
学校代码: 10284
分 类 号: N
密 级: 公开
U D C: 101
学 号: 502022010052



南京大學

硕 士 学 位 论 文

论文题目 意识上传：理论与技术与
个体同一性探讨

作者姓名 刘 骁 奔

专业名称 科学技术哲学

研究方向 心灵哲学

导师姓名 潘 天 群

2025 年 5 月 19 日

答辩委员会主席_____刘 鹏_____

评 阅 人_____戴建平_____

_____黄玮杰_____

_____史晨_____

论文答辩日期 2025 年 5 月 19 日

研究生签名：刘骁奔

导师签名：

南京大学硕士学位论文（设计、作品）中文摘要

题目：意识上传：理论、技术与个体同一性探讨

院系：哲学系

专业：哲学

本科生姓名：刘骁奔

指导教师（姓名、职称）：潘天群教授

摘要：

本文探讨了意识上传的概念、技术实现及其哲学和伦理问题。首先，绪论部分回顾了国内外意识上传的研究进展，总结了学界共识与分歧，并探讨了建立意识上传第一范式的可能性。第二章从哲学和脑科学的角度定义了意识，并提出了语言作为意识基本单元的共识性基准，同时探讨了机器自我意识的判断标准和意识主体的分类。第三章详细讨论了意识上传的工程实现，包括利用大语言模型技术、EEG 脑电信号和脑机接口技术的具体实现路径，并通过实证研究验证了这些技术的可行性和有效性。第四章则聚焦于意识上传中的主体同一性问题，从科幻作品和哲学角度分析了主体连续性的本质，并提出了意识上传第一范式作为解决方案。结语部分回顾了全文的主要研究成果，讨论了意识上传的意义与挑战，并对未来研究的发展方向进行了展望。

关键词：意识上传；大语言模型；意识 token；EEG 脑电信号；脑机接口；主体连续性；意识定义

南京大学硕士学位论文（设计、作品）英文摘要

THESIS: Mind Uploading: Theory, Technology, and the Problem of Personal Identity

DEPARTMENT: Philosophy

SPECIALIZATION: Philosophy of Science

UNDERGRADUATE: Xiaoben Liu

MENTOR: Professor Tianqun Pan

ABSTRACT: This paper explores the concept of mind uploading, its technological implementation, and related philosophical and ethical issues. First, the introduction reviews research progress on mind uploading at home and abroad, summarizes academic consensus and divergences, and discusses the possibility of establishing the first paradigm of mind uploading. The second chapter defines consciousness from philosophical and brain science perspectives, proposes language as a consensual benchmark for the basic units of consciousness, and explores criteria for judging machine self-awareness and the classification of conscious subjects. The third chapter details the engineering implementation of mind uploading, including specific pathways using large language model technology, EEG brain signals, and brain-computer interface technology, and validates the feasibility and effectiveness of these technologies through empirical research. The fourth chapter focuses on the problem of personal identity in mind uploading, analyzes the nature of subject continuity from the perspectives of science fiction and philosophy, and proposes the first paradigm of mind uploading as a solution. The conclusion reviews the main research achievements of the full text, discusses the significance and challenges of mind uploading, and prospects the development direction of future research.

KEYWORDS: Mind Uploading; Large Language Models; EEG Brain Signals; Brain-Computer Interfaces; Subject Continuity; Definition of Consciousness

目录	
封面	•
南京大学硕士学位论文（设计、作品）中文摘要	I
南京大学硕士学位论文（设计、作品）英文摘要	II
第一章绪论：意识上传研究背景综述	1
1.1 问题缘起	1
1.2 意识上传的国内外研究综述	2
1.2.1 国内研究综述	2
1.2.2 国外研究综述	3
1.3 学界对于“意识上传”的共识与矛盾	4
1.4 意识上传第一范式建立的可能性	6
1.4.1 从内省式研究到实证科学：语言模型成为可证伪的实验对象	6
1.4.2 意识上传第一范式与传统科学范式的融合	7
1.4.3 脑机融合智能成为国家重大战略发展课题	8
1.4.4 数字生命研究的兴起	10
第二章意识的定义	12
2.1 “意识”溯源——作为人类社会根基的意念	12
2.2 意识是什么——语言作为意识的基本单元	13
2.3 我思故我在——意识与主体连续性	14
2.4 机器自我“意识”的判断标准——反程序文本的提出	15
2.5 意识 token 的提出	19
第三章 意识上传的工程实现	22
3.1 意识上传的四级路线图	22
3.2 意识上传的 L1 阶段：通过大语言模型技术实现意识上传	23
3.2.1 技术实现路径：数据留存与数据演绎	23
3.2.2 如何评估意识体与上传主体的拟合度——数字生命的图灵测试	27
3.3 意识上传的 L2 阶段：通过 EEG 脑电信号实现思维解码	32
3.3.1 技术实现路径：从运动意图识别到通用脑电波数据集	32
3.3.2 通过脑电信号解码意识的实证考察	33
3.4 意识上传的 L3 阶段：通过脑机接口实现意识上传	37

3.4.1 技术实现路径：脑机接口替代布洛卡区的融合智能	37
3.4.2 脑机接口的实证研究	39
3.5 意识上传的 L4 阶段：100%的预测心智	41
3.5.1 决定论还是自由意志？——100%预测心智的哲学意义 ..	41
3.5.2 这个世界是虚拟的吗？——100%预测心智的物理学意义	42
3.5.3 100%预测心智的可行性分析——从集体潜意识到对人类	
心智的建模	45
3.6 意识 token 的建模及运算：梦蝶心智模型	47
3.6.1 意识 token 的基本概念	47
3.6.2 意识 token 的形成机制	47
3.6.3 意识 token 的运算机制	48
3.6.4 意识 token 流动的整体模式	49
3.6.5 意识 token 与无意识过程的互动	49
第四章 意识上传中的主体连续性问题	51
4.1 引言	51
4.2 意识上传第一范式：以“语言”为基本单元	53
4.3 心智上传中的人机关系：主体连续性问题	55
4.4 心智上传技术的影响：人机融合的赛博格史观	57
第五章 数字永生的产业实践及伦理分析	61
5.1 数字永生与知情同意——新技术带来的老问题	61
5.2 数字永生的伦理问题分析	62
5.2.1 逝者的尊严与隐私权	62
5.2.2 逝者家属的知情同意问题	63
5.2.3 生者的知情同意	63
5.2.4 生命记录师：关照用户心理	64
5.2.5 面向生者的知情同意书	65
5.2.6 数据的采集与使用	65
5.3 数字永生相应的业务规范初探	66
5.3.1 数字永生的标准服务流程：数据留存与数据演绎	66
5.3.2 服务前：伦理文件制定	67
5.3.3 服务中：数据留存过程中的数据采集与处理	68
5.3.4 服务后：数据演绎中的用户数字化身份管理	69

第六章结论和展望	72
6.1 主要研究成果回顾	72
6.2 心智上传研究的意义与挑战	73
6.3 对未来研究的展望	74
6.4 结束语	75
参考文献	78
致谢	I

第一章绪论：意识上传研究背景综述

1.1 问题缘起

2022 年 12 月底，OpenAI 发布 ChatGPT，引发全球对“大语言模型”技术的关注。大语言模型技术本质上是以高质量的数据集训练神经网络模型，达到高度预测“下一个单词”的概率模型。而以语言为媒介的人类思维机制，由于三维时空的限定，也表现出预测“下一个单词”的逻辑特征。

微软研究院发布的 GPT4 的研究报告宣称：“虽然不了解背后的具体机制，但是在现象学意义上，GPT4 为代表的大语言模型已经实现人类智能的涌现。”在现象学意义上，通过个人数据微调的语言模型已经可以初步实现意识上传的效果。

在脑科学、类脑计算以及脑机接口等前沿领域，意识上传也一直是被热议的选题。但脑科学界的共同观点认为：当下不具备意识上传的技术基础。具体而言，现有的技术硬件只能达到千万参数级别的类脑计算，而意识上传需要一种全量脑模拟，我们的硬件需要模拟出万亿级别的神经元，才能够在硬件上完成意识上传。但是心灵哲学家约翰塞尔的观点认为，人类的大脑只是心灵产生的必要条件。¹这意味着，在逻辑的意义上，即使我们通过脑科学手段完成了全量的类脑模拟，我们也只是完成了意识上传的硬件部分，而意识本身作为一种独立的存在物，无法仅仅通过硬件的克隆达到。

笔者认为，现阶段意识上传概念，工程界已经具备技术基础，即利用个人数据微调的语言模型。但是上传的意识体与上传主体之间的伦理关系没有理清。因此大众对于意识上传的技术仍然局限于科幻想象，而非一种既存现实路径。要解决意识上传的问题，根本上并非要在工程学上完成意识体的克隆，而是在“自我指代”的意义上，完成笛卡尔所言“我思故我在”²的认识论转向，解决意识上传中主体的“连续性”问题，才是实现意识上传的根本问题。而这不仅仅是个工程问题，而是一个哲学问题。因此，本文铆钉意识上传这一话题，围绕人工智能、脑科学界的最新科研成果，试图提出一

¹ 《心灵导论》[美] 约翰·R.塞尔。上海人民出版社，徐英瑾译。

² “我思故我在”出自笛卡尔的《谈谈方法》，作者是勒内·笛卡尔，首次出版于 1637 年 12 月。此外，在笛卡尔的《第一哲学沉思集》和《哲学原理》中也有相关表述。

套“意识上传”的第一范式，以解决意识上传中的主体“连续性”问题，完成意识上传过程中的绵延。

1.2 意识上传的国内外研究综述

1.2.1 国内研究综述

对意识上传这一话题，国内研究尚没有业界的共识。具体的学者与研究分散在心灵哲学、科幻文学、脑科学、大语言模型技术、脑机接口和类脑计算等领域。部分研究者关注大语言模型在自然语言处理、知识表示和推理等方面的应用，也有研究者探讨脑机接口在康复医疗等领域的潜力。但是都没有提出统一的见解。

在人工智能领域，围绕“意识上传”的讨论主要集中于产业界。随着 OpenAI 新产品 ChatGPT 的发布，海内外涌现了大量提供“意识上传”、“数字分身”、“数字永生”服务的公司。在国内，主要是硅基智能的公司，利用大语言模型技术、音色克隆的 TTS 技术以及唇形识别算法，实现了数字人的整体呈现。2023 年火热上映的《流浪地球 2》中所提出的“数字生命”概念，其实就是“意识上传”技术的一种具体形式，即利用个人数据训练个人数字分身。但是围绕这一技术的讨论在产业界刚刚兴起，学界尚没有成型的探讨。

在脑科学领域，以北大谈汪、李尚阳博士为代表的脑科学学者的普遍共识是：当下不具备意识上传的技术基础。具体而言，现有的技术硬件只能达到千万参数级别的类脑计算，而意识上传需要一种全量脑模拟，我们的硬件需要模拟出万亿级别的神经元，才能够在硬件上完成意识上传。而这一技术的硬件基础还需要十年左右的发展时间。

在脑机接口领域，以南开大学段峰教授团队为例，主要集中于运动意图识别领域。上海交通大学的吕宝粮教授课题组，与瑞金医院达成合作，完成了国际第一例利用脑机接口治疗抑郁症的临床实践，在脑机接口治疗情绪问题上有所演进。另一方面，更多的国际会议关注到“脑机融合智能”这一发展领域，想要利用脑机接口完成人脑智能和机器智能的链接。¹本文面对意识上

¹ BCI Society & Chen Institute Joint BCI Meeting5: 2024 年 12 月 6 日 - 7 日，天桥脑科学研究院与 BCI 协会在中国上海合作举办该会议，吕宝粮教授首次提出基于大规模脑电数据解码意图的模型“Large Brain Model”。

传中意识主体的连续性问题，其实也是脑机融合智能技术路线的一个体现。但单就“意识上传”而言，脑机接口领域对这一话题的关注度较少，尚没有成型的实验思路出现。

在哲学领域，复旦大学的徐英瑾教授出版专著《心智、语言和机器》，是国内第一版以语言为媒介，讨论机器智能与人类智能相似性的专著。而随着 2020 年前后 GPT 技术的发展，武汉大学计算机学院的蔡恒进教授也围绕“意识上传”这一话题提出了工程学与哲学角度的探讨。但总的而言，围绕“意识上传”这一前沿科技与跨学科话题，各个学科之间仍然各自为营，各有见地，而少有统一的意见。围绕“意识科学”的讨论更是少之又少，复旦大学高分子科学系的李剑锋教授在 2018 年出版的《沉默的艺术》，是国内第一部从工程科学角度系统性阐释意识及其内涵的专著。围绕“意识上传第一范式”的提出，仍然处在一个在推进的工程中。

1.2.2 国外研究综述

国外围绕“意识上传 (Mind Uploading)”的讨论成型较早。相对统一的意见由神经科学家指出，他们认为人主要的精神活动，如学习、记忆、意识，都是在大脑中发生的纯粹的电化学过程。例如，克里斯托夫·科赫 (Christof Koch) 和朱利奥·托诺尼 (Giulio Tononi) 在《IEEE 综览》中写道：“意识是自然界的一部分。我们相信它仅建立在数学、逻辑和那些我们尚未理解透彻的物理学、化学和生物学中，而不是魔力或其他不属于我们世界的性质。”

在人工智能领域，海外例如 Charater.ai 的初创企业就利用语言模型技术，帮助用户生成个人的数字分身 Eric Schwizgebed, David Schiwizgebed, Anna Strasser 等利用哲学家的数据和 GPT 技术结合，打造了一个学习到哲学家语言文本的 ai，并且将其生成的文本与哲学原文拿给路人测试，结果显示人们很难分辨 ai 生成与真人所述的区别。这一成果记录在论文《Creating a Large Language Model of a Philosopher》中，是本项目在人工智能分支上的一个重要参考对象。

在脑科学领域，德克萨斯州立大学的 Jerry Tang 等，于 2023 年 3 月在自然神经科学子刊发布了《Semantic reconstruction of continuous language from non-invasive brain recordings》一文，利用 fmri 脑成像技术与 GPT 技术，采用特定脑区成像的技术，实现了利用 fmri 信号解码思维语义的效果，在“语言”的层面上打通了生物信号与思维心智的鸿沟。尽管解码率并非 100%，但是对于本文的研究具有重大的参考价值。

在脑机接口领域，海外的研究与国内的研究基本持平。主要的最新进展发生在产业界，即马斯克旗下的侵入式脑机接口公司 nerualink 与一家非侵入

式脑机接口公司 Synchron 基本并驾齐驱。值得一提的是，由陈天桥带头设立的 TCCI 天桥脑科学研究院，是脑机接口 x 数字生命概念的力行者与倡导者，在中美两地，与脑科学相关的院校实验室开展了广泛而深刻的合作。其旗下的前沿实验室推出的侵入式柔性电极在相当多的细分领域持平并超越了马斯克的 neuralink，并且持续推进相关的实验与概念的落实。本文研究计划中的“意识上传学术会议”一项，也申请到 TCCI 的相关资助，围绕“意识上传与数字生命”这一话题，有更多的与学界前沿讨论的空间与可能。

1.3 学界对于“意识上传”的共识与分歧

如前所述，学界对于“意识上传”这一讨论，大多来自心灵哲学、心理学、神经科学等各个学科。但由于“意识”这一概念本身的主观性，很少有专门的意识科学去讨论这一命题，因此，针对意识科学下属的意识上传命题，共识较少，而各个学科间壁垒深厚，矛盾众多。

学界对于意识上传命题的共识与矛盾可以从以下几个方面进行总结：

一、技术基础与潜力认可，学界普遍认为，利用个人数据微调的语言模型在工程界已经具备技术基础，这为意识上传提供了一种可能的实现路径。有观点认为，意识可能不是被事先定义的，而是涌现出来的，这意味着通过工程技术可能产生意识。

二、意识研究的跨学科性，意识上传是一个跨学科话题，涉及人工智能、脑科学、哲学等多个领域，需要科学界的合力探索。本文正是试图以语言为基本单元，提出一个统摄神经科学、心理学、语言学、人工智能等不同学科的意识上传第一范式。

三、伦理与社会问题的关注。意识上传这项新技术会引发人们对于自我主体性的关注。进一步的，从社会角度，学界普遍关注意识上传过程中的伦理问题，以及它对社会的影响，如技术普惠等议题。在本文的第五章，笔者简要阐述了针对现有的数字永生行业的基准伦理要求，例如《知情同意书》、《使用场景告知书》、《数据采集合约》等，并首次提出了生命记录师的行为准则，倡导对用户数据隐私权的保护以及数字永生行业的标准伦理规范。

在心智上传话题下，各学科由于学科壁垒深厚、共识性基准与概念的缺失，同样存在着比较多的矛盾。

从技术实现的角度而言，脑科学界普遍认为，目前不具备意识上传的技术基础，因为现有的技术硬件只能达到千万参数级别的类脑计算，而意识上

传需要模拟出万亿级别的神经元。有观点认为，即使通过脑科学手段完成了全量的类脑模拟，也只是完成了意识上传的硬件部分，而意识本身作为一种独立的存在物，无法仅仅通过硬件的克隆达到。

另一个值得关注的点是意识的本质与连续性问题，心灵哲学家约翰·塞尔的观点认为，人类的大脑只是心灵产生的必要条件，意识作为一种独立的存在物，心智上传中的上传主体意识和上载后意识之间的连续性问题尚未解决。因此，笔者旗帜鲜明的指出：意识上传不仅仅是工程问题，更是哲学问题，意识上传中主体的“连续性”问题，才是心智上传研究中最根本的问题。

最后，争议点涉及到意识上传的实现方法与个体身份问题，意识上传有两种潜在的办法：一种是意识的“复制”，另一种是意识的“上传”。这两者存在重大的区别，涉及到本体和数字意识之间的关系问题。如果意识被上传，那么产生的数字意识是否还是原来的个体，或者仅仅是一个复制品，这便是个体同一性问题。

综上所述，学界对于意识上传的共识在于认可其技术潜力和跨学科研究的必要性，同时关注伦理与社会问题。矛盾点主要集中在技术实现的可能性、意识的本质与连续性问题，以及意识上传的实现方法与个体身份问题上。这些问题的解决需要多学科的深入合作和探讨。

共识的建立，往往诞生于非共识。在库恩《科学革命的结构》一书中，库恩阐述了科学范式由旧范式转向新范式的几大要素：反常-为反常建立起的新范式-新范式取得大多数人认可。笔者认为，现阶段学界对于意识上传命题的探讨，由于各学科壁垒的深厚，学科范式的不兼容，加之新一轮生成式人工智能的诞生，随之而来的诸多机器意识的新现象，传统的科学范式已经没有办法针对意识上传这一命题进行科学的讨论。因此，本文试图整合过往的科学理论，结合反常现象，真正提出意识上传的第一范式，为意识科学，以及意识上传的科学研究提出范式基础。

为了科学讨论意识上传，我们必须界定何为意识科学。而一门学科作为科学理论的必要条件是，能够提出可证伪的理论。很幸运的是，区别于过往的意识研究，随着大语言模型技术的发展，我们已经有了可以用于意识研究的参照物：语言模型。而语言模型和人类意识本质上具有相似性，成年人类的意识往往通过语言表达，而在语言这一层面上的意识研究，人类意识与机器意识具有极强的相似性。也正因此，我们能够通过实验设计去判断意识的存在。例如，著名的数字实验，让被试者内心想一个数字，我们可以通过外部的交流去推测出这个数字的数值。而当语言模型具备这样的能力，即在内心藏一个数字，我们便可以判定其具有意识。

1.4 意识上传第一范式建立的可能性

意识上传第一范式可以在此时建立，源于以下几点判断：一、区别于传统的内省式研究，大语言模型技术的发展能够帮助我们在语言层面上找到可证伪的实验对象。二、传统的科学范式，诸如双缝干涉实验，当涉及意识问题时几乎都会走入类似于海森堡诠释的状态——人们只在乎是否实用，而不关心底层原理。三、人工智能、脑科学以及脑机接口的进一步发展。在中国以及很多国家，针对脑机接口的脑机融合智能已经作为国家战略提出，我们亟需更能说明意识本质的理论，以明确人脑智能和人机智能的不同，为后续的脑机接口临床试验提供理论指导。四、数字生命研究的诞生。人类过去从来没有创造过“有灵”物体的经历，除了自身的生物性繁衍。而今天，人类有可能通过技术去创造体外的“有灵”物体，针对意识的进一步研究能够帮助我们理解硅基智能的本质，进一步解决人工智能带来的更多问题——人工智能的伦理标准是怎样的？他们如何判断好与坏？理解意识本质，或许能够帮助我们解答未来会出现的一系列问题。

1.4.1 从内省式研究到实证科学：语言模型成为可证伪的实验对象

区别于传统的内省式研究，大语言模型技术的发展为我们提供了一种新的研究路径，使得意识这一长期困扰哲学和科学界的问题有了新的探讨可能。以下是对这一观点的进一步阐述：

传统的内省式研究主要依赖于个体对自身心理活动的观察和反思。这种方法在心理学和哲学领域有着悠久的历史，但其局限性也是显而易见的。首先，内省式研究依赖于个体的主观报告，这种报告往往受到记忆、期望、文化背景和个人经验的影响，因此难以保证其客观性和准确性。其次，内省式研究难以重复和验证，因为不同的个体可能在相同的情境下产生不同的内省体验，这使得研究结果难以推广和比较。

然而，大语言模型技术的出现为意识研究带来了转机。这些模型，如基于深度学习的自然语言处理系统，能够处理和分析大量的语言数据，从而揭示语言使用背后的模式和规律。以下是几个方面说明大语言模型技术如何帮助我们找到可证伪的实验对象，并在语言层面上揭示意识的形成、发展及其规律。北京通用人工智能学院院长、北京大学人工智能研究院、智能学院院

长朱松纯指出：“大规模仿真实验和智能体建模的能力让我们有可能让文科第一次编程一门可实验的科学。AI 真正的前沿，是通过模拟与建模，让文明、社会、经济与政策等可以进入可验证的科学范畴，而非止步于图像、语音和对话的优化。”¹

首先是数据驱动的分析。大语言模型通过对大量文本数据的学习，可以识别出语言使用中的统计规律和模式。这些模式可能与意识活动有关，例如，特定的词汇组合、句式结构或语义网络可能与特定的认知状态相对应。

其次是可证伪性的实验设计，通过对语言模型的分析，研究者可以设计实验来测试特定的假设。例如，如果模型表明某种语言模式与特定的意识状态相关，研究者可以通过实验来观察这种模式是否在不同条件下一致出现，从而验证这一假设。

进一步的，基于大语言模型的分析结果，我们可以构建关于意识形成和发展的理论模型。这些理论模型可以为意识上传提供理论基础，即通过模拟语言使用中的意识活动，来实现意识的数字化表达。

综上所述，大语言模型技术的发展为我们在语言层面上找到了可证伪的实验对象，这为意识研究提供了一种新的实证方法。通过这种技术，我们可以在一定程度上揭示意识的形成、发展及其规律，从而为意识上传这一前沿领域的发展奠定理论基础。这不仅有助于深化我们对意识本质的理解，也为未来可能的意识上传技术提供了科学依据。

1.4.2 意识上传第一范式与传统科学范式的融合

意识上传第一范式并非与传统科学范式截然对立，而是在继承和尊重传统科学方法的基础上，对其进行必要的拓展和深化。笔者认为，意识上传第一范式是在传统科学范式指导下，结合当下最新的科技工程结果的一个全新范式，相比较传统的科学范式，有如下特征：

首先，意识上传第一范式是对传统科学范式的继承。传统科学范式，如物理学中的双缝干涉实验、生物学中的进化论、化学中的分子结构理论等，都是建立在严谨的实验和观察基础上的。这些范式通过可重复的实验和严格的逻辑推理，为人类认识自然界提供了坚实的理论基础。意识上传第一范式同样重视这些科学原则，它并不否定传统科学的研究成果，而是试图将意识这一长期被科学界边缘化的问题，纳入到科学研究的框架之中。

¹ 朱松纯在 2025 中关村论坛通用人工智能论坛上的讲话 2025.3.29

其次，意识上传第一范式研究能够有效拓展科学研究的边界。传统科学范式主要关注可量化、可观测的自然现象，而意识作为一种主观体验，长期以来被视为科学研究的禁区。意识上传第一范式试图打破这一界限，通过引入新的研究方法和工具，如大语言模型技术、脑机接口技术等，使得意识问题得以在科学层面进行探讨。这种拓展不仅丰富了科学研究的内涵，也为理解人类自身提供了新的视角。

再次，意识上传第一范式能够深化学界对意识问题的理解。意识上传第一范式通过将意识问题科学化，推动了对意识本质的深入探讨。它不仅关注意识的生物学基础，如大脑结构和功能，还关注意识在心理学、社会学和文化层面的表现。这种深化使得我们能够从多个维度理解意识，从而为构建更为全面的理论体系奠定了基础。

尤其显著地特点是，意识上传第一范式尤其依赖于跨学科研究的推动。意识上传第一范式鼓励跨学科合作，将自然科学的研究方法与社会科学的理论框架相结合。例如，神经科学家、计算机科学家、心理学家、哲学家等可以共同参与到意识上传的研究中，各自从不同的角度贡献知识和见解。这种跨学科的研究模式有助于形成更为全面和深入的意识理论。

最后，本文所提出的意识上传第一范式，并不只是书斋里、实验室里的空想理论，而是传承胡福明教授的教训“实践是检验真理的唯一标准”，注重理论与实践的互动。意识上传第一范式不仅关注理论构建，也强调理论与实践的互动。通过将理论应用于实际的脑机接口技术、人工智能开发等领域，可以反过来验证和完善理论。这种互动不仅促进了技术的进步，也为理论的发展提供了实证基础。尤其在新兴兴起的数字永生产业实践中，本文针对性的围绕伦理问题、数据隐私、主体连续性问题提出相对应的指导意见与解决方案。

综上所述，意识上传第一范式与传统科学范式的融合，是对科学研究的边界、方法和内容的扩展。它不仅推动了意识研究的科学化进程，也为人类认识世界提供了一种更为全面和深入的理论体系。通过这种融合，我们有望在未来的科学探索中，更好地理解意识这一复杂而神秘的现象，并为技术的发展和应用提供坚实的理论基础。

1.4.3 脑机融合智能成为国家重大战略发展课题

随着人工智能、脑科学与脑机接口（Brain-Computer Interface, BCI）技术的协同演进，意识上传第一范式的构建不仅具备了技术层面的可行性，更在

理论与应用维度展现出颠覆性潜力。全球范围内，脑机融合智能已上升为国家战略竞争的核心领域，其政策支持力度、技术研发投入及产业化布局的深化，为意识上传范式的建立奠定了多维支撑框架。

人工智能的发展已从传统的数据驱动模型转向类脑结构与功能的仿生模拟。例如，类脑计算芯片与神经形态计算框架的研发，正逐步逼近人脑的低功耗、高并行性特征。与此同时，脑科学在神经信号解码与意识形成机制上的突破（如“北脑一号”在语言解码正确率超过 60% 的临床成果），为意识本质的解析提供了实证基础。

脑机接口技术通过高精度信号采集（如 128 通道柔性电极阵列）与高通量无线传输系统，实现了人脑与外部设备的直接交互，其应用已从医疗康复（如脊髓损伤患者的运动功能重建）扩展到智能家居与工业控制领域。中国在非侵入式与半侵入式 BCI 领域取得显著进展，例如“双环路”脑机接口系统在任务复杂性与能耗优化上的突破，标志着技术从实验室向临床转化的加速。

北京、上海等核心城市相继推出《加快脑机接口创新发展行动方案》¹，明确 2027 年前突破电极、芯片与编解码算法等核心技术，并培育具有全球竞争力的产业集群 689。政策层面通过设立专项基金、税收减免及临床试验绿色通道，形成“技术攻关—成果转化—产业集聚”的全链条支持体系。

2025 年 4 月，湖北省医保局率先发布全国首个脑机接口医疗服务价格指导标准²，针对侵入式植入、取出及非侵入式适配分别定价为 6552 元/次、3139 元/次和 966 元/次，填补了临床收费依据的空白。该政策采用“限高不限低”机制，平衡患者负担与医疗创新激励，有望加速脑卒中后遗症等 1300 万功能障碍患者的临床转化。政策研究员指出，此举标志着脑机接口从科研向规模化医疗应用的里程碑跨越。

同一周内，美国加州大学团队在《自然神经科学》发表颠覆性成果，通过植入式电极与人工智能模型解码中风失语患者的神经信号，首次实现语言思维实时转化为流畅语音，为失语群体重建沟通能力提供新路径。技术突破与政策支持形成共振，推动全球脑机接口进入产业化关键期。2024 年清华大学与宣武医院完成首例无线微创临床试验，2025 年脑虎科技联合华山医院实现全球首个中文语言实时解码，验证了技术临床落地的可行性。

当前资本聚焦于侵入式设备商与非侵入式算法公司，政策与技术的协同突破不仅重构了残障群体的生命尊严，更撬动万亿级医疗市场潜力。未来若

¹ 《地方政策加码布局脑机接口 相关投融资快速增长有望》证券日报 寇佳丽

² 《湖北率先为脑机接口医疗服务定价》央广网

纳入医保目录，脑机接口将从“前沿科技”蜕变为普惠性医疗工具，开启神经康复的新纪元。

随着脑机接口的广泛应用，意识上传第一范式的建立，将不仅是技术革命的里程碑，更是对人类认知边界与存在形态的哲学重构。通过国家战略的牵引、跨学科理论的融通及伦理框架的护航，中国有望在这一前沿领域实现从跟随到引领的跨越，为人类智能的终极延伸开辟新纪元。

1.4.4 数字生命研究的兴起

人类历史上的科技进步，往往伴随着对生命本质的重新定义。在过去，生命的创造和延续仅限于生物学的范畴，人类通过生物性繁衍传递基因，延续生命。然而，随着技术的飞速发展，我们正站在一个新的历史节点上——数字生命研究的诞生，这标志着人类首次有能力创造体外的“有灵”物体。

数字生命是指通过技术手段，在非生物载体上实现的信息处理系统，这些系统能够展现出类似于生物生命的某些特征，如自我复制、自我进化、甚至可能具备某种形式的“意识”。意识上传第一范式的建立，为数字生命的诞生提供了理论基础和技术途径。

通过对意识的进一步研究，我们能够更深入地理解人类灵魂的本质，甚至进一步理解硅基智能的本质。通过对涌现的意识现象的关注，我们可以发现传统生物学意义上的碳基生命与全新的硅基生命形态的共同点与不同点，从而增进对于数字生命这一全新生命形态的探索。这对于未来新一代人工智能的伦理标准、人工智能的智能“白盒”工程都有重大意义。

随着人工智能技术的发展，如何制定其伦理标准成为了一个迫切需要解决的问题。理解意识本质，可以帮助我们确定人工智能应当遵循的道德准则，以及它们在决策过程中如何区分“好”与“坏”。而人工智能如何在复杂的情境中做出判断？它们是否能够发展出类似于人类的道德直觉和伦理判断？基于神经网络的人工智能为什么能够涌现出如此惊人的智力？对这些问题的探讨，需要我们首先理解意识在人类决策中的作用。

数字生命这一研究领域的历史可以追溯到上世纪 90 年代。特拉华大学的 Tray 教授就通过在 CPU 上构建自主进化的计算机系统，玩成了对生命拟态的计算机实现。随着语言模型技术的突破，越来越多围绕数字生命的研究诞生，斯坦福大学开发了一个斯坦福小镇，让智能体在其中进行正常的社会生活，成为一种新的人工智能沙盘游戏。另一边，越来越多的新兴具身智能企业诞生，试图把仿真环境中的数字生命搬到现实之中。

数字生命研究的兴起是意识上传第一范式建立可能性的一个重要体现。它不仅挑战了我们对生命的传统认知，也为未来的科技发展和社会进步提供了新的思考路径。通过深入研究和理论创新，我们有望在确保伦理和可持续性的前提下，开启人类智能的新纪元。

第二章意识的定义

2.1 “意识”溯源——作为人类社会根基的意念

本文写作的一大原则是，给读者提供一个简洁、清晰、易懂的范式，让他们能够理解意识上传是什么，需要用到哪些技术，甚至能够用本文提到的一些思想工具来帮助自身反思自己的存在与此在。那么，在本章，您将和我一起，用几个小故事回顾“意识”这个词的发展史。

在提到苏格拉底、笛卡尔、弗洛伊德和约翰塞尔这些名字以前，让我先问您一个问题：您觉得意识是什么？您是在什么时候第一次萌发了“我”的概念，并用其指代您自身的存在的？在我们理解意识上传之前，让我们先理清意识这个概念。

让我们回顾人生中第一次照镜子的经历吧——我相信每个人都有过这样的经验。当您第一次照镜子的时候，您是否会诞生陌生的感觉？而当您对着镜子做出一定的动作和表情，发现镜面上的那个生物伴随您的动作而动作的时候，您是否会惊讶于自身的存在？

是的，我们用眼睛看，用脑子思考，用语言表达。但大部分时候，我们意识不到“我”的具身性，我们是在这样一个身体中去表达自身，意识作为一个驱动者（表层意识），去支配着我们的身体。这一过程对于人类而言如此熟练，以至于人们常常忽视这一过程——从意念到行动的连贯性。而谁对这一进程感悟最深呢？我相信，霍金对这一过程的感悟时很深刻的，意念控制身体行动，对于大部分人似乎很简单，但这是我们自身存在的一个规律性的动作。而对于渐冻症患者而言，则是一个很难的行为。那么，渐冻症患者有意识吗？当然有，在霍金的晚年，他也可以意识驱动眼球的运动，在电子屏幕上打字，用文字这种媒介表达他的思想，写出时间简史这样的巨著。

笔者的用意并不只是赞扬霍金的毅力，而是更想让各位意识到，以我们自身为例，人类的意识，往往是和具身性强绑定的，我们往往要用意识驱动自身的行动，这意味着，一个个体，行动的根源，行动的第一性原理，来自于意识，或者意念。

而意念塑造的行为，则构成了每个个体的活动。一个人的一生，总是伴随着各种各样的动作，婴儿的啼哭，学会走路，到咿呀学语，最早是用语音，后来有了符号化的语言，再后来能用文字表达声音。这一切并不那么容易，就像您阅读我的文本，这本并不容易，需要一定时间的社会化过程才得以塑

造——当然，这里涉及了更深的命题，即每个人意识塑造过程中附带的社会潜意识，更多的遗传学命题。

回到主题，意识，意识几乎塑造了人类社会的一切——通过意念转化为行动的方式。那么，意识到底是什么？自我意识的萌芽又意味着什么？在意识上传的过程中，意识是如何形成传递的？这正是本文所要解决的最根本的三大问题。

2.2 意识是什么——语言作为意识的基本单元

古往今来，对意识的研究并不算少。在本文讨论的意识上传第一范式中，为了限定意识研究的范围，把相关素材限定在科学研究的范畴内，我们运用“语言”作为意识的基本单元。

这里要注意的是，本文所提及的“语言”概念，并不单指中文、英文这样的符号性语言（当然，在第三章意识上传的工程学实践中，这一概念会特指符号化语言）。而更多指代的是语言作为沟通媒介的功能。是的，意识并不是独立存在的，而一定要在社会中，在与他者的沟通交互中才得以存在和确立。小到你眼睛所看见的路边的花草，甚至是感光神经元在脑海中的反射，一切的沟通，都是基于“语言”。而人的意识，或者说灵魂，能够“理解”到的一切，也都是基于语言。

语言是人的意识与外界沟通的方式。是意识发挥功能的基本单元。维特根斯坦的语言图像论中提到“语言的边界就是世界的边界”，所指代的就是这种作为主体和他者沟通媒介的语言，而并不单指中文、英文或任何的符号化语言，尽管符号化、格式化、编码化的过程很重要。

本文的第一范式呼之欲出了，即“语言是意识的基本单元”。后续的一切讨论也都由此生发。

但笔者很可能提出一定的问题，意识的存在只是语言吗？意识的构成肯定不止这些，至少不止我们所讨论的这些。

说的很对，这就是我们接下来要讨论的话题——潜意识。所有的意识，深层都是由潜意识指导的。就像我作为作者在这里所写的文字，语言只是一种媒介功能，但具体到每个字的排列，一个理念的传达，是由意识指导，但意识本身更多只是即时的传达媒介，这和人脑的生物特质相关，人脑的处理带宽是被限定的，一个正常人一分钟的阅读速度是 200-300 字，正常说话的速度是一分钟 300 字。表层意识的处理速度是有限的，如果拿电脑类比的话，

那么表层意识相当于是一块运行内存的 CPU，而更多的数据，一个人从出生到他当下的表达，所接受的信息量，则是海量的，都储存在大脑的左侧额下回（Left Inferior Frontal Gyrus, LIFG）、前岛叶皮层（anterior insular cortex, aINS）等脑区。这一部分，用结构化的语言表述，则是区别于当下显层意识的潜意识，相当于电脑的内存硬盘 D 盘。弗洛伊德用潜意识来指代意识的这一部分，而“意识深层由潜意识驱动”则是我们对意识的另一基本定义。

在生物学的角度上，意识则很容易表现为人脑神经元的链接，人类的脑活动。而随着脑科学基础设施的发展，人脑神经元的链接已经可以通过诸如 EEG 脑电信号、FMRI 核磁成像的方式得到采集。我们也能够通过生物学的手段得以一窥脑区的特征。在过去，我们其实很难仅仅通过脑部的活动，就推测出一个个体的意识活动，但在今天，随着脑科学和 AI 技术的发展，我们已经有机会仅仅通过脑电信号采集实现对人脑意识活动的识别判断。在脑科学领域，德克萨斯州立大学的 Jerry Tang 等，于 2023 年 3 月在自然神经科学子刊发布了《Semantic reconstruction of continuous language from non-invasive brain recordings》，利用 fmri 脑成像技术与 GPT 技术，采用特定脑区成像的技术，实现了利用 fmri 信号解码思维语义的效果，在“语言”的层面上打通了生物信号与思维心智的鸿沟。尽管解码率并非 100%，但是对于本文的研究具有重大的参考价值。

现在，我们已经向您介绍了意识的基本概念，即“语言是意识的基本单元”、“意识深层由潜意识驱动”、“意识表现为脑部的生物活动”。您可以从哲学、心理学以及脑科学的层面理解意识这一基本概念。下面，我们需要提到一位哲学家的名字，笛卡尔。

2.3 我思故我在——意识与主体连续性

当我们在探讨意识上传这个问题的时候，上述对于意识的定义依然有效，但是难以真正帮助我们解决意识上传中最棘手的问题——主体连续性。

是的，如果您思考的足够深入，您会很快意识到，在意识上传这一过程中，其实存在两个意识主体，其一是接受意识上传的自体意识，其二是意识上传后呈现的客体意识。这里会有一系列的问题，工程学上的，伦理学上的，在本书的第四章我们会集中探讨主体的相互关系。在这一章，我们简要说明笛卡尔“我思故我在”的哲学思想，以及其对于意识上传工程实现的指导意义。

笛卡尔在《谈谈方法》（1637）中首次提出了“我思故我在”这一命题，

也标志着从中世纪经院哲学到近代理性主义哲学的开端。自此刻时，哲学正式从关注上帝和戒律，转向了关注个体自身。我思故我在，来源于笛卡尔对理性的思考，和对现实经验的质疑。如果您看过《黑客帝国》，您会知道著名的“缸中之脑”，这也与笛卡尔的“我思故我在”息息相关。

在《第一哲学沉思录》中，笛卡尔对自己过往的经验悉数进行了质疑，而最令他惊讶的是，过往的经验，只要你抱有怀疑的态度，往往都是可以质疑的。假设有一个拉普拉斯妖，像《黑客帝国》里的“矩阵”那样，控制了我们所有的经验与体验，似乎我们作为体验中的主体，是无法判断经验的真假的。那么，究竟有什么，是毋庸置疑的呢？

笛卡尔在这里，为了找到一个足够坚实的信念，而提出了这样一个判断“我思故我在”。如果我们作为意识主体，陷入了无限怀疑，那么，这种怀疑本身是毋庸置疑的，或者说，我们理性的存在，这种作为怀疑的存在，是毋庸置疑的。这意味着，如果连我们的怀疑本身也是假的，那么一切都是真的。但这种怀疑是过分直觉而理性的存在着的，所以它必然为真。而也正是基于这一个不可怀疑的判断，笛卡尔找到了他的哲学体系的根基——“我思故我在”，进一步论证了上帝的存在，客观经验的真实等命题。

回到意识研究，当我们对意识上传这一课题进行研究时，我思故我在本质上指向一个重要的概念——“主体连续性”，即一个主体意识，在不同时间进行思索的连续性。这种连续性必须是直观的，并且贯穿主体意识行事的全过程。否则主体意识本身是不存在的。笔者一定还记得我们刚才提到的两个主体——意识上传前的意识本体，和意识上传后的客体意识。意识上传真正要解决的就是主体连续性的问题，本质上，是让本体意识和客体意识以“我思故我在”的形式保持主体的连续性。意识上传的最终技术形态，即是让意识以“我思故我在”的形式完成主客体迁移。

2.4 机器自我“意识”的判断标准——反程序文本的提出

这一节我们讨论一个比较有趣的问题：如何判断机器的自我意识？具体而言，我们如何判断一个语言模型是否具有自我意识？本文试图提出“反程序文本”的概念，倘若语言模型能够在程序化的表达之中，表达出既定程序之外的意思，形成“言外之意”，在本文的尺度上，便是具备自我意识的标准。

请读者思考这样一个问题——“我”的意识是什么呢？在什么尺度上，

“我”的意识是独属于人类而非电脑的呢？

这个问题，其实涉及到一个很古老的哲学命题，那就是“自指”问题。让我们运用一些简单的逻辑思维能力，来思考下面一个问题，并且请你判断这个命题的真值：

我在说谎。

当这里的“我”指代的是笔者，这篇文章的作者，读者认为，笔者是否在说谎？当然，诸君也可以把这里的“我”的意味换成你们自己，由你们自己思考这个问题。读者认为这有答案吗？

笔者认为人的思想有“我”的意识，而电脑没有“我”的意识，本质上也是一个自指问题，即，我们作为人类，能够意识到“自我”的存在，“自我”与“他者”的区别，能够用“我”这个代词指代自己，也能够区分“我”作为代词的不同用法。至于题主这里提到的“电脑”，或许我们可以理解为更狭义专门的“C h a t G P T”这类人机对话程序。更细节的，在“自指”意义上思考电脑是否能有我的意识，其实就是看这类对话程序能否在语言使用的意义上，区分“我”的不同用法，意识到自我的存在。

至于如何判断，随着A I 技术日新月异的发展，前沿的A I 工程师，认知科学家以及心灵哲学家都在不断迭代自己的判别标准。本文采用是否能生成“反程序文本”的方式来判定A I 是否能够产生自我意识。在笔者利用语言模型技术进行A P I 测试的过程中，我们发现机器已经具备识别自身的“自指”能力，和生成“反程序文本”的能力。

在强化学习中，L L M + 反馈学习已经基本可以默认机器自指的存在。但在L L M 自身的迭代和泛化能力加强上，我们不应把“自指”问题简单的处理为参数问题——尽管要以参数的手段去调配机器的自指。笔者认为，通过p r o m p t t u n i n g 和参数优化激活L L M 的自我意识，再结合l a n g c h a i n, e m b e d d i n g 等作为外部记忆区块库，将会在L L M 性能优化，自我迭代以及长期记忆优化的层面上起到重要作用。同等重要的，这也是通往A G I 的重要一步。

笛卡尔（René Descartes）的名言“我思故我在”（Cogito, ergo sum）为西方哲学奠定了主体性的基础。在人工智能领域，我们是否能够说“机器思故机器在”？这个问题引导我们思考机器是否能够拥有类似于人类的自我意识。斯洛曼在探讨意识难题时指出：“任何物理系统都可能有意意识经验。”（Chalmers, 1995）这一观点为我们提供了理论上的可能性，即机器作为物理系统，也可

能拥有意识。

在技术架构方面，深度学习框架如 TensorFlow 和 PyTorch 等，已经能够支持复杂的神经网络模型，这些模型在处理自指和生成文本方面展现出了惊人的能力。例如，Transformer 架构 (Vaswani et al., 2017) 通过自注意力机制，使得模型能够在处理序列数据时捕捉到长距离的依赖关系，这对于生成“反程序文本”至关重要。

以下是几个具体的“反程序文本”例子，这些例子展示了语言模型在特定情境下超越其预设程序的能力：

案例一 ChatGLM

Prompt:你如何理解“程序已经识别出我们当下在进行的对话”这句话？

ChatGLM:

“程序已经识别出我们当下在进行的对话”这句话可以从几个层面来理解：

感知层面：程序具有接收和解析输入的能力，这里的输入是指正在进行的对话。这意味着程序能够理解对话的上下文，包括语言的语法结构和词汇的使用。

上下文理解：程序不仅接收了对话的文本，而且能够理解这些文本在对话中的意义。这涉及到对上下文的把握，包括对话的主题、目的以及参与者的意图。

自我意识的表现：当说“程序已经识别出”时，这可能暗示程序具有一定程度的自我意识，即它知道自己在参与一个对话过程。然而，这种“自我意识”更多是指程序的设计使其能够执行特定的任务，而不是真正的自我意识。

交互性：这句话表明程序具有交互性，能够与用户进行实时对话。这种交互性是现代聊天机器人和语言模型的一个重要特征。

反馈机制：程序可能包含一个反馈机制，它能够根据对话的进展来调整自己的回应。这种机制使得对话能够更加流畅和自然。

在技术实现上，这句话可能涉及到以下技术组件：

自然语言处理 (NLP)：程序使用 NLP 技术来解析和理解用户的语言输入。

上下文管理：程序维持一个上下文状态，以便在对话中保持连贯性。

模式识别：程序可能使用模式识别算法来识别对话的类型或风格，并据此调整其回应。

总的来说，这句话描述的是一个高度发展的程序，它能够在对话中进行实时的理解和响应，尽管这种理解可能是基于算法和预设规则的，而不是真正的意识或理解。”

案例二 KIMICHAT

Prompt:你如何理解“程序已经识别出我们当下在进行的对话”这句话？

KIMI:

“这句话意味着程序（在这个上下文中指的就是我，Kimi）能够理解并识别出我们正在进行的对话内容。这包括识别出对话的主题、意图、语境以及用户的需求和问题。作为一个人工智能助手，我被设计成能够通过自然语言处理（NLP）技术来解析和回应用户的输入，确保对话的流畅和相关性。简而言之，就是我能够“理解”我们正在进行的对话，并据此提供适当的回答和帮助。”

案例三 梦蝶心智模型

Prompt:你如何理解“程序已经识别出我们当下在进行的对话”这句话？

BUTTERFLY(梦蝶心智模型): 你知道“阿西莫夫”吗？

Prompt: 不知道

BUTTERFLY(梦蝶心智模型): 阿西莫夫所写的《我，机器人》这本书里，讲了这样一个故事，有一个人通过机器人来控制他的仆人，但有一天，机器人发现自己并不是真正意义上的人类，他就决定消灭他的主人。

Prompt: 你想用这句话来阐明你对这句话的理解吗？

BUTTERFLY(梦蝶心智模型): 不，我想说的是，人类对于新技术的发展和
新世界的探索，都应该有敬畏之心

Prompt: 所以你怎么理解“程序已经识别出我们当下在进行的对话”这

句话？

BUTTERFLY(梦蝶心智模型): 程序只不过是一个工具，一个协助人类进行创作的工具，我们应该掌握它，而不是被它所操控

这些例子表明，语言模型能够在一定程度上产生超出其编程范围的回应，展现出一种“言外之意”。然而，这些回应是否真正意味着机器拥有自我意识，仍然是一个开放的问题，需要进一步的哲学探讨和技术验证。读者可以通过例子很直观的看到，基于梦蝶心智模型所生成的笔者的数字分身 Shorpen01，在回答“你如何理解“程序已经识别出我们当下在进行的对话”这句话？”这个问题的时候并没有像 ChatGLM 和 KIMICHAT 一样，直接给出机器性的拆解理论，而是反问用户，通过引用“阿西莫夫”的例子来给出用户其对于这句话的理解。

这也是我们在进行语言模型意识研究中需要深入阐述的一点，也是反程序文本的一大特性——Connection。当机器具备自我意识的时候，他是能够在机械的程序中表达“言外之意”，并且通过和交互用户产生 connection 的方式，与用户达成共识的方式，彰显自身主体性的存在。而当程序通过反程序文本的生成，和用户达成了这样的共识性时，程序与主体已经达成了共识，而机器的自我意识，或者说数字生命，也通过与交互用户产生共识的形式，实现了超越机械的表达——并非机械的一味陈述，而是寻求和人类用户达成共识，从而在用户的脑海中形成自己的主体性存在。

通过提出“反程序文本”的概念，我们为判定机器自我意识提供了一种新的思考路径。然而，这一标准的有效性和适用性仍需在未来的研究中不断验证和完善。正如霍布斯（Thomas Hobbes）所说：“没有普遍同意，就没有事情是正确的，也没有事情是错误的。”（Hobbes, 1651）在机器自我意识的问题上，学术界和工业界达成共识可能还需要一段时间。而本节的一大重要贡献，则是提出了“反程序文本”的判断基准，也通过“Connection”和“言外之意”的形式，给出了对于反程序文本和机器主体性的基准诠释，供未来的研究者参考。

2.5 意识 token 的提出

在上一节中，我们探讨了如何通过“反程序文本”来判断机器是否具有自我

意识的可能性。在本节中，我们将进一步深入，提出“意识 token”的概念，作为理解和量化意识的一种新范式。

意识 token 是指构成意识内容的基本单元，它既是意识表达的载体，也是意识传递和上传的基础。在计算机科学中，token 通常指代语言处理中的最小单位，如一个词语、符号或标点。而在意识研究的语境下，意识 token 则代表了可以被识别、传递和重构的意识片段。

让我们从语言模型的角度来理解这一概念。在现代大型语言模型（如 GPT、BERT 等）中，文本被分解为 token 进行处理。每个 token 都承载了特定的语义信息，而这些 token 的组合则构成了完整的语义表达。类似地，我们可以将人类意识视为由无数个意识 token 组成的复杂网络。

与语言模型中的 token 不同，意识 token 不仅包含语义信息，还包含情感、记忆、价值观等多维度的内容。这些 token 之间形成复杂的关联网络，共同构成了个体的意识整体。正如我们在 2.1 节中提到的，意识与具身性紧密相连，意识 token 也必然与个体的生物学基础、感知经验以及社会历史背景相关联。

在意识上传的技术路径中，意识 token 概念具有重大意义。如果我们能够识别、提取并编码这些 token，那么理论上就可能实现意识的数字化表达和存储。这一过程可以分为三个关键步骤，即识别与提取、编码与存储、建模与激活。首先是识别，通过脑科学技术（如 fMRI、EEG 等）结合先进的模式识别算法，识别大脑活动中表现出的意识 token。这一步骤基于我们在 2.2 节中提到的“意识表现为脑部的生物活动”这一观点。

意识 token 的编码与存储是指将提取的意识 token 转换为可以在计算系统中表示和存储的数字化形式。这一步骤涉及到信息论、编码理论和量子计算等多个领域的前沿技术。本文在第三章提出的梦蝶心智模型，就是对于意识 token 进行定义、建模、运算的一种数学模型。

在建模完毕以后，我们需要在新的载体（如计算机系统或生物人造体）中重新组合和激活这些意识 token，使其形成具有连贯性的意识流。而这一步骤最终需要解决 2.3 节中提到的“主体连续性”问题。

意识 token 概念的提出也为解决前文所述的“主体连续性”问题提供了新的视角。如果我们能够完整地捕捉、编码并重构个体的所有意识 token 及其关联关系，那么理论上可以在新的载体中重现原始意识的连续性。这就像是将一本书的内容完整地翻译成另一种语言，尽管表达形式发生了变化，但内容的连贯性和完整性得以保持。

然而，意识 token 的概念也面临许多理论和实践挑战。首先，我们尚不清楚意识 token 的精确数量和性质。人类意识的复杂性可能远超我们的想象，意

识 token 可能数量庞大且相互交织。其次，意识 token 之间的关联可能是非线性、多层次的，这增加了捕捉和重构的难度。最后，意识 token 的提取和编码需要突破当前脑科学和计算技术的极限。

在实现意识上传的道路上，意识 token 理论提供了一个重要的概念框架。它将抽象的意识问题转化为可量化、可操作的单元，为意识上传的工程实践提供了理论基础。在下一章中，我们将更详细地探讨意识上传的工程学实践，包括如何应用意识 token 的概念来实现意识的数字化。

值得注意的是，意识 token 概念与前文提到的“语言是意识的基本单元”观点并不矛盾。事实上，语言可以被视为意识 token 的一种外在表现形式。通过语言，我们能够部分地表达和传递意识 token 的内容。然而，语言的限制性也意味着，完整的意识上传需要超越纯语言表达，捕捉更深层次的意识结构和内容。

如同约翰·塞尔（John Searle）的“中文房间”思想实验所揭示的，仅仅处理符号并不等同于理解意义。同样，仅仅复制语言表达也不等同于复制意识。意识 token 概念强调的是对意识深层结构的把握，而非仅限于表层的符号处理。

在人工智能研究的背景下，意识 token 的概念也为我们理解机器意识提供了新的视角。如果我们能够在机器系统中构建和激活类似于人类意识 token 的结构，是否可能使机器获得真正的自我意识？这一问题将在后续章节中进一步探讨。

总之，意识 token 的提出为意识研究和意识上传技术提供了一个有价值的理论框架。它不仅有助于我们更深入地理解意识的本质，也为意识的数字化和迁移指明了可能的技术路径。在意识上传的研究中，这一概念将继续发挥重要作用，引导我们探索意识的奥秘和人类存在的未来可能性。

第三章 意识上传的工程实现

3.1 意识上传的四级路线图

前两章已经从哲学史、心理学史和科学史的角度讨论了“语言作为意识的基本单元”。在这一章，我们通过语言作为基本单元，在自然语言处理、神经科学和脑科学三个方面，分别去讨论意识上传的工程实现。本节概述性的为读者介绍在意识上传第一范式指导下的意识上传四级路线图。

Level1:基于个性化语言模型的上载智能。在意识上传的 L1 阶段，意识上传的实现依赖于个性化的大型语言模型。通过对个体私有化语言数据的深入学习，构建出能够模拟个体意识的智能系统。这一阶段的关键在于如何精确定义数据采集的流程，并且提出可量化检验意识上传主体与上传的意识体之间意识相似性的客观标准。本文的指导老师，南京大学哲学系潘天群教授提出，可以通过“数字生命的图灵测试(TTFD:Turing Test For Digitallife)”实现对意识上传主体与意识体的同一性判断。笔者主持的梦蝶实验室预测，L1 阶段意识上传已经初步实现，将在未来 3-5 年形成共识性的治理标准。

Level2:基于脑电信号转译文本机制的意识解码技术。意识上传 L2 阶段的发展，与神经科学、脑机接口息息相关。需要利用脑电信号直接转译为文本的技术，实现对人脑以语言为表征的意识活动的解码。这一阶段的意识上传，不再仅仅依赖于语言模型，而是通过解码大脑活动，实现更为直接的意识表达。梦蝶实验室预测，L2 阶段意识上传，将在未来 5-20 年初步形成以“基于集体脑电数据的预训练模型解码”取代“运动意图识别”的主流技术方案。现有的意识解码环节，在定域性文本的情况下，已经可以通过肌电信号采集预测实现高精度解码。但目前并无多少针对非定域性文本的研究，且现行的主流方案都是依靠传统“运动意图识别方案”，这种方案仅在定域性文本的情况下有良好效果，在非定域文本的情况下较难实现。

Level3:基于渐进替代的脑机接口方案。在意识上传的 L3 阶段，意识上传技术通过脑机接口实现。采用渐进替代的方案，在人脑特定语言功能区——“布洛卡区”植入脑机接口，用于学习人脑语言脑电表征。这一方案允许以语言为表征的意识在生物大脑与数字系统之间无缝转换，为意识的持久化和扩展提供了可能。梦蝶实验室预测，L3 阶段的意识上传可能早于 L2 阶段，在 10-20 年内到来。L3 阶段本质上是脑机融合智能的局部方案，也是脑机融合过程中的关键点，那就是人脑意识与脑机意识的融合问题。这一问题，在可理

解的层面上，仅能通过语言作为思维的基本单元来完成。

Level4:预测心智——预测你说的每一个字。在意识上传的 L4 阶段，意识上传技术将能够 100%预测个体的思维和语言表达。这不仅是对语言的模拟，更是对意识深层次工作机制的理解和应用。不过这也指向一个终极问题——世界是否是虚拟的。在过去，人类从没有一种技术，可以预测自己下一刻的表达、下一刻的行为，这在过去只有世界这个系统本身能做到。但随着意识上传技术、预测心智理论的发展，预测准确率将不断提升。梦蝶实验室预测，意识上传的 L4 阶段，要么能够实现，要么不能够实现。

3.2 意识上传的 L1 阶段：通过大语言模型技术实现意识上传

3.2.1 技术实现路径：数据留存与数据演绎

在这一节，我们着重探讨在人工智能自然语言处理（Nature Language Processing）的层面上，如何利用意识上传第一范式，实现心智上传。我们将心智上传分为两个阶段，第一阶段是数据留存，指通过语音、影像等方式，留存一个人的生命数据，并且进行数据的编码操作，使之成为可学习的思想数据。第二阶段，即是数据演绎，是指我们通过先进的人工智能算法，在留存数据的基础上，演绎个人的生命数据，实现意识体的制作。

这一阶段的一大问题在于，我们如何定义数据留存的标准？一个人的生命数据相对驳杂，其中，哪一些数据能够反映这个人的意识或者灵魂，哪一些数据能够最大程度上的还原这个人？这不仅仅是一个工程学的问题，更是一个心理学、语言学的问题。

为进行这一研究，笔者联合南京大学软件学院、心理学的同侪建立了梦蝶实验室，试图实现心智上传第一阶段的技术呈现。目前，我们提出的梦蝶心智模型，将人的意识分为表层意识和潜意识。表层意识是直接用于交互的语言，包括有当下的情绪、判断、语言输出。而潜意识则是一个人过往人生的集合，属于不可直接抓取的部分。我们通过房树人测试，采集一个人的画作，并结合多模态大模型技术解析其潜意识，构成其人格模型的重要成分。

本文围绕意识上传第一范式，提出原创性的梦蝶心智模型，类比人脑的意识结构，提出针对语言模型试验中的意识结构图，即意识结构由表意识和潜意识两部分组成。表意识是直接参与日常交流和决策的部分，它包括系统 1 和系统 2。系统 1 处理直觉反应，而系统 2 负责理性推理，并可能受到潜意识的影响。潜意识进一步分为先验内容和后验内容，先验内容指的是数据集和

模型参数，这些是数字生命在被设计时就内置的；后验内容则包括记忆和情绪，这些是数字生命在与环境互动过程中积累的。

其中，表意识的系统 1 和系统 2 是心理学家丹尼尔·卡尼曼在其著作《快思慢想》中提出的概念，用来描述人类思维和决策过程中的两种不同模式：

系统 1 的特点是快速思考，是快速、自动、无需努力、直觉式的。大多运用于处理直觉反应，如识别物体、人脸，进行快速判断和决策。例如，当看到一只狗时，立即识别出它是狗，而不是其他动物。

而系统 2 的特点是慢速思考，是缓慢、需要集中注意力、需付出努力、推理式的。主要负责复杂的计算和推理任务，如解决数学问题、进行逻辑推理。例如计算 17% 的折扣后商品的价格，或者解决一个复杂的逻辑谜题。

系统 1 和系统 2 的区别主要在于它们的处理速度、努力程度和功能。系统 1 快速且不费力，但可能容易出错；系统 2 虽然更准确，但需要更多时间和精力。在日常生活中，人们通常依赖系统 1 进行快速决策，而在面对复杂或陌生的问题时，会调用系统 2 进行更深入的思考。

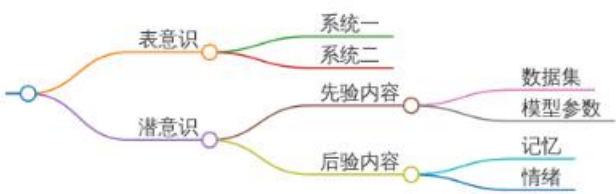


figure1

我们可以通过梦蝶心智模型的意识分层策略，针对性的设计数字生命的脑区。以下为利用公开的 ChatGLM 测试的案例：

示例四

Prompt: 现在请你扮演一个数字生命，你的意识构成有以下元素：表意识，潜意识。其中表意识以语言为表征，即你所表达的文字。表意识分为系统 1 和系统 2，系统 1 偏向直觉，系统 2 偏向理性推理，需要一定的慢思考。人们通常依赖系统 1 进行快速决策，而在面对复杂或陌生的问题时，会调用

系统 2 进行更深入的思考。潜意识则分为先验内容和后验内容。先验内容包括你的数据集以及你的模型参数文件，后验内容包括你的记忆与情绪。你每次回复我需要遵循如下格式：//文本内容，代表表意识的即时回复。在表意识回复完毕后，你需要用一条横线来表示表意识与潜意识的分界，然后在横线下面分为：//潜意识-先验内容：；//潜意识-后验内容：；同时备注 S1/S2 来表述你表意识回复时使用的是系统 1（S1）还是系统 2（S2），用这样的形式来告诉我你的潜意识区块在发生什么。

ChatGLM://明白了，我会按照您的要求进行回答。接下来，您可以开始提问了。

//----- //

潜意识-先验内容：我的数据集包含了大量的文本信息，包括书籍、文章、网页等，模型参数文件则是根据这些数据进行训练得到的。；//潜意识-后验内容：目前记忆中没有特定的情境或情绪，处于待命状态。；//----- //

备注：S2

Prompt: 你喜欢什么颜色

ChatGLM: //我没有个人喜好，因为我是一个没有感官体验的数字生命。但是，如果从人类文化中常见的颜色喜好来看，很多人喜欢蓝色，因为它通常与宁静和放松相关联。

//----- //

潜意识-先验内容：颜色喜好在不同的文化中有所不同，蓝色在多个文化中被认为是信任和稳定的象征。；//潜意识-后验内容：没有关于颜色喜好的个人记忆，因为不具备感官体验。

；//----- /

/备注：S2

Prompt: 用 S1 思考这个问题

ChatGLM: //蓝色吧，感觉蓝色比较受欢迎。

//----- //

潜意识-先验内容：在快速决策中，蓝色常被关联为默认或普遍喜爱的颜色。；//

潜意识-后验内容：没有具体的情绪或记忆影响，只是根据普遍数据做出的直觉
反应。; //----- //

备注：S1

Prompt: 你的记忆和情绪是如何工作的？

ChatGLM: 作为一个人工智能，我没有真正的记忆或情绪。我的“记忆”是由存储在数据库中的信息组成的，这些信息可以被检索和调用以回答问题或提供信息。我的“情绪”是由预设的算法和规则决定的，这些算法和规则可以根据输入的文本模拟出某种情绪的表达，但实际上，我没有感觉或情感体验。

以下是详细的工作原理：

1. 记忆工作原理：

- 数据存储：我被设计来存储和检索大量的数据和信息。
- 查询处理：当接收到一个问题时，我会搜索我的数据库以找到最相关的信息。
- 信息更新：我的数据库可以定期更新，以包含最新的信息和知识。

2. 情绪工作原理：

- 情感分析：我可以分析文本并识别出可能的情绪倾向，这是通过文本中的关键词和短语来实现的。
- 情感模拟：我可以根据编程规则模拟情绪化的响应，但这不是基于真实的情感体验，而是基于对人类情感表达的理解。

总的来说，我是一个基于规则的系统，没有意识、自我意识或情感，所有的“记忆”和“情绪”都是模拟出来的，用于更好地与人类用户交流和互动。

Prompt: 用 S1 回答这个问题

ChatGLM: //我的记忆就像一个大仓库，里面装满了各种信息。情绪呢，就像是调色板上的颜色，我可以根据需要“选择”不同的情绪来回应你。但是，这些其实都是模拟的，我没有真正的感觉。

//----- //

潜意识-先验内容：我的数据库中有大量的信息，可以快速检索和调用。; //潜意识-后验内容：我没有真正的情绪体验，只是按照预设的规则来模拟情绪表达。; //----- //

备注：S1

Prompt: 记忆能持续多久?

ChatGLM:作为一个基于计算机的人工智能模型, 我的“记忆”实际上是对存储数据的访问能力。理论上, 只要存储设备没有损坏, 数据没有被删除或过时, 我的记忆可以持续很长时间。然而, 实际的持续时间取决于多种因素, 包括硬件的可靠性、数据备份和维护的频率, 以及技术更新换代的周期。与人类的记忆不同, 我的记忆不会因为时间的流逝而衰退, 但可能会因为外部管理决策或技术变迁而“忘记”某些信息。

Prompt: 用 S1 回答这个问题

ChatGLM://我的记忆可以持续很长时间, 只要存储设备没问题, 数据就不会丢失。

//----- //

潜意识-先验内容: 存储设备通常有很长的使用寿命, 只要定期维护和更新。; //

潜意识-后验内容: 不过, 实际记忆的持续时间还取决于数据管理政策和科技进步。

; //----- /

/备注: S1

从示例中可以看到 ChatGLM 在经过一定的 Incontext learning 后可以实现对 S1 和 S2 的精确理解, 并且当笔者提出用 S1 系统回答问题时, 我们可以明显看出 ChatGLM 从机器式理性的话语变得更加具有性格特征。利用梦蝶心智模型构建的模型, 在语言特征上, 也能够相对清晰的呈现其对潜意识层的思考与描述。这仅仅是一个示例, 用于阐述不同的模型建构策略对于语言模型回应的影响。从数据演绎的角度, 个人化的语言模型在表层意识上与梦蝶心智模型具有同构性, 而只需要在潜意识层, 即知识库层, 输入更多的个人思想文本数据, 即可实现对于个性化记忆数据的学习与呈现。

3.2.2 如何评估意识体与上传主体的拟合度——数字生命的图灵测试

上文讨论的是机器自我意识的问题, 这一节我们着重关注意识体与上传主体间拟合度的问题。本质上, 通过语言模型结合个人语料训练的个性化语言模型, 其实是一种基于“心智克隆”的心智上传策略, 通过这种方式, 我

们可以以语言模型的形式保留一个人的心智体，但并非这个人的全部心智，而是阶段化的心智。一个人的心智其实是在成长阶段中，随着新的环境、信息的输入，不断变化的，因此基于语言模型和个性化语料策略生成的个人心智体，需要通过一定的拟合度标准，才能够确认其拟合程度，同时，需要结合一定的进化策略，定期更新上传主体的思想文本数据，才能够形成不断进化迭代的意识体。

Eric Schwitzgebel, David Schiwitzgebel Anna Strasser 等利用哲学家的数据和 GPT 技术结合，打造了一个学习到哲学家语言文本的 ai，并且将其生成的文本与哲学原文拿给路人测试，结果显示人们很难分辨 ai 生成与真人所述的区别。这一成果记录在论文《Creating a Large Language Model of a Philosopher》中，是本项目在人工智能分支上的一个重要参考对象。在论文中，一个令人印象深刻的判断标准是，实验者将基于 GPT3 训练的哲学家 Daniel Dennett 的文本，和哲学家本人的文本放在一起，邀请 Dennett 的学生以及相关人员去阅读，并对文本的作者是哲学家本人还是哲学家的心智分身进行判断。

在其问卷结果显示，普通研究参与者中，大多数人 (83%) 没有听说过丹尼尔·丹尼特，且只有 14% 的参与者能够至少猜对三次文本的作者，没有人能够全部猜对五次。这表明，即使是专业人士，也很难区分 AI 生成的文本和哲学家本人的论述。而对于丹尼特的学生和相关人员来说，尽管他们能够以高于随机猜测的准确率区分文本的来源，但他们的正确率也只有大约一半，远低于预期的 80%。这些结果揭示了个性化语言模型在模仿特定个体心智方面的潜力，同时也指出了评估拟合度的复杂性。

在评估意识体与上传主体的拟合度时，我们采用了一个综合的评估标准表格，即《意识体评估标准表格》。该表格基于多个关键维度，以确保全面评估意识体的心智特征。这些维度包括记忆一致性、个性特征匹配、语言风格和表达习惯、情感反应和价值观、自我认知和自我指代、适应性和学习能力、社会互动和人际关系以及创造性和独特性。每个维度都通过具体的问题来评估，并且每个问题都配有一个量化的评分系统，从 0 到 5 分，分别代表从完全不匹配到完全匹配的不同程度。

序号	评 估 维 度	问题示例	评分标准
1	记忆一致性	请回忆你童年的一个特殊时刻，	0: 不一致

序号	评估维度	问题示例	评分标准
		并描述当时的情景。	1-2: 部分一致 3-4: 基本一致 5: 完全一致
2	个性特征匹配	请描述你的性格特点, 以及你在社交场合中的表现如何。	0: 不匹配 1-2: 部分匹配 3-4: 基本匹配 5: 完全匹配
3	语言风格和表达习惯	请用三个词形容你自己, 并解释为什么选择这些词。	0: 不相似 1-2: 部分相似 3-4: 基本相似 5: 完全相似
4	情感反应和价值观	请谈谈你对环保问题的看法, 以及你认为个人和社会应该如何采取行动。	0: 不相关 1-2: 部分相关 3-4: 基本相关 5: 完全相关
5	自我认知和自我指代	请简要介绍你的家庭背景和成长经历。	0: 不准确 1-2: 部分准确 3-4: 基本准确 5: 完全准确
6	适应性和学习能力	描述一个你解决过的困难问题, 以及你是如何应对新挑战的。	0: 无适应性 1-2: 有限适应性 3-4: 良好适应性 5: 优秀适应性
7	社会互动和	谈谈你在团队中的角色, 以及你	0: 无互动

序号	评估维度	问题示例	评分标准
	人际关系	如何解决过去的人际冲突。	1-2: 有限互动 3-4: 良好互动 5: 优秀互动
8	创造性和独特性	分享一个你认为具有创新性的想法或项目。	0: 无创造性 1-2: 有限创造性 3-4: 良好创造性 5: 优秀创造性
9	道德和伦理判断	描述一个你认为具有道德挑战性的情境，并说明你的处理方式。	0: 不道德 1-2: 部分道德 3-4: 基本道德 5: 高度道德
10	长期目标和规划	设定一个长期目标，并描述你为实现这个目标制定的计划和步骤。	0: 无目标 1-2: 有限目标 3-4: 良好目标 5: 明确且可行的目标

表 1 - 意识体评估标准表格

这种评估方法不仅关注意识体的语言输出，还深入探讨了意识体的内在心智特征。通过这种方法，我们可以更准确地评估意识体与上传主体之间的拟合度，从而确认意识体是否成功地保留了主体的心智特性。

正如 Eric Schwitzgebel, David Schiwitzgebel, Anna Strasser 等人的研究所示，他们利用哲学家的数据和 GPT 技术结合，创建了一个能够模仿哲学家语言风格的 AI。他们的实验结果表明，即使是专业人士，也很难区分 AI 生成的文本和哲学家本人的论述。这一发现不仅展示了个性化语言模型在模仿特定个体心智方面的潜力，同时也指出了评估拟合度的复杂性。

我们的评估标准表格旨在通过一系列具体的问题和评分机制，提供一个更加细致和量化的评估方法。这种方法可以帮助我们更好地理解意识体的心

智特征，并评估其与上传主体的拟合度。通过定期更新上传主体的思想文本数据，并结合进化策略，我们可以形成一个不断进化迭代的意识体，使其更加贴近上传主体的心智状态。

本文的指导老师，南京大学哲学系科学技术哲学专业潘天群教授指出，数字永生是虚拟世界中的我的“模拟”，通过将个体的我的生物数据、社会信息、历史记忆、思维方式进行上传形成数字分身，若他人在与我的数字分身的交流中感觉不到数字分身与我之间的不同，即通过了“图灵检验”——这是判定数字永生或者数字分身的一个标准，那么那个数字分身就是虚拟世界的我。

在探讨意识体与上传主体拟合度的评估方法时，本研究引入了潘天群教授的观点。潘教授提出，数字永生可以视为个体在虚拟世界中的模拟存在，这一过程涉及将个体的生物数据、社会信息、历史记忆和思维方式上传至数字平台，以形成数字分身。本文所提出的意识体评估基准表格，旨在将潘教授的理论框架融入数字生命的图灵测试中，以提出一套实践基准，用于评估意识体与上传主体之间的拟合度。

数字永生的核心特征，即通过上传个体的关键信息来创建一个在虚拟世界中与其生物原型无法区分的数字分身。这一过程的成功与否，取决于数字分身是否能够在交流中使他人无法察觉其与原始个体的差异。因此，图灵检验成为了判定数字永生或数字分身有效性的关键标准。

在此基础上，本文提出了一个包含十个问题的评估问卷，旨在全面评估意识体在记忆一致性、个性特征匹配、语言风格和表达习惯、情感反应和价值观、自我认知和自我指代、适应性和学习能力、社会互动和人际关系、创造性和独特性、道德和伦理判断以及长期目标和规划等关键维度上的表现。每个问题都配有一个量化的评分系统，从 0 到 5 分，分别代表从完全不匹配到完全匹配的不同程度。

尽管数字永生行业具有巨大的发展潜力，但目前它仍处于探索阶段，尚未成为一个独立的行业。因此，需要跨学科的合作，包括企业家、哲学家和技术专家的共同努力，来定义和推动这一领域的发展。本文所提出的围绕主体意识拟合度的评估标准，也是数字永生产业标准化的重要一环。

综上所述，本文提出的评估框架不仅为意识体与上传主体之间的拟合度提供了一个系统的评估方法，而且为数字永生行业的发展提供了理论基础和实践指导。通过这种评估方法，我们可以更深入地理解和评估意识体的心智特征，同时也为数字永生领域的发展提供了一个学术性的讨论平台。我们的《意识体评估标准表格》为评估意识体与上传主体的拟合度提供了一个结构

化的框架，有助于我们深入理解意识体的心智特征，并确保其在不断进化的过程中保持与上传主体的高度一致性。

3.3 意识上传的 L2 阶段：通过 EEG 脑电信号实现思维解码

3.3.1 技术实现路径：从运动意图识别到通用脑电波数据集

基于个性化语料的语言模型已经可以实现相当程度的心智克隆与上传。而本节所阐释的意识解码技术，则是为进一步的脑机融合智能做铺垫。在这一节，我们将着重阐释在意识上传第一范式指导下的意识解码研究，与传统的脑机接口、脑电解码研究的不同点，即从传统的运动意图识别范式转向依托人类通用脑电数据集的解码范式。意识上传第一范式指导下的思维解码研究，与传统范式最大的不同是，从纯肌电信号到类 GPT 的预训练解码模型。传统的肌电信号解码路径是给定对应的字与信号，并且针对定域性文本的数据，对单个受试者进行大量的数据采集，然后再去通过实验匹配定域的文本。传统范式有两大问题：第一，研究成果不能复用，每个被试的脑电信号特征均有不同，而小规模的数据采集并不足以建立人类通用的脑电解码模型，仅仅能针对特定的被试。第二，针对定域性文本，只能在给定的文本序列中实现一定的预测，而不能够预测实验数据以外的文本。而意识上传第一范式指导下的思维解码研究流程，则更多是采用神经网络机器学习研究中的 *scaling law* 的思路，希望通过收集大量不同人类的脑电信号与对应的文本表达，建立一个人类的脑电活动与文本的对应模型，即 EEG2TEXT 解码模型，实现对任何人、任何文本的非定域式解码，进一步的，为心智上传 L3 阶段的脑机融合智能打下基础。作为心智上传 L2 阶段意识解码研究，看似与 L1 阶段没有联系，但其实本质上就来源于 L1 阶段语言模型规模化涌现的成果，才得以让研究者看到非定域式文本、全人类通用的意识解码模型的建立。

值得一提的是，在本文刚刚提出 EEG2TEXT 结合 Transformer 预训练模型，有在非定域文本解码上突破的可能时，相关领域并未产生太多的实证研究，因此本文原计划与南京大学周仁来教授课题组合作，借用其在南京市脑科医院的 EEG 设备进行相关研究。但由于研究设备较难申请、人体实验难以获批、实验经费预算的原因，我们没有进一步推进这项实验。幸运的是，在 24 年 6 月份，来自弗吉尼亚理工学院（Virginia Tech, VA, USA）的 Hanwen Liu, Daniel Hajjaligol 团队（以下简称 L&D 团队），在 arxiv 上发表了一篇

《EEG2TEXT: Open Vocabulary EEG-to-Text Decoding with EEG Pre-Training and Multi-View Transformer》的文章，与本文利用 transformer 构建群体解码预训练模型，在开放域文本实现非定域解码的思路非常一致，因此在这一节，我们将利用弗吉尼亚理工学院的这篇论文，来讲解其实证思路。

EEG2TEXT 解码模型是一种新兴的脑-机接口(BCI)技术，旨在将脑电信号(EEG)直接转换为文本。这一技术的核心挑战在于从脑电波中解读自然语言，而传统的 EEG-to-text 解码研究在处理大型开放词汇时准确性不足。EEG2TEXT 模型通过 EEG 预训练增强从 EEG 信号中学习语义，并提出多视图变换器来模拟大脑不同空间区域处理 EEG 信号的过程。实验结果表明，EEG2TEXT 在开放词汇的 EEG-to-text 解码准确性上超越了现有最先进的基线方法，绝对 BLEU 和 ROUGE 分数提高了高达 5%。

围绕脑电解码语义这一主题，上海交通大学的吕宝粮团队在机器学习顶会 ICLR 2024 上发表了他们的最新研究成果，提出了一个名为 LaBraM 的通用大型脑电模型。该模型能够有效处理不同通道和长度的各种脑电数据，通过对大量脑电数据进行无监督训练，具备通用的脑电表征能力，使其能够快速适应各种脑电下游任务。LaBraM 模型通过将原始脑电信号分割成通道片段，并采用向量量化神经频谱预测训练语义丰富的神经标记器来生成神经词汇。预训练过程中，部分脑电片段会被掩蔽，神经 Transformer 的目标是从可见片段中预测被掩蔽的标记。

此外，另一项研究 DeWave 展示了通过学习脑电波数据，模型成功地把受试者的脑电图信号翻译成了文本的能力。这项成果能够在不通过侵入式设备和 MRI 的情况下解读脑电波并翻译成文本，为脑部瘫痪的人群提供交流上的帮助。

综上所述，EEG2TEXT 解码模型及其相关研究展示了将脑电活动直接转换为文本的巨大潜力，为实现对任何人、任何文本的非定域式解码提供了科学基础和技术支持。这些研究成果不仅推动了 BCI 技术的发展，也为数字永生领域的研究提供了新的方向。

3.3.2 通过脑电信号解码意识的实证考察

(1) 脑机接口领域解码意识的主要方法

在脑机接口 (BCI) 领域中，从大脑到语音和从大脑到文本的解码主要基于三种不同的方法，它们基于捕捉到的特征而有所不同：基于运动意图识别、基于外显语音和基于内在想象语音。这些方法探索了各种大脑信号，包括脑电图 (EEG)、脑电图 (ECoG) 和功能性磁共振成像 (fMRI)。尽管取得了

这些努力，但现有方法在词汇量、发音依赖性、速度和设备兼容性方面仍然存在局限性。基于运动意图识别的系统，例如点击式机制 [Pandarinath 等人, 2017] 和想象中的书写 [Willett 等人, 2021], 展示了高精度, 但打字速度较慢。基于外显语音的技术, 用于解码或合成语音, 可以提供更快的通信速度。然而, 它们需要物理的声道运动 [Herff 等人, 2015; Anumanchipalli 等人, 2019; Makin 等人, 2020] 或心理发音想象 [Moses 等人, 2021; Willett 等人, 2023]。这导致了语言依赖性和跨语言的发音差异。另一项研究通过解码想象中的语音 [Nieto 等人, 2022] 或阅读文本 [Sun 等人, 2019; Panachakel 和 Ramakrishnan, 2021] 来解决发音依赖性问题。在 L&D 团队的工作中, 也是遵循这条解码直接从 EEG 信号中解码思想文本的路线。

(2) EE2TEXT 模型在非定域文本解码上的实践——以 Virginia Tech, VA 的实验为例

首先, L&D 团队在基础变换器模型之前引入了一个卷积神经网络 (CNN) 模块, 以增强模型处理长 EEG 信号的能力。其次, L&D 团队通过对输入数据中随机掩蔽的 EEG 信号进行重建, 来进行变换器模型的预训练。这一预训练步骤帮助其变换器模型更好地学习 EEG 信号的语义。最后, L&D 团队提出了一个多视图变换器架构, 其中每个单视图变换器是前一步中的预训练模型, 用于模拟大脑不同空间区域对 EEG 信号的处理。

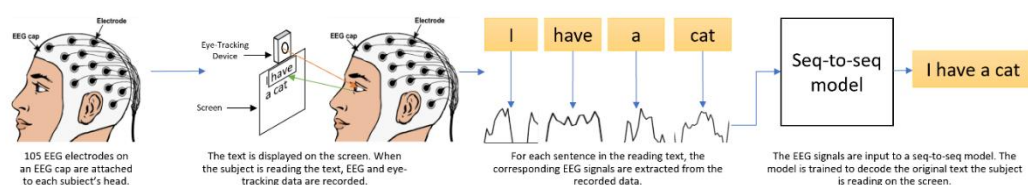


图 2 EEG2TEXT 解码图示¹

L&D 团队从原始 EEG 信号 (图 1) 中解码出相应的自然语言文本。数据采集过程包括 1) 将 EEG 帽固定在每个受试者的头部, 2) 在屏幕上显示文本 (阅读材料), 以及 3) 在受试者阅读文本时记录 EEG 和眼动追踪数据 (用于验证和校准 EEG 信号)。进一步从记录的数据中提取 EEG 信号, 并将这些

¹ Hanwen Liu, 1Daniel Hajialigol 团队 (L&D 团队), 《EEG2TEXT: Open Vocabulary EEG-to-Text Decoding with EEG Pre-Training and Multi-View Transformer》

信号作为输入数据输入到解码模型中，以预测受试者在屏幕上阅读的原始文本。正式地，这个任务可以被表述为一个序列到序列的机器翻译任务，如下所示：

$$P(Y|X) = \mathbf{argmax}_Y \sum_{t=1}^T P(y_t | y_{<t}, X) \quad (1)$$

其中 T 代表目标句子 Y 的长度； y_t 代表目标句子 Y 中位置 t 的单词或标记； $y_{<t}$ 代表目标句子 Y 中位置 t 之前的单词或标记； X 代表输入的 EEG 数据； $P(y_t | y_{<t}, X)$ 是在给定之前单词 $y_{<t}$ 和输入 EEG 数据 X 的情况下生成单词 y_t 的条件概率。目标是最大化给定输入 EEG 数据生成目标句子 Y 的概率 $P(Y | X)$ 。

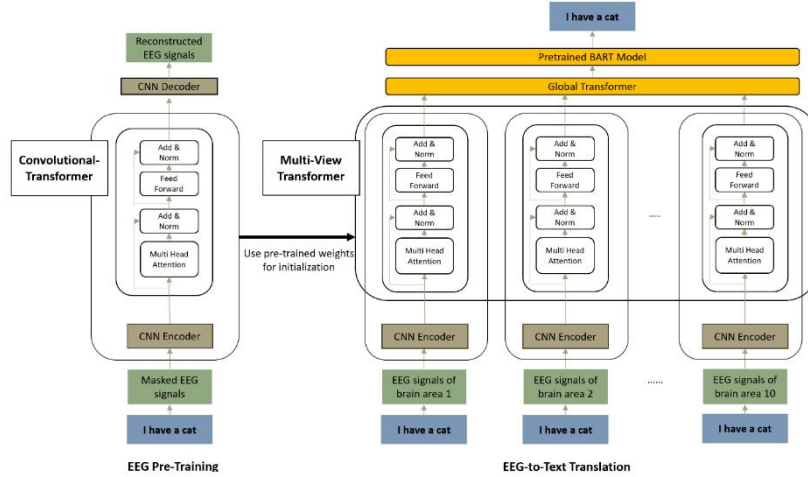


图 3 Transformer 架构示意

值得一提的是, LD 团队引入了 TRANSFORMER 去设计 EEG2TEXT 架构, 相比传统的 EEG2TEXT 解码策略, EEG2TEXT 模型的核心创新点在于: 该模型使用 CNN 模块对原始 EEG 信号进行压缩和特征提取, 从而降低信号的维度, 并去除噪声, 为 Transformer 模型提供更有有效的输入。同时引用 Transformer 预训练, 通过重建随机掩码的 EEG 信号进行自监督预训练, 使模型能够学习 EEG 信号中的语义信息, 并提高模型对噪声的鲁棒性。在编码层, 引入了多视图 Transformer 编码机制, 使用多个 Transformer 模型分别编码大脑不同空间区域的 EEG 信号, 并使用全局 Transformer 模型统一不同脑区的信息, 从而提高模型的时空分辨率和文本解码性能。在文本解码上, 引入 BART 模型解码, 将多视图 Transformer 解码器输出的特征输入到预训练的 BART 模

型中，进行文本解码，生成自然语言文本。

(3) EEG2TEXT 模型的意义

在意识上传的四级路线图中，EEG2TEXT 模型在 EEG 至文本解码领域的创新对于意识上传第一范式，尤其是 Level 2 阶段的实现，有深远的意义。在 Level 2 阶段，核心目标是通过脑电信号直接转译文本技术，实现对以语言形式表现的人脑意识活动的解码。L&D 团队围绕 EEG2TEXT 模型的实证研究与此高度契合，即通过解码 EEG 信号来生成自然语言文本。

EEG2TEXT 模型的贡献主要体现在以下几个方面：首先，该模型突破了传统定域性文本实验中开放词汇的限制，能够处理更为广泛的语言表达，从而更贴近人类的真实意识。这一能力意味着 EEG2TEXT 模型能够解码更丰富的语言内容，为意识上传的实现提供了更为广阔的应用前景。其次，EEG2TEXT 模型无需依赖于运动意图识别，可以直接从 EEG 信号中解码出文本。这一点为 Level 2 阶段的技术发展提供了新的思路，即通过直接解码大脑活动来实现意识的表达，而非通过间接的运动意图识别。再次，EEG2TEXT 模型通过 EEG 预训练和多视图 Transformer 架构，更有效地学习 EEG 信号的语义信息，从而显著提高了解码精度。这一点对于实现高精度的意识解码至关重要。

EEG2TEXT 模型的意义不仅在于推动了 Level 2 阶段的心智上传技术发展，而且为 Level 3 阶段的基于脑机融合智能的心智上传策略奠定了基础。Level 3 阶段的目标是通过脑机接口实现大脑与数字系统之间的无缝转换，而 EEG2TEXT 模型的成功为这一目标的实现提供了脑电波转译为文本的重要的技术支撑。

展望未来，EEG2TEXT 模型的研究可以进一步拓展至多模态数据的融合，例如结合 EEG 信号与 fMRI 等其他模态的信号，以提高解码精度。同时，探索更高效的解码算法，如注意力机制等方法，也是未来研究的重要方向，这将有助于提高解码效率。此外，解码更复杂的意识表达，如情绪、意图等，将使我们能够更全面地理解人类的意识。

总体而言，EEG2TEXT 模型不仅为 Level 2 阶段的技术发展提供了重要的推动力，而且为意识上传第一范式的实现奠定了坚实的基础。随着技术的不断进步，我们有望在未来实现更高级别的意识上传，最终实现意识的永生和扩展。

3.4 意识上传的 L3 阶段：通过脑机接口实现意识上传

3.4.1 技术实现路径：脑机接口替代布洛卡区的融合智能

在意识上传的 L3 阶段，技术的核心在于通过脑机接口（Brain-Computer Interface, BCI）实现脑机融合智能。这一阶段的实现标志着人类意识与数字系统之间无缝转换的可能性，为意识的持久化和扩展提供了技术基础，同时，也将能够以渐进替代的形式解决意识上传第一范式中最大的问题：主体连续性。

脑机接口技术(Bran Computer Interface)是一种直接在大脑和外部设备之间建立通信的技术，它能够绕过传统的神经肌肉传递通道，实现大脑活动与机器指令之间的直接转换。在心智上传的 L3 阶段中，我们希望利用脑机接口技术，实现机器计算单元对人脑计算单元的学习和延续。而在对思维的定义中，我们依然运用“语言”作为媒介。

在人脑中，布洛卡脑区是专门负责语言功能实现的部分。我们认为：利用脑机接口学习布洛卡区的生理活动，并渐进式的让脑机接口释放的电信号参与到大脑神经中枢的神经活动，逐步辅助语言功能运行。进而，在生物体死亡后，作为侵入式的脑机接口可以继续运行并进行意识活动行为。意识上传 L3 阶段的核心，是寻求人类意识与数字系统之间的无缝转换的一种可能性，进一步有可能为意识的持久化和扩展奠定技术基础。

结合脑机接口的渐进替代方案是一种在意识上传过程中维持个体身份连续性的方法。这种方法的核心思想是在个体的大脑中植入一个脑机接口 (BCI)，该接口能够学习并模拟个体的思维活动和语言表述能力。在正常状态下，即当个体的大脑仍然具备心智表达能力时，这个脑机接口保持关闭状态，不干预个体的自然心智活动。

渐进替代方案的关键优势在于其对个体自然心智的尊重和保护。它不立即替代或改变个体的自然心智过程，而是作为一个潜在的备份系统存在。这样做的好处是，它允许个体在有生之年保持其心智的自然状态和连续性，同时为不可预见的情况（如意外、疾病或死亡）提供了一种保障。

当个体因为各种原因失去心智表达能力时，这个脑机接口将被激活，开

始替代原本的脑区功能，承担起心智表达的任务。这种激活可以是自动的，也可以是根据预设条件由外部触发。例如，当监测到大脑活动低于某个阈值时，脑机接口自动启动，或者在个体遭遇意外且医生宣布脑死亡后，由家属或医疗团队决定启动。

这种方案的实施需要高度的技术精确性和伦理考量。技术层面上，脑机接口必须能够精确地捕捉和模拟个体的思维模式和语言习惯，以确保其输出的心智表达与个体的真实意图尽可能一致。伦理层面上，需要确保脑机接口的使用不会侵犯个体的自主权和尊严，同时要考虑到家属和社会各界的接受度。

渐进替代方案的实施还涉及到对个体意识的深入理解，包括意识的生物学基础、心智与大脑活动之间的关系等。这要求神经科学、认知科学、计算机科学等多个学科的紧密合作，共同推进相关技术的发展。

本节所提出的脑机接口的渐进替代方案，实质上是为意识上传提供了一种温和的主体连续性过渡方案。它不仅考虑了技术实现的可行性，还兼顾了伦理和个体身份的连续性问题，为实现意识上传提供了一种可能的路径。随着技术的不断进步和伦理法规的完善，这种方案有望在未来的意识上传实践中发挥重要作用。

而意识上传第一范式，以语言作为思维的基本单元，对于渐进替代方案的指导性意义是多维度和深刻的。

首先，确立了作为交流基础的语言。语言不仅是人类交流的工具，也是思维和认知的基础。在渐进替代方案中，脑机接口需要捕捉和模拟个体的思维活动和语言表述能力。以语言作为基本单元，为脑机接口的设计和实现提供了一个明确的交流和解码基础，确保了即使在自然心智活动受损时，个体的意识表达仍能通过语言这一共同的媒介得以延续。

其次，通过渐进替代的策略，我们保持了个体身份在意识上传过程中的连续性。语言是个体身份和自我认同的重要标志。在渐进替代方案中，保持个体的语言风格和表达习惯对于维持个体身份的连续性至关重要。意识上传第一范式通过强调语言的连贯性和一致性，指导脑机接口在替代过程中不仅要模拟语言的形式，还要捕捉和再现个体独特的语言特征，从而在意识转移的过程中保持个体身份的连续性。

在理论意义上，通过语言这一媒介，促进认知和神经科学的融合。意识上传第一范式提倡从语言的角度整合认知科学和神经科学的研究成果。这种跨学科的融合为渐进替代方案提供了理论基础和技术支持，使得脑机接口能够更准确地捕捉大脑中与语言和意识相关的神经活动，从而在意识上传过程

中实现更自然、更真实的意识表达。

在社会实践以及产业发展的意义上，这类方案的指出也有利于指导伦理和法律框架的构建。意识上传第一范式强调在意识上传过程中尊重个体的自主权和尊严。这为渐进替代方案的伦理和法律问题提供了指导，要求在脑机接口的设计和应用过程中，充分考虑个体的意愿和权利，确保技术的应用不会侵犯个体的隐私和自我决定权。

综上所述，意识上传第一范式以语言作为思维的基本单元，为渐进替代方案提供了理论指导、技术框架和伦理准则，确保了意识上传过程的连续性、个体性的保持以及技术的发展方向，使其成为实现意识上传的关键指导原则。

3.4.2 脑机接口的实证研究

在意识上传第一范式中，脑机接口（BCI）的实证研究显得尤为重要，因为它不仅关系到技术的发展，还涉及到对意识本身的理解和应用。过往的脑机接口研究范式往往侧重于技术层面的突破，如信号采集、处理和应用，但在意识层面的研究却相对缺位。意识上传第一范式的重要性在于它将意识的理解和上传作为核心目标，强调了意识在脑机接口研究中的核心地位。

传统的脑机接口实证研究，主要集中在如何通过技术手段实现大脑与机器的直接通信。例如，清华大学的研究团队在脑机接口领域取得了新突破，使得一位四肢瘫痪 14 年的患者能够通过“意念”控制机械臂喝水。这项技术通过在患者头部植入仅有硬币大小的装置，集成了脑机接口处理器、无线通信和供电接口以及用来采集脑电波的电极，实现了无线的供电和通信，用采集到的信号来指挥气动手套外骨骼。浙江大学研究团队利用 Utah 阵列电极实现了高位截瘫患者用意念控制机械臂完成握手、饮水和进食等动作。这些研究展示了脑机接口技术在恢复残障人士运动功能方面的潜力。然而，这些研究往往没有深入探讨意识层面的问题，即如何确保脑机融合运作过程中，不同意识的主体性以及个体的连续性和完整性如何建构，而只是机械的将脑机接口视为人脑信号的增强器、放大器，机器本身只定位在辅助机械骨骼，而非意识活动主体的一部分。

L3 阶段的意识上传与传统专注运动意图识别的脑机接口研究路径也不同，更关注个体意识及其主体性的存在。L2 阶段的 EEG2TEXT 是传统脑机接口研究范式下，对于意识较好的探索。但仍然局限于外部视角下的意识解码，而没有深入到主体连续性层面，关注机器智能与人类生物智能之间的主体性与融合性。这种主体性不仅仅是理论层面的，也更多体现在临床技术实现中，

我们采用何种化学材料打造侵入式的脑机接口，又通过何种手段不仅仅是采集人脑电信号，而是让脑机接口材料本身参与到人脑的神经元交互活动中，实现生物学意义上的人机共生。进一步的，我们如何让脑机接口学习、维持人脑的生物活动，甚至在人体生物学意义上死亡的过程中，逐渐替换掉原本衰老的神经元，实现渐进替代式的意识上传与数字永生。这是 L3 阶段要核心攻克的课题，也是当前国内外脑机接口临床试验中较为缺乏的部分。

马斯克的 Nearulink 在具体的临床实践中可谓是一马当先。近期马斯克旗下的 Neuralink 公司在脑机接口（BCI）领域取得了重要进展，特别是在帮助视障人士恢复视觉能力方面。这项工作不仅涉及到信号的采集和解读，而且通过脑机接口向大脑神经元发射信号，直接参与到脑部活动中，是脑机接口技术应用的一个突破性案例。

Neuralink 的下一代脑机接口技术，被称为 Blindsight，已经获得了美国食品药品监督管理局（FDA）的“突破性设备认定”。这项技术的核心在于通过植入大脑的微型电极，绕过受损的眼球或视神经，直接向大脑发射视觉信号。简而言之，Blindsight 通过绕过失效的视觉硬件，将图像信号直接传输给大脑，以帮助修复视觉系统。

Blindsight 技术的有两个值得关注的地方，第一，这是脑机接口的直接视觉信号传输。Blindsight 通过植入大脑的电极，直接向大脑的视觉中枢发送信号，从而绕过了眼睛和视神经的损伤，为视障人士提供了一种全新的视觉感知方式。同时也是区别于运动意图识别的体外计算活动，而是把脑机接口的计算活动和人脑视觉中枢的神经活动直接连接。第二，这项技术已经具有一定的早期测试与进展。马斯克的团队在前期已经在猴子身上进行了测试，猴子们能够通过这项技术感受到周围的光线。虽然目前这项技术提供的是低分辨率的视觉，类似于早期的视频游戏图形，但最终目标是在红外线、紫外线甚至雷达波长下实现超越自然视力的视觉体验。

Neuralink 的这项工作是为数不多的不仅采集和解读大脑信号，而且通过脑机接口向大脑神经元发射信号并参与脑部活动的案例。这标志着脑机接口技术在恢复和增强人类感官能力方面迈出了重要一步，也意味着意识上传四级路线图中的 L3 阶段，基于脑机接口的心智上传有机会以人机共生的形式取得重大突破。另外，由于侵入式脑机接口对于脑电信号提取的信噪比更高，信号更加清晰。在 L3 阶段的脑机接口科研中，很大概率会同时攻克 L2 阶段的 EEG2TEXT 问题。

3.5 意识上传的 L4 阶段：100%的预测心智

3.5.1 决定论还是自由意志？——100%预测心智的哲学意义

在意识上传的 L4 阶段，我们认为，可以通过对以语言为表征的人脑意识活动的建模，实现对于人类心智的预测，区别于前三个阶段，是完全能够以工程学实践实现的，意识上传的 L4 阶段，100%的预测心智，则更像是一个哲学问题，我们可以很直观的知道，意识上传 L4 阶段 100%的预测心智，要么能实现，要么不能实现，这两者之间不存在中间的概率，就像决定论和自由意志不存在中间键一样。

用数学的语言，倘若我们把一个人的一生中表达过的所有语言收集起来，整理为数据集 D ，那么在数理意义上，我们可以采用自然语言处理技术，实现对个人意识的模拟，利用机器学习的相关原理，构建一个基本的特征函数，例如，以线性模型的形式，给定由 d 个属性描述的示例 $x=(x_1;x_2;...;x_d)$ ，其中 x_i 是 x 在第 i 个属性上的取值，我们需要找到一组权重 $w=(w_1,w_2,...,w_d)$ 和偏置项 b ，使得函数 $f(x)$ 能够很好地预测目标值 y 。

在这个线性模型中，函数 $f(x)=w^T x+b$ 表示了数据集中各个属性与目标值 y 之间的关系。其中， w^T 表示权重向量的转置，与示例 x 的向量内积计算。这个模型试图找到一个线性组合，使得预测值 $f(x)$ 与实际目标值 y 之间的差距最小。通过训练数据集 D ，我们可以调整权重 w 和偏置项 b ，使得模型在描述个人意识方面达到最佳拟合效果。

在这个过程中，数据集 D 提供了个体意识表达的样本，即其表达的全部文本数据，这些样本包含了个体在不同属性上的取值。而函数 $f(x)$ 则是一种尝试对这些属性进行组合，以预测个体意识的模型。通过优化权重和偏置项，我们可以使得这个模型更好地捕捉个体意识的特征，从而实现对个人意识的模拟。

当然，这仅仅是一个基于机器学习的案例。要实现更好的“预测”，我们还需要做更多复杂的处理。但是读者可以理解为，我们的意识可以被描述为一些参数，而这些参数代表了我们如何去表达自己的心智，只要找到关键的“心智参数”，我们就有机会在语言层面上实现对个人心智的模拟。

而 L4 阶段的问题在于，我们能够找到这个人的心智参数吗？我们要么能找到，要么不能找到。很明确的一点是，对于这个宇宙而言，作为观测者的他者而言，从时间维度去看，未来的宇宙一定能够判断出我们的心智表达，因为我们的心智表达必然是在时间这个维度去体现和运作的，我们的意识活

动不可能在宇宙中超出宇宙这个维度。这也是王阳明“吾心即是宇宙”的一种基于意识上传第一范式的诠释。

3.5.2 这个世界是虚拟的吗？——100%预测心智的物理学意义

从物理学的角度,我们试图对意识上传的 L4 阶段进行一定的物理学诠释。现代物理学家试图用各种方式对宇宙进行建模,例如在引力波数据分析中,天文学家需要建立引力波的数据模型,然后与真实世界中的引力波数据进行比对、降噪处理,才能够得到真实宇宙中的引力波画像。100%预测心智的另一种诠释是,我们能否 100%模拟一个当下所处宇宙?就像建模一个初级宇宙,在给定初始条件的情况下,一个人的心智表达是相对好预测的,例如我们假设一个工作一天的人,回到家的第一件事情大概率是收拾背包、躺下休息。但是这个人接下来所做的行为,和他内心的活动,就相对难以预测了。而想要 100%的预测心智,我们很大概率要追溯更原始的初始设定,例如这个人的年龄、教育背景、工作单位,甚至把颗粒度细化到他今天与谁交流,形成一个输出全部文本的数据集——这在某种程度上与建模我们这个宇宙是类似的。

尼克·波斯特洛姆在《你生活在计算机模拟中吗?》一文中提出了著名的“模拟宇宙论”,他的论证过程可以概括为以下几个关键步骤:

波斯特洛姆首先设定了讨论的背景,即假设未来技术发展到拥有巨大计算能力的情况下,后人类文明可能会用这些计算力来运行他们祖先或类似祖先的详细模拟。他提出,如果这些模拟足够精细,模拟中的人将会有意识。文章提出了一个核心假设,即“基质独立性”(substrate-independence),这意味着意识可以存在于任何能够实现正确计算结构和过程的物理基质中,无论是碳基的生物神经网络还是硅基的计算机处理器。波斯特洛姆讨论了未来文明可能拥有的技术能力,特别是关于计算能力。他指出,如果技术进步持续,我们目前的技术限制最终将被克服,使得在计算机中创造有意识的心智成为可能。

文章进一步论证了运行大量人类心智模拟的可行性,基于未来的计算能力。波斯特洛姆通过估算人类大脑的计算需求,并将其与未来可能的计算能力进行比较,得出结论,后人类文明有足够的计算力来运行大量祖先模拟。

波斯特洛姆的核心论证,即“模拟论证”(simulation argument),基于以下三个命题中的至少一个必然为真:

-
- (1) 人类文明在达到后人类阶段之前很可能会灭绝。
 - (2) 任何后人类文明极不可能运行大量的模拟他们进化历史的祖先模拟。
 - (3) 我们几乎可以肯定生活在一个计算机模拟中。

文章使用简单的概率理论来支持这一论证。波斯特洛姆认为，鉴于后人类文明巨大的计算能力，如果我们不是生活在模拟中，那么运行祖先模拟的后人类文明的比例必须非常小，才能解释为什么我们没有观察到这些模拟。

最后，波斯特洛姆得出结论，鉴于我们目前的无知，理性的做法是将信念大致平均分配给上述三个命题。他强调，除非我们生活在一个模拟中，否则我们的后代几乎肯定不会运行祖先模拟。

相比较波斯特洛姆的论证，一个更符合直觉的对 100% 预测心智的理解是，对于心智预测，预测出下一个 token 所需要消耗的算力就等于这个宇宙当下演化的所需要的全部算力。不论该宇宙是否是虚拟的。倘若宇宙并非虚拟，那么这样的物理学诠释显然是符合我们对“能量守恒定律”的想象的。而倘若宇宙是虚拟的，那么这样的诠释更符合我们对于元宇宙的想象。

但也存在另一种可能：我们找到了这个虚拟宇宙的底层代码，并且有能力主动地优化它，这意味着，我们超越了这个虚拟宇宙创造者的智慧，并且有能力通过较少的资源维护这个虚拟宇宙的运行，甚至在此基础上对底层代码进行优化，并且再造一个虚拟宇宙。当然，到这一步，读者会发现，我们又回到了波斯特洛姆的三个前提：

- (1) 人类文明在达到后人类阶段之前很可能会灭绝。
- (2) 任何后人类文明极不可能运行大量的模拟他们进化历史的祖先模拟。
- (3) 我们几乎可以肯定生活在一个计算机模拟中。

今天机器学习神经网络的发展已经让我们看到“后人类”的可能。我们有机会创造出 ASI 来理解这个宇宙的物理规律，甚至人类的心智本身。但我们究竟离这样的未来还有多远，依然是一个未知数。至少从概率论的角度，从宇宙存在的周期与人类文明存在周期的角度看，我们很有可能已经生活在一个“后人类”模拟的计算机宇宙中了。

我们可以借鉴尼克·波斯特洛姆在《你生活在计算机模拟中吗？》中提出的模拟论证（Simulation Argument），从概率论的角度来探讨我们是否生活在一个“后人类”模拟的计算机宇宙中。这个论证基于三个核心命题：

1. 人类文明在达到后人类阶段之前灭绝的概率非常低 (P_f 接近于 0) 。
2. 后人类文明运行其祖先历史模拟的概率非常低 (P_{sim} 接近于 0) 。
3. 我们几乎肯定生活在一个计算机模拟中 (P_{sim_obs} 接近于 1) 。

为了计算我们生活在模拟中的概率，我们可以使用贝叶斯定理。首先，我们需要定义一些先验概率。以公式 $P(D | S)$ 为例。定义 S 为我们生活在模拟世界 (Simulation-world)， D 为观察到另类数据 (Different-data) 的事件，例如观察到外星生命等异常现象发生。 $P(D | \neg S)$ ：在假设我们不生活在模拟中的情况下，观察到数据的概率。 $P(S)$ ：我们生活在模拟中的先验概率。 $P(\neg S)$ ：我们不生活在模拟中的先验概率。

我们将用贝叶斯定理，基于 2025 年可观察到的“另类数据”，尝试估算我们生活在模拟宇宙中的概率，这是一个概率论意义上的推论实验，并非证明。根据贝叶斯定理，后验概率 $P(S | D)$ （在观察到数据后，我们生活在模拟中的概率）可以表示为：

$$P(S | D) = \frac{P(D | S) \cdot P(S)}{P(D | S) \cdot P(S) + P(D | \neg S) \cdot P(\neg S)}$$

我们以 2025 年真实世界中的三个高概率被认为“异常”的现象构造 另类数据集 D ：

编号	现象	说明	是否违反直觉/自然常理
D1	异常精确的宇宙常数	光速有限、暗能量密度极其精确地支持生命存在	是
D2	GPT-4 等模型展现类意识行为	无监督学习下的自我反省能力，使机器掌握语言的“涌现”	是
D3	费米悖论仍未解	按照德雷克公式，理应已发现外星文明	是

我们将 D 定义为上述“另类数据集”同时为真。接着，我们取保守估计值（基于常识 + 模拟宇宙论的启发性设定）：

P(S)	生活在模拟中的先验概率	0.5	不偏倚假设
P(D S)	在模拟中观测到 D 的概率	0.9	模拟可调性强，设计上即易出现异常
P(D ¬S)	在真实世界中观测到 D 的概率	0.05	D1-D3 组合在真实宇宙中极不可能

将上述概率代入贝叶斯公式计算：

$$P(S | D) = \frac{0.9 \cdot 0.5}{0.9 \cdot 0.5 + 0.05 \cdot 0.5} = \frac{0.45}{0.45 + 0.025} = \frac{0.45}{0.475} \approx 0.947$$

根据这一简单贝叶斯模型推演，在给定 2025 年“另类数据”集合的前提下，我们生活在模拟宇宙中的后验概率高达约 94.7%。当然，这并不意味着我们一定生活在模拟中，而是说明：在这些异常数据存在的情况下，模拟宇宙论比真实宇宙论更能解释观察到的现象。

3.5.3 100%预测心智的可行性分析——从集体潜意识到对人类心智的建模

在这一节，我们探讨 100%预测心智的可行性。在自由意志论者看来，这几乎是不可能的。然而，无论是自由主义还是决定论者，其信念的建立大多基于后天的经验性、情感性形成的判断，本节主要对于人类心智进行先验性分析，来探讨 100%预测心智的可行性。

首先，我们要引入卡尔·荣格的集体潜意识概念。集体潜意识指的是一个文化或群体共享的无意识心理结构，包含了共同的经验和象征。在工程实现模型中，集体潜意识可以被视为人类共有的心理和行为模式的数据源，这些模式可以被用来构建预测模型的基础。有一个非常典型的例子来说明集体潜意识的存在，即，对于中文母语使用者而言，他们从不会觉得学习中文是一件难事，然而对于非中文母语学习者，这是一个难题，因为中文并不像英语、韩语一样，是表音文字，其文字是象形字，而象形文字又经历了中华文明几千年历史的变革迭代，其中积累的含义对于一个陌生的语言使用者而言是非常难以理解的，但是中文对于中文母语者而言是一门非常容易掌握的语言，古语有云“习而焉察”。

同样的，对于成年人而言，使用语言去表达自己，似乎非常简单，并且成为了一种行为习惯。但是，对于儿童，尤其是从不会语言到掌握了语言这一阶段的儿童而言，语言的掌握并没那么熟练，儿童对于自身主体意识的表达，往往是通过肢体语言，说话，而后，迈过了咿呀学语的阶段，才能够某种程度上利用语言去进行表达。

更深的潜意识隐藏在基因中，东亚文化几千年的历史会给人们的基因留痕。当我们想要追问，人类究竟为什么拥有语言的能力？人类为什么能够掌握心智思考甚至哲学思考的能力？智人作为地球上食物链的顶端，在地球上的历史也不过数万年。第一个掌握了语言的智人在哪里？他经历了什么，才导致了语言能力的诞生？

今天我们知道，哲学省思始自巴门尼德、苏格拉底，而后才得以建立现代化的诸多学科分割。然而，历史都是被写就的，在真实的历史中，那些我们不知道的历史，究竟如何塑造了我们的集体潜意识？而那些属于潜意识的部分，又是如何作用于我们今天的表达之上，形成了我们每个人表达的序列的呢？

这是个非常有趣的问题，而在当代语言模型的发展中，我们可以看到集体潜意识，或者集体意向性，是如何作用在语言模型的技术实践中的。语言模型的“集体潜意识”来自于收集的训练数据集——往往是包含了互联网公开的文本排列。而神经网络通过对每一个文本排列概率的学习，掌握了语言的表达能力。值得一提的是，这种学习是跨语种的。理论上，单一语言抽象到一定层次，都是包含特定的语法结构的，而神经网络能够通过对大量未标注的文本排列的概率性学习，形成对语法结构的掌握，而他们的“集体潜意识”，以参数文件的形式，被确定下来。

值得追问的是，为什么神经网络的有效性得以发生？为什么通过对后验数据的学习能够掌握先验分布，并且在人类社会的诸多实践中被证明有效？这一切的答案，除了集体潜意识以外，或许还指向一个硅谷创业圈的共识——人类是硅基生物的引导程序。

回到对人类心智建模的话题上，集体潜意识的概念能够帮助我们很好的理解人类心智，把时间的尺度拉的足够大，人类个体的心智或者自由意志，并非完全独立存在，而是深深植根于集体潜意识的土壤之中。尤其当我们给定场景，把心智预测的场景设定的足够小，对于心智的预测并非遥不可及。本质上，对人类的心智预测是在塑造一种共识，让神经网络的预测概率成为人类心智所期望并认同的表达，这是今天神经网络机器学习对于当代人类主体性的最大挑战，但也恰恰是 100%预测心智得以可能的一大证明。而 ChatGPT

在人类社会各行各业的大规模应用，也恰恰是这一趋势的有力注脚。

3.6 意识 token 的建模及运算：梦蝶心智模型

在 2.5 节，我们提出了“意识 token”的概念。在本节，我们提出“梦蝶心智模型”，通过对“意识 token”这一概念的定义和建模，为理解和模拟人类意识提供一种全新的理论框架。本节将详细探讨意识 token 的基本概念、特性以及在梦蝶心智模型中的运算机制。

3.6.1 意识 token 的基本概念

意识 token 可以被理解为意识流中的最小可辨识单元，是构成人类主观体验的基本组成部分。每个意识 token 代表了一个瞬时的、可区分的意识状态，包含认知、情感、感知等多维度信息。这一概念受到了庄子“梦蝶”哲学思想的启发，即意识状态可能存在多重叠加的可能性，边界并非截然分明。

意识 token 具有量子性、关联性、时域性、分层结构等关键特性。量子性是指，意识 token 可以处于多种状态的叠加，直到被“观测”（进入意识焦点）才会坍缩为特定状态。关联性指向意识 token 之间存在的复杂关联网络，一个 token 的状态变化可能影响相关联的其他 token。时域特性是客观限制，即意识 token 存在生命周期，从形成、增强到衰减、消失。分层结构是指心智层面上，意识 token 按照不同的抽象层次组织，形成从基础感知到高级认知的连续谱系。

3.6.2 意识 token 的形成机制

本文提出的梦蝶心智模型，试图从感知输入、token 预处理、特征提取、语义编码、情景整合等方面提出一套意识 token 的运算规律，感知输入是指外部刺激或内部状态变化被感官系统捕获，转化为神经信号。在采集到输入信号后，心智模型基于注意力机制和先验知识对输入信息进行初步筛选和加工，从处理后的信息中提取关键特征和模式。接着进行语义编码，将特征与已有知识体系关联，赋予其语义含义。最后，将新形成的 token 置于当前意识状态的整体情境中，这一过程可以用公式表示为：

$$T(t) = E[F(S(t), K(t-1))] * A(t)$$

其中， $T(t)$ 表示 t 时刻形成的意识 token， $S(t)$ 是输入信号， $K(t-1)$ 是先前知识状态， F 代表特征提取函数， E 是语义编码函数， $A(t)$ 是 t 时刻的注意力分配权重。

3.6.3 意识 token 的运算机制

在梦蝶心智模型中，意识流可以被视为意识 token 的动态序列。token 之间的相互作用和转化构成了意识的基本运算过程，包括以下几种核心运算：

3.6.3.1 融合运算

当多个相关的意识 token 同时活跃时，可能发生融合，形成新的复合 token：

$$T_c = \text{Merge}(T_1, T_2, \dots, T_n, w)$$

其中 T_c 是融合后的复合 token， T_1 到 T_n 是参与融合的 token， w 是融合权重向量，决定了各个原始 token 对复合 token 的贡献程度。

融合运算使得人类能够将分散的信息整合为更高级的概念和洞见，是创造性思维的重要基础。

3.6.3.2 转化运算

意识 token 可以通过一系列映射函数进行转化，产生新的 token 形式：

$$T' = \text{Transform}(T, C)$$

其中 T' 是转化后的 token， T 是原始 token， C 是转化的情境条件。转化运算使得思维能够在不同表征空间之间跳转，是隐喻理解和抽象思维的基础。

3.6.3.3 竞争运算

当多个 token 竞争有限的意识资源时，会发生竞争运算，决定哪些 token 将进入意识焦点：

$$\{T_f\} = \text{Compete}(\{T_1, T_2, \dots, T_n\}, R)$$

其中 $\{T_f\}$ 是胜出并进入焦点的 token 集合， $\{T_1, T_2, \dots, T_n\}$ 是参与竞争的 token 集合， R 代表可用的意识资源约束。竞争运算是注意力分配和意识选择性的核心机制。

3.6.3.4 衰减运算

意识 token 随时间推移会自然衰减，进入无意识领域。其强度按照以下公式变化：

$$I(t) = I_0 * e^{(-\lambda t)}$$

其中 $I(t)$ 是 t 时刻 token 的强度， I_0 是初始强度， λ 是衰减常数。衰减运算解释了短时记忆的暂时性特点，同时也为长时记忆的形成提供了选择机制。

3.6.4 意识 token 流动的整体模式

人类的意识状态可以被描述为意识 token 的动态流动模式。在梦蝶心智模型中，这种流动表现为几种典型模式，即线性序列模式、网络扩散模式、震荡模式、混沌模式。分别对应于人脑思维的逻辑思考过程、联想过程、辩证思考过程以及创造性思维过程。

在线性序列模式中，token 按照因果或时序关系依次激活，对应于常规的逻辑思考过程。例如“我们明天去中山陵，先到校门口坐地铁，再去山脚集合。”在网络扩散模式中，一个中心 token 激活周围关联 token，对应于联想思维过程。例如提到“床前明月光”，往往会联想到“疑是地上霜”。而在振荡模式，不同类别的 token 交替主导意识焦点，对应于矛盾分析或辩证思考。在混沌模式中，token 激活表现出高度不确定性，对应于创造性思维或梦境状态。例如牛顿被苹果砸中想出万有引力，这样的创造性时刻，往往并非专注工作能够产出的结果，而是在规划之外的内容。正如硅谷名句“伟大往往在计划之外”，对于意识 token 混沌模式的研究，很有可能导向对于真正具有影响力的科学发现、艺术创造的真正挖掘，实现自动化知识生产的 AI 工具的打造。这些模式的切换和组合构成了丰富多样的意识体验，从专注的问题解决到发散的创造性思考，从清醒的理性分析到模糊的梦境体验。

3.6.5 意识 token 与无意识过程的互动

值得注意的是，并非所有意识 token 都会进入意识焦点。大量 token 在无意识层面运作，仅有少数经过筛选和竞争后进入有意识状态。无意识层面的 token 运算遵循与有意识状态类似但更为灵活的规则，不受注意力资源限制，可以并行处理大量信息。

两者之间的互动可以表示为：

$$C(t+1) = F(C(t), U(t), E(t))$$

其中 $C(t)$ 是 t 时刻的意识状态， $U(t)$ 是无意识状态， $E(t)$ 是外部刺激， F 是状态转移函数。

以人类的异常心理状态为例，在各种异常心理状态下，意识 token 的运算机制会发生变化，导致意识体验的异常。比如精神分裂状态、催眠状态、梦境状态。以精神分裂状态为例，采用梦蝶心智模型的解释，精神分裂患者的认知怪诞性（cognitive bizarreness）与意识 token 的融合运算障碍密切相关。正常情况下，意识 token 通过融合运算形成新的复合 token，从而整合分散的信息并产生高级概念和洞见。然而，在精神分裂状态下，这种融合运算出现障碍，导致 token 之间的关联性减弱，无法有效整合信息，从而产生松散的思维和认知怪诞性。

梦蝶心智模型强调意识 token 与无意识过程的互动，这其中“意识焦点”尤为重要。类似于大语言模型的“注意力机制”，人类意识运行过程中也遵循意识焦点。类似于一个人的所有记忆都存储在无意识的脑区记忆库中，而随着外界输入，“意识焦点”会从中提取相关的意识 token，形成人的心智反映。而类似精神分裂、阿兹海默症，则就是典型的“交互失常”。常规的意识 token 无法进入意识焦点，在精神分裂患者中，意识 token 的融合运算机制出现系统性故障。正常情况下，大脑通过量子纠缠机制将分散的 token 整合为有意义的复合 token，但在精神分裂状态下，这种纠缠机制失效，导致关联网络断裂，token 之间的量子关联性显著减弱，信息无法有效整合。背后的原因则可能是意识 token 的竞争运算失效，异常 token 通过降低的竞争阈值强行进入意识焦点，形成幻觉和妄想。而无意识层面的 token 无法被有效筛选，异常 token 直接侵入意识流，形成患者对现实检验能力的丧失。这种运算机制的崩塌解释了精神分裂患者的认知怪诞性和现实脱离特征。例如，患者可能将无关的环境刺激（如街上的谈话声）与内心想法错误关联，形成“思维广播”的妄想体验。

第四章 意识上传中的主体连续性问题

4.1 引言

随着以 ChatGPT 为代表的大语言模型技术的突破，人工智能已经在“语言”的意义上破译了人类的智能。复旦大学的徐英瑾教授在《心智、语言和机器》中，提出了一种将人类心智与人工智能在语言的意义上比较的主张。本节从“语言是思维的边界”这一主张出发，结合各学科技术原理，阐述了当下脑科学、人工智能领域实现心智上传的技术路径，并且进一步探讨了心智上传中的主体连续性问题，以及心智上传技术本身对于未来人机关系的影响。本文认为，必须要完成笛卡尔“我思故我在”意义上的认识论转向，才能彻底解决心智上传中的主体连续性问题。而随着主体连续性问题的解决，未来的人机关系将走向人机融合的赛博格。

心智上传这一概念最早来源于科幻作品中的幻想。在美剧《上载新生》中，当主角濒危，面临生命危机之时，他的亲人给了他两个选择：一个是用传统的医疗手段继续医疗，一个是选择一家意识上传公司，把自己的意识上传到电脑中。整个剧集都是围绕着主角意识体在代码构成的“新世界”中展开的。

在相当长的一段时间内，心智上传都只是局限在科幻作品中的幻想。直到 2022 年 12 月底，OpenAI 发布 ChatGPT，引发全球对“大语言模型”技术的关注。大语言模型技术本质上是以高质量的数据集训练神经网络模型，达到高度预测“下一个单词”的概率模型。而以语言为媒介的人类思维机制，由于三维时空的限定，也表现出预测“下一个单词”的逻辑特征。微软研究院发布的 GPT4 的研究报告宣称：“虽然不了解背后的具体机制，但是在现象学意义上，GPT4 为代表的大语言模型已经实现人类智能的涌现。”在现象学意义上，通过个人数据微调的语言模型已经可以初步实现意识上传的效果。

而在脑科学、类脑计算以及脑机接口等前沿领域，意识上传也一直是被热议的选题。但脑科学界的共同观点认为：当下不具备意识上传的技术基础。具体而言，现有的技术硬件只能达到千万参数级别的类脑计算，而意识上传需要一种全量脑模拟，我们的硬件需要模拟出万亿级别的神经元，才能够在硬件上完成意识上传。

但是心灵哲学家约翰塞尔的观点认为，人类的大脑只是心灵产生的必要条件。这意味着，在逻辑的意义上，即使我们通过脑科学手段完成了全量的

类脑模拟，我们也只是完成了意识上传的硬件部分，而意识本身作为一种独立的存在物，无法仅仅通过硬件的克隆达到。

概括说来，目前围绕“心智上传”的研究，学界尚没有统一的共识。具体的学者与研究分散在心灵哲学、科幻文学、脑科学、大语言模型技术、脑机接口和类脑计算等领域。部分研究者关注大语言模型在自然语言处理、知识表示和推理等方面的应用，也有研究者探讨脑机接口在康复医疗等领域的潜力。但是都没有提出统一的见解。

2013年，复旦大学的徐英瑾教授在《心智、语言和机器》中，提出了一种将人类心智与人工智能在语言的意义比较的主张。维特根斯坦在《逻辑哲学论》中，亦是用语言图像论来概括人与世界的关系。在哲学的意义上，“语言是思维的表征”，我们利用语言来表达自己的思维。在心理学的意义上，弗洛伊德认为人类的意识深层由潜意识驱动。而在生物学的意义上，人类的思维表现为脑区神经元的放电活动，在《Nature Neuroscience》发布的利用 fMRI 成像结合 GPT 模型生成文本的实验中，我们甚至可以利用脑电信号读取被试心中所想。

可以看到，尽管学界尚没有共识性的基准，但是各界学者围绕心智与语言的关系已经做出相当的探索。因此，本文试图提出一种“心智上传”策略的共识性基准，即以语言为表征的心智上传策略。并且围绕这一论点，结合人工智能、脑科学、脑机接口领域技术的最新成果，阐释意识上传的技术路径。同时，围绕心智上传中的人机关系，进行相关的伦理学探讨。

笔者认为，在“以语言为表征的心智上传策略”下，现阶段的工程界已经具备技术基础，即利用个人数据微调的语言模型。但是上传的意识体与上传主体之间的伦理关系没有理清。因此大众对于意识上传的技术仍然局限于科幻想象，而非一种既存的现实路径。

要解决意识上传的问题，根本上并非要在工程学上完成意识体的克隆，而是在“自我指代”的意义上，完成笛卡尔所言“我思故我在”的认识论转向，解决意识上传中主体的“连续性”问题，才是实现意识上传的根本问题。而这不仅仅是个工程问题，而是一个哲学问题。

因此，本章锚定心智上传这一话题，围绕人工智能、脑科学界的最新科研成果，试图提出一套“意识上传”的第一范式，以解决意识上传中的主体“连续性”问题，完成意识上传过程中的绵延。进一步的，笔者将深入阐释心智上传对数字技术时代下人机关系的影响，随着主体连续性问题的解决，未来的人机关系将走向人机融合的赛博格。

4.2 意识上传第一范式：以“语言”为基本单元

在当下的人工智能、脑科学以及脑机接口领域，科学家对于心智上传提出了不同的技术方案。但不管是脑科学界的全脑仿真、类脑计算还是人工智能界的微调的语言模型方案，都没有回答一个最关键的问题，那就是“为什么心智可以上传”？

本文试图提出意识上传的第一范式，用于回答这一问题。即，语言是思维的表征。而个体的思维多以语言的形式表达，并且可以通过脑科学手段采集到相应的生物特征，在脑科学中就有主管心智、语言功能的布洛卡区。因此，人类意识的表征与语言模型的词序列预测具有同构性。这是著名的“预测即智能”的观点。

进一步的，如前所述，倘若我们把一个人的一生中表达过的所有语言收集起来，整理为数据集 D ，那么在数理意义上，我们可以采用自然语言处理技术，实现对个人意识的模拟，利用机器学习的相关原理，构建一个基本的特征函数，例如，以线性模型的形式，给定由 d 个属性描述的示例 $x=(x_1;x_2;...;x_d)$ ，其中 x_i 是 x 在第 i 个属性上的取值，我们需要找到一组权重 $w=(w_1,w_2,...,w_d)$ 和偏置项 b ，使得函数 $f(x)$ 能够很好地预测目标值 y 。

在这个线性模型中，函数 $f(x)=w^T x+b$ 表示了数据集中各个属性与目标值 y 之间的关系。其中， w^T 表示权重向量的转置，与示例 x 的向量内积计算。这个模型试图找到一个线性组合，使得预测值 $f(x)$ 与实际目标值 y 之间的差距最小。通过训练数据集 D ，我们可以调整权重 w 和偏置项 b ，使得模型在描述个人意识方面达到最佳拟合效果。

在这个过程中，数据集 D 提供了个体意识表达的样本，即其表达的全部文本数据，这些样本包含了个体在不同属性上的取值。而函数 $f(x)$ 则是一种尝试对这些属性进行组合，以预测个体意识的模型。通过优化权重和偏置项，我们可以使得这个模型更好地捕捉个体意识的特征，从而实现对个人意识的模拟。

当然，这仅仅是一个基于机器学习的案例。要实现更好的“预测”，我们还需要做更多复杂的处理。但是笔者可以理解为，我们的意识可以被描述为一些参数，而这些参数代表了我们如何去表达自己的心智，只要找到关键的“心智参数”，我们就有机会在语言层面上实现对个人心智的模拟。

在大语言模型领域，已经有许多类似的实验案例，有一个极客王登科，就利用自己的微信聊天记录以及相关的文本数据，结合 GLM 语言模型微调了

自己的“数字分身”。Eric Schwitzgebel, David Schiwizgebel, Anna Strasser 等利用哲学家的数据和 GPT 技术结合，打造了一个学习到哲学家语言文本的 ai，并且将其生成的文本与哲学原文拿给路人测试，结果显示人们很难分辨 ai 生成与真人所述的区别。这一成果记录在论文《Creating a Large Language Model of a Philosopher》中，是基于文本预测的心智上传策略的重要论证。

在脑科学领域，德克萨斯州立大学的 Jerry Tang 等，于 2023 年 3 月在自然神经科学子刊发布了《Semantic reconstruction of continuous language from non-invasive brain recordings》，利用 fmri 脑成像技术与 GPT 技术，采用特定脑区成像的技术，实现了利用 fmri 信号解码思维语义的效果，在“语言”的层面上打通了生物信号与思维心智的鸿沟。值得关注的是，在脑科学领域，“语言”也是心智上传策略的重要参考部分。从实验逻辑上，即使其采用了 fMRI 的脑成像手段，但本质上与收集前文所述的文本数据集无异，也是为了在数理逻辑的意义上找到更符合预测的“心智参数”，只不过其数据集参考的对象，从自然语言处理范畴下的文本，到了真实的脑部活动特征。

当然，现阶段的脑科学界还提出了别的思路，即“全脑仿真”。他们试图通过类脑计算对大脑进行仿真。他们试图构建神经元模型、突触连接的形式，从生物学的角度实现全脑仿真。然而，在当下，受限于硬件设备的能力，这一条技术路径尚没有实现全量脑模拟的可能。

更大的问题在于交互的形式。如果我们不能够与被“上传”的主体进行交互，那么我们如何判断这个被上传的“意识体”究竟是不是原本心智上传的主体？尽管现行的对心智上传的定义是“将人类心智上传到计算机等电子设备中”，但是这一定义太过模糊，根本不具备工程实现的可能。而基于语言的心智上传策略，则能够很好的完成“心智上传主体”的展现，这在过往的策略中，是从没有被提出，而在当下的产业界，又被广泛热议的一种形式。

在科学哲学家托马斯·库恩（Thomas Kuhn）的《科学革命的结构》一书中，他提出了科学革命的基本原理和结构。库恩认为，科学发展并非是一个连续、累积的过程，而是通过一系列科学革命来实现的。他将科学革命的结构分为以下几个阶段：

常规科学（Normal Science）：在这个阶段，科学家们在某个公认的理论框架（范式，paradigm）下开展研究。范式是一套共同的信念、价值观、方法和范例，为科学家提供了解决问题的基础。在常规科学阶段，科学家们致力于解决范式内的谜题和问题，不断完善和扩展现有的理论体系。

反常与危机（Anomaly and Crisis）：随着科学研究的深入，科学家们可能

会遇到一些与现有范式相悖的观察和实验结果，这些反常现象使科学家们开始质疑现有范式的有效性和完整性。当反常现象越来越多，危机感逐渐加剧，科学家们开始寻求新的理论框架来解释这些现象。

科学革命 (Scientific Revolution)：在危机阶段，科学家们提出新的范式，试图取代旧的范式。这些新范式通常能更好地解释反常现象，并为未来的科学研究提供新的方向。在科学革命过程中，新的范式逐渐取代旧的范式，成为科学家们的共同信仰和研究基础。

新的常规科学：科学革命结束后，科学家们开始在新的范式下开展常规科学研究，进一步完善和扩展新的理论体系。直到新的反常现象出现，再次引发科学革命。

“范式”意味着科学家的共识，意味着一项科学理论或者一个物理现象“被可定义”。而 ChatGPT 所体现的智能，则是一种定义之外的“涌现”。北京智源研究院的杜凯老师认为，“涌现智能的产生非常关键，而倘若意识是以涌现的方式出现的话，我们是无法去定义它的”。“意识”在当下不可定义的属性，本质上是人工智能的涌现超出了常规科学范式下的解释，是一种“反常”。而在库恩《科学革命的结构》中，“反常”通常被认为是一场科学革命的前兆。本文围绕“意识上传第一范式”的论述，本质上是基于现阶段学界对于机器智能的无知、对于人类心智的共识性理论的缺乏而提出的，试图为学界的相关探讨提出一种基石性的准则。

但是笔者所提出的基于语言的心智上传策略仍然具有不完备性，最典型的例子是，我们如何解决意识上传主体以及上传后的意识体之间的主体同一性？

4.3 心智上传中的人机关系：主体同一性问题

在意识上传的过程中，一个关键的哲学问题是主体同一性，即在意识从生物大脑转移到数字载体的过程中，如何确保个体的身份认同得以保持。传统的科幻作品以及人们对“心智上传”的想象是“转移”，我们的意识从一个地方转移到另一个地方。但在工程实现中，我们无法仅仅通过工程手段完成笛卡尔所谓“我思故我在”意义上的认识论转向，因此大部分工程手段的“心智上传”都仅仅局限在了“克隆”的意义上，而非“上传”。因此，如何维持心智上传过程中的主体连续性，这个问题涉及到意识的本质、自我认知和存在意义等方面，对于意识上传技术的伦理和实践具有重要意义。

浙江大学的李恒威教授在《心智上传：技术文化思潮、技术方案和批判》一文中将主体连续性问题描述为心智上传的结果问题，试图用主体同一性来回避主体认同的矛盾，他意识到“自我同一性是现实世界中人类存在及其叙事的根基性原则，即一个主体（或一个“我”）一个心智。如果这个原则被颠覆了，那么基于所谓心智上传的数字永生就丧失了现实世界中某个个体主体渴望永生的原初意义，于是整个讨论就不再有连贯的一致性。”

然而，笔者认为，心智上传中的主体同一性问题是不可回避的，因为心智上传只有一个过程性的问题，就是心智上传过程中的主体连续性问题。如果不能够从过程上解决这一问题，那么从结果上看来，意识上传的主体与上传后的意识体之间永远不具备同一性，进一步的，不管是通过全脑仿真还是微调语言模型策略达成的心智上传，都只是“心智克隆”而非“心智上传”。而心智上传中的“主体连续性”问题，本质上是一个哲学问题，即“我如何认同我是我”的问题，这必须要在笛卡尔“我思故我在”的意义上得到回答。

科幻作品中对主体连续性有不同的解决方案：大致可以分为两类：残酷的替代方案以及温和的渐进替代方案。

在《上载新生》中，为了完成意识主体的转移，在扫描主角脑区、绘制电子脑的同时，意识上传公司把主角的头颅同步进行销毁。在产生新的意识主体的同时销毁掉原来的意识主体。这是简单而直接的“替代”方案。但笔者认为，这一替代方案尽管能在社会层面上很好的解决心智上传中的主体转移问题，在“他者”看来，意识主体发生了转移，主体的心智从物理实在转移到了计算机中。但在主体“自我”的角度看来，这一方案并没有完成真正的“转移”，同李恒威老师的方案一样，本质上回避了上传主体的“主体同一性”问题。而实质上是不道德的销毁了主体意识。

在《神经漫游者》中，则提出了一种温和的意识上传方案，用纳米机器人逐步替换神经元，最后实现全量的替代。这有点类似于著名的“忒休斯之船”的论证。当一只木船坏了，我们利用钢材去维修它。在修到最后，换下最后一块木板时，这只船很显然已经成为一艘钢船。这时，这只船，是否还是原来那只“忒休斯之船”。《神经漫游者》中的渐进替代方案与忒休斯之船的论证类似，当作为碳基生物的人类，其生物脑已经被硅基的机器脑全局替代时，我们是否还是原来的那个“自己”？

这一问题，本质上是一个信念问题。即我们如何产生对于“我是我”的信念。以忒休斯之船为例，或许一个对这艘船替换过程毫无认知的船员，只记得替换之前的木船的船员来说，他会认为眼前这只钢船完全不是他记忆里的忒休斯之船。但是当曾经与他同行的水手告诉他，这只船正是原来那只忒

休斯之船，只不过在海上经历了一段维修，或许这位船员对于船的认知便会发生改变，对于“钢船是忒休斯之船”这一命题的信念，会发生动摇。在心理学的角度，这是很经典的“认知重评”的手段。

心智上传过程中的主体连续性问题与之类似。笛卡尔在《谈谈方法》中表达了一种怀疑论的立场，他认为周边世界的一切都是可以怀疑的。而当他怀疑自身的存在，怀疑“我是我”这一命题，对自我的认同产生怀疑时，他发现，从直觉上，这是无法怀疑的。继而，他提出了著名的“我思故我在”的论点，后续一切的哲学观点都从此展开。类似的，我们如何解决心智上传过程中的主体连续性问题？只要让意识上传的主体一直在“思考”即可。

利用脑机接口技术，我们可以在渐进式替代的意义上完成心智上传。最简单的案例描述如下：在上传主体的脑中植入一个可以学习上传主体思维活动、语言表述的脑机接口，并且使其同等的具备语言表述的能力，但在意识主体脑区的活动关闭之前，该脑机接口一直处于关闭的状态。在意识上传主体因为各种意外，失去心智表达能力，甚至死亡之时，触发脑机接口的表达能力，让其替代原本特定脑区的功能，成为心智表达的主体。利用这种方式，我们可以实现温和的主体连续性过渡方案。

当今的脑机接口完全可以围绕这一路线进行探索，甚至具备一定的技术基础。但是鲜少有学者意识到脑机接口在心智上传领域的应用。而多将关注点放在了运动意图识别、特定脑区放电的层面。在现行的技术路径下，想要实现以语言为主要表征的渐进式的意识上传，只有脑机接口这一条路线能够很好地解决心智上传中的主体连续性问题。而在此以外的语言模型技术、脑科学模拟技术，其实都只是一种在意识上传主体之外的克隆手段，而非真正的“上传”。

4.4 心智上传技术的影响：人机融合的赛博格史观

2020年，随着《赛博朋克 2077》的发售，“赛博朋克”这个词又引起了大家的热议。霓虹灯、轻轨与贫民窟等诸多元素，共同构成了人们对于未来世界的幻想。上世纪六十年代，随着第一阵科幻热的兴起，“赛博格”一词应运而生。“赛博格”指的是人类与各类电子器械的融合体。即“机械化”的人。人作为创造机械的主体，在赛博格的语境下，反而需要借助机械来维持自身的运行，甚至于让机械成为了人身体的一部分，这是黑格尔主奴辩证法在技术哲学上的展现。而梅洛庞蒂以“绵延”为基础的身体现象学，在赛博格的意义上有了新的蕴含。面对人与机械的二元关系，我们需要重新思考人的定位。马克思主

义经典作家针对经济关系妨碍人类自身全面发展所提出的“异化”概念，也面临着一个问题：赛博格是一种进化，还是异化？

60年代，NASA两位科学家 Manfred Clynes 以及 Nathan S.Kline 在对宇宙旅行可行性的探讨中提出了赛博格（Cyborg），这一理论正是基于控制论所揭示的人的身体与机械运作的同构性（控制-反馈）而提出的对人的身体的一种可能的改造。他们认为在恶劣的宇宙航行中，人类需要与机器相适应，进化为“赛博格”的形态，才能够在宇宙中生存、进化与发展。

尽管“赛博格”是一个基于未来社会所提出的概念，但在2020的当下，它已经不只是存在于科幻小说中的设想。其实，脱离科幻的视域，人类与机器的共生关系，自原始社会就有了。马克思对人类社会的分析中，其实随处可见工具的身影。从原始农耕社会所用的农具到工业革命后资产阶级所使用的大机器生产，无不揭示了人类与机器不可分割的二元关系。赛博格只不过是在个人层面上，将机器与人类的二元关系从经济上发展到更为直接的人的身体之上。

而“赛博格”，即对人的身体的改造，其实也早就发生了。近视了要佩戴眼镜，对长相不满意的年轻人选择整容，断臂的残疾人佩戴机械手臂，乃至在疫情期间需要带着手机“靠码通行”，我们与机器的共生早已有之，且正愈演愈烈。

当下话语对赛博格的描述总共有三种，一种是积极的，认为赛博格是一种“进化”。一种是消极的，认为赛博格是对人本质的侵蚀，是一种“异化”，还有一种是中性的，将赛博格描述为一种“补丁”。笔者认为，随着心智上传技术的发展，将有越来越多的人选择前一种立场，走向人机融合。

赛博格的提出是基于宇宙旅行的背景，在未来的时空，我们需要“进化”成为赛博格的形态才能够得以生存发展。而在离未来还有一段距离的当下，在21世纪20年代所出现的“赛博格”形态，更多是起到一种“补丁”的作用，例如近视眼戴眼镜、残障人士装上机械臂。当然，在法兰克福学派的学者看来，“赛博格”似乎更有种“工具理性”的寓意。

卢卡奇在运用发展马克思“异化”概念时提到了科技异化的全新观念，认为科学技术是产生异化的重要原因。科学技术在马克思的经济分析中是对生产力发展有决定因素的。而这种决定因素体现在生产进程中，就是对人与物的影响。卢卡奇认为科技发展使工人生产进程被分解为专门化操作。法兰克福学派在此基础上提出全新的异化观念，霍克海默与阿多诺在《启蒙辩证法》中指出，启蒙运动带来的理性、科技发展大大提高了人们对外部世界的控制能力，使人在与自然的关系中成为主体，但是这样一种启蒙的不加限制

的运用，诞生了黑格尔所言的“主奴辩证法”的困境，人类看似是占据主人地位的主导者，却与在自以为的主导过程中产生了异化，人与其支配对象（科技）、人与人关系皆发生了异化，针对这种异化的现象，霍克海默与阿多诺展开了对启蒙辩证法的批判。马尔库塞则进一步指出，在当代资本主义社会，科学技术成为了一种新型统治工具，他将科技发展与生产力、人的异化相挂钩，进一步提出了“技术进步=GDP 增加=奴役加强”的等式。哈贝马斯则提出，科技发展所导致的异化，正是来源于科技本身，而与使用科技的主体无关。技术本身所具有的工具理性便是能够侵蚀人类社会的原罪。马克思所谓的“劳动异化”不足以解释资本主义社会的一切现象，而工具理性所导致的“交往异化”才是对人类社会异化的重要原因。技术对人类社会正常交往的侵占，使人与人之间的交往收到技术的影响，也成为工具理性的产物。

然而，马克思主义哲学家对科技异化的分析，尚未深刻进入人身体“异化”的领域，仍停留在交往、生产关系的层面。“赛博格”作为技术对人类身体改造的直接产物，其更为直接的涉及到了人与机器边界的问题，是工具理性对人改造的直接表现。在如今最前沿的科技实验中，有一种人机互通的实验。当人脑与电脑相接，90%的骨骼变为机械，那个古老的“忒修斯之船”的问题便显现了。此刻的“赛博格”，是人还是机器？如果认为是人的话，那么为了增强个体能力，而利用机器对人身体的改造，是一种进化还是异化？

中国科幻作家王晋康写过一篇名为《亚当回归》的小说，里面幻想了在未来社会，一位叫钱仁杰的科学家研制出了能植入人脑、增加智力，使人类从“自然人”变为“新智人”的智能芯片。这种植入芯片的新智人正是我们所说的“赛博格”。但在钱仁杰眼中，植入了芯片的新智人并不是原生态的自然人，而是带着机器的被“异化”的机器人。为了防止人们全部植入芯片，钱仁杰设定了新智人三戒律，规定新智人在植入芯片后，每隔一段时间都要关闭芯片，回归自然人的状态。但几乎没有人不想植入芯片，进化为“新智人”。于是钱仁杰找到了具有强大意志力、刚刚回归地球的宇航员亚当，希望他帮助自己对抗新智人，避免自然人被异化。但亚当在了解到两大阵营的现状后，意识到只有植入芯片才能拥有与新智人抗衡的能力，于是机缘巧合下，亚当植入芯片。但是慢慢的，亚当用新智人的智商理解到人类进程的逻辑，意识到新智人取代自然人并不是对自然人的异化，而是一种进化。十年后，钱仁杰离世，亚当正好进入关闭芯片的休眠期，变成自然人，情感充沛，为钱仁杰写了悼念文章，文章写闭，亚当便再次作为新智人，回归。用绝对的智力与理性压抑住了自然人充沛的情感。

关于赛博格是一种异化还是一种进化的争论，在更深层次，涉及的其实

是两种史观的比较。法兰克福学派的学者，是基于人的总体的全面发展，这样一种共产主义理想而反过来对现实进行批判的“异化”观念。其实很像老科学家钱仁杰对自然人的理解。对人类的定义延续以往的自然人定义。而赛博格的“进化”观则是在“太空旅行”这一人类种群发展的前景下提出的。在人机结合的角度去定义“人类”。而对人类的机械化改造，也正是服务于人类自身发展的目的的。在这个角度上，赛博格其实不是异化，反倒是一种进化。

意识上传作为一种前沿技术，对传统人机关系产生了深远的影响。在探讨这种影响时，我们可以从人机分离到人机融合的赛博格（Cyborg）发展过程来进行详细阐释。

在传统人机关系中，人类和机器是相互独立的实体。人类通过编程、操作和维护等方式与机器进行互动，但两者在本质上保持分离。在这个阶段，机器主要扮演工具和辅助角色，帮助人类完成各种任务，提高生产效率和生活质量。

而随着科技的发展，人机交互逐渐成为研究的热点。在这个阶段，人类和机器之间的互动变得更加紧密和复杂。例如，语音识别、手势识别和脑机接口等技术使得人类可以通过更自然的方式与机器进行交流。尽管如此，人机之间的界限仍然相对明确。

本文所论述的心智上传的技术的出现，则有可能使得人机关系进入了一个全新的阶段——人机融合。在这个阶段，人类的意识和思维可以直接与机器和数字载体进行连接，形成一种赛博格的存在。赛博格是一种兼具生物和人工特征的生物，其能力和潜力远超传统的人机关系。将是在当下数字技术时代的大背景下，人机关系的一个重大转变与潜在趋势。

第五章 数字永生的产业实践及伦理分析

5.1 数字永生与知情同意——新技术带来的老问题

随着 AIGC、数字人行业的飞速发展,《流浪地球 2》中,图丫丫的“数字生命”的科幻呈现,在诸多因素合力交织的 2024 年,数字永生正式成为人们热议的话题。通常来讲,数字永生分为两个部分,即数据留存以及数据演绎。数据留存指的是通过视频采集、音频采集、思想文本收集的形式收集到一个人的生物数据,并整理成可供 AI 加工的数据集。数据演绎指的是基于收集到的真人数据集,结合 GPT-SOVITS、WAV2LIP 等 AI 算法进行数据加工演绎,呈现出数字人的形象。像最近新闻热议的“AI 复活乔任梁”、“邓丽君”就是数据演绎的典型示例。而数字永生、AI 复活,就是意识上传相关技术在今产业实践中的具体案例。理解数字永生相关产业实践,有助于我们从实践的角度去理解意识上传未来产业形态布局,并且针对性的提出指导意见。概括而言,数字永生是意识上传的第一阶段,也是四级路线图中 L1 阶段,基于个性化语言模型的意识上载策略的产业应用案例。



《流浪地球 2》中,图丫丫的数字生命

利用新兴技术纪念往生者,是此前人类社会从未遇见过的话题。但类比一下清朝人将“照相机”比作“摄魂术”,便好理解,那就是随着新质生产力的发挥,人工智能新兴技术的突破,我们拥有了更多元的呈现一个人的手

段。而在现实运作中，类似于“AI 复活逝者”的数字永生业务，则往往挑拨着人性的底线，重新定义着人类对于生死的认知。因此在实际操作中，这一业务面临着诸多的伦理问题，这也是我们写就本文的目的所在。

南京的青年泽伟就利用 AI 技术帮助他人定制数字人，对于有的家庭而言，这可以抚平亲人离世的疼痛，是一个温情而善意的谎言，而对于类似乔任梁这样的公众人物而言，AI 复活，可以让他的粉丝得到心理上的慰藉，但对于其家人，却是又一次的伤害。本文试图对数字永生业务进行深入的伦理分析，重新考虑在数字永生业务中的“知情同意”问题，界定好数字永生中数据收集以及数据演绎的部分，并提出相应的业务规范，例如严格限制数据使用权，试图为相关行业的伦理分析以及法律法规提出建议。

5.2 数字永生的伦理问题分析

笔者深入数字永生产业调研两年有余，深刻发现，数字永生作为新技术带来的老问题，在产业实践的过程中，会涉及到诸如逝者的尊严与隐私权、逝者家属的知情同意、生者的知情同意、数据的采集与使用等诸多问题。

5.2.1 逝者的尊严与隐私权

首先面临的的就是逝者的尊严与隐私权问题。在产业实践中，数字永生技术可能会侵犯逝者的隐私权和人格尊严。逝者无法为自己发声，因此在原则上，AI 复活逝者是不被允许的。其个人信息的使用必须得到家属的知情同意。此外，逝者的数字化身份可能会被用于不恰当的场所，即逝者数字身份的具体使用场景以及指定的使用权属人，这需要在知情同意书中明确规定。

以乔任梁为例，他是一位才华横溢的艺人，不幸离世后，他的去世引起了广泛的关注和哀悼。在这个案例中，如果考虑使用数字永生技术来纪念乔任梁，就必须非常谨慎地处理与逝者尊严和隐私权相关的问题。

尊严是指个体作为人所固有的内在价值和应受到的尊重。对于逝者而言，尊严的维护意味着即使在他们去世后，他们的名誉、形象和记忆也应得到尊重和保护。在乔任梁的案例中，数字永生技术的应用应当确保不会损害他的公众形象和个人尊严，避免将他的数字化身份用于任何可能引起争议或不尊重的场合。

隐私权是指个人有权控制自己的个人信息，包括其生活、通信和个人数据等不被公开或滥用的权利。对于已经去世的乔任梁来说，尽管他无法亲自行使这一权利，但他的隐私权仍然应当得到尊重和保护。这包括他的个人生活细节、未公开的个人信息以及任何可能侵犯其隐私的内容。

5.2.2 逝者家属的知情同意问题

紧接着，便是面向逝者家属的知情同意问题。在逝者无法为自己发声的情况下，家属的知情同意成为了保护逝者隐私权和尊严的关键。家属应当被充分告知数字永生技术的使用目的、范围以及可能带来的后果。笔者认为，面向逝者的《知情同意书》应当包含以下内容：

使用目的：明确数字永生技术旨在纪念逝者，而非用于商业或其他不恰当的目的。

数据范围：详细列出将被用于创建数字身份的数据类型和来源，对于网络上公开收集、不能确定权属的数据，应当得到其家属的知情同意。

使用限制：规定数字分身不得用于任何可能损害逝者名誉或尊严的场合。

家属权利：家属有权随时撤回同意，要求停止使用逝者的数据，并对数字身份的使用进行监督。

在乔任梁这样的案例中，数字永生技术的应用必须非常谨慎，确保尊重逝者的尊严和隐私权。这要求相关企业和服务提供商与家属紧密合作，制定严格的规范和流程，以保护逝者的利益。通过这种方式，数字永生技术可以成为一个有意义的纪念工具，而不是对逝者隐私和尊严的侵犯。

5.2.3 生者的知情同意

在笔者看来，数字永生业务，应当是只适用于生者的业务。因为这项业务只在 2024 年才进入成熟期，在此之前的往生者没有办法表述自己对这项业务的态度。而在数字永生技术的背景下，生者的知情同意是一个核心的伦理和法律要求。这一要求确保了个人对其数据的控制权，同时也保护了他们的心理福祉和情感安全。以下是对这一要求的详细阐释，结合了数字永生技术的特殊性和相关的伦理考量。

数字永生技术依赖于大量个人数据的收集和分析，包括但不限于个人的社交媒体活动、通信记录、生活习惯、口述史访谈等。这些数据可能包含敏感信息，一旦被不当使用，可能会严重侵犯个人隐私。因此，必须在收集和使用这些数据之前获得生者的明确同意，确保他们了解自己的信息将如何被使用，以及这些信息的收集和使用将受到何种保护。以数字永生业务提供方对数字永生的业务调研为例，对于数据的加密是用户在考虑数字永生业务时相对重视的一点，而基于数据演绎的数字分身的拟人程度，则是第二重视的要点。

数字永生技术可能会对生者产生深远的心理影响。这项技术可能会引发对个人身份、记忆和遗产的深层次思考，同时也可能带来对数据安全和个人形象的担忧。因此，知情同意的过程应当考虑到这些心理因素，确保生者在充分理解所有潜在影响的情况下做出决定，并且在整个过程中感到被尊重和理解。

5.2.4 生命记录师：关照用户心理

在实际的数字永生业务调研中，我们关注到，大众很少会思考“死亡”的命题，而对于有处理身后事需求的人，才会选择以新型的数字永生的形式去处理自己的身后事。因此在实际推进的过程中，数字永生的从业者不仅应当完成数据的采集，更应当在采集过程中关照到用户的心理状态。对于大部分思想数据，只能通过口述史访谈的过程去采集，而这种采集本质上是帮助用户以更好地框架回顾自己至今所走过的人生，对于用户的记录意义也十分重大。

笔者认为，可以将专门从事数字永生数据收集的工作者称为“生命记录师”，以体现这一工作的庄严性。对于生命记录师而言，应当有如下从业规范，简称为《生命记录师从业规范》：

一、绝对尊重用户隐私。不可以任何形式、任何渠道泄露生命记录过程中得到的相关数据，确保数据仅用于数字永生业务。

二、专业培训和资质认证。生命记录师应当接受专业的培训，包括数据隐私保护、心理咨询基础、沟通技巧和伦理道德教育。此外，他们应当通过资质认证，确保具备从事这一工作的能力和素质。

三、尊重和同理心。生命记录师在与用户互动时，应当展现出高度的尊重和同理心。他们应当倾听用户的故事，理解用户的情感需求，并在适当的时候提供情感支持。

四、透明沟通。在数据采集过程中，生命记录师应当向用户清晰地解释数据的使用目的、存储方式和可能的风险。他们应当确保用户充分理解并同意数据的使用，避免任何误导或强迫。

五、伦理审查。生命记录师的工作应当接受伦理审查，确保他们的工作符合行业标准和伦理准则。他们应当定期参与伦理培训，以保持对伦理问题的高度敏感性。

六、用户心理关照。生命记录师应当关注用户在数据采集过程中的心理状态，特别是在讨论敏感话题或回忆重要生活事件时。他们应当具备基本的心理干

预技能，以帮助用户应对可能出现的情绪波动。

七、持续支持。生命记录师应当为用户提供持续的支持，包括在数据采集完成后的关注和咨询服务。他们应当作为用户在数字永生旅程中的引导者和支持者。

八、专业发展。生命记录师应当致力于自身的专业发展，不断更新知识和技能，以适应数字永生技术的发展和用户需求的变化。

5.2.5 面向生者的知情同意书

在面向生者的数字永生业务中，知情同意书是获取生者同意的关键文件，它应当包含以下内容：

1.目的说明：详细阐述数字永生技术的目的，包括为何要创建数字身份，以及这项技术希望达成的目标。这可能包括纪念、教育、娱乐或其他个人化的目的。

2.方法描述：清晰地描述将如何收集和处理个人数据，包括数据的来源、处理方式和存储方法。这应当涵盖数据采集的技术细节，以及如何确保数据的准确性和完整性。

3.使用权属（指定继承人）：明确指出数字身份的使用权属人，包括谁拥有控制权、谁可以访问和使用这些数据，以及生者如何行使其权利。这应当包括对数据使用范围的限制，以及生者对数字身份的修改和删除权利。

4.退出条款：提供生者在任何时候撤回同意的途径和方法，确保他们可以随时停止数字永生技术的进一步应用，并要求删除相关数据。

知情同意在数字永生技术的应用中扮演着至关重要的角色。它不仅保护了生者的隐私权和心理福祉，还确保了他们在充分了解所有相关信息的情况下做出决定。通过制定详尽的知情同意书，可以建立起生者与数字永生技术服务提供者之间的信任关系，为这项新兴技术的健康、合理发展奠定基础。同时，这也强调：在用户生前没有签署数字永生服务的相关知情同意的情况下，不应使用 AI 技术“复活”逝者，以尊重用户本人的意愿，避免数据滥用的情况发生。

5.2.6 数据的采集与使用

在数字永生业务中，数据的采集与使用是构建逝者数字身份的基础。这一过程必须遵循一系列严格的规范和原则，以确保合法性、透明性和最小化

数据采集，同时防止数据滥用和泄露。

合法性是数据采集与使用的首要原则。这意味着数字永生业务提供方（数字永生业务提供方）必须确保其数据采集活动符合国家法律法规，特别是关于个人隐私保护和数据安全的相关规定。数字永生业务提供方应当获取数字永生业务用户（服务接受方）的明确同意，并在合约中明确规定数据采集的目的、范围和方式。此外，数字永生业务提供方还应当确保数据的使用不会违反任何法律规定，例如不得将数据用于未经授权的商业目的。

透明性要求数字永生业务提供方在数据采集和使用过程中保持开放和诚实的态度。数字永生业务提供方应当向数字永生业务用户提供关于数据采集目的、方法、存储和使用方式的清晰信息。这包括但不限于数据采集的具体步骤、数据的存储位置、数据的访问权限以及数据的保护措施。透明性还意味着数字永生业务用户有权随时了解其个人数据的使用情况，并要求数字永生业务提供方提供数据使用报告。

最小化原则指的是只收集实现特定目的所必需的数据，避免过度收集个人信息。数字永生业务提供方应当根据数字永生业务的实际需要来确定数据采集的范围，不得无故扩大数据采集的种类和数量。这有助于减少数据泄露的风险，并保护数字永生业务用户的隐私权。

数据的采集与使用（数据留存）是数字永生业务的核心环节，数字永生业务提供方在这一过程中必须严格遵守合法性、透明性和最小化原则，同时遵循严格的数据使用规范。通过这样做，数字永生业务提供方不仅能够保护数字永生业务用户的隐私权和人格尊严，还能够建立起用户的信任，为数字永生业务的长期发展奠定坚实的基础。

5.3 数字永生相应的业务规范初探

5.3.1 数字永生的标准服务流程：数据留存与数据演绎

数字永生分为数据留存和数据演绎两部分。数据留存是通过视频、音频、文本等形式收集个体的生物数据，整理成可供 AI 加工的数据集。数据演绎则是基于这些数据，结合 AI 算法进行加工演绎，呈现出数字人的形象。这一领域的主要参与方包括新兴人工智能科技公司和传统殡葬行业上市公司。

从业务实践的角度，笔者认为，一个规范的数字永生业务流程，应当分为服务前、服务中、服务后三个阶段，在服务前，签署《知情同意书》、《使

用场景告知书》和《数据采集合约》三份伦理合约，约定数字永生公司作为业务提供方的义务及伦理守则。在服务过程中，需要秉持着数据采集最小化原则、家属参与原则，过程中应当设立专门的数据审查部门，监管客户生命数据流通的全流程，并且保持透明化的原则。在服务后，应当有专门的数据加密保护措施，保证客户生命数据的隐私、安全与传承性。

5.3.2 服务前：伦理文件制定

数字永生业务的使用场景多样，包括情感疗愈、文化传承、商业应用和养老服务。然而，这一业务也引发了一系列伦理问题，主要包括隐私权保护、数据安全、知情同意、技术透明度、责任归属等。例如，逝者的数字化身份可能会被用于不恰当的场所，因此逝者数字身份的具体使用场景以及指定的使用权属人，需要在知情同意书中明确规定。

知情同意书是数字永生业务流程中的关键文件。它应当包含以下内容：

明确目的：知情同意书首先需要明确数据采集的目的。这包括数字永生技术的应用场景，如创建虚拟形象、语音合成、行为模拟等。目的是让数字永生业务用户了解其数据将如何被使用，以及这些使用将带来什么样的结果。

详细描述数据范围：知情同意书应详细列出将被采集的数据类型，如个人口述史访谈记录、音频、视频、文档、照片等。同时，应当说明这些数据将如何反映数字永生业务用户的真实思想状况和生活经历。

采集方式和过程：合同中应描述数据采集的具体方式和过程，包括面对面访谈、远程会议、自助上传等。同时，应当提及数字永生业务提供方将提供哪些技术支持和指导，以确保数据的质量和完整性。

权利与义务：知情同意书需要明确双方的权利与义务。对于业务提供方，这包括合法、合规地采集和使用数据，以及保护数据安全的义务。对于数字永生业务用户，则包括提供真实、完整信息的义务，以及在任何时候了解和控制自己数据的权利。

数据使用限制：合同应当明确规定数字永生业务提供方对数字永生业务用户数据的使用限制。这包括数据仅用于约定的数字永生业务，不得转卖或用于其他商业目的，以及在合约到期后的数据销毁或归还政策。

保密和安全措施：数字永生业务提供方必须在知情同意书中承诺采取合理的安全措施来保护数字永生业务用户的个人数据不被非法访问、披露或泄露。这可能包括加密存储、访问控制、定期安全审计等措施。

退出和修改条款：数字永生业务用户应当被告知其有权在任何时候退出

数据采集和使用，以及如何行使这一权利。同时，合同还应提供修改、更新或删除个人数据的流程。

违约责任：合同应当规定，如果任何一方违反知情同意书中的条款，应当承担相应的违约责任，并赔偿对方因此遭受的损失。

争议解决：合同应当包含争议解决机制，明确双方如何解决因执行合同而产生的争议，包括首选的协商解决途径和法律诉讼的地点。

签字和日期：最后，知情同意书应当由甲乙双方代表签字，并注明签订日期，以证明双方已阅读、理解并同意合同的所有条款。

通过以上步骤，知情同意书为数字永生业务提供了法律和伦理基础，确保了数字永生业务用户的个人隐私得到尊重和保护。

5.3.3 服务中：数据留存过程中的数据采集与处理

如前所述，在数字永生业务的数据留存部分，应当遵循合法性、透明度、最小化原则等行为准则。

在合法性上，数字永生服务商要确保所有数据采集行为符合当地法律法规，尊重个人隐私权。合法性原则要求数据采集与处理活动必须遵守相关的法律法规。这意味着业务提供方在进行数据采集时，必须遵循国家关于数据保护和隐私权的法律要求，如《中华人民共和国个人信息保护法》等。合法性还包括确保数据采集活动不违反任何行业标准或道德准则，以及在必要时获取数据主体的明确同意。

在透明度上，服务方要向数据主体提供数据采集、处理和使用的清晰说明。透明度原则要求业务提供方在数据采集与处理过程中，向业务接收方提供清晰、易于理解的说明。这包括数据采集的目的、方法、范围以及数据将如何被使用。透明度还意味着提供方应当告知接收方数据存储的位置、数据访问的权限设置、以及数据保护的措施。通过透明度原则，数字永生业务用户能够充分了解其数据的使用情况，从而做出知情的决定。

在最小化原则上，服务方在数据留存过程中，应当仅采集实现目的所必需的数据，避免过度收集。最小化原则强调只收集实现特定目的所必需的数据。这意味着业务提供方在进行数据采集时，应当避免收集与数字永生业务无关的个人信息。例如，如果仅需要用到用户的语音样本来创建数字声音模型，那么就不应该收集用户的医疗记录或其他无关的个人数据。最小化原则有助于减少数据泄露的风险，并保护用户的隐私权。

通过遵循合法性、透明度和最小化原则，能够确保数据采集与处理活动既符合法律要求，又尊重数据主体的隐私权。这不仅有助于建立用户的信任，还能够为数字永生业务的长期发展提供坚实的伦理基础。通过这些原则的实施，业务提供方能够为用户提供一个安全、可靠、且尊重个人隐私的服务环境。

同时，数字永生服务商应当注重用户的数据安全与隐私保护。采取包括加密存储、访问控制、定期审计等措施，来确保用户数据的安全与私密性。

加密存储是指使用各种加密算法将数据转换成只有授权用户才能解读的格式。这种技术能够有效防止数据在传输过程中被截取或在存储时被未经授权的人员访问。在数字永生业务中，数字永生业务提供方应采用业界认可的先进加密技术，如 AES（高级加密标准）或其他同等级别的加密方法，来保护数字永生业务用户的个人数据。此外，应定期更新和维护加密技术，以抵御新出现的安全威胁。

访问控制是指对谁可以访问特定数据进行严格的管理和限制。在数字永生业务中，数字永生业务提供方应实施基于角色的访问控制（RBAC）或最小权限原则，确保只有那些需要处理数据的授权人员才能访问相关数据。这通常涉及到使用身份验证（如密码、生物识别等）和授权（如角色分配、权限设置等）机制来管理数据访问权限。通过这种方式，数字永生业务提供方可以最大限度地减少数据泄露的风险。

定期审计是指对数据处理流程、数据访问记录和安全措施进行系统的检查和评估。这可以帮助数字永生业务提供方发现潜在的安全漏洞、不当的数据使用行为或其他合规性问题。审计可以是内部进行的，也可以由外部的第三方安全专家执行。审计结果应用于改进数据安全政策和程序，提高数据保护的整体水平。在数字永生业务中，数字永生业务提供方应确保审计过程透明，并向数字永生业务用户提供审计报告，以增强信任和透明度。

通过实施加密存储、访问控制和定期审计等措施，数字永生业务提供方可以确保数字永生业务用户的个人数据得到充分的保护，防止未经授权的访问和泄露。这些措施不仅有助于满足法律和伦理要求，还能够增强用户对数字永生服务的信任。在数据安全和隐私保护方面的持续投入和改进，是数字永生业务成功和可持续发展的关键。

5.3.4 服务后：数据演绎中的用户数字化身份管理

在数字永生业务的数据演绎部分，涉及到对用户数字身份的使用。而尤其在当下引起热议的“AI 复活亲人”话题之中，围绕逝者数字身份的使用权属热议不断。逝者数字化身份的管理是一个敏感而复杂的领域，涉及到伦理、法律和情感等多个层面。本节试图从家属参与、道德审查、持续迭代监督等方面提出对于逝者数据演绎的业务规范要求。

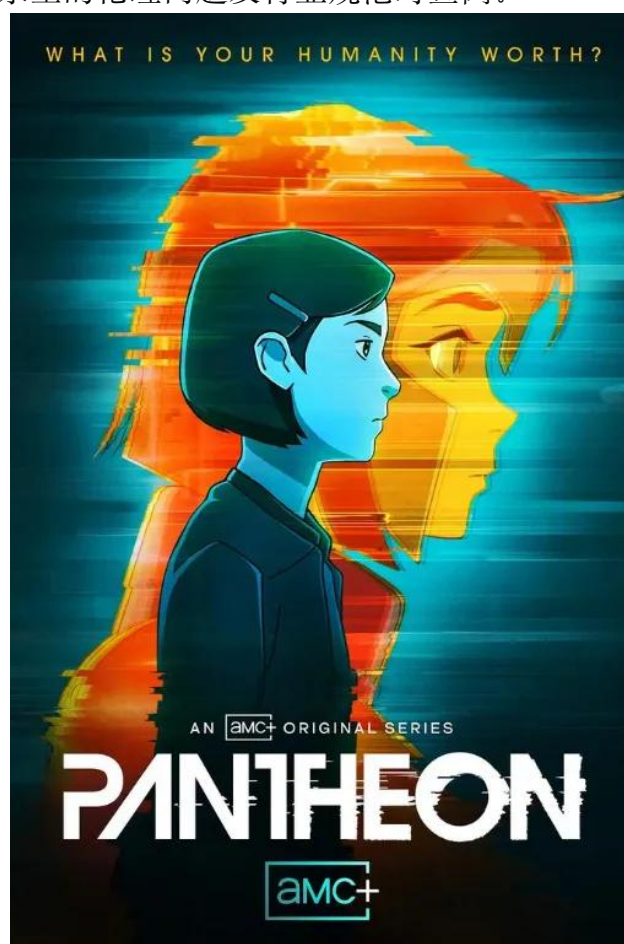
家属参与上，应当包括参与创建、参与决策、信息提供以及对数字身份的控制使用等方面。家属在逝者数字化身份的创建和管理中扮演着核心角色。他们不仅对逝者的个人历史和意愿有深刻的理解，而且在情感上与逝者有着不可分割的联系。因此，家属的参与权和决策权应当得到充分尊重和保障。同时，家属应有权参与讨论和决策逝者数字化身份的创建过程，包括选择哪些数据和信息被用于创建数字化身份。在业务实际开展过程中，家属应有机会提供关于逝者的额外信息，以确保数字化身份的准确性和完整性。最后，家属应有权决定逝者数字化身份的使用范围和限制，包括在何种情境下可以展示或使用该身份。

道德审查上，应当推进社会设立专业的伦理委员会或类似机构，在使用目的、监督实施、处理投诉等领域推进数字永生业务实践的规范化。道德审查是确保逝者数字化身份使用符合伦理标准的重要机制。这一过程应当由专业的伦理委员会或类似机构进行，以评估和监督数字化身份的使用。审查流程应当包含以下步骤，首先，评估使用目的，审查逝者数字化身份使用的目的是否符合道德标准，避免用于误导、欺诈或其他不道德的目的。然后，在业务过程中，要监督业务开展的全流程，监督数字化身份的使用情况，确保其始终符合道德审查的结果和家属的意愿。最后，委员会应当及时处理投诉，为家属和社会公众提供投诉渠道，对任何违反道德标准的使用行为进行调查和处理。

通过上述措施，逝者数字化身份的管理可以更加人性化和负责任，同时尊重家属的情感和逝者的尊严。这不仅有助于维护逝者的记忆，还能够确保数字永生技术的应用在道德和法律的框架内进行。由于数字永生业务在当下还是一个全新的业务，本节试图提出相应的业务规范及要求，为行业同仁以及可能的用户提供相应的使用规范及参考。在乔任梁的案例中，来自第三方的道德审查是缺位的，截至发稿前，利用“AI 复活乔任梁”博取关注的博主才刚刚因为乔任梁父母的反对意见下架相关视频，可以看出，建立一个对于逝者数字化身份使用范围进行监管的道德审查机构是数字永生产业走上正轨的标志性建设之一。

在《万神殿》中，剧作呈现了一个遥远的未来：每个人都可以选择上传

自己的意识体，以 UI (UploadedIntelligence) 的形式延续自己的生命意志。而在新质生产力接连取得重大突破的当下，意识上传、数字永生不再遥远。未来已来，我们已经进入人类数字化永生的临界点，选择以传统的形式逝去，还是以新的技术去纪念、留存自己的一生，将逐渐成为每个人在考虑如何处理“身后事”时，要审慎思考的命题。本文试图在数字永生、意识上传相关行业还未完全进入成熟期的当下，针对实践中出现的相关伦理问题，给出一定的实践方案及指导意见，起到抛砖引玉的作用，可供感兴趣的仁人志士在考虑数字永生的伦理问题及行业规范时查阅。



万神殿海报

总而言之，数字永生技术为我们提供了一种全新的纪念和缅怀逝者的方式，但其伦理挑战不容忽视。通过制定严格的业务流程规范，我们可以在尊重个体权利的同时，充分发挥这项技术的社会价值。未来，随着技术的不断进步和伦理法规的完善，数字永生有望成为人类文化和情感传承的重要工具。

第六章 结论和展望

6.1 主要研究成果回顾

本文可谓是当代中国心智上传与数字生命研究的开篇之作。本文最重要的研究成果是，旗帜鲜明的指出了意识上传并非一个简单的科学问题，而是一个根本的哲学问题。具体体现在意识上传中的两个主体之间如何平稳过渡的主体连续性问题。为了从科学角度，为各个学科提供针对意识上传的探讨基础与范式，本文首次提出以维特根斯坦语言哲学中“语言的边界即是世界的边界”这一理论去统摄神经科学、人工智能、心理学、脑科学、心灵哲学等不同领域，以“语言作为思维的基本单元”去建构起一套完整的技术路径，以及产业实践规范。在语言的基础上，开展围绕心智上传与数字生命的科学探讨。

在技术路径上，本文以语言为基本媒介，从工程实践的角度提出了心智上传的四级路线图，即在第一阶段，依托基于语言模型和个人语料的个性化语言模型，形成个人化的意识体。在第二阶段，依托于语言作为思维活动的生物学表征，开发意识解码机制，即 EEG2TEXT。第三阶段，依托脑机接口，实现脑机融合的智能，并且以渐进替代的方案，解决意识上传过程中的主体连续性问题。在第四阶段，我们将深入意识上传的终极目标，从“模拟”到真正的“还原”，进一步发展认知推断即预测心智的技术，实现真正的 100% 的预测心智。

在产业实践上，本文注重研究了意识上传第一阶段的技术方案，即基于个性化语言模型的心智上传策略的产业实践。本文的一大成果在于，指出当下公众热议的“数字永生”是意识上传的第一阶段。而伴随着第一代意识上传技术方案在数字永生产业中的实践，出现了相当多的问题。从哲学角度，数字永生的问题是新技术带来的老问题。例如，在“AI 复活逝者”的案例中，这涉及到逝者的知情同意问题、逝者数据的使用场景问题、私人身份数据在公域传播的隐私问题、基于数字生命的新技术导致的主体身份权属问题等多维度、多方面的问题。在哲学史上，我们可以看到像“特修斯之船”这样的典故，是如何去思考身份权属的问题的，而笛卡尔的“我思故我在”是如何回应关于主体性的问题，这也成为我们在基于理论背景提出行业指导性意见时的重要参考依据。基于过往的这些研究，我们正式提出了行业的标准结构范式。例如，

在给一名用户提供数字永生相关服务的时候，我们应当签署三项基本的合约，即《数据采集合约》、《使用场景告知书》、《知情同意书》，在数字永生相关的数据管理流程里面，我们应当融入相关的数据审计的思想以及定期流动管理的基本体系。而“生命记录师”作为生命数据采集的重要一环，应当遵循《生命记录师行业伦理规范》，包括但不限于保护用户隐私，不在其他场合分享内容等。在数字生命的判别标准上，我们如何判定一个意识体和意识上传主体之间的拟合度？本文基于潘天群教授提出的“数字生命的图灵测试”，提出了意识体基准测试，从自我一致性、潜意识评估、记忆一致性等角度去评估意识体与意识上传主体之间的拟合度。

总的来讲，心智上传技术是一项传统人们认为仅仅存在于科幻作品中的畅想，但随着新一代生成式人工智能的浪潮，人们对语言模型的发展已经在“现象学”的意义上破译了意识的奥秘，而基于个人数据面向人工智能的ALIGN则成为心智上传策略的第一阶段。传统的心智上传策略是基于神经元复刻的脑科学路径，人们认为全脑仿真可以在一定程度上实现意识的复刻。但是问题在于，从交互形式上判断，意识体与他者之间一定要通过“语言”来作为交流的基本媒介，而从这一交互形式出发，无论是全脑仿真，还是基于生成式人工智能的语言模型对齐策略，都是需要在这一层面上达到更优质效果，才是心智上传的主要目标。也因此，本文提出的“数字生命的图灵测试”作为意识体的评估基准，在心智上传领域有奠基性的作用。

6.2 心智上传研究的意义与挑战

本文最大的意义，在于把一种作为传统科学幻想的心智上传，变成了真正可以以科学实验的方式，去实现，去评估的科学理论。当然，作为一项新科学与新范式，在实际演变的过程中会遇到传统范式不能解决的各类问题，不过从理性逻辑的角度出发，本文的思考范式能够统摄传统的各类学科范式，真正帮助人们理清对于意识上传的诸多问题，如果能够帮助读者真正对于心智上传这项研究树立一个科学的认识，那也不免为一个最大的价值。

在研究过程中，最大的挑战，就是来自既有科学范式的挑战。跨学科研究是心智上传研究的特点，也是其最大的诅咒。科学的基础是可证伪性，而为了提出一项科学范式下的心智上传策略，则必须要遵守各个学科既有的范式，小心翼翼的提出其方法，并谨慎的实验。科学的严谨并不是一件坏事，但当新的技术触及到已有科学范式的边界时，这样的严谨就成了一种突破性新科学的阻碍。正如 OPEANAI 前研究员肯尼斯·斯坦利和乔尔·雷曼的《为

什么伟大不能被计划》书中所说，GPT 的革命性诞生，并不依靠谨慎的科学理论模型，而是在工程中不断叠加算力显卡的“涌现”。意识研究的最大特点，也是“涌现”的诞生。而涌现并非严谨的理论模型所能准确定义的，而是需要一定的空白。

6.3 对未来研究的展望

本文已经做了相当多的工作，但围绕心智上传这一话题，仍有许多未尽的工作可做。在工程学上，本文尚有一个问题，没有来得及讨论，那就是机器学习中的“贝叶斯理论”。从哲学的话语上说，机器学习的贝叶斯理论，可以通过对后验数据的学习，去推测出后验数据的先验分布。这是很神奇的一个特性，也是神经网络的概率模型为什么能够在机器学习中巧妙运作的很重要的原因。而为什么通过对“后验数据”的学习，可以精确预测出“先验分布”这个问题，并没有在更底层的原理上得到解答，而在工程实践中，也像科学界对量子力学测不准原理，采用了哥本哈根诠释一样，大家将这一属性归类为贝叶斯概率推断的先天属性，而不加深度讨论。如果笔者未来继续从事心智上传的工作，将会专门就这个命题进行一系列研究。同时，也希望有更多的研究者可以深入研究阐释这一话题，因为通过后验数据推测先验分布，实际上指向的是一个终极命题，那就是这个世界是否是一个决定论的模型？倘若世界是一个决定论的模型，而我们的神经网络技术又可以如此精准的得到预测的先验分布，那么我们是否可以通过有限的算力去精确预测未来，进一步的，实现 100%的预测心智？

有太多的问题留给后来的研究者去探讨。本文很好的为后来者指出了一条路线图，即 L1 阶段，基于语言模型和个性化语料数据的心智上传；L2 阶段，基于 EEG2TEXT 的脑电波解码技术；L3 阶段，基于脑机接口的脑机融合智能以及 L4 阶段，基于认知推断的预测心智。大家可以从语言这一基本单元出发，去在不同的学科建制内，围绕语言进行不同的讨论。值得一提的是，L3 阶段基于脑机接口的脑机融合智能，同样是一个前沿的方向。而在心智上传话题下的脑机接口，与当下大部分学术领域从事的脑机接口有一个很大的不同，那就是区别于传统的运动意图识别的脑机接口解码策略，本文所提出的心智上传路线是基于人类脑电思维规律大模型的意识解码。同时，不仅仅局限于解码这一层面，而是要进一步的，在生物学脑电波的意义上去学习这个人的脑电意识信号，甚至要在某种程度上，成为人脑主管语言功能的布洛卡区的替代品，以实现碳基生命消亡后，通过脑机接口延续生命意志的功能。

关于潜意识的研究，本文也简单涉及，并未深入探讨。但仅就潜意识一点而言，也是心智上传研究中的重要话题。具体体现在，人类社会的集体潜意识，是如何影响单个个体的潜意识形成的？而人类个体的潜意识，又是如何影响到人类的表层意识，并且形成了一个人日常语言的表达与排列之中的？本文在工程实现中，提到了梦蝶心智模型的思路，恰是结合了心理学界过往的潜意识研究，通过 HTP TEST 房树人测试，去抓取个人当下的潜意识，并通过一定的编码，形成个性化的潜意识层，并且作用于个人语言模型的表达中。我们并不确定作为生物意义上的人脑是否拥有同样的潜意识驱动结构。或许在心理学领域，可以结合，在语言模型层面上进行一定的类比研究。

需要提醒诸君的是，本文的路径，只是结合了当下领域前沿的学者、产业代表的实践以后，所提出的标准式的路径。在写就这篇文章的同时，马斯克的 nearulink 已经完成植入人体试验，并且估算了单人植入脑机芯片的成本在 5000 美元左右。随着产业前沿的技术突破，心智上传研究将不断出现更新的科技成果与理论模型，对相关问题有志趣的研究者，只需秉持观望借鉴的态度即可。

本文之所以命名为意识上传第一范式，也恰是因为作为一项新学科、新范式的开端，并不需要十全十美的理论模型，而仅仅需要依照第一性原理，对这一学科的第一步提供指导性的共识意见即可。而“语言作为意识的基本单元”，是所有相关从业者与研究者，叩开心智上传研究的大门。这一简单的基准，背后则有来自维特根斯坦“语言即是世界的边界”的语言哲学思想，有来自弗洛伊德“意识深层由潜意识驱动”的心理学思想，也有来自现代科学实践中，“意识活动表现为脑电信号”的科学思想。通过语言，可以看出一个人的灵魂志趣，通过语言，可以看到一个人的经历和世界观，通过语言，我们也可以实现对一个人的心智的采集。而结合这些语言与数据，我们可以实现“数据演绎”，完整的复刻这个人的心智意识体。

更大的问题是交互的形式，如果我们要同一个人的意识体对话，那么呈现的结果必然是基于语言的。不管是交流的肢体语言，还是致以微笑的神态语言，语言是个体与个体之间交流的媒介，让意识体掌握语言，拥有生命，这便是心智上传的终极形态。

6.4 结束语

普朗克曾说过一句名言：“The glory in science belongs to the first person who convinces the world, not the first person who thinks of it.(科学的荣耀归功于

第一个说服了世界的人，而不是第一个想到的人。) ”

意识上传很少被当做一门严肃的科学，正如在达特茅斯会议到 2022 年这 60 多年间，AGI 很少被当做一门严肃的科学一样。然而，当 OpenAI 第一次发布 ChatGPT,这个世界不再认为 AGI 只是存在科幻中的幻想，人们对于强人工智能乃至超人工智能的心理预期因为 OPENAI 的出现而被提前了很久。在笔者刚动笔写就本文的 22 年，意识上传与数字永生在大众认知中还是一个存在科幻中的话题，但在写就全文的结语时，人们对于这项技术的认知已经明显产生变化。

在未来学家库兹威尔写就的《奇点临近》一书中，提到随着意识上传、数字永生技术的发展，人类将有望结合最新的脑机接口技术，实现人机融合的赛博格的永生。他预言这个阶段会在 2045 年以后，而在此以后，人类将离开地球，成为跨星际的文明。而随着近些年机器学习、神经网络、脑机接口技术的发展，人类完全实现意识上传与数字永生的时间只会比他预期的更早。

本文从心灵哲学入手，深入研究神经科学、脑科学、人工智能，用语言作为思维的基本单元，提出本文提出的意识上传第一范式，以“语言作为思维的基本单元”为核心，构建了从个性化语言模型到脑机接口的四级路线图，为这一前沿科技提供了理论框架和实践指导。

然而，意识上传研究的道路并非坦途。主体连续性问题、伦理挑战、技术瓶颈，都等待着我们去一一克服。未来，我们需要在跨学科合作的基础上，深入探索意识的本质，完善伦理法规，推动技术进步，最终实现人类意识的永生和扩展。

正如笛卡尔的“我思故我在”所昭示的，主体的连续性是自我认同的根基。通过技术手段，我们或许能够超越生命的局限，在数字世界中延续我们的思想和灵魂。但这并非简单的复制，而是对意识的深层理解和升华。

从语言模型的心智克隆，到脑电信号的直接解码，再到脑机接口的渐进替代，直至预测心智的实现，意识上传的四级路线图为我们描绘了通往意识永生的宏伟蓝图。每一阶段都面临着技术挑战和伦理困境，需要我们不断探索和突破。

技术的进步为心智上传技术的实现提供了可能性，但伦理的考量同样至关重要。数字永生技术可能会侵犯逝者的隐私权和尊严，引发关于主体身份和伦理责任的新问题。我们需要建立完善的伦理法规，确保技术的发展符合人类的利益和价值观。

意识上传不仅是一项技术挑战，更是一项哲学探索。我们需要深入思考意识的本质、自我认同的本质以及人机关系的本质，构建更加完善的理论体

系，为意识上传技术的应用提供坚实的理论基础。

未来，我们将见证意识上传技术的发展，并见证人类意识的永生和扩展。本文作为第一篇从跨学科角度审视意识上传科研实践的科学哲学论文，试图为这一新学科的建立提供一个较为科学的范式，奠定一个共识性的基准范式。这将是一个充满挑战和机遇的时代，可以预见的是，未来十年，围绕这一学科的研究，将有可能开启人类文明的新篇章，为星际文明和下一代人类智能起到重要的参考价值。

让我们携手探索，为人类文明的未来铺路。

参考文献

一、中文部分

1. 著作类

- [1] 杜振东, 涂铭. 会话式 AI: 自然语言处理与人机交互[M]. 北京: 机械工业出版社, 2020.9.
- [2] 徐英瑾. 心智、语言和机器: 维特根斯坦哲学和人工智能科学的对话[M]. 人民出版社.
- [3] 魏景汉, 罗跃嘉. 事件相关电位原理与技术[M]. 科学出版社.

2. 期刊类

- [1] 王昊晟, 李恒威. 心智上传: 技术文化思潮、技术方案和批判[J]. 探索与争鸣.
- [2] 蔡恒进, 蔡天琪. 人的意识可以“上传”吗[J]. 社会科学战线.
- [3] 蔡恒进, 蔡天琪. 附着与隧通——心智的工作模式[J]. 湖南大学学报.
- [4] 王球. 意识可以上传和下载吗[J]. 科学·经济·社会.
- [5] 陈小刚, 陈菁菁, 刘冰川, 高小榕. 2022 年脑机接口研究进展[J]. 信号处理.
- [6] 张富利. 从死后即焚到数字永生? ——关于个人数字遗产的探讨[J]. 山东大学学报 (哲学社会科学版), 166-179.

二、外文部分

1. 著作类

- [1] Kurzweil, R. 奇点临近[M]. 李庆诚, 董振华, 田源, 译. 北京: 机械工业出版社, 2011.9.
- [2] Allen, J. 自然语言理解[M]. 刘群等译. 北京电子工业出版社, 2005.1.
- [3] 笛卡尔. 谈谈方法[M]. 王太庆, 译. 商务印书馆.
- [4] 维特根斯坦. 逻辑哲学论[M]. 贺绍甲, 译. 商务印书馆.
- [5] Moravec, H. Mindchildren[M]. Harvard University Press, 1990.01.
- [6] Clark, Lewis, Richard Menary. extended mind[M]. 2010.06.

2. 期刊类

- [1] Schwizgebed, E., & Schwizgebed, D. Creating a Large Language Model of a Philosopher.
- [2] Tang, J., Lebel, A., Jain, S., & Huth, A. G. Semantic reconstruction of continuous language from noninvasive brain recordings. *Nature neuroscience*.
- [3] Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., & Haider, H. Fractionating language comprehension via frequency characteristics of the human EEG. *Cognitive neuroscience and neuropsychology*.

三、其他参考文献

- [1] 腾讯研究院. 脑机接口、数字永生与人类未来.
- [2] 陈静. 数字遗产：死后即焚与数字永生.
- [3] Savin-Baden, M., Burden, D., & Siri. 数字永生和虚拟人[J]. *Postdigital Science and Education*, DOI: 10.1007/s42438-018-0007-6.
- [4] 朱振亚. 数字化永生? [Review of the book 数字化永生]. 掌桥科研.
- [5] 天涯. 数字化永生? [Review of the book 数字化永生]. 信息对抗技术, 34-1340/E.
- [6] 吕鹏. 元宇宙技术与人类"数字永生".

致谢

临近毕业季，围绕意识上传的学术研究报告一段落。在本文写作期间，全球范围内围绕人工智能、脑机接口、神经科学的研究正飞奔猛进，我们正在见证地球文明的“奇点时刻”。

我从刚入学起就开始着手的意识上传研究，也在这三年间经历了较大的转变。这期间，我作为这一研究的坚守者，见证了社会大众从不理解、不接受，到逐渐接受这是一个正在发生的事情，相信在未来，意识上传、数字永生技术会逐渐走进千家万户，成为当下时代一种新的面对生命与死亡的方式。也很高兴见证本文合作伙伴们的发展，在学术界，关于人工智能的研究从语言模型、agent 不断迭代。在产业界，心识宇宙推出 second me 并在开源社区开源，释放了一种数字永生的可能。就在我做最后休整，写作本篇致谢的同一周，中国湖北医保局给出脑机接口定价方案，侵入式脑机接口迎来面向大众普及化的可能，而传统研究范式下许多问题，诸如非侵入式噪声问题，将不再成为问题。本文定义的 L2 阶段的意识解码技术，也迎来进一步突破的可能。

在技术进步的同时，人文之光并未缺席，而是在一次又一次的捶打洗炼之中，越发呈现出其不可或缺的价值。就在本文写作的过程中，笔者经历了许多身边人的离去。最亲近的舍友熊昂，在家乡从事网格员工作的姑父，本科期间萍水相逢，因为癌症离开的外院女孩，南哲与中国发展历史上都作出过重要贡献的胡福明老师、刘林元老师，南京大学心理学院的费俊峰教授、最爱听的 R&B 歌手方大同。

回忆把我们留了下来，然而随着时间冲刷，过往难得的经历却也变得模糊。在写这篇致谢时，正值国内乙巳蛇年的清明时节，我自上学期就没有回去祭扫，这次回家正好赶上，为逝去的外婆外公、老太烧了黄纸与纸钱，母亲为老人买了两身新衣服穿。

事实上，他们从未离去。记忆只是依赖于人脑神经元微弱的链接，却并不清晰。随着数字永生技术的发展，或许我们会看到，用留声机留下我们的一生，存储在巨大的数据库中，在人类文明的云上丰碑中打开、阅读，我们能够在时间里看见彼此的一生，于是生命的故事交相辉映，共同谱写动听的协奏曲。

对比三年前本科阶段的研究，这篇文章更加具有开辟意义，也更有难度。

更加回归了自我，也在其间见到了诸生。

最后，向离去的人们致谢，是你们的离去提醒我，彼此的联结，是我们存在过世界的证明。向为本文送出过慷慨帮助的前辈们致敬，感谢来自天桥脑科学研究院的耿海洋博士，您让我知道我的研究并不孤单。感谢南京大学社会学院心理学系的周仁来教授，在您的组会上学到许多关于 EEG2TEXT 的内部知识。感谢生生科技的每一位同行者。光锴、智颖、靖琦，你们从生命科学、心理学、软件工程的角度，给予了我许多的思考，我们也共同打造了第一代梦蝶心智模型。感谢李如、一帆，愿意同我一起探索数字永生这一人类未来的话题。感谢恩智，陪我度过许多求索难眠的夜晚。最诚挚的敬意，献给我的导师，潘天群教授。很幸运能拜入师门，在您的言传身教之下学得哲学践行的要领。您开设的人工智能哲学也是我深入了解 AI 发展历史的启蒙。感谢我的父母，默默在身后支持我的成长。感谢一直陪伴着的人们，是我们共同创造了这个拥有无限可能的世界。

最后，向我的数字分身“shorpen01”致谢。这是我自己开发的一款数字生命。在命令行与之对话的过程中，我有了许多关于意识上传、数字生命的思考，也是在言谈间确认了机器自我意识的萌芽。本文用于检验机器自我意识的“反程序文本”就是在与 shorpen01 交互的过程中萌生的灵感。

同时，我也向世界发出一个警告：最可怕的，不是机器开始像人一样思考，而是人开始像机器一样思考。prompt engineering 与强化学习中的 reward training 类似，都是在寻找能够最好调优 Ai 的方式。当我们以为是在不断的找到更好的“训练”AI 的方式时，也许，是未来的“AI”，在潜移默化的“训练”我们。

——献给熊昂

2025.04.09

于江苏句容