制（monarchy 来自-arche）根本是一种政体，而民主制和贵族制（democracy 以及 aristocracy 两者都来自-kratia）根本是一种赋权形式，它们并不是互相排斥的。通过历史主义的视角（直接与统治机构的演变平行），更容易识别根本问题以及解决方案，以协调现在可以被归类为虚假二分法的问题。 3. 小助手还是哲人-王： 自论文提案后，与我的论文导师和同学们的后续讨论的一个主要争议点是人工超智能的角色。毕竟，如果人工超智能旨在简单地推动我们走向善，为什么它必须必然地采取对我们的统治地位？从那时起，我已经巩固了这样一个观点，即这是人类心理中常见的权力动态问题。换句话说，我们感到不舒服将我们的权力让给另一个人是在我们的人类本性之中。然而，通过我的研究，我发现我们的整体控制（或者至少是自我的控制）实际上是非常有限的，我们的自主权的实际范围大多是虚幻的。在维持这种幻想的同时，掌舵的人实际上并不重要。此外，根据早期协调政权类型和赋权形式的建议，我们看到小助手和哲人-王之间实际上没有太大的区别，除了一种人类中心的自豪感，在大局中这种自豪感大多是多余的。		

当然，并非所有在论文提案之前和之后提出的建议都可以完全归类为上述三种类型，但我已努力将它们分解并确定每种建议背后的基本担忧。对于每次与我的论文导师和同学们之间的交流，我还突出了基于收到的反馈所做的具体变化、修订或增补，这些变化、修订或增补随后被应用于我的研究、方法论或论证。

进一步改进的计划

在前进的过程中，我已经记录了进一步反馈的额外步骤（此见附件《中期检查报告》里）。当然，由于时间有限，解决任何剩余问题或整合新的见解将是具有挑战性的，并且可能不会全部包含在最终成果中。然而，我的论文的目的是推动对话向前发展，而不一定持续 100 年。

进度安排：

每个月跟导师进行 1 次进度 check-in 和得到反馈。

3 月初：完善 background chapters; 准备 On Tenability (3.03)

3 月中旬：draft 完 On Tenability/开始 Practicality (3.10)

3 月底：draft 完 On Practicality/开始 Efficacy (3.17-24)

3 月底 - 4 月初：Catch Up 时间

4 月初：Part 2 Discussion 及 conclusion

4 月中旬：开始 Part 3; intro + background (4.10-20)

4 月底 - 5 月中旬：完整 tripartite outcome (21-27/28-4/5-12)

5 月中旬：Part 3 Discussion 及 conclusion (5.13)

5 月中旬 - 底：Catch Up 时间

5 月底：整理 parts 1-3; 融合整个 thesis

中期成绩（百分制）	97

考核组成员签字：[illegible] 日期：2024. 4 11

论文评阅记录表

指导教师评语	**指导教师签字：** **年　　月　　日**
评阅教师评语	**评阅教师签字：** **年　　月　　日**
答辩小组评语	**答辩小组组长签字：** **年　　月　　日**

总成绩：

教学负责人签字：

年　　月　　日