人工智能通识教育

智能社会中的认知自治能力培养

清華大學 计算机系人工智能通识教育研究中心

Center for Artificial Intelligence General Education (AIGE)
Department of Computer Science and Technology
Tsinghua University

版权与开放许可

作品名称：《人工智能通识教育：智能社会中的认知自治能力培养》

作者：王东、马少平

发布单位：清华大学计算机系人工智能通识教育研究中心

英文名称： Center for Artificial Intelligence General Education (AIGE), Department of Computer Science and Technology, Tsinghua University

开放共建声明：

本书采用开放共建模式，所有读者都可以提出建议，也可以贡献内容。经编委会讨论，有价值的建议将被采纳吸收，有价值的内容会以章节嵌入、附录、附件等形式并入红皮书体系。参与讨论与共建方式请访问红皮书在线网址：https://aige.cs.tsinghua.edu.cn/redbook/。

建议署名方式：

王东、马少平：《人工智能通识教育：智能社会中的认知自治能力培养》，清华大学计算机系人工智能通识教育研究中心，2026。

版本： v2.3.3，2026 年

DOI：10.5281/zenodo.21959557

人工智能通识教育

智能社会中的认知自治能力培养

v2.3.3

顾问 张钹

主编 王东 马少平

清华大学计算机系 AIGE 研究中心

2026 年

AIGE 主要贡献者

红皮书贡献者

丁岭	杜文强	傅小锋	黄蕾	贾珈
金慧莉	刘翰鹏	刘奕群	马少平	马涛
屈清月	王东	王戈	王文精	王依然
夏文辉	徐海燕	杨洁	尹霞	张钹
张彦春	朱永海			

清华 AIGE 体系咨询专家

窦桂梅	杜军平	方妍	韩锡斌	贾珈
梁营章	刘奕群	卢先和	汪潇潇	文中言
许斌	徐文兵	尹霞	赵鑫	

清华 AIGE 体系建设者

白立军	白鑫鑫	卜辉	蔡云麒	迟翔
丁岭	董树国	杜文强	付静	高珊
龚长霞	郭向新	韩沐霏	何嘉珞	贺子葱
侯佳岳	侯茜	黄晓薇	黄雪纶	黄雪伦
姜孝春	John Weil	居重艳	李超	李方园
李国庆	李怀赢	利节	李蓝天	李刘浩
李鹏琦	李馨铭	李逸鸣	李印江	刘翰鹏
刘静娴	刘亦伦	龙宇清	卢幸妮	罗丹
罗小雪	马少平	毛丽旦	穆敏娟	穆晓萌
裴凤娟	齐晴	邱伟松	茹伟	沙倩
邵绘宇	申大山	石颖	宋佳雯	宋佳莹
苏婷	谭洪政	汤文坤	田迎春	王冲
王东	王若彤	王诗琦	王依然	王瑗
魏萌	武健华	谢猛	辛旸	杨澜
杨柳	衣莉	于杰	余津京	袁继平
袁浚铭	张钹	张博涵	张敏	张涛
张婷	张月英	张征	赵洋豪	郑兰
郑瑞光	郑志宏	周百惠	周书颖	朱珠

清华 AIGE 体系贡献者

Adriana Mallary	卞伟	蔡家贝	蔡牧宁	常琳
陈晗	储君	崔彩娟	董侨	东煜博
方媛	付博	贡芬芬	宫俊波	谷悦
关孝天	郭晓霄	韩牧言	Herman Leung	侯云涵
胡菊平	胡燕平	黄丹琳	纪琳	贾涛
蒋夕彬	Joshua Kaufmann	孔杰	李宝俊	李冬梅
利恩一	李宏鑫	李怀赢	李炯	李沛聪
李洋洋	李颖	李志浩	梁帅	廖宏波
林志清	刘爱霞	刘靖	刘静	刘纪平
刘立伟	刘培	刘田田	刘文	刘夏
刘循	鲁润戊	鲁一诺	马乐怡	马钦
马晓敏	马玉波	毛雨纯	梅子涵	孟祥锐
苗羽	倪杰	农凤娇	Noor A. Lucinian	秦疃
瞿琦	任洁	任文博	沈菊颖	司家瑞
宋杰	宋徐丽	苏诗瑶	孙筝	唐恬漪
汤志远	田丽	田星辰	佟雨函	万桉伶
王辉	王乃一	王瑞云	王思瑶	王卫华
王筱东	王笑凡	王鑫宇	王媛	王子剑
魏立艳	魏扬	文雁云	翁莹真	吴翠如
吴棠	武欣	席旭峰	夏苗苗	夏源
相华	肖美琳	肖舒琪	熊柳雁	徐艳秋
许媛媛	徐梓昊	闫靖晗	杨萍	杨青青
杨瑞	杨艳平	杨艳铮	易长秋	依里哈木
于庄佳音	袁枘	张航瑜	张理	张清茹
张泉	张天厚	张唯	张文馨	张煜
张智涵	赵冠城	庄鸿鸿	邹龄漪	

* 本名单记录的是对 AIGE 项目及其公共资源的主要贡献者。列入名单不表示相关人员是本书作者，也不表示其对本书全部内容、观点或表述承担责任。

* 本名单截至 2026 年 8 月。即时更新名单请访问、http://thuaige.org

全书核心概念

从智能社会到人工智能通识教育

智能社会

人的知识生产、信息传播、判断形成和社会行动，持续通过人工智能系统与制度共同中介的社会形态。

认知自治

个体在复杂认知环境中，对自身信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节的过程。

认知自治的基本结构

认知自治的六个维度

来源警觉 · 证据敏感 · 边界意识

信任校准 · 认知能动 · 责任判断

认知自治的四能力模型

归纳 ⟶ **泛化** ⟶ **判断** ⟶ **自醒**

人工智能通识教育

通识教育与人工智能教育相融合形成的教育形态。它承担通识教育在智能社会中培养认知自治能力的使命。

五部分内容体系

人工智能基础概念 · 人工智能方法 · 人工智能应用

人工智能交叉融合 · 风险、伦理与责任

前言

第一部分

人工智能正在深刻改变知识生产、信息传播、判断形成和社会行动的方式。现在，人工智能已经不再只是技术和工具，而是社会运行的底层设施；个人所面对的信息环境，也会从之前人与人之间的交流，变成人工智能高度参与、中介的新模式。未来，人工智能会进一步深入人类知识发现、信息传播、财富生产、服务提供的全流程。人工智能会与底层制度深度耦合，共同推动人类进入一个崭新的社会形态，可称之为“智能社会”。

智能社会的形成给教育带来新的压力：既然人工智能已经成为社会基础设施，学习人工智能知识、熟练使用人工智能就成为每个人必须完成的学习任务。

然而，这还远远不够。智能社会最重要的改变，不是拥有了大量高度智能化的设备，而是人类的认知环境发生了根本性变化：你看到的信息，有可能经过 AI 多重过滤，或者就是 AI 生成的；你做出的选择，有可能包含了系统对你潜移默化的引导；你发出的声音、做出的动作，很可能通过 AI 系统影响到你的朋友和家人。更严重的是，你在使用 AI 工具提高效率的同时，可能正在外包自己的思考能力；你在 AI 系统里听到的暖心的安慰与肯定，正在让自己陷入偏执。未来的人工智能，可能会预判你的想法，左右你的感情，控制你的心智。这不是某个人的处境，而是全人类要面对的认知困境。

由此，教育必须回应一个关键问题：在未来的智能社会，如何保持每个公民的“认知自治能力”。

认知自治并非简单的认知清醒；它的核心在于，每个人能做自己的主人，不被复杂的认知环境所裹挟，包括智能系统，也包括人。

认知自治的关键是“自主”：对智能时代的工具、设备、运行方式，主动接纳、主动学习、主动参与、主动运用，拥抱人工智能、驾驭人工智能。同时，还要能够合理评估自己的处境，理解信息、选择、动作背后的风险，并做好承担责任的准备。

从整个社会来看，认知自治是现代社会的基础，如果公民个体丧失了认知自治能力，整个社会的运行基础都将空心化。

那么，什么样的教育才能承接培养认知自治能力的时代使命呢?

1945 年，哈佛大学发布了一份名为《General Education in a Free Society》的报告，因封皮是纯红色，通常被称为“哈佛红皮书”。这份报告回应了当时教育面临的一个严峻问题：当社会分工越来越细致，专业化程度越来越强，如何建立整个社会的共同知识、共同判断和共同责任理解。报告详细阐述了哈佛大学的“共同教育方案”，通过设置文

化、艺术、科学等通识课程重建全社会的公共认知。这成为现代通识教育的重要源头之一。

今天，我们面对类似的情境：智能社会的运行基础和认知环境发生了巨大变化，人类如果无法保持认知上的自主性，不仅个体会丧失主体能力，整个社会的运行机制也会失去基础。我们需要另一本红皮书，为培养公民在智能社会的认知自治能力设计方案。

这正是这份报告诞生的时代背景。报告从通识教育的使命和人工智能的学科特性出发，构建了从人工智能教育到认知自治能力训练的可行路径，基于此建立起人工智能通识教育（AI General Education, AIGE）的理论基础和实现方案。报告还比较了人工智能通识教育与现有基础学科（数学、语言文学、物理、历史、计算机）在心智训练上的区别，分析了人工智能通识教育在培养认知自治能力上的特别优势。人工智能通识教育独立而独特的学科属性由此确立。

这份报告承袭了哈佛红皮书在关键历史节点对教育使命的思考传统，也承袭了其以纯红色为封面的风格，是人工智能时代的通识教育红皮书。下表列出了两本红皮书所面对的社会问题，特别是两者在维系社会运行基础上的共同贡献。这也再次明确了人工智能通识教育承担的历史使命：它既为个人的发展提供认知保障，也为整个社会的健康运行塑造公民基础。

哈佛红皮书面对的问题	**清华红皮书面对的问题**
专业分化和社会多样化削弱共同教育基础	算法中介、生成内容和认知外包削弱公民的判断主体性
现代社会需要能够思考、交流和判断的公民	智能社会需要能够调节信念、证据、信任、判断与责任的公民
通识教育维护现代社会的公共对话基础	人工智能通识教育维护现代民社会的认知主体基础
防止知识专业化造成公共判断的分裂	防止认知主体性削弱造成公共判断的空心化

第二部分

这本红皮书源于清华大学计算机系人工智能通识教育（AIGE）研究中心数年的持续探索，是一份关于智能社会人工智能通识教育使命的学术报告。中心成立于 2025 年 1 月，但在此之前，中心成员已经完成了大量相关工作。早期工作集中在大学人工智能教育和公众科普，出版了《人工智能导论》(林尧瑞，马少平，1989)、《人工智能》（马少平、朱小燕，2004）、《机器学习导论》（王东，2021）、《艾博士：深入浅出人工智能》(马少平，2023) 等专业著作和《图解人工智能》（王东、马少平，2023）等科普读物。2024 年以后，中小学人工智能通识教育工作逐渐展开。

2024 年 1 月，AIGE 中心开始整理人工智能通识教育方案。当时关于“人工智能教育如何开展”有各种声音。一些学者主张动手实践，通过软、硬件编程学习人工智能，

这是项目式学习的外延；一些学者主张大模型体验，教写提示词，这是当时大模型热潮在教育领域产生的影响；一些学者主张机器人项目，以赛促学，这是传统竞赛项目的思维顺延；一些学者主张将大学人工智能、机器学习课程简化，使之适合非计算机专业和中小学，这是大学里跨专业教学的常见方式。这些方式都有一定的合理性，但都无法达到“通识教育”的目标：或者覆盖程度不够，没有考虑到全体学生；或者课程内容偏置，不适合通识教育培养思维、训练认知能力的目标。

我们选择了一种全新的方案，就是这本红皮书给出的设计：以培养认知自治能力为统筹目标，以人工智能学科特性和思维方式为根基，建设面向全社会公民、面向未来智能社会的 AI 通识教育体系，即清华 AIGE 体系。

我们强调基础概念，建立科学的概念体系；关注基本方法，建立人工智能的知识图谱；介绍人工智能应用案例，解析背后的技术原理；讨论人工智能跨学科交叉成果，建立人工智能前沿视野。人工智能的科学精神、思维方式、伦理责任等内容贯穿于所有章节，引导认知自治能力的养成。

2025 年 4 月，基于清华 AIGE 体系的《清华大中小学人工智能通识教育》小、初、高三册读本出版，同步开源讲义、教案、学案、实践手册。我们召集了 300 多所实践校，基于这套教学资料展开教学研究，优化教学方法；我们与 200 多位认证讲师一起工作，共建课程资料；我们与腾讯支教平台合作，在全国 2000 余所乡村小学开展远程双师教学。

2026 年 8 月，在前期探索基础上，清华 AIGE 体系推出《清华大中小学人工智能通识教育》分年级版，包括小学三年级到六年级，初中一、二年级，高中一、二年级，每学期一册，共 16 册（2025 年已开始在河南省和安徽省率先实践）。这一版本形成了小学、初中、高中三轮螺旋式培养方案，针对认知自治能力培养的学习内容和练习内容大幅增加。

上述这些学习资料的设计思路、基本准则在本书的第三、第四篇具体阐述；本书的附录也展示了这些资料的总体结构。需要强调的是，人工智能通识教育是一个开放领域，在保证“培养认知自治能力”这一总体目标的前提下，可以容纳多种设计方案。因此，清华 AIGE 体系所给出的方案只是众多可能方案中的一种，只不过它已经经过了大量真实课堂的检验，可靠性会高一些。

最后，回到这本红皮书。张钹教授是清华 AIGE 体系的领路人，也是报告中众多重要思想的提出者，包括人工智能的定义、认知自治的完整内涵、基于人工智能科学属性的教育学论证路径等。没有张钹教授就没有这本红皮书。还要感谢参与教学实践活动的老师们，他们的一线教学经验让清华 AIGE 体系走向严谨、系统、可行、可靠，也让我们有信心把过往的实践探索总结成这本红皮书。

总之，这本红皮书不是哪个人的作品，而是清华 AIGE 体系的全体参与者对人工智能时代的集体贡献。

编者
于清华大学
2026 年 8 月

目录

总论：人工智能通识教育 – 智能社会中的认知自治能力培养

人工智能已经进入学习、生活、工作、公共治理等广泛领域。搜索与推荐系统影响人们首先看到什么，生成式模型参与交流与创作，自动评价系统进入组织决策，智能体开始调用工具并连续执行任务。人工智能由此具有两种相互关联的身份：它既是人们日常使用的基础工具，也在持续参与人的认知过程。

这种变化首先向教育提出了新的要求。社会成员需要知道人工智能是什么，了解它如何形成能力，能够使用相关系统解决现实问题，也需要理解人工智能进入社会以后带来的风险、伦理和责任问题。然而，知识普及、工具操作和使用规则还不足以完整回答人工智能时代的教育问题。更深的变化发生在人的认知活动之中：答案越来越容易获得，依据却未必清楚；任务越来越容易完成，理解却未必同步形成；系统能够提供预测和建议，责任仍需由人来承担。

教育因而需要回答一组更根本的问题：当人工智能参与问题提出、材料选择、证据组织、观点表达和行动决策时，人如何积极利用人工智能成果，同时保持自己的理解与判断？当模型输出形式完整、语言流畅的内容，却可能包含偏差、错误或虚构的资料时，人应当怎样分配信任？当一种方法从熟悉场景进入新对象、新群体和新环境时，人如何识别其成立的条件和失效的边界？当人工智能参与高风险行动时，谁应负责核查、解释和补救？

这些问题共同指向一个更基本的问题：智能社会需要培养什么样的人。人工智能通识教育的出现，正是对这一问题的历史回应。

0.1 历史参照：从共同教育基础到认知自治

进入现代社会以后，社会分工逐渐深化，教育也趋向专业化。专业教育提高了职业训练的水平，同时也带来知识分散、公共判断弱化和教育目标窄化等问题。通识教育因此诞生，目的是重建共同教育基础，使受教育者在专业身份之外仍具有共同理解、公共判断和共同规范。

1945 年出版的《自由社会中的通识教育》（*General Education in a Free Society*）集中体现了这一历史关切 (Harvard Committee, 1945)。这份报告通常被称为“哈佛红皮书”。它试图回答：在专业分工不断加强的现代社会中，学习者还需要接受怎样的共同教育，才能形成有效的思考、清晰的表达、清醒的判断和共同的价值理解。它所提出的

课程方案显然带有特定时代的文化背景，但其提出的命题至今仍具有基础意义：现代社会需要每个成员关注个人专长，同时也需要全体成员共享基本知识、心智能力和责任意识。

人工智能时代延续了这一历史命题，并改变了问题呈现的社会场景。过去的通识教育主要面对学科分化和共同文化基础的问题；今天，智能化系统深刻改变了信息呈现的方式、知识形成过程和社会行动的规则。个体既依赖教师、专家、机构，也越来越多地依赖搜索、推荐、生成和智能决策系统。共同教育基础因而需要延伸到人如何理解这些中介系统、如何主动探索合理的使用方法、如何校准对它们的信任、如何在信息依赖中保持判断和责任，从而保持人的主体性。认知自治正是对这一问题的概括。

0.2 认知自治：通识教育的现代命题

0.2.1 认知自治的定义

认知自治可以定义为：**个体在复杂认知环境中，对自身信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节的过程**。认知自治能力，则是个体发起、维持和调整这一过程，使其能够跨越不同任务与情境持续展开的能力。

认知自治建立在开放依赖之中。现代社会的知识高度分工，任何人都需要依赖教师、专家、机构、媒体、数据库、技术工具和社会协作 (Hardwig, 1985; Hutchins, 1995)。认知自治所强调的自主，体现为个体在充分利用外部知识和智能系统的同时，仍能追问、解释、比较、核验、暂停和修正。个体需要知道自己依赖了什么，理解这种依赖在何种条件下合理，能够发现依赖的边界与风险，并在必要时调整或收回信任。

开放依赖中的认知主导，主要体现为个体仍然掌握认知活动的发起、审查和统摄。提出什么问题，决定认知活动的目标与意义；如何组织人、工具和资料，决定外部能力在过程中承担什么角色；所得结果能否接受，则需要由人依据证据、边界和风险加以核查。人工智能可以帮助发现问题、扩展问题、搜索信息、执行计算、生成方案和验证结果，人仍需主导问题的选择与界定，并对结果及其行动后果作出最终判断。人的主导地位体现为对认知活动方向、调节和有效性的掌握，与亲自完成多少步骤并无直接对应关系。

认知自治包含六个相互关联的分析维度。来源警觉关注结论由谁或什么系统产生，以及其专业能力、激励结构和传播路径；证据敏感关注结论受到何种材料、数据、实验或论证支持；边界意识关注样本、经验和模型是否适用于当前对象、群体、时间与场景；信任校准关注应当根据证据强度、任务风险和错误后果给予多大程度的信任；认知能动关注个体能否从自身问题与目标出发，主动探索和调用人工智能以扩展学习、创造、判断与行动；责任判断关注结论进入行动以后，谁应当核查、决定、监督、解释和补救。

为了进入课程、教学和评价，本书进一步提出归纳、泛化、判断与自醒四种能力。归纳使学习者从有限材料中形成可检验的理解；泛化使学习者把已有理解迁移到新情境并识别边界；判断使学习者在证据不足、后果不确定和价值冲突中作出有理由的选择；自醒使学习者持续觉察自身的认知状态、信任状态和责任状态。四种能力形成循环：归

纳产生假设，泛化检验边界，判断做出选择，自醒校正全过程。

六个维度与四种能力处于不同层次。六个维度提供分析认知自治过程的基本框架，四种能力提供组织课程任务、发展进阶和评价证据的教育框架。两套结构相互支撑，共同构成全书对认知自治及认知自治能力的完整表述。

0.2.2 认知自治的社会意义

认知自治既关系个体能否主导自身的认识与行动，也关系社会能否拥有真正能够参与公共生活的公民。现代社会把越来越多的公共事务交由公民讨论、判断和共同决定，这一制度安排以公民能够形成、表达、检验和修正自身判断为前提。公民需要理解事实与主张的来源，评价证据及其边界，合理分配对专家、机构和技术系统的信任，并对自己的表达、选择和行动承担责任。认知自治因此具有超越个体发展的公共意义，是公民参与公共事务的前提。

人工智能使这一问题变得更加迫切。生成模型、推荐算法和自动化决策系统正在进入议题选择、信息获取、理由组织、意见表达和行动实施等环节。它们能够扩展公民接触信息、分析问题和参与公共生活的能力，也可能使人的判断在不易觉察的情况下受到生成、筛选和引导。如果公民无法追踪信息来源、核查证据、识别适用边界、校准系统信任并承担最终责任，社会中流通的意见仍可能十分丰富，其背后的公民主体却可能逐渐空心化。保持认知自治，既是个体在智能社会中保持主体性的必要条件，也是现代社会持续运行的认知基础。

0.3 人工智能通识教育

0.3.1 人工智能通识教育的定义

认知自治确立了智能社会通识教育的时代使命，也为具体教育形态规定了目标。在这一目标之下，接下来需要回答什么样的教育能够承担这一使命。人工智能是智能社会的直接塑造者，通过学习人工智能来培养认知自治能力是一种自然选择。通识教育与人工智能教育在这里交汇，成为人工智能通识教育的共同源头。本书将人工智能通识教育定义为：

人工智能通识教育是通识教育与人工智能教育相融合形成的教育形态。它以人工智能的知识、方法、系统及其社会运行为学习内容，承担通识教育在智能社会中培养认知自治能力的使命。

从通识教育的角度看，人工智能通识教育是一种以教授人工智能的方式培养认知自治能力的特殊通识教育。从人工智能教育的角度看，人工智能通识教育又是一种以通识教育目标引领的特殊人工智能教育。

认知自治在这一融合中发挥统摄作用。学习人工智能使学习者能够理解数据、模型、目标、评价、泛化、误差和反馈等系统环节，认识人工智能的能力来源与适用边界；分

析人工智能的社会运行，则使学习者看到平台、组织和制度如何借助人工智能进行信息筛选、判断形成和行动实施。知识学习、能力发展和责任意识由此围绕培养认知自治能力得到组织。

0.3.2 人工智能通识教育的内容基础与认知训练机制

确立认知自治的教育目标以后，还需要回答一个基础问题：人工智能通识教育为什么能够发展认知自治能力？这一问题的答案蕴含在人工智能学科本身及其进入社会活动的过程之中，即智能活动的计算化与社会活动的计算化。

人工智能首先把学习、推理、判断、创造和行动等抽象的智能活动，转化为可以研究和构建的计算过程。在这一过程中，现实经验需要被选择和表示，任务目标需要得到明确，系统需要依据数据、知识或反馈形成能力，所得结果还需要在新的对象和环境中接受检验。经验从哪里获得，对象怎样表示，规律怎样形成，结论能否迁移，误差如何发现，方法在什么条件下成立，都由隐含的认知问题转化为可以观察、比较和讨论的系统环节。

人工智能通识课可以把样本、表示、目标、模型、证据、反馈和边界组织为可以分析和检验的系统环节。学习者由此可以观察一个人工智能系统如何从有限经验中形成结论，如何把已有能力用于新的情境，如何依据目标作出选择，以及如何暴露错误和边界。重要的是，学习这些过程不仅可以理解人工智能系统，也可以用来反观人的认知特性。这些计算实现与人的内部认知机制可以存在显著差异，却共享若干关于经验、证据、迁移和校正的问题结构。智能活动的计算化因此为反观人的认知提供了构建性的参照，使认知形成中的基本问题获得外在、可分析的形式。

人工智能形成计算能力以后，还会进一步进入知识生产、信息传播、组织判断和社会行动。现实对象转化为数据，决定系统能够看见什么；历史经验转化为模型，影响系统依据什么形成判断；社会目标转化为指标和约束，规定系统趋向什么。模型输出随后接入工作流程，并通过平台、组织和制度获得现实效力，在规模复制、系统连接和社会反馈中形成持续影响。理解了这一过程，就可以进一步确定：哪些现实进入了数据，哪些经验沉淀为模型倾向，谁设定了系统目标，模型输出获得了多大的行动权限，受影响者能否核查、否决、修正和申诉，相关责任又怎样在开发者、部署者、使用者和制度之间分配。

社会活动的计算化使人工智能背后的来源、目标、分工、权力、制度和责任成为可以追踪的结构。它为认知自治教育提供了具体的社会化材料与证据，使学习者能够分析计算系统如何介入人的认识和行动，以及一种计算结果如何在特定组织和制度条件下转化为现实后果。

两种计算化共同构成人工智能通识教育的基本学习对象。智能活动的计算化主要揭示结论如何形成，社会活动的计算化进一步揭示结论在什么社会条件下形成、传播并获得现实效力。前者显化认知形成中的基本问题，后者显化群体认知活动的社会组织方式。二者相互连接，使人工智能教育能够同时进入个体认知与社会认知两个层面。

重要的是，这些内容需要经过有意识的课程组织，才能转化为认知自治能力。课程

既要帮助学习者理解人工智能如何形成能力，也要引导其追问数据来源、证据强度、目标设定、适用条件、社会效力和责任结构，并要求学习者在使用人工智能的过程中进行比较、验证、解释、修正和反思。人工智能由此同时成为学习对象、实践工具和反思材料，使认知自治获得可理解、可体验、可检验和可评价的训练过程。

0.3.3 人工智能通识教育在认知训练上的独特性

不同学科都承担着发展学生认知能力的任务，其培养重点因研究对象和知识结构而有所差异。数学突出抽象、演绎、形式化；物理学突出观察、实验、归纳和因果解释；语言与人文学科突出意义理解、表达、证据和历史情境；计算机科学突出问题分解、算法化、可执行性和系统思维；伦理教育突出价值辨析、规范意识和责任判断。

人工智能通识教育承担通识教育在智能社会中培养认知自治能力的使命，尤其关注个体在模型中介、来源混合和结果不确定的信息环境中，能否保持对自己认识活动的理解和调节，能否在合理的安全边界内保持积极探索，能否合理分配对外部系统的信任，能否识别结论边界，并能否对最终判断和行动承担责任。它把智能社会中的认知主体地位确立为集中的教育目标。

各学科的认知训练相互联系。人工智能通识教育同样需要数学推理、科学证据、语言表达、历史理解和价值判断，其他学科也会促进学习者形成自主思考和责任意识。人工智能通识教育的特殊性，来自其学习对象同时贯通认知活动与社会活动：学习者既要理解智能行为怎样被转化为可以构建和检验的计算过程，也要理解这些计算过程怎样进入社会并影响知识、判断和行动。这为培养认知自治能力提供了体系化的知识内容和针对性的思维训练。

人工智能通识教育由此在智能社会中承接了通识教育维护共同认知基础的历史使命。它与数学、语言文学、物理、历史、计算机科学和伦理教育共同发展人的认知能力，同时集中回应人工智能参与知识形成、信息传播、社会判断和现实行动以后出现的认知问题。

0.4 从认知目标到五部分内容体系

智能活动的计算化与社会活动的计算化，规定了人工智能通识教育需要理解的基本对象。学习者既要认识人工智能是什么、怎样形成能力、如何构成现实系统，也要理解人工智能怎样进入不同学科和社会活动，以及人应当如何使用、评价、监督和治理这些系统。两种计算化提供内容来源，认知自治则规定这些内容应当怎样选择和组织。

为了把这些基本对象转化为稳定、连贯并适合教学的课程体系，人工智能通识教育需要回答五类相互衔接的问题。

1. 人工智能是什么，从何而来，边界在哪里；
2. 人工智能如何表示、学习、推理、生成和行动；
3. 人工智能如何形成现实系统，在什么条件下有效或失效；
4. 人工智能如何进入不同学科并参与知识生产；

5. 人应当如何使用、监督和治理人工智能，并承担相应责任。

这五类问题分别导出人工智能基础概念、人工智能方法、人工智能应用、人工智能交叉融合以及风险、伦理与责任五部分内容。治理作为把价值和责任转化为制度、流程与技术机制的方式，纳入第五部分展开。

基础概念建立人工智能学科的概念坐标，帮助学习者区分人工智能与自动化、机器人、算法、互联网、大数据、机器学习、深度学习和生成式人工智能，理解当前技术在长期发展中的位置。

人工智能方法建立解释系统能力的知识骨架。学习者需要理解基于知识和基于学习的两条基本路径，理解目标、模型、算法、数据和知识之间的关系，理解训练、测试、模型选择、部署监测和反馈更新的完整过程，并将深度学习、大模型、多模态模型和智能体放入这一体系中定位。

人工智能应用把抽象原理还原为现实系统。机器视觉、机器听觉、语言处理、人机对战与具身行动、搜索与推荐等典型应用，可以沿“任务与场景—数据与表示—模型与方法—系统集成—输出与评价—失效条件—风险与责任”的框架进行分析。学习者由此形成解释新应用的共同方法。

人工智能交叉融合讨论人工智能如何进入数学、工程学、物质科学、生物与医学、地球与空间以及社会、文化与艺术等领域。课程需要呈现“领域问题—形式化与表示—数据和知识—方法迁移—领域验证—解释与责任”的完整链条，使学习者理解通用方法进入不同学科时仍需接受领域知识、证据标准和专业责任的约束。

风险、伦理与责任既构成独立内容，也贯穿其他四部分。风险描述在不确定条件下可能发生的伤害，伦理讨论应当维护的价值，责任明确谁应预防、监督、解释和补救，治理则把这些要求转化为制度、流程和技术机制。隐私、公平、真实性、安全、人的主体性和社会责任需要进入数据选择、模型评价、应用分析和领域验证的全过程。

0.5 分层实施：从学校教育到终身学习

人工智能通识教育面向全体学习者，共同目标需要通过符合认知发展、专业背景和现实需要的课程形态来实现。各阶段共享五部分内容和认知自治目标，内容深度、任务复杂度、实践开放性和责任要求逐步提高。

0.5.1 中小学：横向五部分内容与纵向三阶段

中小学课程采用“横向五部分内容、纵向三阶段递进”的结构。基础概念、人工智能方法、人工智能应用、交叉融合以及风险、伦理与责任贯穿小学、初中和高中，保证课程内容的完整性；兴趣培养、体系构建和知识拓展构成三个学段的主导任务，保证学习过程的连续性。

小学阶段以兴趣培养和科学精神为主。课程通过故事、生活场景、观察、体验和讨论，使儿童愿意接近人工智能，同时接受初步认知训练。学生需要知道人工智能由人设计，人工智能可以帮助人也可能出错，重要信息需要检查，个人信息需要保护，使用人

工智能应当遵守安全、隐私和诚实原则。复杂算法和开放式工具的使用应当服从儿童发展与安全需要。

初中阶段以体系构建和全局视野为主。课程把零散体验组织成人工智能的历史脉络、概念关系和方法结构，以“数据—模型—输出—评价”为技术主线。学生逐步能够解释简单系统、分析数据影响、识别输出误差、评价结果并完成项目化问题解决。编程可以作为拓展路径，系统理解、证据意识和责任意识构成共同基础。认知自治在这一阶段表现为机制化的判断框架。

高中阶段以知识拓展和创新实践为主。学生可以深入理解深度学习、生成式人工智能、多模态模型、智能体和人机协同等关键技术，分析典型应用背后的机制与演进，进入不同学科的前沿问题，并通过研究性项目和复杂案例形成跨域迁移、证据分析、专业兴趣和公共责任。认知自治在这一阶段表现为负责任的认知行动。

三个学段形成由亲近并质疑人工智能、理解并评价人工智能，到负责任地使用和创造人工智能的发展路径。同一概念和案例可以螺旋重访，其抽象程度、系统复杂度和判断要求随学段提高。

0.5.2 大学：方法理解、专业迁移与跨学科协作

大学阶段在基础教育形成的基本认知之上，进一步发展人工智能方法理解、专业迁移和跨学科协作。学生需要系统认识人工智能的基础思维方式和研究范式，理解计算、数据归纳、知识与数据结合以及实验验证在人工智能中的作用，并把人工智能放入本专业的问题体系、证据标准和责任结构中。

大学课程可以由人工智能基础方法和人工智能与学科交叉融合两个主要部分构成。基础方法提高理论深度，交叉融合围绕数学、工程学、物质科学、生物与医学、地球与空间、社会文化与艺术等方向组织代表性案例。每个案例都需要分析人工智能进入领域的环节、成立条件、验证标准和责任边界。

大学阶段的认知自治继续沿用归纳、泛化、判断与自醒四种能力，其训练对象从日常信息和一般问题扩展到专业知识的形成、检验和应用。当前学生的人工智能基础差异较大，过渡阶段可沿用高中阶段“概念—方法—应用—融合”的课程结构，参考高中知识安排并提高分析深度，逐步把更多学习时间转向方法深化、论文研读、专业案例、跨学科实践和创新研究。

0.5.3 终身学习：共同核心、情境模块与前沿更新

终身学习贯穿成年以后的职业发展、家庭生活、社会参与和退休生活。它涵盖在职人员、职业转型者、家庭学习者、老年学习者、普通公众以及数字能力薄弱或具有特殊需要的学习者，具有多次进入、路径弹性和持续更新的特点。

终身学习课程采用“共同核心—情境模块—前沿更新”的结构。共同核心建立人工智能与智能社会、身边的人工智能系统、基础方法、大模型与智能体、风险、伦理与责任等稳定认识；情境模块连接学习与信息获取、工作与职业发展、家庭生活、健康养老

与照护、文化创造和专业领域融合；前沿更新吸收新方法、新产品、新应用和新的社会问题，帮助学习者修正已有认识。

成人学习可以从现实场景进入，以基本原理支撑操作和判断，通过“现实问题—人工智能任务—基本原理—实际操作—结果核查—迁移与反思”形成持续学习过程。职业群体需要加强工作流程、人机协作、组织数据保护和专业责任；老年群体需要兼顾生活便利、健康照护、防诈、社会联系和主体性；数字基础薄弱和具有特殊需要的学习者需要低门槛材料、线下支持、无障碍设计、本地语言支持和人工服务保留。科普活动可以成为公共入口和动态更新机制，并与连续课程、社区学习和实践活动相连接。

0.6 全书结构与论证路径

全书按照“通识教育的历史使命如何在智能社会中转化为新的现代任务—认知自治如何概括这一任务—人工智能通识教育如何作出教育响应—人工智能为什么能够训练相应能力—课程应当教什么—不同阶段与群体如何实施”的顺序展开。

第一篇从通识教育在智能社会中的现代任务出发，逐步说明认知自治何以成为第一性的教育目标，以及人工智能通识教育如何在此基础上形成。第一章回到现代通识教育的历史使命，说明通识教育的共同基础如何回应专业化和社会复杂化；第二章从相关研究出发界定智能社会及其基本特征，并从七个方面分析其认知环境；第三章界定认知自治与认知自治能力，提出来源警觉、证据敏感、边界意识、信任校准、认知能动和责任判断六个分析维度，说明认知自治何以成为智能社会通识教育的时代使命；第四章界定人工智能通识教育及其与人工智能教育、人工智能素养、认知自治和认知自治能力的关系，分析现实实践中的主要偏差与政策、能力框架中的认知转向。四章内容构成了对人工智能通识教育历史定位的完整论证。

第二篇说明人工智能通识教育为什么能够并且应当承担培养认知自治能力的任务。第五章考察人工智能的研究对象、科学体系和基本思维方式，揭示人工智能如何把抽象、内隐的智能活动转化为可观察的任务、可计算的表示、可构建的机制和可检验的系统；第六章分析现实对象、历史经验和社会目标如何分别进入数据、模型和指标，模型输出又怎样经由工作流程、组织制度、规模复制、系统连接和社会反馈转化为社会力量。两章分别讨论智能活动的计算化与社会活动的计算化，共同建立人工智能通识教育的科学基础。第七章通过理解智能机制、参与计算建构和分析社会嵌入三条路径，把这一基础转化为归纳、泛化、判断与自醒四能力模型及其课程、评价方式；第八章比较数学、语言与文学、物理、历史和计算机科学的心智训练，说明人工智能通识教育把这些认知资源组织到“数据—模型—输出—行动—反馈”的完整链条中，由此形成相对独立的心智训练价值。

第三篇回答人工智能通识教育应当教什么。第九至第十三章依次讨论人工智能基础概念、人工智能方法、人工智能应用、人工智能交叉融合以及风险、伦理与责任。每一部分都围绕教育功能、内容原则、共同主干、拓展内容和认知自治展开，使快速变化的技术被组织进相对稳定的知识结构。

第四篇讨论不同阶段和群体如何实施。中小学课程采用横向五部分内容与纵向三阶

段递进的结构；大学课程突出方法理解、专业迁移和跨学科协作；终身学习采用共同核心、情境模块和前沿更新的组合结构，并根据生活阶段、现实任务、数字基础和特殊需要提供差异化支持。

在附件部分，本书提供了清华大中小学人工智能通识教育研究组为中小学、大学和终身学习三个阶段设计的教材结构与实践项目。这些内容汇集了已有课程资源和实施形态，为后续按照认知自治目标重组知识、任务、能力与评价证据提供参考。

由此，全书形成一条完整的逻辑主线：

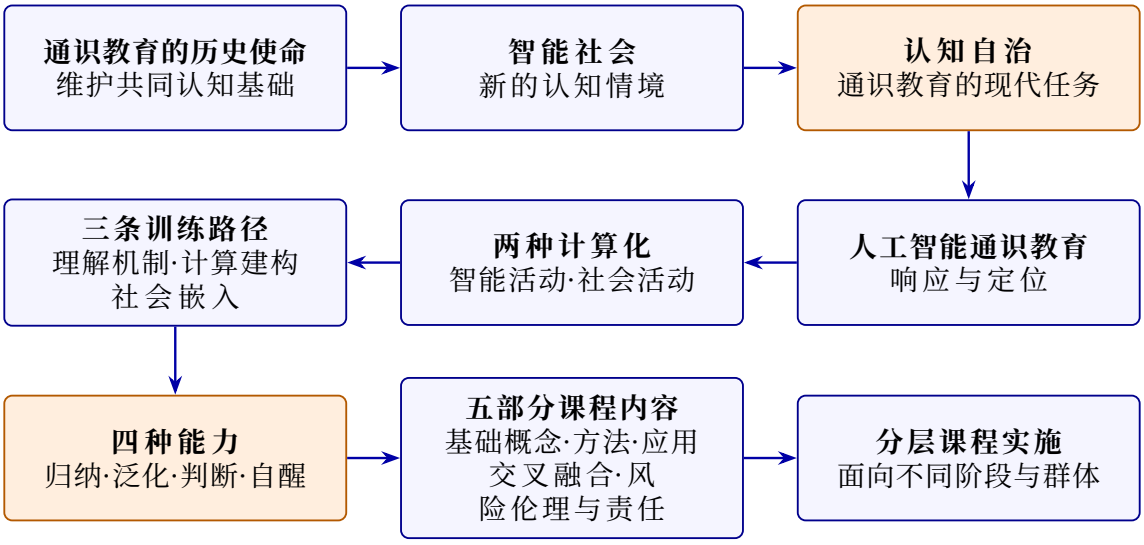

图 1: 人工智能通识教育的总体逻辑

0.7 结语：智能社会中的自由人格

人工智能正在扩展人的能力，也在重新组织人的认知环境。教育需要帮助学习者理解和使用这些新能力，同时守住人在智能社会中的认知主体地位。一个能够熟练使用人工智能却无法解释输出的人，仍可能处于深度认知依赖之中；一个能够迅速生成作品却不能重建证据和思路的人，尚未完成相应的学习；一个把重要选择交给系统却无法说明理由和后果的人，也难以承担现实责任。

人工智能通识教育最终培养的是具有认知自治能力的人。这样的人能够从有限材料中形成可检验的理解，能够把认识迁移到新情境并识别边界，能够在证据不足和价值复杂的条件下作出有理由的选择，也能够持续觉察自己的信任、依赖、偏见和责任。他愿意借助人工智能扩展能力，也能够在需要时核查、修正、暂停或拒绝；他能够主动参与人机协作，也能够说明最终判断的依据并承担行动后果。

从现代分工社会中的共同教育到智能社会中的认知自治，通识教育的历史使命在新的技术条件下得到延伸。人工智能越深入地参与知识与行动，教育越需要把人的理解、判断、反思和责任置于中心。人工智能通识教育由此成为面向全体社会成员的基础性教育：它帮助人认识、理解人工智能，也帮助人在智能社会中保持独立思考、审慎信任和负责任的公共参与，从而塑造不被外部环境引导和裹挟的自由人格。

第一篇

时代背景：为什么人工智能通识教育成为必需

本篇导言

通识教育始终承担着一项基础使命：在知识分化和社会变化中，为社会成员建立共同的知识基础、心智能力、价值意识和公共责任。1945 年的《哈佛红皮书》正是在战争、科学进步、专业化发展和公共秩序重建相互交织的背景下，重新提出现代社会需要怎样的共同教育。今天，人工智能的发展使这一历史问题获得了新的时代内容。

人工智能技术与系统正在持续进入知识生产、信息传播、教育学习、劳动组织、公共治理和日常生活。数据、模型、平台和制度共同参与现实的表征、信息的筛选、内容的生成、判断的形成与行动的实施。知识生产、信息传播、判断形成和社会行动等基础社会活动越来越多地通过人工智能系统与制度共同中介，智能社会由此逐渐形成。

智能社会扩展了人的认知能力，也改变了人所处的认知环境。信息和知识更容易获得，来源混杂、证据不清、生成内容失真、算法中介、自动化评价和责任边界模糊等问题也更加突出。个体需要在开放使用人工智能和其他外部资源的同时，保持对自身认识和行动的主动调节。本书将这一要求概括为认知自治：个体在复杂认知环境中，对自身信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节的过程。

培养公民的认知自治能力由此成为智能社会通识教育的重要目标，而人工智能教育是实现这一目标的天然途径。通识教育和人工智能教育因此汇合，形成“人工智能通识教育”独特的学科定位：它以人工智能的知识、方法、系统及其社会运行为学习内容，承担通识教育在智能社会中培养认知自治能力的使命。

本篇沿着这一逻辑展开。第一章回到通识教育的发展历程，说明其重建共同教育基础的历史使命；第二章界定智能社会，分析人工智能带来的新型认知环境；第三章界定认知自治，说明其作为智能社会通识教育时代使命的基本内涵；第四章界定人工智能通识教育及其与人工智能教育、人工智能素养、认知自治和认知自治能力的关系，并分析现实实践中的主要偏差与发展转向。

第一章　通识教育的历史使命：从专业化危机到共同教育基础

摘要：本章讨论通识教育何以在现代教育史中反复成为核心议题。通识教育的出现和重构，常常对应社会结构、知识制度和大学功能的深刻变化。现代专业化教育提升了知识生产效率和职业训练能力，同时带来了知识割裂、公共判断弱化和教育目标窄化等问题。通识教育由此承担起重建共同教育基础的任务：帮助受教育者在专业身份之外形成共同语言、共同理性、共同的价值判断。本章以博雅教育传统、现代大学制度、杜威的公民教育思想、怀特海的教育节奏观、芝加哥与哥伦比亚的核心课程实践、1945 年哈佛红皮书等为线索，分析通识教育在现代社会中的历史使命。章末进一步指出，人工智能时代的通识教育需要继承这一传统，并将共同教育基础从工业社会的文化共同体维度推进到智能社会中的认知自治问题。

关键词：通识教育；博雅教育；专业化；共同教育基础；哈佛红皮书；认知自治

1.1　通识教育何以成为现代教育问题

通识教育之所以成为现代教育史上的反复议题，根源在于社会对“受教育者应当成为什么样的人”的持续追问。教育制度一旦走向规模化、专业化和职业化，就必须处理一个基本矛盾：社会需要大量具有专业能力的人，又需要这些人在具体职业之外具备共同语言、共同理解、共同价值判断。前者推动学校不断细分学科、设立专业、强化技能训练；后者要求教育保留对人的完整发展和公共理性的关注。通识教育正是在这一矛盾中获得了自身的独特使命。

从历史上看，通识教育的形态多次改变。古典博雅教育强调语法、修辞、逻辑、数学、音乐、几何与天文等训练，目标在于培养自由人参与公共生活的能力。近代大学兴起之后，科学研究、专业训练和国家建设逐渐进入大学的核心体系，传统博雅教育开始受到现代学科制度的重塑。十九世纪以来，研究型大学、专业学院和职业训练不断扩张，教育越来越多地承担起培养工程师、医生、律师、教师、管理者和科研人员的任务。现代教育因此获得了强大的社会功能，也付出了知识割裂和目标窄化的代价。

通识教育的出现，正是为了回应现代教育体系的这种内部张力。它关心受教育者能否在专业身份之外理解世界，能否将局部知识放入更广阔的历史、科学、社会和伦理背景中，能否与不同知识背景的人进行理性交流，能否在事实、价值和行动交织的公共问题中作出判断。Newman (1996) 在十九世纪关于大学理念的讨论中，将大学教育理解为

心智的训练和理智习惯的形成，而非某种功利性的职业准备。Dewey (1916) 则把教育与公民生活联系起来，强调学校应当培养个体参与公共生活的能力。Whitehead (1929) 批评“惰性知识”对教育的损害，强调知识必须与理解、想象和应用结合起来。这些思想虽然产生于不同历史语境，却共同指向一个核心判断：教育的目的不能被短期技能和孤立知识所全盘占据。

现代通识教育的特殊处境还在于，它总是在两种压力之间寻找平衡。一方面，社会生产和知识发展要求学生尽早进入专业领域，掌握可验证、可操作、可迁移到职业场景中的能力；另一方面，社会复杂性越高，个体越需要跨越单一专业的共同判断能力。专业教育可以使人有效地处理局部问题，通识教育则帮助人理解局部问题在整体社会中的位置。专业知识提供工具，通识教育提供方向、边界和责任意识。二者之间的关系应当被理解为互相支撑，而非互相排斥。

在本书的论证中，第一章承担一个基础任务：我们将努力阐明，通识教育远远超出课程表上的附属安排，它是现代社会在专业化、公民化和技术化条件下对人的共同能力的制度性回应。只有先理解这一点，才能进一步理解人工智能时代为什么需要一种新的通识教育，以及为什么人工智能通识教育的核心目标应当指向智能社会中的认知自治。

1.2 专业化的胜利及其教育后果

现代大学的兴起与专业化密不可分。Humboldt (1810) 关于大学的构想强调研究与教学的统一、学术自由和知识探索的内在价值，对十九世纪以来的研究型大学产生深远影响。此后，现代学科逐渐形成稳定制度，学术期刊、实验室、研究生培养、专业协会和职业资格制度共同塑造了现代知识生产体系与人才塑造流程。专业化提升了知识增长速度，也强化了高等教育对工业化、国家治理和现代职业体系的支撑。

这种专业化具有明确的历史合理性。现代医学需要系统的解剖学、病理学、临床训练和证据标准；现代工程需要数学、物理、材料、机械、控制和设计知识；现代法律、金融、教育、公共管理等职业，也各自形成复杂的专业规范和制度体系。若没有专业化，现代社会难以获得高水平的技术能力和组织能力。Flexner (1930) 对大学的分析表明，现代大学必须服务于知识的高水平探究，同时避免被短期实用主义完全支配。这提醒我们，专业化本身并非教育问题的根源，关键在于专业化是否压缩了人对整体、价值和公共责任的理解。

Veblen (1918) 在对美国高等教育的批判中已经指出，大学容易在商业力量、职业利益和制度扩张中偏离学术使命。Kerr (2001) 在“多元大学”的描述中进一步揭示，二十世纪大学已经成为研究、教学、政府项目、产业合作和社会服务等多重功能汇集的复杂组织。这种多功能结构扩大了大学的社会影响，也使教育目标变得更加分散。学生在大学中获得越来越多选择，大学也提供越来越多专业路径；与此同时，共同学习经验和共同判断基础容易被削弱。

专业化造成的第一个教育后果，是知识经验的碎片化。学生可能在某一专业中掌握高深知识，却缺少理解其他领域问题的基本语言。工程学生可能熟悉技术设计，却缺少社会影响和伦理责任的训练；人文学生可能具备文本解释能力，却缺少基本科学和数据

素养；社会科学学生可能能够分析制度和行为，却对技术系统内部机制理解不足。知识碎片化使跨领域交流变得困难，也削弱了个体面对复杂问题时的整体判断。

第二个后果，是教育目的的工具化。专业化一旦被就业逻辑单独驱动，学校很容易把教育目标收缩为职业准备和能力证书。学生也可能把学习理解为获取进入某一职业通道的资本，而非形成稳定的心智能力和人格结构。Nussbaum (2010) 在关于人文教育和社会活动的讨论中提醒，过度功利化的教育会削弱同情、想象、批判和公共责任等公共生活所需能力。这种批评对人工智能时代同样具有意义：技术能力越强，教育越需要守住人的判断力和责任感。

第三个后果，是公共判断的弱化。现代社会中的许多问题都无法被单一专业解决，例如气候变化、公共卫生、人工智能治理、能源转型、城市规划、教育公平和生物伦理。这些问题同时涉及科学事实、统计证据、制度安排、文化意义、伦理价值和政治选择。若受教育者只接受高度专门化训练，他可能在自己的领域内精确有效，却难以在复杂公共问题中理解多方证据和价值冲突。通识教育因此具有公共理性训练的意义。

教育研究也支持这种担忧。Arum and Roksa (2011) 对美国大学生学习结果的研究曾指出，相当一部分大学生在批判性思维、复杂推理和写作能力方面提升有限。Pascarella and Terenzini (2005) 关于大学影响的大规模综述也显示，大学对学生发展的影响取决于课程结构、师生互动、挑战性任务和整合性学习经验等多重因素。这些研究并不否定专业教育价值，却提示我们，专业课程本身不会自动生成通识能力。共同学习、反思、表达、证据评估和跨情境迁移需要有意识的课程设计。

因此，通识教育的现代使命可以初步概括为：**在专业化的社会中维持人的整体性，在知识分化的制度中重建共同理解，在个性化发展之外保留公共判断，在技术效率之外强调价值和责任**。这一使命构成后来很多教育改革的统一背景。

1.3 博雅教育、公民教育与共同学习的思想谱系

现代通识教育并非单一理论的产物。它继承了博雅教育、公民教育、科学教育、核心课程等多条思想谱系。不同传统之间常有张力，但它们共同构成通识教育的思想资源。

博雅教育传统强调心智的自由发展。Newman (1996) 认为，大学教育的价值在于形成一种普遍的理智训练，使人获得清晰、准确、审慎和有秩序的思维习惯。这一传统重视语言、逻辑、哲学和历史等训练，强调知识之间的内在联系。其优势在于维护教育的非功利维度，提醒大学不能完全服从即时职业需要。但其局限也很明显：如果博雅教育过度依赖精英式的经典阅读和人文训练，就容易忽视现代科学、社会多样性和大众教育需求。

公民教育传统强调教育与共同生活之间的关系。Dewey (1916) 把公民活动理解为共同经验的交流方式，而学校则是组织经验、培养反思和参与能力的重要场所。在这一视角下，教育的意义不仅在个人修养，也在公共生活能力的形成。学生需要学习如何提出问题、协商方案、理解他人、检验证据、承担后果。通识教育由此获得强烈的社会维度：它培养人参与公共生活的能力，而公共生活需要比专业知识更广泛的理解和判断。

科学教育传统强调现代公民必须理解科学方法和科学精神。Whitehead (1929) 指

出，教育不能停留在惰性知识的堆积，知识必须进入理解、想象和应用的循环。Bruner (1960) 则强调任何学科都有其基本结构，教育应当帮助学生理解学科的核心观念和探究方式。这种思想对通识教育具有重要意义。通识教育的重点不在各学科的浅层介绍，而在帮助学生理解不同学科如何提出问题、形成证据、建立解释和检验结论。

核心课程传统则试图用制度化课程回应共同学习问题。哥伦比亚大学的“当代文明”课程起源于一战后的 1919 年，其目标在于让学生面对现代社会的紧迫问题，并在阅读、讨论和写作中形成思想表达能力 (Columbia College, 2026)。芝加哥大学的核心课程于 1931 年秋季实施，其“新计划”试图通过共同课程抵抗过度分散的选课制度，维持本科教育中的共同智识经验 (The College, University of Chicago, 2026)。这些实践说明，通识教育既是一种理念，也是一种课程制度。

可以将上述思想谱系概括为表1.1。这一概括无法穷尽通识教育的全部起源，但可以显示其内部的多重来源。

表 1.1: 通识教育的若干思想谱系

思想谱系	核心关切	代表性表达	对人工智能通识教育的启示
博雅教育传统	心智训练、理智习惯、知识整体性	纽曼关于大学理念的论述，强调理智训练和知识关联	AI 教育需要超出工具操作，指向人的理解能力和判断习惯
公民教育	共同生活、经验反思、公共参与	杜威将教育与公民经验联系起来	AI 教育需要服务于智能社会中的公共理性和责任参与
科学教育传统	学科结构、证据标准、探究方法	怀特海反对惰性知识，布鲁纳强调学科基本结构	AI 教育需要讲清数据、模型、证据、误差和方法边界
核心课程传统	共同学习经验、共同文本、共同问题	哥伦比亚和芝加哥核心课程实践	AI 教育需要建立面向全体学生的共同课程与共同问题意识

从这些谱系可以看到，通识教育并没有单一固定形式。它可以采取经典阅读、科学素养、社会问题、跨学科项目、写作讨论、核心课程等不同制度形态。其稳定内核在于：教育需要帮助个体越过局部训练，形成面向整体世界、公共生活和自我发展的能力。

1.4 哈佛红皮书：现代分工社会中的共同教育方案

1945 年出版的 *General Education in a Free Society*，通常被称为“哈佛红皮书”，是现代通识教育史上最重要的文件之一 (Harvard Committee, 1945)。这份报告由哈佛委

员会完成，詹姆斯·布莱恩特·科南特在序言中阐述了报告的教育关切。其历史背景十分特殊：第二次世界大战即将结束，美国社会正在面对战后公共秩序、科技发展、大众教育和大学扩张等问题。教育需要回答一个关系到社会建制的基础问题：在专业化和社会分工不断加强的社会背景下，受教育者需要怎样的共同教育，才能保持人的独特性和共同的公民责任。

红皮书的重要之处，首先在于它将通识教育界定为现代社会中的共同教育问题，而非大学内部的课程调整问题。报告区分了专门教育与通识教育。专门教育服务于特定职业、特定学科和特定技术能力；通识教育服务于每一个人，培养个体作为社会成员的公民责任。这样的区分具有深刻的教育哲学意义。它说明现代教育必须同时处理两类任务：一类是使个体能够胜任某种社会分工，另一类是使个体能够理解并参与共同社会生活。

红皮书的第二个重要贡献，是把通识教育目标落到心智能力和价值判断上。报告强调受教育者应当发展有效思考、沟通思想、作出相关判断和辨析价值的能力。1946年Wilson (1946) 在书评中概括红皮书时，也指出该报告关注思考能力、交流能力、相关判断。这些目标与今天关于批判性思维、沟通能力、伦理判断和公民素养的讨论高度相通。

红皮书的第三个重要贡献，是将中学和大学放入同一教育连续体中讨论。报告既关注美国中等教育，也关注哈佛学院本科教育，尝试在二者之间建立连贯的共同教育框架。对今天的大中小学人工智能通识教育而言，这种中学与大学衔接的思想仍有启示：共同教育既要尽早进入学生体验，也要随着学段提升而形成更系统的理解。

红皮书提出的人文、社会科学和自然科学一体的共同教育结构，也具有启示意义。它试图把文化传统、社会制度、科学方法和公共责任结合起来，使学生既理解人类经验，也理解现代科学和社会生活。尽管其课程内容带有明显时代特征和西方中心色彩，红皮书提出问题的方式仍值得继承：现代社会不能只依赖分散的专业能力，它还需要某种共同学习经验，使不同背景的受教育者拥有可以彼此交流、共同判断和承担责任的基础。

当然，红皮书也受到许多批评。它所依赖的文化传统、经典范围和社会想象，难以充分回应后来的多元文化、性别平等、种族正义、全球化和后殖民批评。Kliebard (2004) 关于美国课程史的研究表明，二十世纪课程改革始终处在不同教育取向之间的争论中，包括人文主义、社会效率、发展主义和社会改造等传统。红皮书体现的是其中一种强烈的人文主义和共同文化取向。它的不足提醒我们，任何共同教育基础都需要在时代变化中重新解释，不能简单固化为特定文明或特定社会阶层的知识清单。

因此，本书引用红皮书，是承袭其提出问题的方式，而非复制具体课程方案。1945年的红皮书面对的是专业化和共同教育基础之间的关系；人工智能时代的红皮书需要面对智能社会、认知中介和共同认知能力之间的关系。二者的历史条件不同，但教育问题具有结构上的连续性。社会越复杂，越需要共同基础；知识越分化，越需要整体理解；技术越强大，越需要人的判断和责任。

1.5 共同教育基础的三重结构

通识教育的核心任务，可以概括为重建共同教育基础。这个共同基础不能理解为所有人学习完全相同的知识清单，也不能理解为用若干公共课程简单覆盖专业教育之外的空白。更加稳妥的理解是，共同教育基础包含知识、能力和责任三重结构。

第一，共同教育基础包含必要的共同知识。任何社会都需要一定程度的共享背景知识，使个体能够进行基本交流。现代社会中的共同知识包括历史理解、科学素养、数学与数据意识、语言表达、社会制度、技术基础、生态环境、文化多样性和伦理问题等各个方面。共同知识的意义在于提供公共讨论的最低背景，使社会成员能够在同一问题上形成可交流的理解。

第二，共同教育基础包含跨领域能力。教育学长期强调，知识的迁移需要深层理解和反思性调节。Bransford, A. L. Brown, and Cocking (2000) 在 *How People Learn* 中指出，迁移依赖于学习者是否理解知识背后的组织结构，是否能够在新情境中识别相关原则。Perkins and Salomon (1988) 关于迁移的研究进一步区分了较为自动的低路迁移和需要有意识抽象的高路迁移。这说明通识教育不能满足于让学生接触多个领域，它需要训练学生能够在不同领域之间建立联系，理解问题结构，迁移概念和方法。

第三，共同教育基础包含责任意识和价值判断。教育不仅培养能力，也塑造能力使用的方向。Biesta (2010) 在关于教育目的的研究中强调，教育不能只以效果评测为中心，还要处理资格化、主体化、社会化之间的关系。这对通识教育具有重要意义。一个人拥有很强的分析能力，却缺少基础判断力，分辨何时应当使用这种能力、能力服务于何种目标、行动后果由谁承担，这样的人在认知和行动能力上都是有缺陷的。共同教育基础因此必须包含伦理、公共责任和自我反思。

这三重结构之间具有内在联系。共同知识提供理解世界的材料，跨领域能力提供组织和迁移知识的方式，责任意识提供使用知识和能力的方向。在专业化教育中，这三者常常被拆分：知识归专业课程，能力归训练项目，责任归伦理课程。通识教育的任务，正是把它们重新组织起来，使学生在学习知识的同时形成判断，在形成能力的同时理解责任。

从课程制度上看，共同教育基础可以通过不同方式实现。核心课程强调共同阅读和共同讨论，分布课程强调跨领域接触，项目课程强调真实问题和实践整合，写作课程强调表达与论证，科学素养课程强调证据与方法，公民教育课程强调公共问题和价值判断。不同课程制度形态各有优劣，关键在于课程是否真正建立了共同问题意识，是否提供了认知挑战，是否促进学生在证据、表达、判断和反思上的发展。

这一点对人工智能通识教育至关重要。课程需要把有关人工智能的共同知识、跨情境认知能力和伦理责任组织起来，使技术理解、判断训练与具体应用相互联系。这样，它才能接续通识教育的三重结构。

1.6 通识教育的内在张力

通识教育始终伴随争议，因为它必须处理若干无法彻底消除的张力。理解这些张力，有助于避免把通识教育简单化为某种课程口号。

第一重张力，是共同性与多样性之间的关系。通识教育需要共同基础，但现代社会具有文化、语言、阶层、性别、地域和知识背景的多样性。共同教育如果缺少开放性，容易变成单一文化传统的再生产；多样性如果完全缺少共同问题，学生又难以形成公共交流的基础。通识教育需要把共同性理解为共同问题和共同能力，而非封闭的单一知识清单。

第二重张力，是知识深度与广度之间的关系。通识教育强调广阔视野，但广度若缺少深度，就会变成浅层浏览。专业教育强调深度，但深度若缺少横向连接，就会造成视野窄化。较好的通识教育追求有结构的广阔，在若干关键领域中让学生理解问题背景、证据标准和方法边界，使其获得进一步学习和跨领域交流的能力。

第三重张力，是自由选择与制度要求之间的关系。现代大学重视学生自主性，选修制度有助于学生发展兴趣和个性。但如果全部依赖个人选择，学生往往会受到专业压力、成绩策略和即时兴趣的影响，难以主动选择通识课程。核心课程等方案虽然限制了选择，却可以为共同学习提供制度保障。关键不是制度要求是否必要，而是制度要求是否具有清晰的教育理由，课程内容是否足够开放，质量是否满足要求。

第四重张力，是职业准备与人的发展之间的关系。职业能力对学生未来生活极为重要，教育不能轻视就业和社会分工。但职业准备需要放入人的长远发展中理解。未来职业环境持续变化，过度狭窄的职业训练可能很快失效；能够持续学习、判断复杂问题、与他人合作、理解技术和社会后果的人，更可能在变化中保持发展能力。通识教育在这里发挥长期、基础作用。

第五重张力，是评价可测量性与教育深层目标之间的关系。知识记忆和技能操作较容易评价，判断力、责任感和自我反思则较难评价。教育制度往往倾向于测量表观内容，进而影响课程设计。Biesta (2010) 指出，教育目标被测量逻辑单独支配时，教育的价值维度容易被压缩。通识教育需要发展更加丰富的评价方式，关注过程性证据、表现性任务、论证质量、反思深度和跨情境迁移。

1.7 面向人工智能时代的历史转接

通识教育的历史使命，在人工智能时代获得新的表现形式。人工智能的快速进步已经超出了工具和技术的范畴，而是深入到社会生产生活的基础运行机制，形成“智能社会”的基础架构。在这样的智能社会中，专业化带来的知识割裂仍然存在，同时又出现了更深层的认知中介问题。过去，学生面对的主要挑战是如何在专业之外获得更广阔知识；今天，学生还需要理解信息、知识和判断如何被数据、模型、平台和自动化系统重组。过去，通识教育主要回应学科分化；今天，通识教育还要回应认知外包、算法推荐、生成内容、自动化决策和模型中介。

本章由此得出一个基本结论：现代通识教育的目的是建立社会的共同基础、公共判

断、心智训练和责任意识；在当前的人工智能时代，它面对的对象和环境发生了新的变化，因此需要新的概念结构和课程设计。1945 年的哈佛红皮书试图回答分工社会需要怎样的共同教育，今天我们要回答智能社会需要怎样的认知能力，而这正是本书的核心任务。

本章小结

本章从现代通识教育的历史使命出发，讨论了专业化教育的成就及其后果，梳理了博雅教育、公民教育、科学教育和核心课程传统，分析了哈佛红皮书在现代通识教育史中的代表意义，并提出共同教育基础的知识、能力和责任三重结构。通识教育的核心任务，是在知识分化和社会复杂化条件下帮助受教育者形成共同理解、公共判断和责任意识。

人工智能时代并没有消减传统通识教育问题，反而使其更加紧迫。下一章我们将具体讨论智能社会的形成以及个人在这样的社会形态中所面对的新的认知困境，由此衍生出“认知自治”这一核心命题。

第二章　智能社会及其认知环境

摘要：人工智能技术体系正在成为现代社会的重要基础设施，催生了智能社会的到来。本章回答两个基础问题。第一个问题是，什么是智能社会，它具有哪些基本特征。学术界围绕信息社会、平台社会、算法社会、通用技术、分布式认知和社会技术系统形成了多条研究路径。本章在这些研究的基础上，将智能社会界定为：智能社会是人的知识生产、信息传播、判断形成和社会行动持续通过人工智能系统与制度共同中介的社会形态。第二个问题是，智能社会形成了怎样的认知环境。本章从信息过载、来源混杂、生成内容、算法中介、边界限制、指标治理和自动化决策七个平行方面展开分析，说明信息丰富与可靠性不明、知识易得与根据难辨、系统高效与责任分散何以同时出现，并由此引出智能社会对认知自治的要求。

关键词：智能社会；社会技术系统；算法中介；认知环境；不确定性；信任校准；认知自治

2.1　人工智能催生智能社会

人工智能已经超越单一技术和具体工具的范畴，而是深入到社会生产生活的各个角落，并与现行社会运行机制相融合，共同构建起一种新的社会形态，我们称之为“智能社会”。

目前，“智能社会”尚未形成一个由各学科共同接受的单一定义。社会学、传播学、经济学、法学、认知科学和科学技术研究分别从不同侧面描述人工智能进入社会以后发生的结构变化。与其从若干应用事实直接归纳这一概念，更稳妥的方式是考察已有研究如何理解技术、信息、知识、组织和制度之间的关系，并从中提取智能社会的共同结构。

2.1.1　从信息社会到平台社会与算法社会

信息社会研究首先揭示了信息生产、处理和网络连接在现代社会中的基础地位。数字化使越来越多的社会活动留下可记录、可传输和可计算的数据，网络则改变了信息、资本、组织和人的连接方式。大数据研究进一步指出，数据分析不仅提高了信息处理规模，也在重组知识生产的认识论和方法结构 (Kitchin, 2014)。这一研究传统为理解智能社会提供了前提：人工智能所处理的对象，来自持续扩展的数据化社会过程。

平台社会研究把注意力从“信息越来越多”转向“信息和社会活动由什么机制组织”。Dijck, Poell, and Waal (2018) 将平台社会概括为社会、经济和人际活动越来越多地通过全球平台生态系统展开，并分析数据化、商品化和选择机制如何进入新闻、交通、

医疗和教育等社会领域。平台由此成为连接使用者、机构、市场和公共部门的组织结构，也参与安排内容的可见性、活动的入口和公共价值的实现条件。

算法社会研究进一步关注算法和人工智能如何参与社会治理。Balkin (2018) 指出，大型数字平台位于国家与个人之间，并通过数据、算法和人工智能治理其使用者群体。这里的“治理”不限于公共权力的正式决定，也包括排序、推荐、分类、预测、审核和规则执行对人的机会、表达和行为所产生的持续影响。平台社会与算法社会研究共同表明，人工智能已经进入社会秩序的形成过程。

2.1.2 从技术系统到社会技术系统

经济学中的通用技术研究解释了人工智能为何能够广泛扩散。Bresnahan and Trajtenberg (1995) 把广泛适用、持续改进并能引发互补创新概括为通用技术的重要特征。人工智能即是一种“通用技术”，它与具体行业、组织流程和专业知识结合以后，会形成大量互补应用，并推动原有任务、制度和分工继续调整。因而，智能社会的基础是人工智能与社会各部分的持续结合。

社会技术系统研究则提醒我们，算法模型只是现实系统的一个组成部分。技术能否实现预期目标，取决于数据、操作人员、受影响者、组织规则、法律制度和使用情境之间的共同作用。Selbst et al. (2019) 指出，把技术模型从社会背景中抽离，容易造成对现实问题的错误抽象。模型的输入、目标和输出均经过人的选择，模型进入组织以后又会改变人的行为与制度流程。因此，智能社会的基本分析单位应当是由人、技术与制度共同构成的系统。

认知科学中的分布式认知研究提供了另一个重要视角。Hutchins (1995) 说明，认知活动可以分布在人、工具、符号和协作程序之间。人工智能进入知识生产和社会决策以后，信息的记忆、检索、归纳、表达、预测和筛选进一步分布到人和智能系统之中。个体仍然是理解、判断和承担责任的主体，其认知活动所依赖的外部结构却发生了显著变化。智能社会由此也是一种新的分布式认知环境。

2.2 智能社会的工作定义

综合上述研究，本书将智能社会界定为：

> **智能社会是人的知识生产、信息传播、判断形成和社会行动持续通过人工智能系统与制度共同中介的社会形态。**

这一定义包含两个关键内容。第一，人工智能参与知识生产、信息传播、判断形成和社会行动的全过程，构成社会活动所依赖的底层条件。第二，人工智能系统要与制度深度耦合，形成互相支撑的社会组织结构。

依此定义，判断一个社会是否进入智能社会，主要依据是人工智能是否已经成为知识与行动的常态化中介，以及人工智能技术与社会制度是否形成深度耦合，而不取决于某一类设备的数量或某一项技术的先进程度。

智能社会是一个描述性概念，不预设技术进步必然带来更为理想的社会。人工智能可以提高信息处理和社会协调能力，也会带来新的权力关系、认识风险和责任问题。

“智能社会”这一概念与数字社会、信息社会、平台社会和算法社会具有连续性。数字社会强调社会活动的数字化，信息社会强调信息生产与网络结构，平台社会强调平台生态对社会活动的组织，算法社会强调数据和算法对群体的治理。智能社会吸收这些研究所揭示的结构，并进一步突出人工智能技术深入社会过程与治理系统以后，人类个体与社会组织结构所发生的联动变化。

2.3 智能社会的基本特征

下面从若干关键维度概括智能社会的基本特征。这些特征虽不能穷尽智能社会的全部面貌，却能够揭示人工智能系统与社会制度共同参与社会运行的基本机制。

- **人工智能的基础设施化：**人工智能从专门工具扩展为可被多种领域调用的通用能力，并通过云服务、平台接口、嵌入式系统和组织流程进入日常运行。基础设施化意味着许多活动开始默认依赖算法搜索、模型生成、智能推荐和自动化分析，使用者却未必直接看见其技术环节。人工智能的社会意义由单项任务的效果，扩展到它为知识、交流和行动提供了什么基础条件。
- **社会过程的数据化与模型化：**智能社会持续把人的行为、关系和环境转化为数据，再通过模型形成分类、预测、生成和评价。数据并非对现实的完整复制，模型也不是现实本身；它们通过变量选择、样本范围、标签规则和优化目标建立对现实的特定表征。数据化与模型化提高了社会过程的可计算性，同时使“现实如何被表示”成为影响知识和决策的基础问题 (Kitchin, 2014; Gillespie, 2014)。
- **认知与行动的人机混合化：**知识形成和社会行动越来越多地由人和人工智能共同完成。人在这一结构中提出目标、解释情境、核验结果并承担责任；人工智能承担检索、匹配、生成、预测和执行等部分认知或行动任务。人机混合并非简单的能力相加，它会重新分配谁提出问题、谁提供证据、谁形成建议、谁复核结果以及谁承担后果。智能社会中的认知主体由此处在更复杂的分布式系统中 (Hutchins, 1995)。
- **社会连接的平台中介化：**平台把数据、算法、服务提供者、使用者、组织和公共部门连接起来，并通过规则、接口和排序机制组织相互作用。人们接触什么内容、获得什么服务、与谁发生连接，越来越多地经过平台选择。平台拥有的并非一般意义上的传播渠道，它还能够设计规则、记录行为、分发注意力资源并形成新的社会协调机制 (Dijck, Poell, and Waal, 2018; Balkin, 2018)。
- **系统运行的反馈演化：**智能系统根据使用数据持续调整，人的行为又会适应系统给出的推荐、评分和机会。新的行为数据重新进入模型和平台，形成“系统影响行为、行为更新系统”的反馈循环。这种循环使智能社会具有动态性：模型性能、使用习惯、组织规则和社会规范会相互推动并持续变化。系统的后果因而不能只根据设计时的功能判断，还需要观察部署后的长期反馈。
- **技术、制度与价值的深度耦合：**人工智能进入社会以后，总是在具体制度中运行。

模型优化什么目标、采用什么指标、允许何种申诉、由谁承担责任，都与组织利益、法律规则和公共价值相关。技术选择会产生制度后果，制度安排也会塑造技术的设计和使用。公平、隐私、安全、可解释性、公共利益和责任由此成为系统属性，需要在人、模型和制度构成的整体中分析 (Selbst et al., 2019; Dijck, Poell, and Waal, 2018)。

以上六项特征共同构成智能社会的基础特征。基础设施化说明人工智能的社会位置，数据化与模型化说明其认识方式，人机混合化说明其行动主体，平台中介化说明其组织形态，反馈演化说明其运行机制，制度耦合说明其价值与责任结构。它们也共同改变了个体获取信息、形成理解和作出判断的条件。

2.4 智能社会的认知环境

智能社会的结构变化最终会落实为每个人所处的认知环境。现代教育长期面对知识和信息不足的问题，数字网络与人工智能则显著降低了信息获取和内容生成的成本。信息丰富会消耗有限注意力，并造成注意力稀缺 (H. A. Simon, 1971)；信息过载研究也表明，信息数量、任务复杂度、时间压力、组织流程和个人能力会共同影响信息处理质量 (Eppler and Mengis, 2004)。

人工智能进一步延长了信息生产和社会判断的链条。内容可以由模型生成，由平台排序，由推荐系统强化，再由组织算法转化为评分和决策建议。个体获得结论更加容易，判断结论的来源、证据、适用范围、可信程度和行动责任却更加困难。这种困难超出了简单的真假辨别：许多结论只在一定条件下成立，证据存在强弱差别，模型在不同群体和场景中可能表现不同，不同风险等级也需要不同的核验标准。

本书将这一认知环境分为七个平行方面。它们分别揭示注意力、来源、表达、信息组织、模型边界、制度指标和行动责任中的不确定性，并共同构成人工智能通识教育需要回应的基础处境。

2.4.1 信息过载：注意力成为稀缺资源

信息过载指个体或组织接收的信息数量、速度和复杂度超过其处理能力，从而影响理解、判断和行动。即使每一条信息都具有一定质量，有限注意力也难以充分处理大量竞争性内容。生成式人工智能降低了内容生产和改写成本，搜索、摘要和推荐又使信息能够持续到达个体，信息供给与人的处理能力之间的差距进一步扩大。

信息过载容易促使人依赖搜索排名、标题、点赞量、熟悉来源或模型摘要快速判断，并以信息的易得性替代证据质量。相同内容的重复出现也可能造成“多方印证”的错觉，而这些内容实际上可能来自同一原始来源或同一生成链条。

这一环境要求个体能够筛选信息、划分来源层级、比较证据强弱并分配核验精力。教育的重点不再只是帮助学习者找到更多材料，还要训练其在材料过多时建立问题边界、确定证据优先级，并根据任务风险决定需要投入多少注意力。

2.4.2 来源混杂：知识生产链条变得不透明

智能社会中的内容常由人和模型共同生产，并经过采集、生成、编辑、筛选、发布、推荐和转述等多个环节。一个文本可能由人提出观点、由模型生成初稿、由编辑修改、由平台推荐，再被搜索引擎摘要。读者看到最终内容时，作者署名只能说明部分责任关系，无法完整呈现知识的生产链条。

Sperber et al. (2010) 把人类评价信息提供者和信息内容可靠性的能力概括为“认识警觉”。认识论认知研究也强调，学习者需要理解知识从何而来、如何得到证据以及何时可以修正 (Hofer and Paul R. Pintrich, 1997; Sandoval, Greene, and Bråten, 2016)。智能社会扩展了认识警觉的对象：除作者和机构外，还要考虑模型、数据、平台和传播者在内容形成中的作用。

来源判断因而需要从识别署名进一步走向追溯生产链。机构报告可能包含模型生成内容，专家材料可能经过团队与工具共同编辑，普通评论也可能来自自动化营销系统。学习者需要区分原始研究、权威报告、新闻转述、商业宣传、平台评论和模型生成内容，并判断不同环节对可靠性和责任产生的影响。

2.4.3 生成内容：流畅性与可靠性相分离

生成式人工智能能够产生语法正确、结构完整、语气自信的文本，也能够生成视觉上连贯的图像、音频和视频。表达质量与事实可靠性由此进一步分离。自然语言生成研究表明，模型可能产生流畅却与事实、输入或外部证据不一致的内容 (Ji et al., 2023; L. Huang et al., 2025)。

加工流畅性研究指出，人们对容易处理的信息产生熟悉感或偏好 (Reber, Schwarz, and Winkielman, 2004)。生成内容很容易使学习者把表达清晰等同于解释正确，把结构完整等同于论证充分，把数字精确等同于证据可靠。语言形式能够增强说服力，却不能独立提供认知证据。

教育需要把表达质量与认识证据分开评价。科学结论要由数据、实验设计、模型假设、可重复性和同行检验支持；历史解释要由史料、语境、时间线和解释一致性支持；公共判断则要考虑来源、利益、价值和行动后果。流畅性有助于理解和传播，真实性仍需外部证据支持。

2.4.4 算法中介：可见性与注意力被持续组织

个体接触的信息总会经过某种组织方式。智能社会中的特殊变化在于，搜索、排序、推荐和个性化机制能够持续依据用户数据调整可见内容。Gillespie (2014) 指出，算法具有公共相关性，因为它们参与判断什么是重要、相关和可见。平台呈现的信息流由算法选择、个人行为、社群关系、内容供给和商业目标共同形成，不能直接等同于现实本身。

算法中介还具有反馈放大效应。用户的点击、停留、转发和购买成为新的数据，系统据此调整推荐，调整后的信息环境又继续影响用户的兴趣、信念和行动。平台研究表明，算法排序、个体选择和社群关系都会影响信息暴露 (Bakshy, Messing, and L. A. Adamic,

2015)；错误信息的传播也受到内容特征、社会动机和平台结构的共同影响 (Vosoughi, Roy, and Aral, 2018; D. M. J. Lazer et al., 2018)。

理解算法中介，需要同时看见技术机制和人的参与。学习者应当追问信息为何出现在眼前，排序依据包含哪些信号，平台试图优化什么目标，个人行为如何反过来塑造信息流。这样的分析有助于把“我看到的内容”与“世界中存在的内容”区分开来。

2.4.5 边界限制：模型能力具有条件性

人工智能系统通常在特定数据上训练和评估，再被部署到新的场景。训练数据、测试数据和真实环境之间的差异，可能导致性能下降、偏差放大或风险转移。Guo et al. (2017) 发现，现代神经网络的准确率与置信度并不天然匹配；Recht et al. (2019) 重新构建经典图像测试集后观察到模型准确率下降；Hendrycks and Dietterich (2019) 则通过常见扰动基准揭示模型在环境变化中的鲁棒性问题。

边界限制也具有社会维度。Buolamwini and Gebru (2018) 对商业人脸分析系统的评估发现，不同性别和肤色群体之间存在显著错误率差异。整体准确率可能掩盖群体差异，平均表现也可能掩盖局部伤害。系统在一个样本、群体或演示环境中表现良好，无法保证它在新的对象和情境中同样可靠。

这类环境要求学习者形成稳定的边界意识：模型基于什么样本建立，评价指标说明了什么，结论能够推广到哪些对象，哪些变化可能导致失效，错误在不同群体和场景中的后果是否相同。理解模型能力的条件性，是管理智能社会不确定性的核心部分。

2.4.6 指标治理：代理变量参与塑造现实

现代组织广泛使用量化指标、评分系统、排名机制和风险模型。人工智能进入这些系统以后，指标不仅用于描述现实，还会成为模型优化和组织行动的依据。现实中难以直接测量的目标，常被转化为可计算的代理变量；代理变量如何选择，会影响系统看见什么、忽略什么以及奖励什么。

Espeland and Sauder (2007) 对法学院排名的研究表明，排名能够改变组织的资源配置和自我理解；Strathern (1997) 对审计文化的分析也揭示，当指标成为目标时，被测量的活动本身会随之变化。算法模型以指标为优化目标，组织又依据模型输出调整行为，描述性数字由此可能转化为塑造现实的制度力量。

教育领域中的学习数据可以帮助教师了解部分学习过程，却无法穷尽理解、思想发展与创造。医疗、金融和公共管理中的风险评分也只能在特定目标和代理变量下表示现实。学习者需要理解指标的构造、代理关系、遗漏维度和行为效应，避免把可测量部分等同于完整现实。

2.4.7 自动化决策：判断与责任发生分离

模型输出进入招聘筛选、教育预警、信贷审批、医疗评估和公共服务分配等流程以后，会影响人的机会和权益。自动化系统可以提高处理效率和一致性，也可能使建议、

决定与责任分布到模型开发者、数据提供者、系统部署者、专业人员和组织管理者之间。决策链条越长，个体越难识别谁能够解释、纠正并承担后果。

Lee and See (2004) 指出，对自动化系统的适当信任取决于系统性能、过程理解和目的匹配；Parasuraman and Riley (1997) 则分析了自动化的误用、滥用和弃用。当组织把模型输出视为客观依据，人的复核容易转化为形式确认，责任也可能被推向“系统给出的结果”。自动化建议由此带来判断权与责任承担相分离的风险。

不同场景需要不同的证据标准、人工监督和问责机制。低风险的内容推荐与高风险的医疗、教育、法律和公共治理决定，不能采用相同的复核强度。学习者需要追问谁设定目标、谁解释输出、谁有权改变决定、谁处理申诉以及谁承担后果。模型可以支持决策，责任链条仍需由制度和人明确建立 (National Institute of Standards and Technology, 2023)。

2.4.8 七个方面的整体关系与教育转接

上述七个方面的认知处境互相关联。信息过载增加筛选压力，来源混杂延长知识生产链，生成内容使表达形式与可靠性分离，算法中介组织可见性与注意力，边界限制限定模型能力，指标治理改变组织目标与行为，自动化决策则把模型输出带入行动和责任。它们相互作用，却分别对应一种不可省略的判断任务。

表 2.1: 智能社会认知环境的七个方面及其教育含义

认知环境	结构来源	主要认知风险	人工智能通识教育任务
信息过载	内容生成成本下降，搜索、推荐与摘要持续供给信息	注意力分散，以易得性和重复性替代证据判断	建立问题边界，训练筛选、分层和证据排序
来源混杂	人、模型、机构与平台共同构成生产和传播链条	只看署名或形式，无法识别内容形成过程	追溯生产链条，辨认来源层级与责任环节
生成内容	模型能够快速生成形式完整、语气自信的内容	把流畅、完整和精确误认为真实可靠	区分表达质量与认识根据，进行事实和证据核验
算法中介	排序、推荐和反馈机制持续组织信息可见性	把经过选择的信息流误认为现实全貌	理解平台目标、排序信号和行为反馈
边界限制	训练、测试与使用环境存在差异	把局部表现推广到新场景，忽视群体差异	判断样本代表性、适用边界和失败条件

认知环境	结构来源	主要认知风险	人工智能通识教育任务
指标治理	组织以代理变量、评分和排名表示并管理现实	把指标等同于目标，忽视遗漏和行为塑造	分析指标构造、代理关系和制度后果
自动化决策	模型输出进入组织行动，判断与责任分布到多个主体	过度信任系统，复核、申诉和问责不足	区分风险等级，明确人机分工与责任链条

面对这一认知环境，普遍怀疑会使知识信赖与社会合作难以维持，过度信任则会削弱判断与责任。关键能力是信任校准：根据信息来源、证据强度、适用范围、场景风险和错误后果，给予结论和系统适当程度的信任。来源、证据、边界、风险和责任共同决定核验强度。

2.5 本章小结

本章围绕两个问题展开：什么是智能社会，以及智能社会形成了怎样的认知环境。综合信息社会、平台社会、算法社会、通用技术、分布式认知和社会技术系统等研究，本章将智能社会界定为“人的知识生产、信息传播、判断形成和社会行动持续通过人工智能系统与制度共同中介的社会形态”。人工智能的基础设施化、社会过程的数据化与模型化、认知与行动的人机混合化、社会连接的平台中介化、系统运行的反馈演化，以及技术、制度与价值的深度耦合，构成了这一社会形态的六项基本特征。

这些结构变化进一步塑造了七个相互关联又各有侧重的认知环境：信息过载使注意力成为稀缺资源，来源混杂使知识生产链条更难追溯，生成内容使表达流畅性与事实可靠性进一步分离，算法中介持续组织信息的可见性，边界限制使模型能力呈现明显的条件性，指标治理使代理变量参与塑造组织行为，自动化决策则可能造成判断权与责任承担的分离。个体面对的主要困难由获得信息逐步转向辨析根据、识别边界、校准信任和承担责任。

智能社会由此对通识教育提出了新的要求：学习者需要理解智能系统如何参与知识与行动，能够根据来源、证据、适用范围和风险后果调节自己的判断，并在人机协作中保持责任意识。这就是下一章所讨论的“认知自治”。

第三章　认知自治：智能社会通识教育的时代使命

摘要：智能社会中，人的认知活动越来越多地受到数据、模型、平台和制度的共同中介。信息来源日益混合，内容可以被规模化生成，算法参与信息的筛选与放大，模型输出也开始进入判断形成和现实行动。通识教育因此需要帮助个体在开放依赖中保持认知主导。本章将认知自治界定为个体在复杂认知环境中，对自身信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节的过程；认知自治能力则是个体发起、维持和调整这一过程，使其能够跨越不同任务与情境持续展开的能力。本章从来源警觉、证据敏感、边界意识、信任校准、认知能动和责任判断六个维度展开认知自治的基本结构，说明其个体表现与社会意义，辨析其与批判性思维、元认知、自我调节学习、认识论认知、信息素养和媒介素养的关系，并回应围绕这一教育目标的主要质疑。认知自治由此构成智能社会通识教育的时代使命。

关键词：智能社会；通识教育；认知自治；认知自治能力；来源警觉；证据敏感；信任校准；认知能动

3.1　认知自治的概念界定

前一章讨论了智能社会带来的认知环境变化。信息生成、筛选、传播和使用方式的改变，对人的认知活动提出了新的要求，本书将这种要求概括为认知自治（Cognitive Autonomy）。

认知自治可以定义为：**个体在复杂认知环境中，对自身信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节的过程**。认知自治能力，则是个体发起、维持和调整这一过程，使其能够跨越不同任务与情境持续展开的能力。这一定义可以从以下几个方面来理解。

- **认知自治是一个主动调节认知活动的过程**。
  认知自治贯穿信息获取、知识理解、问题求解、判断形成和行动选择的全过程。个体需要在这一过程中不断识别信息来源、评价证据质量、调整信任程度、把握判断边界，并对依据判断采取的行动承担责任。认知自治具有持续性和动态性，会随着任务目标、证据状况、外部环境和潜在后果的变化不断展开和调整。
  认知自治与数据驱动、实践检验等学科思维方式处在不同层面，也不能等同于人工智能工具的设计、操作和使用。学科思维方式提供认识问题的具体路径，工具

使用体现解决任务的具体行为，认知自治则贯穿个体理解和选择这些路径、评价工具结果以及根据新证据修正判断的过程。个体能否主动发起、有效维持并及时调整这一过程，体现为其认知自治能力。这种能力可以跨越不同学科、任务和生活场景，具有通用性和迁移性。

- **认知自治面向现代社会普遍存在的复杂认知环境。**
  复杂认知环境是指信息来源多样、证据质量不一、知识持续更新、不同判断相互竞争，并且信息的生成、筛选、传播和放大越来越多地受到算法、平台、组织和制度中介的认知环境。在这种环境中，个体接触到的内容往往已经经过多重选择和加工，信息数量的增加也不会自然带来更可靠的认识。
  人工智能进一步提高了这种环境的复杂程度。它可以快速生成内容、组织知识、提出建议并参与决策，同时也可能产生错误、遗漏和缺乏依据的结论。个体面对的不再只是某一条信息是否真实，还需要判断信息如何产生、证据能支持什么结论、系统适用于什么情境，以及应当给予其多大程度的信任。
  不同国家和制度中的具体表现、风险分布与治理方式存在差异，但信息过载、来源混杂、算法中介、生成内容和知识快速更新等结构性变化具有广泛性。因此，主动开展认知自治超越特定技术、平台和社会制度，构成个体进入现代认知环境的普遍要求；支撑这一过程的认知自治能力，也成为现代社会所需要的基础能力。
- **认知自治的核心在于主动调节。**
  认知自治所强调的主动性，是个体能够意识到自己的认识正在受到哪些信息、工具和社会机制的影响，并根据任务目标、证据状况和可能后果，主动调整自己的认知活动。个体既可以接受可靠信息、借助专家知识和使用人工智能，也能够在条件变化、证据不足或风险升高时重新核验、修正判断和调整行动。
  这种主动性以对人工智能技术及其社会作用方式的基础理解为前提。理解人工智能怎样从数据中形成能力、怎样受到目标和评价方式的影响，以及怎样通过平台和组织进入现实生活，可以帮助个体形成有根据的信任与行动信心。在此基础上，人们能够坦然面对人工智能，自信地选择和使用工具，主动探索新的学习、创造和问题求解方式，同时持续检查自己的理解、判断和责任是否仍然处于自主调节之中。
  因此，认知自治包含风险意识，却不以风险提示为终点。它所追求的是个体在复杂环境中持续理解、选择、行动和修正的积极过程；能够稳定开展这一过程，体现了个体的认知自治能力。
- **认知自治是在开放依赖中保持人的认知主导。**
  人的认知活动始终建立在开放依赖之中。语言、知识、教育、专家、制度和技术工具共同参与人的认识过程。人工智能进一步扩大了这种依赖的范围，使信息检索、材料整理、模式识别、计算推演、方案生成和内容表达等活动都可以得到外部能力的支持。认知自治允许人根据任务需要，有选择地调用和组织外部认知资源，无需由个体独立完成全部认知任务。
  这种认知依赖也是现代知识社会的基本状态。Hardwig (1985) 关于认识依赖的讨论指出，现代知识共同体中的个人常常需要基于他人证言和专家劳动形成信念。

Hutchins (1995) 关于分布式认知的研究进一步表明，复杂任务中的认知活动往往分布在个人、工具、符号系统和组织环境之间。人工智能使这种依赖更加密集、快速和隐蔽，也使人对依赖关系进行主动组织变得更加重要。

认知自治强调在这种开放依赖中人的认知主导，这种主导主要体现为保持对整个认知过程的发起、审查和统摄。其中，有两个角色尤其需要保留在人类一侧：提出问题和核查结果。提出什么问题，决定了认知活动朝向何种目标，也包含着人对现实需要、价值取向和问题意义的理解；核查结果，则要求人判断所得结论是否有充分依据、是否适用于当前情境、是否遗漏重要因素，以及能否据此采取行动。这两个环节共同确定了认知活动的方向、边界和最终有效性。

人工智能可以帮助人发现问题、扩展问题、生成候选问题，也可以参与证据检索、交叉比较和结果验证，但问题的选择与界定仍需由人主导，结果能否接受也需经过人的核查。至于两个环节之间的具体认知活动，例如搜索信息、调用知识、执行计算、识别模式、生成方案和模拟后果，则可以根据任务性质、工具能力和风险程度，有选择地交由人工智能完成。不同任务中的人机分工可以灵活变化，人的主导地位并不取决于亲自完成了多少步骤，而取决于是否仍然掌握问题的提出、过程的调节和结果的核查。

因此，利用人工智能扩展人的能力与保持认知自治可以同时实现。个体可以充分调用人工智能的记忆、计算、生成和分析能力，同时持续理解任务正在怎样展开、人工智能承担了哪些环节、所得结果依据什么，以及哪些结论仍需进一步验证。认知自治所维护的，是人对认知活动整体方向和有效性的主导，使人工智能的能力始终被纳入由人发起、由人核查并由人负责的认知过程。

3.2 认知自治的六个维度

认知自治过程可以具体展开为来源警觉、证据敏感、边界意识、信任校准、认知能动和责任判断六个维度。其中，来源警觉、证据敏感、边界意识、信任校准和责任判断，分别对应信念形成、证据使用、判断边界、信任分配和行动责任五类需要调节的对象；认知能动则体现“主动调节”所包含的内在倾向。它使个体愿意面对认知环境的变化，敢于探索人工智能带来的新可能，并主动寻求扩展自身能力的路径，其具体实现需要多种认知能力共同支撑。六个维度共同构成认知自治过程的基本结构，也为课程设计、教学观察和过程诊断提供了分析框架。

- **来源警觉**。个体需要追问信息由谁或什么系统产生。来源可以是教师、专家、论文、媒体、机构、平台、模型、商业系统，也可以是人机共同生产的混合来源。来源警觉要求个体避免根据身份直接接受或拒绝，并进一步分析产生机制、专业能力、激励结构、可追责性和传播路径。Sperber et al. (2010) 提出的认识警觉概念强调，人类依赖他人交流，同时必须评价交流者和交流内容的可靠性。AI 环境把这一问题复杂化，因为模型输出具有类似证言的语言形式，却缺少普通人类交流中的责任关系。
- **证据敏感**。个体需要区分结论的表达形式和支持它的证据基础。一个结论可以来自

个人经验、统计数据、实验研究、专家判断、历史材料、模型生成、基准测试或商业宣传。不同证据具有不同强度和适用范围。Chinn, Buckland, and Samarapungavan (2011) 关于认识论认知的研究指出，学习者需要理解知识主张如何被证据支持，以及不同学科如何建立正当性。在人工智能环境中，证据敏感意味着学生不把语言流畅、图像逼真或数字精确直接等同于可靠性。

- **边界意识**。人工智能系统从数据中学习，数据来自特定人群、场景、语言、文化和时间。模型在某一分布上有效，并不能保证在新的场景中同样可靠。边界意识把机器学习中的泛化问题转化为日常判断习惯：当前案例是否超出了证据边界？样本是否代表总体？经验是否可以迁移？这类边界问题在人工智能领域极为常见。例如，Recht et al. (2019) 对 ImageNet 和 CIFAR-10 测试集的再评估显示，模型在新测试集上可能出现显著性能下降；这类研究提示教育者，性能结论需要与数据来源和使用场景联系起来理解。
- **信任校准**。个体需要根据证据强度、来源质量、任务风险和错误后果，给自己的信念和行动选择分配合适的信心。心理学与神经科学研究表明，人对自己判断的信心并不总与准确性一致。Fleming and Dolan (2012) 关于元认知神经基础的综述指出，元认知准确性可以与任务表现相区分。AI 研究中，Guo et al. (2017) 发现现代神经网络的预测概率也可能出现校准不足，准确率和置信度之间并无天然一致。人和模型都可能表现出过度自信，教育需要训练学习者调节信心强度。
- **认知能动**。个体在理解人工智能基本原理、能力边界及其社会影响的基础上，能够以开放而自信的态度看待人工智能，将其视为可以认识、探索和调用的认知资源，主动思考它能够怎样参与学习、创造、判断和问题求解。认知能动主要是一种认知倾向，表现为面对新技术时愿意理解、敢于尝试，并对新的认知方式保持开放；在具体任务中，它还表现为能够从自身问题和目标出发，发现人工智能可能发挥作用的环节，设想借助外部能力拓展自身认知范围的可能路径。它为后续的工具选择、任务组织和人机协作提供内在动力，也使来源、证据、边界、信任与责任的调节服务于积极的学习、创造和行动。
- **责任判断**。认知自治最终要进入行动。一个结论如果用于头脑风暴，所需证据标准较低；如果用于医疗、法律、金融、教育评价或公共治理，所需标准就高得多。Lee and See (2004) 关于自动化信任的研究指出，人与自动化系统之间的信任必须与系统能力、过程透明度和使用目的相匹配。NIST 人工智能风险管理框架也强调，AI 风险要结合具体情境和受影响者来管理 (National Institute of Standards and Technology, 2023)。责任判断要求学习者理解：不同场景具有不同的核验标准、人工复核要求和问责方案。

表 3.1: 认知自治的六个维度

维度	核心问题	AI 场景中的表现	社会场景中的表现
来源警觉	我的认识受到哪些信息来源和生成机制的影响?	关注模型、训练数据、提示词、检索系统、平台接口和部署机构怎样影响输出	关注作者、机构、媒体、专家身份、利益关系和传播路径怎样影响信息
证据敏感	当前结论依靠什么证据，证据的支持强度如何?	区分模型生成、检索证据、基准测试、人工评审和事实核验，判断输出是否具有充分依据	区分个人经验、统计数据、实验研究、历史材料、专家判断和商业宣传,评价其证据效力
边界意识	当前结论在什么条件和范围内成立?	检查训练分布、测试场景、群体差异、语言文化差异和模型适用条件	分析样本代表性、经验迁移、政策推广、历史类比和地域差异,避免超出证据范围作出判断
信任校准	我应当给予这一信息、主体或系统多大程度的信任?	根据任务性质、证据质量、模型表现和潜在风险，调节对 AI 答案、概率评分、风险预测和生成内容的信任	根据来源可靠性、证据强度和具体情境,调节对媒体、权威、同伴、机构排名和个人直觉的信任
认知能动	我能否主动面对认知环境的变化，并积极探索扩展自身认知能力的可能?	在理解人工智能能力与边界的基础上，坦然面对、主动接近并积极探索其在学习、创造、判断和问题求解中的可能作用	面对新知识、新技术和新环境保持开放、自信与探索意愿，主动发现和利用能够支持自身认识与行动的外部资源
责任判断	结论将用于什么行动，可能产生什么后果，责任应当由谁承担?	根据任务风险判断是否采用、复核或拒绝 AI 结果，明确人机分工中的决定权和责任归属	区分讨论、学习、出版、医疗、法律、金融和公共政策等场景,判断行动后果及相应责任

3.3 认知自治的个体表现

认知自治首先表现为个体在具体认知任务中持续开展的主动调节。六个维度并非六项彼此分离的技能，它们共同贯穿提出问题、运用外部资源和核查结果的完整过程。个体能否主动发起、有效维持并及时调整这一过程，体现其认知自治能力的发展水平。

提出问题体现了个体对认知活动的发起和定向。学习者需要从现实需要和自身目标出发，判断什么值得了解、解决或创造，并把模糊处境转化为可以探索的问题。人工智能能够帮助发现线索、扩展问题和生成候选方向，认知能动使人愿意调用这些能力；问题的选择、界定及其价值意义仍需由人把握。

运用外部资源体现了个体对认知依赖的主动组织。写作、学习、编程、翻译、设计和研究中，人工智能可以承担资料整理、模式识别、计算推演、方案生成、模拟比较和错误检查等功能。学习者需要根据问题性质、工具能力和任务风险，选择是否使用人工智能、把哪些环节交给系统，并在过程中保持来源警觉、证据敏感和边界意识。人的主导地位不取决于亲自完成多少步骤，而取决于能否用自己的问题和目标统摄人机分工。

核查结果体现了个体对认知过程的审查和最终判断。学习者需要区分事实、解释、评价和建议，追问结论的证据基础、成立条件和遗漏因素，并根据使用场景校准信任。低风险、探索性和创造性任务可以容纳较高的不确定性；高风险、专业性和公共性任务需要更严格的证据、人工复核与责任安排。Lee and See (2004) 强调，适当信任比盲目信任或完全不信任更重要，这一思想应成为智能社会通识教育的重要原则。

认知自治还要求个体持续觉察外部工具对自身认知状态的影响。人工智能可能提高效率，也可能削弱耐心；可能激发想法，也可能让表达趋同；可能帮助理解，也可能掩盖理解不足；可能提供多种视角，也可能强化已有偏好。学习者需要说明自己的判断如何形成，人工智能在哪些环节提供了帮助，哪些内容经过核验，哪些部分仍然不确定，并据此调整后续使用方式。

这些要求可以落实为一组稳定问题：我真正要解决的问题是什么？人工智能能够扩展哪些能力？应当把哪些环节交给它？信息由谁或什么系统产生？证据是什么？结论在什么条件下成立？我应给予多高信任？据此行动可能产生什么后果？对这些问题的持续练习，是智能社会通识教育从知识学习走向心智训练的关键。

3.4 认知自治的社会意义

认知自治同时具有明确的社会意义。智能社会中的判断很少只影响个人。算法推荐影响公共讨论，AI 生成内容影响知识传播，自动化评价影响教育机会，风险模型影响医疗资源，信用评分影响金融服务；与此同时，人工智能也在降低知识获取、内容创造、技术开发和专业协作的门槛。个体能否主动而审慎地使用人工智能，与公共制度、组织责任、社会信任以及参与新型学习和生产活动的机会紧密相连。

首先，认知自治关系到公共讨论质量。D. M. J. Lazer et al. (2018) 关于虚假新闻的研究强调，错误信息传播是技术、认知、社会动机和平台结构共同作用的结果。Vosoughi, Roy, and Aral (2018) 对 Twitter 传播的研究显示，虚假新闻在传播速度和范围上可能超过真实新闻。当生成式 AI 降低内容生产成本，公共讨论更需要参与者具备来源警觉和证据敏感。公民如果无法判断内容的生产机制和证据基础，公共理性就会被流畅叙事、情绪传播和平台放大削弱。

其次，认知自治关系到组织决策质量。自动化系统进入招聘、信贷、医疗、教育评价和公共服务后，组织成员需要理解模型输出的适用边界。Parasuraman and Riley (1997)

关于自动化使用的研究指出，自动化可能产生误用、滥用和弃用。组织中过度依赖系统，可能造成责任外包；完全拒绝系统，也可能错失效率和质量提升。认知自治要求专业人员在组织流程中保持解释、复核和问责能力。

再次，认知自治关系到个体参与学习、生产与创造的能力。人工智能可以帮助人快速进入陌生领域、比较多种方案、跨越部分专业表达门槛，并把初步想法转化为可以继续检验和改进的对象。认知能动使个体愿意发现和探索这些可能，来源警觉、证据敏感、边界意识、信任校准和责任判断则使能力扩展建立在理解与审慎调节之上。认知自治由此既维护人的主体位置，也扩大人能够参与的实践范围。

同时，认知自治关系到教育公平。若智能社会中的教育只关注先进工具的操作，资源丰富学校可能更快获得平台和实践机会，资源不足学校则被排除在外。将认知自治作为通识教育的重要目标，低成本工具、开放资源、纸笔模拟、生活场景分析和人机协作任务也能承担重要作用。学生不一定需要最先进设备，仍然可以学习如何提出问题、选择工具、组织任务，同时理解数据偏差、模型边界、生成内容核验、推荐机制和责任判断。这样的定位有助于使教育从装备条件的差异转向所有学习者都能参与的基础能力培养。

最后，认知自治关系到人的尊严和主体性。Floridi et al. (2018) 提出的 AI4People 原则强调有益、不伤害、自治、公正和可说明性，并把 AI 伦理放在人类福祉与社会治理的背景中讨论。Selbst et al. (2019) 从社会技术系统角度批判抽象化的公平观，指出算法系统总是嵌入制度和社会背景。这些研究共同提示，智能社会的通识教育需要让学生理解技术与人、制度和社会之间的关系。认知自治使个体能够在技术环境中保持主体位置，同时理解个体判断必须进入公共责任结构。

总体看，维护公民的认知自治能力，关系到现代社会的运行基础。现代社会以能够独立判断、参与公共讨论并承担行动责任的公民为主体。公民只有能够自主形成信念、审查证据、调节信任、辨认判断边界，公共协商、共同决策和制度运行才具有真实的主体基础。如果公民的认识和判断越来越受算法筛选、生成内容和自动化系统支配，认知自治能力持续受到削弱，支撑社会运行的公民基础也会逐渐动摇。因此，维护认知自治既是在维护个体对自身认识与行动的主导，也是在维护公共生活和社会正常运行所依赖的基本秩序。

3.5 与相关教育概念的关系

认知自治并非凭空产生的新口号。它与批判性思维、元认知、自我调节学习、认识论认知、信息素养和媒介素养等传统概念密切相关。本书使用“认知自治”这一概念，目的在于把这些传统能力放入智能社会的共同处境中重新组织。

批判性思维强调有理由的判断。Facione (1990) 主持的 Delphi 报告把批判性思维概括为解释、分析、评价、推论、说明和自我调节等核心技能，并强调理性、开放和审慎的心智倾向。Kuhn (1999) 也把论证、证据和反思判断看作批判性思维发展的核心。认知自治继承批判性思维的理由取向，同时更突出模型中介、平台传播和自动化决策造成的新判断环境。

元认知强调个体对自身认知过程的认识和监控。Flavell (1979) 早期提出元认知包含关于认知任务、策略和自身能力的知识，以及对认知过程的监控。Nelson and Narens (1990) 的框架把元认知区分为监控和控制两个方向。认知自治把元认知扩展到人机环境之中：学习者需要监控自己是否理解 AI 答案、是否被流畅表达影响、是否过度依赖工具、是否在不确定情境中给出过高信心。

自我调节学习强调学习者围绕目标主动规划、监控和调整学习过程。Zimmerman (2002) 将自我调节学习描述为一个包含前摄、行为表现和自我反思的循环过程。认知自治继承这种目标导向和循环调节，同时把调节范围从学习过程扩展到信息获取、信念形成、工具使用、判断选择与行动责任。

认识论认知关注学习者如何理解知识的性质、来源、确定性和辩护方式。Hofer and Paul R. Pintrich (1997) 认为，学生关于知识的信念会影响学习过程。Sandoval, Greene, and Bråten (2016) 进一步指出，认识论认知研究需要关注学习者在不同情境中如何思考知识来源和证据辩护。认知自治把认识论认知推进到智能社会场景中：知识来源可能是人、机构、模型、平台或混合系统，知识辩护也可能包含数据、算法、检索证据、模型评估和人工判断。

信息素养和媒介素养提供了另一种视角。信息素养强调识别信息需求、获取信息、评价信息和合乎伦理地使用信息。媒介素养强调理解媒介生产、符号表达、传播结构和权力关系。智能社会把数据、算法、生成模型和自动化评价进一步纳入信息环境。学生需要理解信息的来源与质量，也需要理解这些技术和系统如何改变信息的生产、呈现与传播方式。

表 3.2: 认知自治与相关教育概念的关系

相关概念	**核心关注**	**对认知自治的支持**	**认知自治的进一步拓展**
批判性思维	理由、证据、论证、推理和评价	支持学生分析 AI 输出和社会信息的证据质量	把理由判断放入模型中介和平台传播环境
元认知	对自身认知过程的监控和调节	支持学生觉察理解程度、信心强度和工具依赖	把自我监控扩展到人机协同学习过程
自我调节学习	围绕目标对学习过程进行规划、监控、调整和反思	支持学生主动组织学习策略并依据反馈持续改进	把循环调节扩展到信念、工具、判断与行动
认识论认知	知识来源、知识确定性和知识辩护	支持学生理解不同知识主张如何获得正当性	把人、机构、模型、平台和混合来源纳入知识判断

相关概念	核心关注	对认知自治的支持	认知自治的进一步拓展
信息与媒介素养	信息获取、媒介生产、传播结构和伦理使用	支持学生判断信息来源、表达形式和传播机制	把数据、算法、生成模型和自动化评价纳入信息环境分析

通过这一比较可以看出，认知自治把已有教育概念放入智能社会的共同处境中重新组织。学生需要批判性思维来分析理由，需要元认知和自我调节学习来监控与调整自身活动，需要认识论认知来理解知识主张的来源和辩护，也需要信息素养与媒介素养来评价信息及其传播机制。认知自治进一步关注这些能力能否共同进入工具选择、人机协作、日常判断和现实行动，并形成持续的主动调节过程。

3.6 可能的质疑与回应

将认知自治确立为智能社会通识教育的时代使命，可能面临若干质疑。这些质疑主要涉及目标的可操作性、概念提出的必要性、教育目标的完整性以及对技术的基本态度。

第一类质疑认为，这一目标过于抽象，难以落实。回应这一质疑，需要用来源警觉、证据敏感、边界意识、信任校准、认知能动和责任判断六个维度观察认知自治过程，并通过人工智能案例、社会案例、真实任务、人机协作、跨情境任务和反思记录加以训练。本书后续各篇将从能力结构、课程内容、教学实施和学习评价等方面进一步建立这种可操作性。

第二类质疑认为，批判性思维、元认知、信息素养等已有概念已经能够覆盖相关教育目标，没有必要再提出认知自治。对这一质疑的回应是，认知自治吸收了这些概念的核心成果，同时回应了智能社会中认知环境的结构性变化。信息来源从人与机构扩展到模型、平台和人机混合系统，认知活动从个人头脑延伸到由数据、算法、工具和组织共同构成的分布式过程，模型输出还可能直接进入现实行动。认知自治以人的主动调节为主线，把来源、证据、边界、信任、能动和责任组织为一个连续过程，从而为这些原本分属不同研究传统的能力提供了共同的问题情境和统摄目标。

第三类质疑认为，认知自治固然重要，但若将其视为人工智能通识教育的唯一目标，可能会遮掩其他目标，比如情感态度、人文精神和伦理意识。单提“认知自治”，是不是会使教育目标窄化？对这一质疑的回应是，每一门学科对思维和人格的训练都是多维的，人工智能通识课也是如此，“认知自治能力”只是其中众多认知能力中的一种。但是，认知训练是有主次之分的，那些只有本门学科能完成、其他学科不能替代的训练，才是这门学科的基础任务。对人工智能通识教育来说，它固然也可以对学生进行多方面的训练，包括情感态度、人文精神和伦理意识，但这些并非它本身独特的作用，而是具有普遍性的，其他学科也可以承担。相对来说，人工智能通识教育对于认知自治的培养具有其他学科难以替代的作用。这种独特性不是人为指定的，而是由人工智能学科本身的特性以及人工智能塑造智能社会的过程共同决定的。这一点我们将在下一篇详细讨论。

因此，我们将“认知自治能力”作为人工智能通识教育的引领性目标，并不意味着这门课程只有这一个目标。事实上，这本红皮书会给出细致的方案，以促进多角度的思维训练和人格养成。我们所坚持的是，一门课必须有它独特的目标；对人工智能通识课来说，这个独特的目标就是“认知自治能力培养”。如果这一目标没有得到落实，那么其他方面的教育成效无论多么丰富，也无法弥补人工智能通识教育核心功能的缺失。反之，对认知自治能力的培养必然会带动证据意识、批判精神、责任判断等相关素养的发展，因为这些素养本身就是形成认知自治能力的重要条件。这就是将认知自治作为“统摄性目标”的意义。

第四类质疑认为，强调来源、证据、边界、信任和责任，可能使认知自治演变为风险防范教育，使学习者对人工智能保持防御姿态。对这一质疑的回应是，认知自治的目标超出风险宣讲，它追求对技术的成熟理解，并形成主动面对、积极探索和扩展自身能力的认知能动。学习者在了解人工智能能力形成方式和适用边界的基础上，能够判断它适合完成什么任务，积极利用人工智能学习、创造和解决问题，同时保持证据敏感、边界意识和责任判断。对机制和边界的理解能够形成有根据的信心，使人避免盲目排斥与盲目依赖，在积极使用、审慎信任和必要接管之间形成成熟的行动方式。

3.7 本章小结

智能社会改变了信息生成、知识组织、判断形成和行动实施的基本条件，也使通识教育获得了新的时代任务。认知自治是个体在复杂认知环境中，对自身信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节的过程；认知自治能力，则是个体发起、维持和调整这一过程，使其能够跨越不同任务与情境持续展开的能力。二者的区分，使通识教育既关注真实发生的认知活动，也关注支撑这种活动稳定开展和跨情境迁移的个体能力。

认知自治建立在开放依赖之中。人可以有选择地调用专家知识、制度资源和人工智能的检索、计算、生成与分析能力，无需独立完成全部认知活动；同时需要保持对认知过程的发起、审查和统摄，尤其保持提出问题和核查结果的角色。来源警觉、证据敏感、边界意识、信任校准、认知能动和责任判断构成认知自治过程的六维度结构。来源警觉、证据敏感、边界意识、信任校准和责任判断分别对应信念形成、证据使用、判断边界、信任分配与行动责任，认知能动则为主动面对技术变化、探索认知可能和扩展自身能力提供内在动力。六个维度共同体现了人在利用外部能力的同时保持认知主导的基本要求，也为课程设计、教学观察和过程诊断提供了分析框架。

认知自治既具有个体意义，也具有公共意义。它帮助个体主动组织外部认知资源、核查结果并觉察工具对自身认知状态的影响，也关系到公共讨论、组织决策、学习与生产参与、教育公平以及人的尊严和主体性。批判性思维、元认知、自我调节学习、认识论认知、信息素养和媒介素养为这一目标提供了重要基础；认知自治将这些传统置于模型中介、平台传播、人机协同和自动化决策的共同环境中，使其汇入持续的主动调节过程。

至此，认知自治作为智能社会通识教育的时代使命得到了基本界定。下一章将在此

基础上转向人工智能通识教育本身，讨论它作为一种教育形态如何形成、现实实践中存在哪些目标偏差，以及如何围绕认知自治重新校准其教育定位。人工智能通识教育为什么能够并且应当承担发展认知自治能力的任务，还需要进一步考察人工智能的科学属性、思维方式及其进入社会运行的基本机制，这将由下一篇展开论证。

第四章　人工智能通识教育：定位、意义与实践偏差

摘要：前一章将认知自治确立为智能社会通识教育的时代使命。本章进一步界定承担这一使命的教育形态，即人工智能通识教育。人工智能通识教育是通识教育与人工智能教育相融合形成的教育形态。它以人工智能的知识、方法、系统及其社会运行为学习内容，承担通识教育在智能社会中培养认知自治能力的使命。在现实实践中，人工智能教育、人工智能通识教育与人工智能素养的概念边界仍较含混，工具操作、技术训练、项目竞赛、前沿专题、伦理提醒和作品评价等局部形式容易被用来代表完整的人工智能通识教育。本章据此分析工具化、技术化、活动化与竞赛化、前沿化与碎片化、伦理附加化、评价浅层化和教师培训表层化等主要偏差，并考察国内外政策与能力框架中正在出现的认知转向。厘清教育定位、学习结果与统摄目标之间的关系，是人工智能通识教育从内容扩张走向目标自觉的前提。

关键词：人工智能通识教育；人工智能教育；人工智能素养；认知自治；教育定位；现实偏差；认知转向

4.1　人工智能通识教育的概念界定

前一章从智能社会的认知环境出发，将认知自治确立为通识教育的时代使命。目标得到界定以后，还需要进一步回答：什么样的教育能够承担这一使命？人工智能通识教育正是在这一问题中获得其教育定位。

直观上，智能社会的认知困境很大程度上是因为人工智能的介入，学习人工智能可以帮助学习者理解人工智能系统运行的基本原理和运行方式，从而建立对智能社会的基础认知，而这是达成认知自治的先决条件。[1]通识教育和人工智能教育在此交汇，成为“人工智能通识教育”的共同源头。

从通识教育的角度来看，“培养认知自治能力”是它必须承担的历史使命，而学习人工智能为完成这一使命提供了可行的路径。因此，人工智能通识教育是一种特殊的通识教育，是以教授人工智能的方式实现“培养认知自治能力”的通识教育。

从人工智能教育角度来看，“人工智能通识教育”是一种特殊的“人工智能教育”。人工智能教育是围绕人工智能开展的各类教育的上位概念，涵盖人工智能的知识、方法、

[1]“智能社会的认知困境很大程度上是因为人工智能的介入”只是一个笼统的直觉，要说明人工智能学习可以实现认知自治能力的培养，需要深入分析人工智能学科的特性和思维方式，以及人工智能进入社会活动的方式。这些将在第二篇展开。

系统、工具、应用及社会影响，也包括面向专业培养、职业发展和一般学习者的不同教育形态。在大多数场景下，它更关注技术学习与技能养成，并不把认知自治作为主要目的。人工智能通识教育是以通识教育目标引领的人工智能教育，是人工智能教育为适应通识教育“培养认知自治能力”的目标而建设的特殊形式。

综合起来，本书将“人工智能通识教育”定义为：

人工智能通识教育是通识教育与人工智能教育相融合形成的教育形态。它以人工智能的知识、方法、系统及其社会运行为学习内容，承担通识教育在智能社会中培养认知自治能力的使命。

这一融合包含两个相互联系的方向。通识教育为人工智能通识教育提供教育使命与价值目标，使人工智能学习面向全体社会成员，关注人的心智发展、价值判断和公共生活；人工智能教育为这种使命提供直接的学习对象、知识内容和实践方式，使学习者能够理解人工智能的能力如何形成、系统如何运行，以及人工智能如何进入信息传播、知识生产、社会组织和现实行动。缺少通识教育的使命，人工智能教育容易收缩为专业知识、工具操作或职业技能；缺少对人工智能的系统学习，智能社会中的认知问题又容易停留在一般性的原则告诫。二者相互融合，才形成完整的人工智能通识教育。需要强调的是，这种融合并不意味着人工智能通识教育可以简单归入二者中的任何一类。它不是人工智能专业教育的浅化版本，也不是传统通识教育增加若干人工智能内容后的扩大版本，而是二者融合后形成的独立教育形态。

认知自治在这一融合中发挥统摄作用。学习人工智能可以使学习者理解数据、模型、目标、评价、泛化、误差和反馈等系统环节，认识人工智能的能力来源与适用边界；分析人工智能的社会运行，则可以使学习者看到平台、组织和制度如何借助人工智能参与信息筛选、判断形成和行动实施。这样的学习既支持学习者主动运用人工智能扩展自身能力，也训练其追问信息来源、分析证据、把握边界、校准信任和承担责任。认知自治因此构成人工智能通识教育的根本目标，并把知识学习、能力发展和责任意识组织为一个整体。

4.2 人工智能素养、认知自治与认知自治能力

人工智能通识教育还需要与人工智能素养、认知自治和认知自治能力区分开来。现有政策与能力框架通常从理解人工智能、使用人工智能、评价人工智能输出、创造性应用以及伦理和责任等方面描述人工智能素养 (Miao, Shiohira, and Lao, 2024; 教育部基础教育教学指导委员会, 2025a; OECD and European Commission, 2026)。本书据此将人工智能素养理解为学习者在人工智能情境中形成的知识、技能、态度与责任等综合学习结果。它回答学习者经过人工智能教育以后，在理解、使用、评价和负责任参与方面形成了什么。

认知自治回答的是另一层问题：个体在复杂认知环境中，能否主动调节自身的信念形成、证据使用、信任分配、判断边界和行动责任。认知自治是一种持续发生的认知过

程，也是人工智能通识教育的统摄性目标；认知自治能力则是个体发起、维持和调整这一过程，使其能够跨越不同任务与情境持续展开的能力。

四个概念在本书中的位置可以概括为表4.1。

表 4.1: 人工智能通识教育相关概念的层次与关系

概念	概念性质	核心内容	在本书中的位置
人工智能通识教育	教育形态与教育过程	面向全体社会成员组织人工智能知识、方法、应用及其社会影响的学习	通识教育与人工智能教育的融合，承担智能社会中的通识教育使命
人工智能素养	人工智能情境中的直接学习结果	理解、使用、评价和负责任参与人工智能所需的知识、技能、态度与责任	反映学习者在人工智能领域中已经形成的综合素养
认知自治	统摄性教育目标与主动调节过程	对信念形成、证据使用、信任分配、判断边界和行动责任进行主动调节	为课程内容、学习活动和评价证据提供共同方向
认知自治能力	可发展、可迁移的稳定能力	发起、维持和调整认知自治过程，使其跨任务与情境持续展开	将认知自治转化为可以培养、观察和评价的能力目标

人工智能素养与认知自治相互联系，范围和功能各有侧重。人工智能素养主要描述学习者在人工智能情境中的直接学习成果，认知自治则贯穿人工智能场景以及一般的信息判断、知识学习、社会参与和行动选择。人工智能通识教育需要形成扎实的人工智能素养，并进一步推动学习者把在人工智能学习中形成的来源意识、证据意识、边界意识、信任调节和责任判断迁移到更广泛的社会情境。第二篇将继续说明，人工智能的科学属性及其进入社会活动的方式，怎样为这种能力发展和跨情境迁移提供基础。

4.3 关于人工智能通识教育的讨论

在完成概念界定以后，还需要进一步说明人工智能通识教育与通识教育、人工智能专业教育和信息科技教育之间的关系。这些关系直接影响人工智能通识教育的课程定位、内容选择和实施方式。人工智能通识教育既继承通识教育的历史使命，又回应智能社会提出的新问题；它以人工智能学科为知识基础，同时具有不同于专业教育和信息科技教育的培养目标。只有明确这些关系，人工智能通识教育才能形成稳定而完整的教育形态。

4.3.1 人工智能通识教育的双重社会意义：共同基础与认知主体

通识教育长期承担着建设社会共同基础的任务。知识分化使社会成员掌握不同领域的知识，职业分工使人们在不同制度和实践中生活，多样化的经验和价值立场又使公共问题难以依靠单一知识传统得到理解。通识教育由此需要帮助社会成员形成基本的共同知识、思维方式、表达能力、价值意识和公共责任，使来自不同背景的人能够理解彼此、交换理由、参与公共讨论并共同处理社会事务。哈佛委员会在《自由社会中的通识教育》中把共同文化、共同公民身份和自由社会的维持联系起来，强调通识教育对于公共生活的基础作用 (Harvard University Committee on the Objectives of a General Education in a Free Society, 1945)。

人工智能通识教育首先继承了这一传统任务。人工智能正在进入知识生产、信息传播、公共讨论、组织管理和社会决策，逐渐成为不同社会成员共同面对的认知基础设施。无论个体是否从事人工智能专业工作，都需要对数据、模型、算法推荐、生成内容和自动化决策形成基本理解。缺少这些共同知识，人们便难以讨论人工智能系统的能力与风险，也难以就其使用边界、责任归属和制度安排形成有效的公共对话。人工智能的基本知识和判断方式因此正在成为智能社会共同知识结构的重要组成部分。

人工智能通识教育同时承担着智能社会的特殊任务，即维护人的主体性。人工智能影响的已经超出人获得信息和使用工具的方式，开始进入问题提出、材料选择、理由组织、判断形成和行动实施等认知环节。个体可以借助人工智能扩展学习、创造和行动能力，也可能在持续依赖中逐渐失去对自身认知过程的觉察和调节。人工智能通识教育需要使学习者理解人工智能如何生成结果、结果受到哪些数据和目标影响、系统能力具有哪些边界，并在此基础上学会调节信任、核查证据、组织人机分工和承担行动责任。Miao, Shiohira, and Lao (2024) 提出的学生人工智能能力框架，也把人的能动性、责任、公民身份、人工智能伦理、技术理解和系统设计置于同一框架之中，说明人工智能教育的公共意义已经超出技术操作。

这两重意义都关系到社会的稳定运行。共同知识和公共语言使社会成员能够展开对话，认知自治则保证参与对话和作出决定的人仍然具有独立判断能力。公民如果无法理解人工智能所中介的信息环境，公共讨论便会失去必要的共同基础；公民如果逐渐把信息选择、理由形成和最终判断交由人工智能系统完成，公共生活的主体基础也会受到削弱。民主制度可以在形式上继续运行，支撑制度运行的公民主体却可能逐渐空心化。建设共同基础与维护认知主体因此相互依存，共同构成智能社会通识教育的公共使命。

从这一意义上看，人工智能通识教育属于通识教育在智能社会中的新形态。它延续了通识教育发展共同知识、公共判断和社会责任的历史任务，同时把通识教育所面对的环境扩展到由数据、模型、平台和人工智能系统共同中介的认知世界。它还把通识教育对独立心智的培养推进到人机协作条件下的认知主导，使学习者能够在开放依赖中保持对信念、证据、信任、边界和责任的主动调节。人工智能通识教育由此拓展了传统通识教育的内容边界、能力边界和实现方式。

4.3.2 人工智能通识教育不应视为专业教育的浅显化、通俗化

人工智能通识教育不能被理解为人工智能专业教育的浅显化或通俗化。两者都以人工智能的概念、知识、方法和系统为学习对象，在培养对象、教育目标、内容组织和学习结果方面具有不同定位。

人工智能专业教育主要面向未来从事人工智能研究、开发和工程实践的专业人员。它需要建立系统而深入的数学、计算和工程基础，使学习者能够理解算法原理，完成模型设计、系统实现、实验评价和技术创新。学习深度、技术完整性和专业实践能力构成专业教育的重要要求。

人工智能通识教育面向所有社会成员。人工智能知识在这里既是需要理解的内容，也是培养认知自治能力的重要载体。学习者通过数据标注理解事实如何被转化为可计算对象，通过机器学习理解结论如何从有限经验中归纳形成，通过模型测试理解已有规律如何迁移到新情境，通过生成模型理解流畅输出与可靠知识之间的差异，通过人机协作理解工具使用、信任分配与行动责任之间的关系。人工智能的专业内容由此进入更广泛的认知训练过程，支持归纳、泛化、判断与自醒能力的发展。

同一个人工智能概念，在两种教育中可以承担不同功能。专业教育学习分类模型，可能要求学习者推导目标函数、实现训练算法并比较模型性能；通识教育学习分类模型，则可以通过样本选择、标签形成、错误分布和应用后果，理解数据如何影响结论以及模型为何具有适用边界。专业教育研究生成模型，重视模型结构、训练方法、推理效率和系统优化；通识教育研究生成模型，侧重理解生成机制、输出不确定性、信息核验、信任校准和使用责任。内容深浅只是表面差异，学习目的和知识组织方式才构成两者的根本区别。

通识教育与专业教育具有各自的教育目标，两者并行发展、相互支持，无高下之分，也不存在固定的先后顺位。人工智能通识教育不只是专业学习的预备阶段，专业教育也无法自然包含通识教育的全部任务。人工智能专业人员同样需要形成认知自治、公共责任和社会判断，其他专业及普通公众也需要理解人工智能并在智能社会中保持认知主导。

两种教育还具有不同的社会作用。专业教育培养能够推动人工智能研究和技术发展的专门人才，人工智能通识教育则面向更广泛的社会成员，维护智能社会所需要的共同知识、基本判断能力和公民主体性。从人才培养的专业深度看，专业教育具有不可替代的价值；从覆盖人群和维持社会共同认知条件看，人工智能通识教育具有更广泛的公共基础性。它的社会价值并不来自知识层级更高，而来自其培养对象覆盖全体社会成员，并直接关系智能社会能否拥有能够理解、判断和负责的公民。

4.3.3 人工智能通识教育不应视为信息科技教育的延展

信息科技教育为人工智能通识教育提供了重要基础。数据处理、算法、程序设计、网络、数字工具和信息社会责任等内容，都与人工智能学习存在密切联系。我国义务教育信息科技课程已经形成以信息意识、计算思维、数字化学习与创新、信息社会责任为核心的课程体系，并逐步吸收算法和人工智能相关内容 (中华人民共和国教育部, 2022)。

信息科技课程吸收人工智能知识，有助于更新自身内容，使学生了解数字技术发展的新阶段。

这种联系并不意味着人工智能通识教育可以简单归入信息科技教育。从学科基础看，经过长期发展，人工智能已经形成相对独立的研究对象、基本问题、概念体系、方法谱系和评价方式。人工智能关注智能活动如何被表示、计算和实现，研究机器如何感知环境、表示知识、从数据中学习、进行推理与搜索、作出决策并与人和环境交互。表示、学习、泛化、推理、搜索、优化、决策和反馈等问题，共同构成人工智能的学科结构。第二篇第一章将进一步说明，人工智能已经形成从基础理论、核心机制到系统应用的完整科学体系。

信息科技教育关注信息的获取、表示、传输、处理、交流和应用，重视个体利用数字技术解决问题和参与信息社会的能力。人工智能与信息科技共享数据、算法和计算系统等基础，但两者对这些内容的组织方向有所不同。人工智能需要进一步追问智能活动如何转化为计算对象，模型如何从经验中形成规律，所得规律为什么能够或不能够适用于新对象，以及系统如何在不确定环境中形成输出和行动。这些问题体现了人工智能独有的学科对象和思维方式。将人工智能完全作为信息科技的局部应用处理，容易使其学科结构被压缩为若干工具、算法或操作活动，削弱学生对人工智能基本思想和运行机制的系统理解。

从教育目标看，人工智能通识教育承担着培养认知自治能力的独特使命。人工智能同时具有认知镜像和社会中介两方面特征：一方面，人工智能把表示、学习、归纳、泛化、判断和决策等智能活动转化为可以观察和分析的计算过程；另一方面，人工智能系统已经进入现实社会的信息传播、知识生产、组织评价和行动决策。学习人工智能因而能够使学习者同时观察机器的认知过程和人自身的认知过程，并进一步理解这种计算过程进入社会以后产生的影响。这种教育使命需要围绕来源、证据、边界、信任、认知能动和责任组织教学内容，也需要通过模型比较、边界测试、输出核验、人机分工和社会情境判断等方式设计学习任务。传统信息科技课程的目标和内容无法完整涵盖这一要求。

从课堂实践看，信息科技课程已经具有相对完整的内容结构、课时安排和教学进阶。在课程结构和课时总量保持不变的情况下，人工智能内容进入信息科技课程，通常只能作为部分单元、案例或拓展活动。这些内容可以丰富信息科技教育，也能够为学生提供初步的人工智能体验，却难以承载从人工智能基本机制、学科思维到社会影响和认知自治的完整培养过程。零散介绍生成工具、模型应用或简单编程，也难以形成连续的能力进阶和评价体系。

人工智能通识教育在学校中的组织方式可以保持开放。它可以独立设置课程，也可以与信息科技、科学、通用技术、社会科学和综合实践活动协同实施。教育部基础教育教学指导委员会发布的《中小学人工智能通识教育指南（2025 年版）》已经提出由知识、技能、思维和价值观共同构成的人工智能素养目标；一些地区的实施方案也同时允许独立设课和与现有课程融合开展 (教育部基础教育教学指导委员会, 2025a; 北京市教育委员会, 2025)。课程在行政安排上采用何种载体，可以根据课时、师资和学校条件灵活决定；但人工智能通识教育的目标体系、内容结构、发展进阶和评价要求需要得到完整体现。

因此，信息科技教育可以吸收人工智能知识，以回应数字技术的发展；人工智能通识教育也可以利用信息科技课程的师资、设备和课程基础。然而，两者各有自身的学科内容、思维方式和教育使命。信息科技课程中的人工智能内容能够成为人工智能通识教育的组成部分，但有限的内容延展难以独立完成维护认知自治和公民主体性的时代任务。

4.3.4 人工智能通识教育的独立教育地位

上述讨论表明，人工智能通识教育的教育性质来自三个方面。它以认知自治为统摄性目标，因而承担通识教育在智能社会中的时代使命；它以人工智能的科学体系为内容基础，因而具有独立而完整的知识来源；它面向全体社会成员，因而区别于以专业研究和技术开发为目标的人工智能专业教育。同时，它可以与信息科技及其他课程共享部分内容和实施条件，但这种课程联系不能消解其独立的教育目标。

人工智能通识教育的独立地位，并不要求所有学校立即建立统一的独立课程名称，也不排斥跨学科融合。真正需要保持的是教育使命的完整性：课程能否使学习者理解人工智能的基本机制和学科思维，能否认识人工智能进入社会的过程及其影响，能否在学习和使用人工智能的过程中发展归纳、泛化、判断与自醒能力，能否在开放依赖中保持对信念、证据、信任、边界和责任的主动调节。只有这些目标得到系统实现，人工智能教育才能超越知识普及和工具训练，真正成为智能社会中的基础认知自治教育。

4.4 人工智能通识教育的现实错配

目前，人工智能通识教育刚刚起步，一些基本概念在教育实践中尚未得到充分区分。专业教育、技能培训、兴趣活动和通识教育经常使用相近的课程名称，局部教育目标随之被带入面向全体学习者的课程，并逐渐取代通识教育的整体目标。后文所讨论的许多偏差，都与这种现实中的概念混淆有关。

“人工智能教育”在政策传播和日常表达中，常被用作概括相关教育活动的领域性名称，专业教育、技能培训、兴趣活动和面向公众的课程都可能被纳入其中。这种宽泛用法便于描述整个实践领域，却不能直接说明一门具体课程的教育性质。问题出现在具体课程设计中：专业培养、技能训练或兴趣活动中的局部实践被直接理解、宣传为人工智能通识教育，并以此为名大范围扩散。

这种混淆首先表现为把学习对象等同于教育性质。只要课程涉及算法、编程、机器人、生成式人工智能或人工智能伦理，就被归入同一类教育，并以“人工智能通识教育”之名传播。然而，同一内容可以承担不同任务。算法和编程可以培养专业实现能力，也可以帮助普通学习者理解模型如何形成结果；生成式人工智能工具可以用于岗位技能训练，也可以成为比较输出、核查证据和反思依赖方式的通识学习材料。课程教授了什么，只能说明其学习对象，课程怎样组织这些内容、希望学习者形成什么能力，才决定其教育性质。

第二种表现是把对象的普遍性等同于目标的通识性。课程面向全体学生，是人工智

能通识教育的重要条件；面向全体学生开设工具课、编程课或一次性体验活动，并不会自然形成通识教育。通识性关键在于课程能否把人工智能知识与共同知识、心智能力、价值意识和公共责任联系起来，能否使不同背景的学习者理解人工智能、主动运用人工智能，并在复杂情境中调节证据使用、信任分配、判断边界和行动责任。

第三种表现是用课程名称和活动形式代替目标判断。名为“人工智能通识课”的课程可能沿用压缩后的专业课程结构；科技节、机器人竞赛和创新项目可能被直接作为学校通识教育成果；带有伦理专题的课程也可能仍以工具操作和作品制作为主体。名称、对象和形式都可以提供初步线索，判断一项实践是否实现人工智能通识教育，还需要考察其目标、任务和评价是否共同指向认知自治。

这种含混有其现实来源。中小学人工智能课程常从信息技术、创客教育、机器人活动、科技竞赛或课后社团发展而来，高校相关课程常从人工智能专业课、计算机公共课和“人工智能 + 专业”课程延伸而来，社会培训则更多从产品使用和岗位能力出发。不同路径为人工智能教育积累了丰富资源，也保留了各自原有的目标和实施惯性。当这些内容进入人工智能通识教育时，如果没有围绕新的教育使命重新选择和组织，原有的专业逻辑、技能逻辑、活动逻辑和成果展示逻辑就会继续主导课程。

因此，本节关注这些局部形式在通识教育中的位置错配：工具操作被用来代表完整课程，便形成工具化；专业技术训练被直接移作全体学习者的共同基础，便形成技术化；活动和竞赛被用来代表普遍培养成效，便形成活动化与竞赛化。前沿化、伦理附加化、评价浅层化和教师培训表层化也具有相同结构。关键问题不是教学内容是否有价值，而是这些内容及其教学方式是否完成了人工智能通识教育的整体目标。

4.4.1 工具化倾向：把通识教育压缩为操作能力

人工智能通识教育中最容易出现的偏差，是工具化。生成式人工智能普及之后，许多课程和培训以“如何快速写作”“如何生成 PPT”“如何用 AI 做图”“如何设计提示词”“如何提高办公效率”为主要内容。这样的训练短期见效快，学习者容易获得成就感，也能服务真实工作需要，因而在技能培训和实际工作中具有明确价值。工具化偏差产生于另一种情形：工具操作被当作人工智能通识教育的主要内容，原理理解、结果核验和责任判断随之退出课程。

工具化倾向有三个成因。第一，工具界面直接可见，教学效果容易展示。相比解释数据分布、模型边界和责任归属，演示一个生成工具更容易在课堂中产生即时反馈。第二，生成式人工智能的输出能力强，能快速产出文本、图像、代码和方案，教师和学习者容易把产出质量当作学习质量。第三，社会对效率的需求会把部分人工智能教育推向工具培训，学校和培训机构也倾向于用可见成果证明课程价值。

工具训练有其正当位置，适当的实践体验有助于学生理解 AI 对生活和学习的影响。然而，工具训练必须有一个更深刻的目标，才能成为通识教育的一部分。UNESCO 关于生成式 AI 在教育和研究中的指导文件也指出，教育系统需要帮助学习者理解并适当使用生成式 AI，同时强调人类能动性、伦理、安全和公平 (Miao and Holmes, 2023)。工具训练需要导向理解、判断和反思，才能进入人工智能通识教育的整体结构。

工具化的教育风险，可以从心理学和人机交互研究中得到解释。Risko and Gilbert (2016) 把“认知卸载”概括为人利用外部工具减少内部认知需求的过程，这一过程能够提高表现，也可能改变人对任务的投入方式。Sparrow, Liu, and Wegner (2011) 关于“Google effect”的研究显示，当人预期信息能够被外部系统保存或检索时，他们对信息本身的记忆会下降，而对信息位置的记忆会上升。这些研究并不支持简单的技术悲观论，却提醒教育者：外部认知工具会重组人的注意、记忆和学习策略。生成式 AI 作为更强的认知工具，可能进一步把问题定义、语言组织、证据检索和初步判断外包出去。

人机自动化研究也提供了相关证据。Parasuraman and Riley (1997) 早在 1997 年就提出自动化的使用、误用、弃用和滥用问题，指出过度信任自动化可能导致监控不足和使用不当。Skitka, Mosier, and Burdick (1999) 关于自动化偏差的研究表明，人们在有自动化建议时可能忽略相反证据或执行错误建议。这些发现说明，工具越强大，教育越需要训练使用者的监控、核验和责任判断。人工智能通识教育若只训练“如何让 AI 给出结果”，学生可能形成更强的操作能力，同时削弱对结果来源、证据强度和适用边界的警觉。

工具化倾向在课堂中常表现为“提示词中心主义”。提示词确实是人与生成式 AI 互动的重要入口，设计清晰任务、提供背景、限定格式和要求自我检查，能够提高输出质量。把提示词作为通识教育核心，则会使课程被工具界面牵引。学生可能学会许多模板，却不能解释模型为何产生幻觉；能够生成漂亮文章，却不能说明证据在哪里；能够让 AI 修改方案，却不能判断修改是否合理；能够在某一模型上得心应手，但当模型更新后可能失去优势。

提示词等技巧性教学可以成为课程的一部分，但应当服务于问题定义、证据核验、输出比较和反思记录，不能成为课程目的本身。

可用一个简单案例说明。教师让学生用生成式 AI 写一段关于“人工智能改变教育”的短文。若任务只要求输出一篇格式完整的文章，学生很可能把 AI 结果稍作修改后提交。若任务要求学生比较三个不同提示词所得结果，标出其中未经证实的事实，查找一条原始依据，说明 AI 答案的可用部分和不可用部分，并写下自己的判断过程，同样的工具使用就转化为认知训练。前者训练调用，后者训练校准。

因此，修正工具化倾向的关键，是把 AI 工具的使用转化为认知训练场。课程可以教学生使用工具，但每一次使用都应伴随四个问题：这个输出从何而来？它有什么证据？它适用于什么情境？我为什么相信或暂时不相信它？当工具训练被置于这些问题之下，它才可以进入人工智能通识教育的结构。

4.4.2 技术化倾向：把共同基础缩减为算法和编程

与工具化相对的另一类偏差，是过度技术化。编程、算法、模型训练和技术实现是人工智能专业教育的核心内容，也能帮助普通学习者理解 AI 系统的内部机制。许多课程从 Python、分类器、神经网络、数据标注、模型调参和机器学习项目切入，本身具有明确的教育价值。技术化偏差发生在专业技术训练被直接移作全体学生的通识方案时：课程门槛、内容深度和评价标准沿用专业培养逻辑，理解技术与社会关系所需的判断、

迁移和责任学习则受到挤压。

技术化倾向有其历史来源。计算机科学教育长期以来以编程、算法、数据结构为核心。Wing (2006) 提出“计算思维”概念，强调分解、抽象、算法化和自动化思维对所有人的重要性。这一思想推动了全球范围内计算机教育和编程教育的发展。人工智能教育继承了计算机科学教育资源，也容易沿用“学习技术即学习实现”的路径。

AI 带来的通识教育问题已经超出传统计算思维的边界。Zeng (2013) 在“从计算思维到 AI 思维”的讨论中指出，人工智能涉及知识、语义、学习、语境和非符号数据等方面，不能完全由传统计算思维涵盖。AI4K12 的“五个大观念”也没有把面向中小学生的 AI 学习限定为机器学习实现，它同时包含感知、表示与推理、学习、自然交互和社会影响 (Touretzky et al., 2019; AI4K12 Initiative, 2019)。这些框架说明，人工智能通识教育需要技术理解，也需要由代码和算法通向系统、应用和社会影响。

技术化课程常有一个隐含假设：只要学生理解算法原理，就能够形成较为完整的人工智能素养。这个假设需要修正。学生可能会训练一个图像分类器，却没有理解训练数据如何反映社会偏差；可能会实现一个推荐算法，却没有理解推荐目标如何影响人的注意力；可能会知道神经网络由层和参数组成，却无法判断一个人工智能医疗建议是否需要医生复核；可能会写出代码，却不懂得在高风险场景中如何说明责任边界。技术知识是人工智能通识教育的重要基础，但技术知识不会自动生成社会判断和认知自治。

这一点在机器学习教育中尤其明显。面向学生的机器学习项目常从“采集数据-训练模型-测试准确率”展开。这样的流程能够帮助学生理解模型从数据中学习，也能产生直观作品。但如果课程只关注完成训练流程，学生容易忽视数据来源、样本代表性、测试集设计、分布变化、错误类型和应用风险。实际 AI 系统中的问题，常常出在这些环节。一个猫狗分类器准确率很高，并不意味着学生理解了算法公平；一个语音识别项目运行成功，也不意味着学生理解了不同口音、方言和环境噪声带来的数据分布问题。

教育学中的“学科教学知识”概念提醒我们，掌握内容与教会内容之间存在差异。Shulman (1986) 提出 pedagogical content knowledge，强调教师需要理解学科内容如何转化为可教学的形式。Mishra and Koehler (2006) 在此基础上提出 TPACK 框架，指出技术整合需要内容、教学法和技术知识的复杂结合。人工智能教育尤其需要这种结合。一个 AI 研究者能够解释深度学习细节，却未必知道四年级学生如何通过生活案例理解数据偏差；一个信息技术教师能够教编程，却未必知道如何把算法判断与社会责任联系起来。

修正技术化倾向，需要建立“通识化技术理解”。通识化要求把技术原理转换为面向所有学习者的关键问题，使学生能够理解数据与模型的基本关系、学习与生成的主要机制、系统的误差与适用边界，以及技术进入真实场景后的责任要求。技术知识由此成为认知自治的基础语言，避免被处理为专业训练的压缩版本。

4.4.3 活动化与竞赛化倾向：用局部成果替代普遍培养

人工智能教育在许多学校首先以活动、社团、竞赛、科技节、夏令营和项目展示的形式出现。这种路径具有现实合理性。新领域进入学校时，常常先通过兴趣活动试点，

积累教师经验和学生案例。项目活动能够激发好奇心，降低入门压力，也能体现人工智能的开放性和创造性。活动和竞赛还可以服务兴趣培养、拔尖教育与创新人才选拔。活动化与竞赛化偏差发生在这些局部项目被用来代表全体学生的通识教育成效时，课程的普遍性、连续性和学段递进因而被削弱。

活动化倾向常见于三个层面。第一，学习以一次性体验为主。学生参加一次 AI 绘画、智能机器人、模型训练或提示词活动，获得新奇感，却没有形成稳定概念。第二，课程以作品展示为中心。学生完成一个看起来完整的项目，但项目背后的数据、模型、评价、错误和责任没有被充分分析。第三，学校以竞赛成绩代表教育质量。少数学生参加比赛并获奖，容易被视为学校人工智能通识教育成效，广大普通学生的基础素养则缺少保障。

好的研究性项目和竞赛能够把概念、方法、应用和伦理结合起来，也能为有兴趣的学生提供深入探索的空间。要使这些经验进入通识教育，它们还需要课程结构支撑。学习科学表明，深层理解和迁移需要概念组织、反馈、反思和跨情境应用 (Bransford, A. L. Brown, and Cocking, 2000; Perkins and Salomon, 1988)。若一个 AI 项目只有制作过程，没有概念抽象和迁移任务，学生很难把一次活动经验转化为长期能力；若少数参赛学生的成果被当作整体教育成效，广大普通学习者的共同基础也难以得到保障。

AI 教育研究中也出现了类似发现。Zhou, Van Brummelen, and P. Lin (2020) 综述 K-12 AI 学习经验设计，指出 AI 教育需要将核心能力、工具资源和学习活动组织成适合儿童和青少年的设计框架。Morales-Navarro et al. (2025) 对一小时编程活动中 AI 与机器学习内容的分析发现，许多活动集中在感知和机器学习等主题，对表示、推理等主题关注有限，动手参与也存在不足。这类研究提示我们，活动资源为课程提供了重要素材，完整的课程结构仍需对这些素材作出选择和组织。

活动化还容易带来“成功样例偏差”。在科技展示中，学校通常呈现运行顺利、效果漂亮、学生积极的项目。模型训练失败、数据采集困难、输出错误、伦理争议和团队分歧等过程很少展示。然而，真正有教育价值的恰恰包括这些不完美过程。学生在失败中理解过拟合，在错误分类中理解样本偏差，在模型不稳定中理解泛化，在项目争议中理解责任。若课程只展示成功作品，学生对 AI 的理解会过于乐观。

可比较两种项目设计。第一种项目要求学生用可视化平台训练一个垃圾分类模型，最后展示分类准确率和应用界面。第二种项目在同一任务上增加五个要求：记录数据来源，比较不同样本量下的效果，分析错误分类案例，讨论真实垃圾站场景中的光照和遮挡问题，说明系统进入校园后可能产生的责任边界。前者是作品项目，后者是通识项目。前者训练学生完成一个系统，后者训练学生理解系统能否进入真实世界。

修正活动化和竞赛化倾向，需要建立三条原则。第一，活动需要回到概念，每个项目都要明确所涉及的数据、模型、反馈、泛化、偏差或责任等核心问题。第二，活动需要进入反思，学生要说明自己的选择、错误、证据、边界和改进。第三，活动需要纳入完整课程。孤立的一两次活动往往停留在短暂兴趣和成果展示，持续课程才能使活动成为知识理解与能力训练的有机环节。

4.4.4 前沿化与碎片化倾向：用热点内容替代稳定结构

人工智能发展极快，大模型、生成式 AI、多模态模型、智能体、具身智能、AI for Science 等概念不断出现。学校和培训机构及时把新进展纳入课程，有助于保持教育与真实技术环境的联系。前沿化偏差发生在热点内容决定课程结构时：学生接触许多新词和新产品，却缺少理解这些新形态所需的稳定概念。与之相伴的碎片化，则把传统内容、新技术名词和零散活动简单并置，难以形成连贯的知识主干。

前沿化倾向常表现为两种情况。一种是把新技术直接搬进课堂。课程标题包含“大模型”“智能体”“多模态”“具身智能”，但学生并不理解模型、数据、训练、生成、环境、行动和反馈这些基本概念。另一种是把前沿应用当成神奇案例。教师展示 AI 写诗、AI 作画、AI 生成视频、AI 做实验设计，课堂气氛活跃，但学生没有机会分析它为何能做、为何会错、错在哪里、如何验证。

课程内容也可能滞后于技术环境。许多课程仍停留在早期人工智能或传统机器学习框架，强调规则系统、简单分类和可视化训练，缺少对当前人工智能发展及其社会环境的回应；有些课程在传统信息技术内容中零散加入新概念和新名词，便标识为人工智能课程；也有课程直接压缩大学人工智能专业课，以删减后的专业知识代替系统化的通识体系。

不论哪种情况，都会导致前沿与基础之间的断裂，使学生既缺少稳定概念，也缺少理解新技术变化的能力。

人工智能通识教育需要处理好“前沿性”和“稳定性”的关系。稳定性来自数据、表示、模型、学习、生成、反馈、评价、系统和责任等核心概念，前沿性则体现为这些概念在新技术形态中的组合与发展。课程应以稳定主干解释前沿变化，使学生能够把已有概念迁移到大模型、多模态模型、智能体、具身智能和 AI for Science 等新情境中，并继续追问输出的证据、边界和责任。

4.4.5 伦理附加化倾向：用风险提醒替代责任判断

人工智能伦理已经进入多数 AI 教育框架。隐私保护、算法偏见、公平、透明、深度伪造、版权、学术诚信和高风险应用等问题，经常出现在教材和课程中。国际框架也普遍强调伦理与负责任使用。UNESCO 的 AI 能力框架把 AI 伦理作为学生和教师能力的重要维度 (Miao, Shiohira, and Lao, 2024; Miao and Cukurova, 2024)；AI4K12 的第五个大观念也明确关注社会影响 (Touretzky et al., 2019)；中国《中小学人工智能通识教育指南（2025 年版）》强调安全可控、伦理审查、技术风险防控以及社会责任意识 (教育部基础教育教学指导委员会, 2025a)。

伦理专题、风险清单和安全提醒可以帮助学习者建立初步意识，也是伦理教育进入课程的重要起点。伦理附加化偏差发生在这些提醒被用来代表完整的责任教育时：技术内容讲完之后安排一节风险提示，项目展示之后附上一页隐私和安全注意事项，教材列出“AI 可能有偏见、可能侵犯隐私、可能被滥用”等条目，却很少把伦理问题解释成技术与社会相互作用的必然结果，也很少将伦理问题拆解在数据选择、模型设计、任务目标、部署场景和使用责任之中。

伦理附加化的后果是，学生把伦理理解为外部规范，而没有理解伦理如何内嵌在技术系统和社会应用中。以图像识别为例，伦理问题并非只在系统上线后出现。数据采集时就涉及同意和隐私；数据标注时涉及类别划分和文化偏见；模型训练时涉及不同群体的错误率；应用部署时涉及场景风险和人工复核；结果解释时涉及责任和申诉机制。若课程只在最后提醒“注意隐私和公平”，学生难以形成系统性的伦理判断。

相关研究已经揭示了 AI 伦理在儿童和学校教育中的特殊性。C. Adams et al. (2023) 分析 K-12 教育中的 AI 伦理原则，指出儿童具有独特权利和发展特征，AI 伦理原则用于学校场景时需要更深入反思。Payne (2019) 开发的中学 AI 伦理课程则通过活动和案例引导学生理解算法偏见、隐私和公平等问题，显示伦理可以通过具体课堂活动进入学生经验。这些工作说明，伦理教育需要情境化和发展适切性。

伦理附加化还会削弱学生对技术价值的平衡判断。若伦理课程只列出风险，学生可能陷入简单恐惧；若课程只展示便利，学生可能走向盲目信任。通识教育需要的是“负责任的信任”。学生应当能够看到 AI 的价值，也能识别风险；能够使用 AI，也能提出边界；能够支持创新，也能坚持人的尊严、公平和责任。

更好的做法是把伦理问题嵌入数据、模型、生成、智能体和应用等学习过程，使学生理解风险和责任如何随技术环节与应用情境产生。伦理由此成为理解 AI 系统的必要维度，并在第三篇的风险、伦理与责任部分获得系统展开。

4.4.6 评价浅层化倾向：用可见结果替代能力判断

人工智能通识教育的另一个突出问题，是与其目标相匹配的评价体系尚未形成。许多课程用概念测验、工具操作、项目作品、竞赛成绩和课堂展示评价学生。这些方式各有功能：概念测验能够检查基础知识，工具操作能够检查实践能力，项目作品能够展示综合应用，竞赛成绩能够体现部分学生的创造性。然而，如果对应通识教育的目标，不足之处就立刻显现出来：这些方式无法评价学生能否解释 AI 输出为何可能错误，能否把一个判断框架迁移到新场景，能否识别证据不足和边界条件，能否反思自己对 AI 的依赖。

评价浅层化首先表现为术语化。学生能够背出监督学习、训练集、神经网络、生成式人工智能、算法偏见等概念，却未必能用这些概念分析真实案例。术语记忆是学习的起点，却不能作为人工智能素养的终点。人工智能概念的教育价值在于帮助学生解释现象和作出判断，名词清单只能提供最初的知识线索。

第二个表现是作品化。学生完成一个 AI 项目后，评价往往集中在界面是否美观、功能是否可运行、展示是否流畅。作品质量容易掩盖学习过程。尤其在生成式 AI 环境中，工具辅助可以显著提高作品的完整度，仅凭作品很难判断学生本身的理解与收获。若评价只看结果，学生可能获得高分，却缺少问题定义、证据核验、错误分析和责任判断。

第三个表现是竞赛化。竞赛能够激励优秀学生，也能推动学校资源投入。但竞赛成绩不适合代表全体学生的通识发展。通识教育强调面向所有学习者的共同基础，评价应覆盖不同能力层次、不同学段和不同学习方式。若评价过度依赖竞赛，人工智能通识教育会向少数学生集中，公平性和普及性受到影响。

教育测量和教学理论为评价改造提供了依据。Biggs and Tang (2011) 提出的“建构性对齐”强调，学习目标、教学活动和评价任务需要保持一致。Shute (2008) 对形成性反馈的综述指出，有效反馈应当支持学习者调整思考和行为，单独给出分数难以发挥这一作用。Sadler (1989) 关于形成性评价的经典研究也强调，学习者需要理解质量标准、比较当前表现与目标之间的差距，并采取行动缩小差距。这些原则用于人工智能通识教育，意味着评价要围绕认知自治能力设计。

评价需要进一步覆盖解释、迁移、判断和自醒等学习结果，考察学生能否说明 AI 系统和输出，能否把所学用于新情境，能否权衡证据、风险和责任，并能否反思自己的 AI 使用与判断过程。具体的评价结构和任务形式将在后文结合课程体系展开。

这样的评价更耗时，也更要求教师的专业能力，但更符合人工智能通识教育的目标。若人工智能通识教育要发展学习者的认知自治能力，评价就必须看到学生如何判断，同时考察学生做出了什么。

4.4.7 教师培训表层化倾向：用工具培训替代教学能力建设

人工智能通识教育的质量很大程度上取决于教师。教师既要理解人工智能基本概念，也要能够把知识转化为适合不同学段的问题和任务，处理诚信、隐私和伦理问题，并评价学生的真实学习进展。工具操作和案例演示能够帮助教师迅速进入新的技术环境，是教师培训的必要内容。教师培训表层化偏差发生在这类培训成为教师发展的主要内容时，概念结构、教学法、评价和伦理治理所需的能力便得不到充分展开。

这一问题提示了教师成长体系的重要性。人工智能是一门新学科，发展迅速，不能只依靠教师通过“自学”完成人工智能通识教育知识框架的搭建。教师需要持续而体系化的支持，包括科学的概念体系、稳定的课程框架和持续更新的资源。

4.4.8 目标错位的共同根源

上述偏差表面上不同，实质上都表现为局部教育形式与通识目标之间的位置错配，背后也有共同根源。第一，人工智能教育领域发展速度很快，课程建设容易被技术更新节奏牵引。第二，教育系统偏好可见成果，工具产出、项目展示和竞赛成绩比认知能力更容易被观察。第三，人工智能本身具有跨学科性质，学校原有学科结构难以自然容纳。第四，教师能力建设、评价工具和资源体系尚未充分成熟。第五，社会对人工智能既充满期待又存在焦虑，教育实践容易在技术乐观和风险警惕之间摆动。第六，“人工智能教育”与“人工智能通识教育”在政策传播和课程实践中时常混用，使专业培养、技能培训、兴趣活动与通识教育的目标边界不够清楚。

这些根源可以概括为表4.2。

表 4.2: 当前人工智能通识教育的主要偏差、后果与修正方向

偏差类型	典型表现	教育后果	修正方向
工具化	以提示词、产品操作和效率提升为核心	学生会调用系统，却缺少来源、证据和边界判断	将工具使用转化为输出比较、证据核验和反思记录
技术化	将专业技术训练直接移作全体学生的课程方案	门槛过高，社会判断和责任意识不足	建立通识化技术理解，讲清数据、模型、生成、反馈与责任
活动化	以短期体验、科技节、社团和展示代表持续课程	学习经验零散，概念结构和迁移不足	将活动纳入学段递进课程，强化概念抽象和反思
竞赛化	以少数学生项目和获奖成果代表普遍教育成效	普及性不足，基础素养被忽视	建立面向全体学生的基础能力评价
前沿化与碎片化	以新词、新产品和零散专题组织课程	学生接触热点，却缺少稳定概念结构	用数据、模型、学习、生成、反馈和系统解释前沿技术
伦理附加化	最后加入风险清单和安全提醒	伦理与技术过程割裂，责任判断弱	将伦理嵌入数据、模型、应用和评价全过程
评价浅层化	重术语、重作品、重展示，轻过程和判断	难以识别真实学习和认知发展	建立解释性、迁移性、判断性和自醒性评价
教师培训表层化	教师培训集中在工具演示	教师难以设计深层学习任务	建设人工智能内容、教学法、伦理和评价整合的教师发展体系

从表中可以看到，当前人工智能教育已经积累了丰富内容，这些内容进入人工智能通识教育时，需要由清晰目标加以统摄，这一目标即是“认知自治”。同样是生成式人工智能写作任务，可以服务工具技能训练，也可以组织证据核验；同样是图像分类项目，可以侧重技术实现，也可以分析数据偏差和泛化边界；同样是人工智能伦理讨论，可以普及风险意识，也可以进入责任判断。课程面向的学习者、预期目标、任务设计和评价方式，共同决定了这些活动属于何种教育并形成何种学习结果。是否以认知自治为统摄性目标，是判断一门人工智能课程能否实现通识教育使命的重要依据。

4.5 政策与能力框架中的认知转向

实践中的错位很大程度上源于理论上的缺失所导致的政策导向上的不稳定性。一段时间以来，人工智能通识教育因而缺少能够统摄全局的高位目标和整体结构，工具、技能、活动等易于实施和展示的目标容易暂时占据课程中心，而这些内容也成为早期政策指向的合理选择。近年来，这一情况正在发生改变，越来越多的政策与能力框架开始转向对认知能力的培养。

例如，在中小学课程建设中，AI4K12 项目提出感知、表示与推理、学习、自然交互和社会影响五个“大观念”，并按照不同学段组织人工智能学习内容 (Touretzky et al., 2019; AI4K12 Initiative, 2019)。这一框架为人工智能知识的课程化和学习进阶奠定了基础，也把人工智能的社会影响纳入学习范围。此后，人工智能教育框架的关注点进一步扩展到学习者如何认识人工智能、判断人工智能输出以及处理人与人工智能之间的关系。

这种变化近年来更加明确地进入官方政策和能力框架。UNESCO 发布的学生人工智能能力框架将“以人为中心的思维方式”置于重要位置，强调人的能动性、人的责任和智能社会中的公民身份，要求学生能够判断人工智能系统是否适合特定目的，其使用是否正当，以及人在使用过程中应当承担何种责任 (Miao, Shiohira, and Lao, 2024)。这里的学习目标已经涉及对人工智能的理解、判断以及对人机关系的主动调节。

中国《中小学人工智能通识教育指南（2025 年版）》提出知识、技能、思维与价值观相互融合的“四位一体”人工智能素养，将思维发展、价值判断、人机协作和社会责任纳入人工智能通识教育的培养目标 (教育部基础教育教学指导委员会, 2025a)。这一设计表明，中小学人工智能通识教育正在从知识普及和技术体验，逐步转向对学习者思维方式、价值意识和责任能力的综合培养。

OECD 与欧盟委员会发布的中小学人工智能素养框架进一步强化了这一方向。该框架要求学习者理解人工智能系统的基本运行方式，批判性评价输出的准确性、相关性和偏差，判断何时以及为何使用人工智能，并通过设置约束和监控结果，使人工智能的使用符合人的任务目标和价值取向 (OECD and European Commission, 2026)。在这一框架中，人工智能素养已经包含对工具使用时机、信任程度、判断边界和人类责任的主动把握，开始接近智能社会中个体保持认知主导所需要的基本能力。

这些文件和框架共同呈现出一条逐渐清晰的发展路径：人工智能教育的关注范围从知识理解和工具使用，扩展到输出评价、伦理判断、人机协作、人的能动性和责任承担。以“通识”为目的的人工智能教育文件尤其重视学习者认知能力和价值意识的发展。这说明，人工智能通识教育的实践已经开始回应智能社会带来的深层认知问题。

然而，现有框架仍存在两个相互联系的局限性。其一，不同概念的层级尚未充分厘清。人工智能素养有时被表述为学习结果，有时承担课程总目标的功能；人工智能能力既可指技术能力，也可覆盖伦理态度和社会参与；人工智能教育则可能同时指整个教育领域和面向全体学习者的课程。政策文件各自服务于特定对象和任务，这种多样性具有合理性，但它增加了课程转化时的解释空间，也使专业训练、技能培养与通识教育容易再次混合。

其二，认知相关要素通常以并列维度分布在框架之中，尚未形成统摄课程内容、学习活动和评价证据的整体目标。理解系统、评价输出、保持人的能动性、伦理使用和承担责任之间已经具有紧密联系，但这些联系往往没有被组织成学习者主动调节自身认知活动的完整过程。来源怎样识别、证据怎样使用、信任怎样分配、判断边界怎样把握、工具依赖怎样反思、行动责任怎样承担，仍可能分散在技术、伦理、态度或社会影响等不同栏目中。

本书的价值在于将现有研究和政策文件中分散的认知要求概括为“认知自治”，并把它确立为人工智能通识教育的统摄性目标。这样的处理保留各框架在知识、技能、伦理和责任方面的积累，同时为课程选择、任务组织和学习评价提供一条贯通的主线。

4.6 本章小结

人工智能通识教育是通识教育与人工智能教育相融合形成的教育形态。它以人工智能的知识、方法、系统及其社会运行为学习内容，承担通识教育在智能社会中培养认知自治能力的使命。人工智能素养是学习者在人工智能情境中形成的直接学习结果；认知自治是统摄课程的教育目标和学习者主动调节认知活动的过程；认知自治能力则支撑这一过程跨越不同任务和情境持续展开。四者分属教育形态、学习结果、统摄目标和稳定能力，相互联系而各有明确位置。

当前人工智能教育正处在快速发展阶段。专业教育培养研究开发与工程实践人才，技能培训回应岗位与应用需要，兴趣活动和竞赛激发探索与创新，人工智能赋能教育拓展教学与学习方式。这些实践共同构成人工智能教育的多样生态，也为人工智能通识教育积累了丰富资源。然而，当任何一种局部形式被用来代表完整的人工智能通识教育时，便会出现目标窄化。

工具化使通识教育停留在操作训练，技术化使共同基础被专业实现遮蔽，活动化和竞赛化使普遍培养缺少持续结构，前沿化与碎片化使课程追逐新词而忽视基本概念，伦理附加化使责任判断停留在风险提醒，评价浅层化使真实学习难以被识别，教师培训表层化则限制了深层学习任务的设计与实施。这些偏差具有共同结构：具有正当价值的局部内容或实施手段脱离整体目标，逐渐替代了人工智能通识教育应当具备的完整学习内容。

国内外政策与能力框架已经开始从知识理解和工具使用扩展到输出评价、伦理判断、人的能动性和责任承担，显示出人工智能教育中的认知转向。认知自治可以把这些分散要求组织为一条贯通的教育主线，使工具、技术、项目、伦理和应用共同服务于学习者对信念、证据、信任、边界和责任的主动调节。至此，第一篇完成了从通识教育的历史使命、智能社会的认知环境和认知自治的提出，到人工智能通识教育定位的完整论证。下一篇将进一步回答，人工智能通识教育为什么能够并且应当承担发展认知自治能力的任务。

第二篇

理论基础：人工智能通识教育何以训练认知自治

本篇导言

第一篇已经提出，在智能社会中，认知自治应当成为通识教育的重要目标，人工智能通识教育由此承担起培养认知自治能力的责任。目标确立以后，还需要进一步回答：人工智能通识教育为什么能够并且应当承担这一任务？“人工智能造成了智能社会的认知困境，因此认知自治能力应该通过学习人工智能来培养”这类主张固然具有直觉上的合理性，但还需要更深入的理性讨论来支撑。

要回答这一问题，首先需要从人工智能学科本身出发，考察它的学科特性以及进入社会活动的途径，两者分别对应智能活动的计算化与社会活动的计算化。这一讨论将为人工智能通识教育培养认知自治能力提供科学基础。其次，需要讨论如何依据人工智能的学科特性及其影响社会的方式，组织课程内容与教学活动，使教学过程始终指向认知自治能力的发展。这一讨论旨在说明人工智能通识教育何以能够培养认知自治能力，构成这一命题的可行性论证。最后，还需要将人工智能通识教育与其他学科的心智训练进行比较，分析它在认知自治能力培养中具有怎样的独特作用，从而说明其在课程体系中保有相对独立内容与结构的必要性。

本篇共四章。第五章考察人工智能的研究对象、科学体系和思维方式，说明智能活动如何获得可计算、可构建和可检验的形式；第六章分析人工智能从计算能力转化为社会力量的内在机制，说明社会活动的计算化及其现实效力；第七章提出归纳、泛化、判断与自醒四能力模型，阐明人工智能学习如何促进认知自治；第八章比较人工智能通识教育与其他学科的心智训练，说明其在认知自治能力培养中的独特作用。四章分别从科学基础、培养机制和学科独特性展开论证，共同回答人工智能通识教育为什么能够并且应当承担培养认知自治能力的任务，并由此确立其在智能社会通识教育体系中的独特地位。

第五章　人工智能：研究对象、科学体系与思维方式

摘要：人工智能通识教育需要建立在对人工智能学科性质的准确理解上。关于智能的讨论长期涉及哲学、心理学、神经科学和计算机科学等多个领域。人工智能之所以能够逐渐发展为一门科学，关键在于把难以直接观察的“智能”转化为可观察的任务、可计算的表示、可构建的机制和可检验的系统，使智能活动获得能够分析、实现、比较和修正的形式。这一过程构成智能活动的计算化，也是人工智能通识教育培养认知自治能力的科学基础。本章据此将人工智能定义为用计算机模拟人类智能行为的科学。在实现智能的探索中，本章进一步区分行为主义与内在主义两条基本道路：前者以实现和检验智能行为为主要目标，后者以复制或逼近人的认知过程及其神经实现为目标，两条道路均需接受科学实践的检验。在此基础上，本章将人工智能的科学体系概括为数学与计算基础、智能实现的机制与方法、问题形式化与系统实现三个相互贯通的层次，并总结人工智能的若干基础思维方式。本章强调，人工智能仍是一门快速发展并保持开放边界的科学，其中既有相对稳定的基础知识，也有正在展开的学术争论和面向未来的研究设想。区分不同知识所处的证据状态，是准确理解人工智能科学及其认知教育价值的重要基础。

关键词：人工智能；科学定义；智能行为；计算模拟；行为主义；内在主义；科学体系；思维方式；知识状态

5.1　人工智能何以成为科学

“机器能否思考”“机器能否拥有智能”最初主要是哲学问题。哲学可以追问智能的本质、意识的性质、心灵与身体的关系，也可以讨论理解、意向和主体经验的含义。这些问题具有基础价值；然而，相关讨论需要进一步转化为可以观察、比较和检验的对象，才能进入以公开证据和可重复检验为基础的科学研究。

人工智能作为一门科学诞生，正是始于把宏大的哲学问题分解为一组可以研究的具体问题：机器能否证明定理、解决问题、识别图像、理解语音、翻译语言、参与对话、规划行动、控制机器人，或者在复杂环境中根据反馈调整行为。每一个问题都可以进一步规定输入、输出、任务条件和评价标准，研究者因而能够提出方法、制造系统、进行实验并比较结果。

图灵关于机器智能的讨论集中体现了这种转变。他没有试图证明机器是否具有与人相同的内在心智，而是通过一种模仿游戏把问题转化为对行为表现的考察（即图灵测试）

(Turing, 1950)。模仿游戏并没有解决智能、理解和意识的全部哲学争论，却提供了一种重要的科学思路：当对象的内部状态难以直接观察时，可以先建立可以公开检验的行为标准。

随着研究的进展，人工智能逐渐形成了相对明确的研究对象、能够交流的方法、可重复、可比较的证据，以及被共同遵守的思维方式和科研习惯。机器对弈、定理证明、模式识别、语言处理和机器人行动构成可研究的对象；逻辑推理、搜索、概率建模、机器学习和神经网络构成可交流的方法；数据集、任务测试、实验对照和真实环境运行提供可比较的证据；学术会议、期刊、实验室和开放基准则使不同研究路线能够在共同问题上竞争和积累。人工智能由此从关于智能机器的设想，逐渐发展为一门可以通过计算实验和系统实践不断推进的科学。

这种科学化并不意味着哲学问题已经消失。机器是否真正理解语言，意识能否由计算产生，主观体验是否可以被模拟，仍然是开放问题。科学化的含义是：人工智能研究不能只依靠概念思辨来证明自己的结论，而要把能够进入研究的部分转化为明确对象，通过模型、系统和实验接受检验。哲学可以提出问题、分析概念和反思前提，科学则必须进一步说明研究什么、怎样实现、如何检验以及什么证据能够支持或反驳结论。

5.2 人工智能的研究对象：智能行为

到目前为止，“智能”没有单一、无争议的定义。日常语言中的智能可以指记忆、推理、理解、学习、创造、适应环境、解决问题或理解他人；心理学、神经科学、哲学和计算机科学关注的侧面也不相同。把不同学科的定义简单合并，反而会使之变得模糊。

相比之下，智能的外在表现更容易观察。本书把这些可以观察和检验的表现称为**智能行为**。从人工智能长期研究的任务看，智能行为可以概括为四个彼此联系的方面：

1. **思考：**包括表示问题、记忆知识、形成概念、推理、学习、规划、解释和创造方案；
2. **感知：**包括从图像、声音、语言、触觉和多种传感信号中识别对象、状态与关系；
3. **行动：**包括根据目标和环境作出选择，控制设备，调用工具，与其他主体协作并根据反馈调整行动；
4. **情感与创造性：**包括识别和表达情绪，生成具有风格、审美或新颖性的内容，以及对人的情感信号作出适当反应。

这里的分类针对可以观察的功能表现。它不预先断言机器是否具有与人相同的理解、情感或主观经验。一个系统能够识别悲伤的语音、生成安慰性的回答，说明它可以实现某些与情感有关的行为；这一行为事实本身不能直接证明系统具有悲伤、同情或关怀的主观体验。类似地，模型能够完成推理题，说明它表现出相应的问题解决能力，但其内部过程是否等同于人的推理，仍需另行研究。

在“智能行为”的定义之下，“智能”可以据此定义为“产生这些智能行为的能力”；能够稳定实现某类智能行为的机器，可以称为具有相应智能的“智能机器”。

5.3 实现智能的两条基本探索道路

人类智能为人工智能提供参照，但机器应当在多大程度上复制人的内部机制，长期以来存在不同的意见。为便于分析，本书根据研究目标与检验标准，将实现智能的探索概括为“行为主义”与“内在主义”两条道路。行为主义以实现和检验智能行为为目标；内在主义把人的认知过程或脑机制作为需要复制或逼近的对象。这里的两个概念是对人工智能研究取向的概括，与心理学史和心灵哲学中的同名术语并不等同。

行为主义道路：以模拟智能行为为目标

行为主义道路首先规定系统需要表现出的行为，再研究如何通过计算实现这些行为。它关心机器能否完成任务、能否适应环境、能否在新的条件下保持表现，而不要求机器内部必须复制人的认知过程。

符号推理通过知识表示、搜索和规则操作完成问题求解；统计机器学习从数据中估计规律；深度学习利用多层网络形成分布式表示；大模型通过大规模预训练、上下文和后训练获得语言、多模态和工具使用能力。这些方法与人的思维可能存在启发和类比关系，但其科学有效性主要由行为表现、实验结果和使用条件检验。机器采用高维向量、梯度优化或大规模并行计算，这些都和人的日常思考过程有所不同，但并不影响这些研究的价值。

行为主义道路的重要特点，是允许研究者绕开暂时无法完整描述的人类内部机制，直接探索模拟智能行为的多种计算可能。当前人工智能在视觉识别、语音处理、机器翻译、科学预测、内容生成和复杂博弈等领域取得的进展，集中体现了这一路线的科学价值。

内在主义道路：以复制人的内部机制为目标

内在主义道路把人的认知过程及其脑实现作为直接模型，目标不仅是在功能上产生相似结果，还要使机器采用与人相同或相近的内部表征、处理过程和神经机制。认知科学中关于“弱等价”与“强等价”的区分提供了相近表述：只产生相同输入—输出行为属于弱等价；强等价意义上的认知模型还需要借助反应时间、中间状态和过程结构等证据，检验模型是否逼近人实际采用的算法和内部过程 (Pylyshyn, 1980)。

沿认知层面展开时，这一路线表现为对记忆、学习、推理和决策等过程的计算建模；沿神经层面展开时，则表现为类脑计算、神经形态计算、脑仿真以及具有生物可信性的神经模型。例如，Eliasmith 等人构建的 Spaun 以约 250 万个脉冲神经元模拟多个脑区，在同一系统中完成视觉识别、记忆、计数、推理和动作控制等多种任务，并同时比较模型的神经活动与心理行为 (Eliasmith et al., 2012)。

内在主义道路面对更严格的证据要求。模型除了能够产生相应行为，还要说明其内部表征、中间状态、处理过程或神经机制为何与人相似，并使用反应时间、认知实验、神经活动或生物约束等独立证据加以检验。人类大脑与身体构成高度复杂的生物系统，认知又同成长、语言、社会关系和文化经验相互作用，现有科学尚未完整解释这些过程如何共同产生人的智能和主观体验。这些困难使强等价很难建立，但不能据此预先判定这条道路必然失败。只要研究提出可检验假设、建立可比较模型并根据证据修正结论，它仍然属于值得探索的科学方向。

两条道路的关系

表 5.1: 实现智能的两条基本探索道路

比较方面	行为主义道路	内在主义道路
主要目标	实现并检验可观察的智能行为	复制或逼近人的认知过程及其神经实现
实现标准	系统能否在规定任务和环境中形成有效行为	系统形成目标行为,其内部表征、处理过程或神经机制还需与人形成可检验的对应
允许的差异	允许机器采用与人不同的计算路径	要求关键内部表征、处理过程或神经机制与人保持相似
当前进展	在大量任务和应用中形成显著成果	在类脑计算、神经形态计算、脑仿真和认知过程建模等方向持续探索
主要困难	行为有效不能直接说明内部机制，也不能直接证明理解、意识和主体经验	人类内部机制尚未得到完整解释，内部对应还需独立证据检验

两条道路可以相互借鉴。行为主义系统可以吸收脑科学启发，但这种启发本身不改变其以任务表现为主要标准的性质；内在主义研究也需要通过外在行为证明模型具有相应功能。它们的区别在于研究目标和证据标准：前者接受功能层面的弱等价，后者进一步追求过程层面或神经层面的强等价。人工智能尚未发展到可以用一条路线排除所有其他路线的阶段。不同学派围绕知识、学习、大脑、身体、环境和意识提出竞争性解释，是这门科学保持活力的重要条件。

5.4 人工智能的定义

根据上述分析，本书从人工智能的研究对象和实现方式出发，对人工智能定义如下：

人工智能是用计算机模拟人类智能行为的科学。

这一定义包含以下几个基本要点：

1. 人工智能是一门科学，既非单指某种技术，也非单指某种“智能”。
2. 人类智能行为提供研究的起点和参照。人工智能源自让机器复现人类思维能力的愿望，许多任务也以人的表现作为比较对象。
3. 机器以人的智能行为为模拟对象，其内部机理和过程可以与人不同。机器可以采用和人不同的机制，形成相同、相近或更强大的功能。

4. 计算构成人工智能的主要实现方式。基于化学、物理、机械等方式也可能模仿人的智能行为，但不属于人工智能的研究范畴。

本书在讨论研究对象、知识体系和学科发展时，主要在学科意义上使用“人工智能”；在讨论具体能力、现实应用和社会作用时，根据语境分别使用“人工智能技术”“人工智能系统”或“人工智能技术体系”。在不引起混淆的前提下，也会以“人工智能”作为人工智能技术、系统、体系的简称。

5.5 人工智能的科学体系

根据前述定义，人工智能以可观察的智能行为作为研究对象。然而，确定研究对象只是学科成立的起点；一门科学还需要围绕研究对象形成相互衔接的基本概念、理论命题、实现机制、研究方法和验证方式。

人工智能首先以能够运行的系统进入人们的视野。机器能够识别图像、理解语言、生成内容、制定规划或者控制设备，这些可观察的行为构成了人工智能研究最直接的现象。面对一个人工智能系统，可以依次提出三个层次的问题：它解决了什么现实问题，系统的任务、边界和评价标准是什么；这些行为由哪些信息加工机制、模型和算法产生；这些机制为什么能够计算、学习和推广，又受到哪些理论条件的限制。

沿着这一思考过程，人工智能的科学体系可以从顶层到底层组织为三个相互贯通的层次：现实任务的问题形式化与系统实现，智能行为的实现机制与方法，以及数学与计算基础。顶层把现实目标转化为可以运行和检验的人工智能系统；中间层解释推理、预测、生成、规划与行动等能力如何形成；基础层说明对象如何表示、计算在什么范围内可能，以及机器学习在什么条件下能够形成可推广的规律。三个层次共同构成从现实问题到系统行为、从行为机制到理论条件的科学结构。

表 5.2: 人工智能科学体系的三个层次

体系层次	直接研究对象	核心科学问题	主要理论构成
现实任务与系统实现	形式化的智能任务及其人工智能系统	现实目标如何转化为可计算、可评价的任务，不同功能和方法如何组成能够在环境中运行并接受检验的系统	任务定义与表示，输入与输出，状态与行动，目标与约束，数据与知识组织，系统结构与功能模块，评价与验证
智能行为的实现机制与方法	产生智能行为的信息加工机制	知识、数据和环境交互如何形成推理、预测、生成、规划与行动	知识表示与推理，搜索，概率模型，机器学习，神经网络，强化学习，生成模型，规划与控制

体系层次	直接研究对象	核心科学问题	主要理论构成
数学与计算基础	形式结构、计算过程与学习规律	智能相关对象如何表示，哪些问题可以计算，机器在什么条件下能够从有限经验中形成可推广的规律	逻辑与离散结构，线性代数，概率与统计，信息论，优化，算法理论、可计算性与复杂性，学习理论

5.5.1 现实任务与系统实现

人工智能科学体系的顶层面对现实任务及其系统实现。人们观察到的人工智能通常表现为一个系统在特定环境中完成某种任务，例如辅助医生诊断、预测交通状况、回答问题、生成图像或者控制机器人行动。要把这些现实目标转化为科学研究对象，首先需要说明系统面对什么环境、接收什么信息、产生什么输出、追求什么目标，以及通过什么证据判断系统是否取得了预期效果。

“帮助医生诊断”“改善交通运行”或“支持学生学习”表达了现实需求，但还没有形成机器能够直接处理的任务。问题形式化需要界定对象、环境与任务边界，把宽泛目标分解为识别、预测、生成、规划或控制等具体任务，规定输入与输出、状态与行动、目标函数与约束条件，并选择符号、向量、图或者概率结构表示现实对象。在此基础上，研究者还需要组织系统所需的数据和知识，确定训练、运行和评价所使用的条件。

问题形式化决定了人工智能系统实际解决的是什么问题。哪些现实差异被保留为变量，哪些条件被纳入模型，何种结果被定义为任务完成，以及采用什么指标评价系统，都会改变系统行为的定义。评价还需要考虑不同错误及其现实代价。例如，在医学诊断中，漏诊和误诊都属于预测错误，但二者可能造成不同后果。如果评价指标没有体现这种差异，技术指标的提高未必能够转化为实际诊疗效果。因此，任务定义、目标设定、约束条件和评价标准共同规定了人工智能系统所追求的行为。

确定任务以后，还需要把信息采集、数据处理、知识组织、模型训练、推理与规划、记忆、工具调用、结果输出和行动控制等功能组织为完整系统。现代人工智能经常以智能体及其与环境的关系描述这一整体结构：系统通过传感器或数据接口取得信息，依据内部状态、知识和模型进行计算，再通过输出或行动对环境产生影响 (Russell and Norvig, 2021)。不同功能模块之间如何传递信息，局部模型的误差如何影响整体行为，系统如何应对环境变化和异常情况，都属于系统层需要研究的问题。

例如，“让机器看懂道路”可以进一步分解为车辆检测、行人识别、车道线分割、距离估计和运动预测等任务。摄像头及其他传感器提供输入，对象类别、空间位置和未来轨迹构成中间结果，识别与预测结果随后进入定位、路径规划和车辆控制模块，最终形成连续行驶行为。系统还需要在不同道路、天气、光照和交通条件下接受评价，以检验各个模块以及整体行为的准确性、稳定性和适用范围。

由此，单个模型的输出只是人工智能系统中的一个环节。只有说明现实对象如何转化为计算对象，局部模型如何进入系统结构，各部分如何共同形成可观察的行为，以及

这些行为如何接受比较和检验，一个人工智能问题才获得完整的科学表达。系统也因此成为人工智能科学的研究对象：理论和方法在系统中被组合起来，关于智能行为的命题则通过系统运行得到观察、测量和检验。

5.5.2 智能行为的实现机制与方法

当任务和系统行为得到明确界定以后，还需要进一步研究如何实现这些智能行为。智能行为的实现机制与方法位于科学体系的中间层，研究信息经过怎样的组织和变化，形成推理、预测、生成、规划与行动等能力。这里的“机制”指信息通过何种过程转化为某类行为，“方法”指构造和运行这种机制所采用的模型、目标与算法组合。模型描述对象、关系、状态变化、概率依赖或者输入与输出之间的规律；算法规定如何在模型上执行搜索、推理、训练、生成、规划或者行动。

从系统能力来源划分，人工智能可以分为基于知识的人工智能方法和基于学习的人工智能方法。

基于知识的人工智能把事实、概念、关系和规则明确表示出来，通过搜索和推理获得结论或方案。物理符号系统研究和专家系统分别展示了符号操作与领域知识在智能实现中的作用 (Newell and Herbert A. Simon, 1976; Feigenbaum, 1977)。在这种机制中，系统的能力主要来自已经建立的知识结构、推理规则和搜索过程。

基于学习的人工智能依据样本调整模型，使系统形成分类、预测、生成或表征能力。概率模型以随机变量及其依赖关系表达不确定知识 (Pearl, 1988)；神经网络则利用参数化模型、目标函数与优化算法，从数据中形成复杂映射和分布式表示 (Bishop, 2006; LeCun, Yoshua Bengio, and G. Hinton, 2015)。在这种机制中，系统的行为规律主要通过基于样本的训练过程形成。

在现代人工智能系统中，这两种方法经常共存：系统可以从数据中学习模型，调用知识库、规则和外部工具，并通过与环境持续交互组织连续行动。例如，一个智能体可以利用神经网络理解输入，通过检索系统获取外部知识，借助规划方法分解任务，再根据工具执行结果调整下一步行动。

机制与方法层由此连接了系统行为与基础理论。向上看，它需要解释系统的识别、预测、生成、规划和行动能力由何种过程产生；向下看，它需要说明所使用的表示、模型和算法，并提供所依赖的边界条件。

5.5.3 数学与计算基础

沿着智能行为的形成机制继续向下延展，就会进入人工智能的数学与计算基础。任何模型和算法都需要使用一定的形式语言表示对象，依照确定的计算过程处理信息，并在特定条件下完成推理、优化或学习。这一层研究的核心，是表示、计算和学习何以可能，在什么条件下成立，以及受到哪些原则性限制。

现实对象需要经过形式表示，才能成为机器可以处理的计算对象。这些表示可以是符号、图、向量、矩阵、概率分布或者动态状态。逻辑与离散数学用于表达事实、规则、关系和推理结构；线性代数为高维数据和参数模型提供统一表示；概率与统计用于描述

不确定性以及样本与总体之间的关系；信息论研究信息的度量、编码和传递；优化理论研究如何在目标和约束下寻找适当的解。这些理论共同决定了人工智能系统能够表达什么，以及能够通过何种形式处理这些表达。

完成形式表示以后，还需要回答一个问题：经由形式化表达的计算问题是否可以通过算法求解，以及求解需要付出多大代价。Turing (1936) 建立的抽象计算模型使“可计算”获得了严格含义，也揭示了某些形式问题无法由算法解决。可计算性理论由此划定计算的原则边界，计算复杂性理论进一步区分理论上能够求解的问题与在现实资源条件下能够有效求解的问题。算法理论则研究计算过程的构造、正确性、效率和资源消耗，为具体人工智能方法提供计算基础。

基于机器学习的方法需要进一步回答：系统能否从有限样本或交互经验中形成对未见情形仍然有效的模型。概率与统计说明如何依据有限样本推断总体规律，优化理论说明如何依据目标调整模型参数，学习理论则研究经验数量、假设空间、模型复杂度与泛化能力之间的条件关系 (Mitchell, 1997; Bishop, 2006)。

这些理论同时揭示学习方法的适用边界。“No Free Lunch”定理表明，在对所有可能的目标函数等量平均的条件下，不存在对所有问题都普遍占优的优化方法；一种方法能够在特定任务上取得优势，需要以其假设与问题结构相匹配为前提 (Wolpert and Macready, 1997)。因此，模型的有效性总是与数据分布、任务结构、表示方式、目标设定和环境条件相关。脱离这些条件讨论一种方法是否“智能”或者是否“有效”，难以形成严格的科学判断。

数学与计算基础由此形成了人工智能科学体系的底层支撑。它并非若干数学工具的简单汇集，而是围绕表示如何建立、计算如何进行、学习如何成立以及能力边界如何确定形成的理论结构。它使人工智能系统中的模型选择、算法设计和行为解释受到明确定义、形式推导与理论条件的约束，也为分析系统失效的深层原因提供依据。

5.5.4 三个层次的贯通关系

从顶层到底层看，人工智能的科学体系形成了一条逐层深入的解释路径：

现实任务与系统行为 $\longrightarrow$ 实现机制、模型与算法 $\longrightarrow$ 表示、计算与学习的基础条件。

在顶层，研究者首先界定现实任务，规定系统的输入、输出、目标、约束、运行环境和评价标准，并通过系统行为提出可以观察和检验的科学问题。在中间层，研究者进一步分析知识、数据与环境交互如何经过模型和算法形成相应行为。在基础层，研究者继续追问这些表示和计算过程为什么成立、能够在什么条件下有效，以及受到哪些理论边界的限制。

三个层次之间具有清楚的接口。现实任务与系统层赋予智能行为以确定的任务含义、环境条件和评价标准；机制与方法层提供推理、预测、生成、规划和控制等能力的实现过程；数学与计算基础则提供形式语言、计算规则、学习条件和理论边界。上层提出的问题不断向下转化为对表示、模型和算法的要求，下层形成的理论与方法又向上支持系统的构造、运行与验证。

人工智能的科学严谨性建立在这种上下贯通的层次关系之上。它从可观察的智能行为出发，把现实目标转化为结构明确、可以运行和接受检验的人工智能系统；通过模型和算法解释行为的形成过程；再以数学和计算理论说明表示、计算与学习成立的条件。智能行为由此成为可以定义、分解、实现、解释和验证的科学问题，人工智能系统则成为连接现实任务、实现机制与理论命题的实验载体。

重要的一点是，这一层次性的科学体系能够容纳不同的理论传统、技术路线和系统形态，其稳定性来自各个层次之间可以明确陈述、逐级追溯和反复检验的关系。随着新的现实任务、系统形态、学习机制和表示方式不断出现，人工智能的具体内容仍将持续扩展，而这一从现实系统深入基础条件的贯通结构，为理解人工智能作为一门发展中的科学提供了基本框架。

5.6　人工智能的学科特征

以智能行为为研究对象：人工智能围绕机器的智能行为组织问题。这一研究对象具有开放性：随着机器能力提高，过去被视为智能的任务可能成为普通计算功能，人们会追求更高级的智能行为。人工智能因此具有持续移动的边界。

以计算为基础方法：实现智能机器的方法并不唯一，人工智能选择的是其中一种：用计算来模拟智能行为，工具是计算机。这种以计算为中心的方案既起源于人类对思维数学化的前期成果，也得益于通用计算机提供的能力基础。

知识与学习构成两类基本能力来源：人工智能能力可以来自人明确提供的知识，也可以来自数据、反馈和交互形成的学习。两类来源各有优势：知识提供结构、约束和可解释关系，学习能够处理复杂模式和不确定环境。检索增强生成、神经符号系统和工具型智能体显示两类方法正在重新融合。

科学研究与工程实践紧密结合：人工智能既需要提出可检验的理论，也需要构建模型和系统。研究结论通常通过数据集、实验、基准和真实应用评价。工程规模、硬件性能、数据管线和部署环境会直接影响科学结果。人工智能的理论、实验和工程很难完全分离。

具有强烈的交叉和基础方法属性：人工智能从多学科中吸收方法，又进入各学科参与数据分析、模型预测、假设生成和实验设计。它逐渐成为一种通用问题求解方法。进入具体领域后，人工智能仍需接受该领域的理论约束和证据标准，通用模型不能替代专业验证。

能力、风险和社会结构共同演进：人工智能系统会进入信息分发、教育评价、医疗诊断、劳动管理和公共决策。技术目标、数据选择和部署方式包含价值判断，也会重新分配机会、权力和责任。因此，人工智能的基础概念不能只描述模型能力，还要说明系统与人、组织和社会环境之间的关系。

5.7 人工智能的思维方式

人工智能在长期发展中形成了一套具有学科特点的思维方式。这些思维方式来自一个共同的研究过程：把现实中的智能问题转化为可以表示、计算、实现和检验的问题，再通过模型、算法和系统构造出解决方案。

面对“如何让机器识别物体”“如何预测未来变化”“如何生成新的内容”或者“如何在环境中作出行动”等问题，人工智能研究通常需要经历几个相互联系的环节：首先明确问题的对象、目标和边界，把现实问题转化为形式化任务；随后从数据和领域知识中提取可用于计算的信息，并对信息不足、环境变化和结果误差所带来的不确定性进行处理；在此基础上选择尽可能简明的方法，在性能、资源和效率之间作出权衡；最后把方法放入实际任务和运行环境中，通过实验与实践检验其有效性。

由此，人工智能的思维方式主要表现为问题形式化、从数据中学习、利用领域知识约束、处理不确定性、权衡性能与效率、优先寻找简单有效的方法，以及重视实践检验。这些方面贯穿人工智能从提出问题、构造模型到形成系统的全过程。

5.7.1 将现实问题形式化为计算问题

人工智能研究首先需要把宽泛、模糊的现实问题转化为结构明确的计算问题。人们可以直接提出“让机器理解语言”“帮助医生诊断疾病”或者“使机器人自主行动”等目标，但计算机无法直接处理这些抽象要求。研究者需要进一步说明系统面对什么对象、接收什么信息、产生什么结果、追求什么目标，以及在什么条件下可以认为任务已经完成。

这一过程称为问题形式化。它通常包括界定任务边界，确定输入与输出，选择描述对象的变量和表示方式，规定目标函数、约束条件与评价标准。例如，“让机器理解一段文字”可以进一步转化为回答问题、提取信息、判断文本关系、生成摘要或者遵循指令等任务。不同的任务定义对应不同的输入、输出和评价方式，也使“理解”获得了可以观察和检验的行为含义。

形式化过程需要对现实问题进行选择和抽象。现实世界包含大量细节，计算模型只能保留其中与当前任务相关的部分。哪些对象被表示为变量，哪些差异受到关注，哪些条件进入模型，都会影响系统最终解决的问题。例如，在预测学生学习状况时，如果模型只使用考试成绩，它所描述的主要是成绩变化；如果同时考虑学习过程、知识结构和任务难度，模型所处理的问题就会发生变化。

经过形式化以后，问题便可以通过数学模型和计算过程加以处理。研究者可以利用方程、概率分布、图结构、向量空间或者状态转移描述问题，再通过搜索、推理、优化等过程寻找解决方案。这里的“求解”并不局限于从公式中推导出解析答案，而是找到解决问题的方法。许多人工智能问题没有可以直接写出的封闭解，需要构造模型、算法或者可运行的系统，通过计算、模拟、训练和交互逐步得到结果。

因此，形式化同时也是一种构造性的思维方式。研究者把现实问题转化为可计算对象，构造能够表现相关行为的模型和系统，再通过这些人工对象研究原问题。一个现实

问题能否得到有效解决，在很大程度上取决于它是否被转化成了适当的计算问题。

5.7.2　从数据中学习规律

传统程序通常由人明确写出处理规则，人工智能则经常需要面对规则难以完整列举的问题。人很容易识别一张人脸、听懂一句话或者判断道路中的障碍物，却很难把完成这些任务所需的全部规则逐条写出。人工智能由此发展出一种重要的解决思路：让机器从数据和经验中学习规律。

从数据中学习，就是依据样本调整模型，使模型逐渐形成输入与输出之间的关系，或者发现数据中潜在的结构。监督学习从带有答案的样本中学习预测关系，无监督学习从数据中发现类别、结构或表示，强化学习则从行动及其环境反馈中形成策略。虽然具体方法不同，它们都把经验看作形成系统能力的重要来源。

数据驱动的关键价值在于，它能够处理难以完全依靠人工规则描述的复杂规律。图像中物体的外观、自然语言的使用方式、蛋白质序列与空间结构之间的关系，都包含大量相互作用的因素。机器学习可以借助大量样本，在高维数据中寻找对任务有用的统计结构。

从数据中得到规律并不意味着模型已经掌握普遍规律。训练数据只是现实世界的有限样本，模型可能记住其中的偶然特征，也可能利用与任务无关的相关性取得较高分数。人工智能研究因此特别重视泛化：模型在已有数据上形成的规律，能否用于尚未见过的对象和情境。

这要求研究者清晰了解数据来自哪里，是否覆盖了重要情形，训练数据与实际应用环境是否一致，数据中是否存在偏差、遗漏和错误。当环境和数据分布发生变化时，过去有效的模型也可能失效。从数据中学习因而包含两个相互联系的任务：发现经验中的规律，同时判断这些规律能够成立和推广的范围。

5.7.3　重视领域知识的约束

数据并非人工智能获得信息的唯一来源。现实问题通常具有自身的对象结构、运行规律和约束条件。医学涉及人体结构、疾病机制和诊疗规范，物理研究受到守恒定律与边界条件的约束，交通系统具有道路结构、车辆动力学和安全规则，教育问题则与知识结构、认知发展和教学目标密切相关。这些领域知识决定了问题中哪些关系具有意义，也为方法选择提供了依据。

领域知识可以通过多种方式进入人工智能系统。研究者可以用领域概念确定数据如何表示，用已知规律约束模型的结构和输出，用因果关系帮助区分表面相关与真实机制，用安全规则排除不可接受的行动，还可以利用专家知识缩小搜索空间、设计评价指标或者解释系统结果。

引入领域知识能够减少模型需要从数据中重新学习的内容。当样本数量有限、实验成本较高或者错误后果严重时，领域知识尤其重要。例如，在材料研究中，已知的物理规律可以帮助系统排除明显不合理的候选结构；在医学应用中，诊疗知识和安全要求可

以限制模型的建议范围；在机器人控制中，运动学与动力学规律可以避免系统探索无法实现或者具有危险的动作。

方法选择也需要与领域特点相匹配。“No Free Lunch”定理说明，在对所有可能问题等量考虑的条件下，不存在对所有问题都普遍占优的方法 (Wolpert and Macready, 1997)。一种算法能够取得良好效果，通常是因为它所包含的假设与特定问题的结构相适应。图像具有空间邻近关系，语言具有序列和语境结构，社会网络具有节点与连接结构，相应的方法需要利用这些不同特点。

因此，人工智能研究不能脱离具体问题单独判断一种方法的优劣。模型是否合适，取决于任务目标、数据条件、领域规律、风险要求和使用环境。理解领域结构、辨认方法假设并使二者相互匹配，是人工智能解决现实问题的重要思维方式。

5.7.4 在不确定性中作出判断

人工智能面对的信息经常是不完整、有噪声和不断变化的。传感器可能产生误差，数据可能存在缺失，同一种现象可能由多种原因造成，未来环境也无法完全预知。即使模型给出了确定的分类、预测或者行动，其依据和结果仍可能具有不同程度的不确定性。

处理不确定性首先需要承认信息和模型的边界。系统需要区分已经获得的证据与尚未掌握的信息，估计不同结果出现的可能性，并说明判断的可信程度。概率模型为描述不确定性提供了重要工具：研究者可以利用概率分布表示多个可能结果，依据新的证据更新判断，并在不同选择之间比较预期收益和风险 (Pearl, 1988)。

不确定性可能来自不同方面。有些不确定性来自对象本身的随机变化，例如天气、市场和人类行为存在难以完全预测的波动；有些不确定性来自知识和数据不足，例如模型很少见到某类样本，或者当前输入已经超出训练数据的范围。前一种不确定性通常只能通过概率进行描述，后一种不确定性则可能通过收集数据、改进模型或者寻求专家判断得到降低。

人工智能系统还需要依据不确定性的大小决定如何行动。当预测较为可靠时，系统可以直接执行相应操作；当证据不足或者风险较高时，系统可以请求更多信息、保留多个候选结果、交由人类判断，或者选择更加保守的方案。医学诊断、自动驾驶和公共决策等领域尤其需要把结果可信度、错误代价与行动规则结合起来。

处理不确定性体现了人工智能中的一种基本理性：在信息有限的条件下保留适当的判断边界，根据证据变化更新信任度，并使行动强度与证据的可靠程度相匹配。

5.7.5 在性能与效率之间作出权衡

人工智能方法需要在有限的时间、数据、算力、存储、能耗和经济成本下运行。一种模型即使能够获得很高的准确率，如果训练成本过高、响应速度过慢或者无法部署到实际设备中，也可能难以形成可用的解决方案。因此，人工智能研究不仅关注系统能够达到什么性能，还关注取得这些性能需要付出多少资源。

性能包含多个方面。分类任务可能关注准确率、召回率和误报率，生成任务可能关注内容质量、多样性和可靠性，智能体系统还需要考虑任务完成率、安全性、稳定性和

环境适应能力。这些指标之间也可能存在冲突。提高模型对某类错误的敏感度，可能增加另一类错误；增加系统能力可能带来更长的运行时间和更高的资源消耗；追求即时响应则可能限制模型能够完成的计算量。

效率同样需要根据应用环境理解。运行在数据中心的模型与运行在手机、汽车、机器人或传感器上的模型，面对的计算和能耗条件并不相同。实时控制要求系统在规定时间内完成计算，边缘设备受到存储和电量限制，大规模公共服务还需要考虑每次调用的成本以及同时服务大量用户的能力。

研究者因此需要在多个目标之间作出权衡。模型压缩、参数共享、近似计算、分层处理和按需调用等方法，都可以用于减少资源消耗。合理的系统还可以把简单情形交给低成本方法处理，把复杂或者高风险情形交给能力更强的模型或人类专家。

这种思维方式体现了人工智能作为计算科学和工程科学的共同特点：有效的解决方案需要在特定资源和环境条件下实现足够好的行为。对性能的追求始终受到时间、空间、能量、成本和风险等条件的约束。

5.7.6 优先寻找简单有效的方法

面对复杂问题，人们很容易直接采用规模更大、结构更复杂的方法。然而，模型复杂度的增加会带来更多参数、更高计算成本和更大的分析难度，也可能使系统更容易学习数据中的偶然特征。人工智能研究因此重视一种朴素而重要的原则：首先寻找能够抓住问题主要结构的简单方法，再根据实际需要逐步增加复杂性。

简单方法可以作为理解问题的起点。一个线性模型、一组明确规则或者一个较小的决策树，往往能够帮助研究者判断哪些变量真正重要，数据中是否存在可学习的规律，以及复杂模型的性能提高来自什么地方。只有与简单基准进行比较，复杂方法所带来的增益才具有清楚含义。

简单性还可以提高方法的稳定性和可检验性。结构较简单的模型通常更容易分析错误、发现数据问题、解释结果和进行重复实验。当数据数量有限时，较简单的模型也可能比复杂模型具有更好的泛化能力。即使最终需要使用复杂系统，简单模型仍然可以承担基准比较、异常检查或者安全保障等作用。

简单性并不意味着始终选择规模最小的模型。现实问题可能包含高度复杂的结构，需要足够强的模型才能表达。这里强调的是，方法复杂度应当具有明确理由：新增的结构、参数和计算需要解决已经观察到的问题，并通过实验表明其确实带来了必要的改进。

5.7.7 通过实践检验解决方案

人工智能是一门高度重视实验的科学。方法的有效性不能仅靠理论推导，还需要把模型和算法实现出来，在数据、任务和环境中观察其行为。一个具有吸引力的理论设想，只有经过实验比较和现实运行，才能判断它是否真正解决了预定问题。

实践检验首先需要建立明确的评价条件。研究者在测试时需要确定使用什么数据，选择哪些比较对象，采用什么指标。为了判断改进来自哪里，还可以通过消融实验去除

某些模块或信息，观察系统性能发生的变化。为了检验系统边界，还需要改变数据分布、增加噪声、构造异常输入或者测试极端条件。

实验结果也需要接受全方位的对比验证。只在单一数据集或少数样本上取得成功，难以说明一种方法具有稳定效果。研究者需要考察结果是否能够在不同数据、参数设置、随机条件和运行环境中重现，并与适当的基准方法进行公平比较。负面结果和失效案例同样具有价值，它们可以揭示方法依赖的条件和尚未解决的问题。

真实环境中的检验还会暴露实验室条件下难以发现的因素。用户行为可能不同于研究者的预期，数据分布可能随时间变化，多个系统模块可能相互影响，技术指标的提高也可能无法直接转化为现实效果。因此，人工智能系统往往需要经历从离线实验、模拟测试、受控试验到实际部署的逐步验证过程。

5.7.8 不同思维方式的内在联系

上述七种思维方式分别回答了人工智能解决问题时的不同问题。问题形式化确定机器究竟要解决什么；从数据中学习说明如何从经验中形成规律；领域知识约束使方法与对象结构相匹配；不确定性处理使系统能够在信息有限的情况下保持适当的判断边界；简单性原则帮助研究者抓住问题的主要结构；性能与效率权衡使解决方案能够在现实资源条件下运行；实践检验则为判断方法是否有效提供最终证据。

从更深层看，这些思维方式体现了人工智能研究的三个基本特点。第一，它把抽象的智能问题转化为可以表示和计算的人工对象；第二，它综合利用数据、知识和环境反馈形成解决方案；第三，它通过构造可运行的模型与系统，使关于智能行为的设想接受实验和现实检验。

人工智能的思维方式由此具有鲜明的构造性和实践性。研究者通过形式化建立问题，通过数据与知识获得依据，通过模型、算法和系统构造可能的解，再在不确定性和资源约束下检验、比较和修正这些解。正是在这一反复循环中，原本抽象的识别、学习、推理、生成和行动问题，逐渐成为可以研究、实现和验证的科学问题。

需要说明的是，本节所列举的思维方式并不能覆盖人工智能领域里的所有思想和方法，而是当前以机器学习为基本特征的人工智能研究所遵守的核心规则。一些细分领域可能有更具体的思维特点，未来新方法的出现也可能带来思维方式的扩充与更新。

5.8 发展中的科学

人工智能的特殊之处在于它仍处于快速发展过程中。传统学科中的许多知识已经经过较长时间检验，而人工智能要面对的是大量未解的问题、不断更新的前沿进展、尚未达成共识的学术争论。理解不同知识所处的状态，是认识这门学科的重要部分。

(1) 相对稳定的知识

人工智能已经形成一批相对稳定的基础知识。例如，形式逻辑和计算理论为机器处理符号与程序奠定基础；人工智能可以通过搜索、推理和学习等多种方法实现；机器学习系统的表现由目标、数据、模型、算法和计算条件共同决定；训练结果需要通过独立

测试检验；模型能力具有任务与环境边界。这些知识可能继续发展，但已经能够构成学科共同基础。

人工智能的发展史也提供相对稳定的认识。符号主义、知识工程、统计机器学习、深度学习和基础模型构成一个围绕知识来源、表示方式、学习机制和计算条件持续演进的方法谱系。

(2) 仍在进行的争论

人工智能的许多核心问题仍无定论：智能是否只需根据行为定义，内部机制在何种情况下不可替代；大语言模型表现出的智能行为应当怎样解释，尺度定律能够走多远；符号与神经方法如何结合；类脑计算方案能否产生新的突破；机器是否可能拥有理解、意识或主观体验。

这些争论不能只靠立场解决。不同观点需要明确概念、提出可以区分彼此的预测，并寻找实验、系统行为和跨学科证据。暂时无法检验的问题可以继续作为哲学或理论问题讨论，但不能把一种概念设想直接宣布为已经证实的科学事实。

(3) 面向未来的设想

通用人工智能、超级智能、具有主体经验的机器和完整脑模拟等概念，包含重要的研究想象，也具有高度不确定性。它们可以帮助研究者提出长期问题，却不能用当前某项任务的性能直接证明已经实现。相反，也不能因为现有系统尚未达到这些目标，就断言任何未来道路必然不可能。

发展中的科学需要同时避免两种倾向：一种把阶段性成功解释为智能问题已经解决，另一种因为系统同人不同或存在明显缺陷，就否定现有道路的科学价值。更加符合人工智能发展实际的态度，是根据现有证据确认已经取得的能力，根据实验说明能力边界，并为尚未解决的问题保留多条探索路线。

5.9 本章小结

人工智能从关于智能机器的哲学设想出发，通过研究对象的形式化、计算方法的建立、模型与机制的建构以及实验事实的积累，逐渐发展为一门科学。本书将人工智能定义为用计算机模拟人类智能行为的科学。人类智能行为为研究提供起点和参照，计算机提供主要实现手段，科学实验提供检验方式。机器对智能行为的模拟可以采用不同于人的内部过程。行为表现、内部实现机制和主观体验属于相互联系但不能相互替代的分析层面。

人工智能的科学体系可以组织为三个相互贯通的层次：数学与计算基础、智能实现的机制与方法、问题形式化与系统实现。三个层次共同连接形式理论、实现机制、现实任务和行为表现，使人工智能的理论依据、运行过程、适用条件和实验结果能够逐级说明和相互检验。

在这一科学体系中，人工智能形成了具有自身特点的思维方式，包括对现实问题的形式化、从数据中学习规律、重视领域知识的约束、在不确定性中作出判断、权衡性能与效率、优先采用简单方法、通过实践检验解决方案。这些思维方式是人工智能研究者共同遵守的，也与人工智能通识课训练认知自治能力的目标形成底层对应。

人工智能同时是一门发展中的科学，既有稳定的知识结构和研究方法，也有尚未解决的问题。这种开放性在其他学科中是少见的。后面我们会讨论，这种开放性并不是缺点，反而是人工智能通识教育中认知训练的重要场景。

本章所揭示的共同主线，是人工智能把原本抽象、内隐且难以直接检验的智能活动，转化为可观察的任务、可计算的表示、可构建的机制和可检验的系统。智能活动由此获得可以分析、实现、比较和修正的形式。这一“智能活动的计算化”构成人工智能通识教育培养认知自治能力的第一重科学基础。下一章将进一步考察人工智能系统如何进入知识生产、信息传播、社会判断和现实行动，使社会活动本身发生计算化。

第六章　从计算智能到社会力量：人工智能影响社会的内在机制

摘要：本章关注人工智能如何进入社会并产生持久而广泛的影响。识别、预测、生成和决策等智能行为获得可计算形式以后，现实对象、历史经验、社会目标便能够进入人工智能系统；模型输出再经由组织流程与制度规则获得行动权限，参与信息传播、资源配置和社会决策；当这些局部作用被大规模复制，在不同系统之间连接，并引发个人与组织的持续适应时，它们又会累积为相对稳定的信息分布、行为规范、机会结构和权力关系。人工智能由此从特定技术逐渐转化为人们获取信息、形成认识和采取行动时共同面对的环境条件。

关键词：人工智能；社会技术系统；计算表示；模型化；目标计算化；制度嵌入；反馈循环；认知环境

6.1　智能行为何以转化为社会力量

人类社会的运行始终依赖智能活动。人们通过感知获得信息，通过记忆和语言积累经验，通过推理、预测与判断形成选择，再通过行动改变环境。教育传播知识，医疗判断病情，市场配置资源，组织选拔人员，公共机构制定和执行政策，这些过程虽然面对的对象不同，却都包含信息获取、意义解释、方案比较、决策和行动等智能行为。这些智能行为正是人工智能的研究对象，因此人工智能的产出，或各种智能系统，也就具有了进入人类社会运行过程的合理途径。

例如，人类长期借助文字、印刷、计算工具和信息网络扩展记忆、传播与推理能力。人工智能带来的进一步变化，是把识别、预测、生成、规划和决策等较为复杂的智能行为转化为计算能力。机器可以读取文本与图像，从数据中形成模型，根据目标比较方案，生成新的内容，并在环境中调用工具或执行动作。它由此能够进入原来主要依靠人的知识、经验和判断才能完成的社会过程。

这一变化的关键，在于智能行为经过人工智能科学的处理以后具有了新的存在形式。人的经验通常依附于特定个体，受到时间、精力、记忆和空间的限制；计算模型可以被保存为程序和参数，在不同设备上重复运行，并同数据库、传感器、网络平台和组织流程连接。同一种能力可以在短时间内服务大量用户，也可以被嵌入许多相似的决策环节。一次局部实验形成的方法，因而可能迅速转化为广泛运行的社会基础设施。

除此以外，人工智能还具有持续调整的能力。系统可以收集新的数据，依据人的评

价或环境反馈修改模型，再把新的输出投入实践。模型的输出会影响人的选择和组织行动，这些变化又会形成下一轮数据。人工智能因而逐渐进入一个由计算、行动和反馈共同构成的循环。

这一过程可以概括为本章的核心命题：

> 人工智能把原本主要依附于个体的智能行为，转化为能够被计算、复制、连接和部署的人工能力（区别于人类能力）；这种能力经由组织与制度进入现实行动，并在规模化运行与社会反馈中不断累积，最终转化为影响社会认知、资源配置和行动秩序的力量。

这一命题也说明，人工智能的社会影响不能由模型性能单独推出。模型能够生成预测、评分、排序和建议，却不能自行决定这些输出在社会中具有什么地位。一个风险分数可以只是辅助信息，也可以成为资格门槛；一个推荐结果可以帮助个人发现内容，也可以持续改变公共注意力的分布；一个生成模型可以用于个人写作，也可以被接入组织流程，批量生产和发布内容。模型能力相同，部署目的、行动权限、覆盖范围和反馈方式不同，社会结果会明显不同。

美国国家标准与技术研究院的人工智能风险管理框架把人工智能明确视为社会技术系统，强调其收益与风险产生于技术要素同使用方式、操作者、其他系统和社会情境之间的相互作用 (National Institute of Standards and Technology, 2023)。因此，理解人工智能如何形成社会影响，需要追踪一条完整的作用链：社会活动怎样进入计算系统，计算结果怎样进入现实行动，局部作用怎样被放大和累积，以及这些过程怎样共同改变人们所处的认知环境。

本章所说的“社会活动的计算化”，涵盖社会活动从进入计算系统到计算系统形成持续中介能力的完整过程。讨论这一过程，一方面是对“智能社会”这一概念补充其形成机制，增强关于“智能社会中复杂认知环境”的理性认识，从而强化“培养认知自治能力”的责任感；另一方面，也为如何设计人工智能通识课程提供认知背景和课程素材。

6.2 社会活动如何获得可计算形式：人工智能的基本转化

人工智能要参与教育、医疗、就业、商业和公共治理，首先需要把这些社会活动转化为机器能够处理的形式。人的状态需要变成数据，过去的经验需要变成模型，社会目标需要变成指标，原有工作需要变成人机共同执行的流程。经过这些转化，人工智能才能读取信息、作出判断、选择结果并影响行动。人工智能的社会影响，正是从这些环节中产生的。

6.2.1 现实对象进入数据：系统能够看见什么

现实世界连续、复杂，充满具体情境，计算机却只能处理经过明确表示的信息。人工智能进入社会活动的第一步，是选择一些可以记录的特征，把现实对象转化为数据。

人的学习状态可以表示为作答记录、知识点掌握概率和成绩等级；健康状态可以表示为医学影像、检验数值、症状编码和风险分数；兴趣可以表示为点击、停留、搜索和购买行为；工作能力可以表示为教育经历、工作履历、任务表现和评价结果。经过这种转化，原本难以直接比较的对象成为变量、类别、关系、向量和指标。

数据使社会现象变得可以汇总、比较和分析。教师可以从大量学习记录中发现学生遇到的困难，医生可以综合更多病例作出诊断，公共机构也可以更快地识别需求和风险。人工智能的许多能力，都建立在这种数据化基础上。

然而，数据只能保留现实的一部分。系统记录什么，就能处理什么；没有被记录的内容，通常也不会进入后续计算。一次点击可以被记录，点击时的犹豫却很难被记录；考试成绩可以被量化，学生的好奇心、努力程度和成长潜力却很难由一个数字完整表达。

因此，数据并不是现实本身，而是对现实作出的选择性描述。选择哪些数据，实际上是在确定系统看见什么。以点击次数表示兴趣，以考试成绩表示学习水平，以履历特征表示工作能力，都意味着用某些可以测量的信号代表更复杂的社会概念。这些信号通常被称为代理变量。

代理变量可以帮助系统处理抽象问题，也可能把复杂概念压缩得过于简单。频繁点击某类内容，可能表示喜欢，也可能来自好奇、焦虑或偶然接触；较低的考试成绩可能反映知识掌握不足，也可能受到语言、健康和家庭环境的影响。如果系统无法看见这些差异，它的判断就只能建立在有限的表示之上。

数据表示因此具有直接的社会后果。信用系统记录哪些信息，可能影响谁能够获得贷款；简历筛选系统设置哪些字段，可能影响哪些经历得到认可；教育系统采用哪些指标，可能影响哪些学生被识别为需要帮助。理解一个人工智能系统，首先要判断：它使用了哪些数据，这些数据代表什么，又遗漏了什么。

6.2.2　历史经验进入模型：系统如何作出判断

人工智能模型通常从过去的数据中学习规律，再用这些规律判断当前情形。模型训练完成以后，过去的经验便被凝结为一套可以反复运行的计算规则。

这种能力扩大了人类可以利用的经验范围。医生可以参考大量历史病例，教师可以从长期学习记录中发现变化，科学家可以从高维实验数据中寻找规律，机构也可以依据过去的风险信号提前采取措施。个人无法逐一阅读和比较的海量信息，可以通过模型进入一次具体判断。

问题在于，历史数据中包含的不只有经验，也包含数据产生时的社会条件。哪些人有机会进入样本，哪些现象受到关注，哪些结果被记录为成功，哪些判断被当作正确答案，都会影响模型学到的规律。例如，如果某一群体过去获得的教育或就业机会较少，那么这一群体在历史数据中的成功记录也可能较少，模型可能学习并延续原有情境的判断，认为同一群体在当前状态下成功的机会也较低。

模型还需要面对相关性与因果关系的区别。某个特征能够帮助预测结果，只能说明它与结果存在统计联系，并不能证明它造成了这个结果。例如，一个人的居住地区可能有助于预测信用风险，但真正产生影响的可能是收入水平、公共资源、就业机会或历史

上的区域差异。如果系统直接针对居住地区采取行动，就可能把表面线索误当成真正原因。

从群体规律推断个体时，还会出现另一层限制。模型能够根据一类人的历史情况，估计某个人出现某种结果的可能性，却无法完整了解这个人的具体经历、当前状态和未来潜力，做出的判断就会出现偏差。

当预测结果进入资源分配时，模型的预测结果会产生自强化式的现实影响。一个学生被判断为学习能力不足，可能因此得不到更有挑战性的课程；机会减少以后，他未来的表现可能真的下降。一个人被判断为信用风险较高，可能因此更难获得资金；资金不足又可能进一步增加风险。模型由此既在预测未来，也可能参与制造未来。

因此，使用模型作出社会判断，需要警醒如下问题：历史数据是在什么条件下形成的，模型学到的是原因还是相关线索，群体规律能在多大程度上用于具体个人。

6.2.3 社会目标进入指标：系统追求什么

人工智能系统还需要知道什么结果更好。分类模型需要规定什么答案算正确，推荐系统需要确定内容如何排序，强化学习需要设置奖励，组织也需要决定系统应当优先提高效率、降低成本，还是改善质量、安全与公平。社会目标只有转化为可以计算的指标，系统才能比较不同结果并选择行动。

现实中的目标通常包含多方面价值。教育关注知识掌握，也关注兴趣、人格、创造力和长期发展；医疗关注诊断效率，也关注患者安全、尊严、知情权和公平可及；公共治理需要同时考虑效果、程序、公平、权利和社会信任。计算系统很难同时处理这些目标的全部含义，通常需要选择其中一部分，用可测量的指标加以代表。

指标使目标变得明确，也可能缩小人们对目标的理解。点击率能够反映用户是否打开内容，却无法完整表示内容是否真正有价值；任务完成速度可以衡量效率，却不能单独说明工作质量；短期成绩能够反映一部分学习效果，却不能充分表示学生的长期发展。

当一个指标成为优化目标以后，系统和人的行为都会逐渐围绕它发生变化。推荐系统持续追求点击率，可能倾向于推送更容易吸引注意的内容；学校过度强调考试成绩，教师和学生就可能把更多时间投入应试活动；平台只考核任务完成数量，工作人员就可能减少对复杂个案的投入。原来用于观察目标的指标，由此反过来改变了被观察的活动。

人工智能的能力越强，目标定义越重要。更多数据、更大模型和更强算力能够加强既定目标的实现，却无法自动补回目标中没有写入的价值。问题形式化需要把高层社会目标转化为具体任务、目标变量和代理指标，其中包含大量判断与选择，也可能产生完全不同的伦理后果 (Passi and Barocas, 2019)。

6.3 计算结果怎样进入现实行动：组织与制度的嵌入

人工智能系统要产生广泛的真实影响，需要与具体的组织流程和制度设计相配合。组织和制度嵌入因此连接了计算能力与社会行动，也决定了技术作用的方向、强度和边界。

同一个医学预测模型可以用于提醒医生补充检查，也可以被用于限制保险服务；同一种生成模型可以帮助学习者理解材料，也可以被组织成批量制造虚假内容的系统。模型能力相同，使用目的、流程位置、行动权限和纠正渠道不同，社会结果会明显不同。这些都有赖于组织流程和制度设计的约束。

6.3.1 组织与制度赋予模型输出以社会效力

模型输出获得社会效力，需要一系列具体安排：谁提出使用需求，系统被放在工作流程的什么位置，输出面向哪些对象，专业人员是否必须采纳，系统能够调用哪些工具，结果怎样被记录，错误如何被发现和纠正。界面设计、默认选项和操作时限也会影响制度效力。即使名义上保留人工决定，如果界面只呈现一个看似精确的分数，流程要求快速确认，模型输出仍可能成为事实上的决定。

因此，模型性能只是社会作用的一个条件。准确率相近的两个系统，如果一个只提供可解释的辅助信息，另一个直接控制重要资源，其社会风险和治理要求并不相同。评价人工智能，需要同时考察模型能够做什么，以及组织允许它做什么。

6.3.2 行动权力在系统链条中重新分配

人工智能模型进入组织流程与制度设计以后，会重新分配谁能够观察、判断、决定和质疑。把现实转化为数据的人，拥有定义类别和指标的权力；训练和部署模型的组织，拥有把特定规律转化为行动的能力；掌握平台入口的一方，还能够决定哪些信息优先出现。受影响者往往只能看到结果，难以了解数据来源、模型假设和制度规则，更难改变它们。

这种不对称会影响人能否质疑决定。人工智能系统的复杂性容易使输出呈现为技术结论，把原本可以讨论的分类方式、代理指标、风险阈值和价值取舍隐藏在数据与流程之中。恢复讨论空间，需要把技术决定重新展开：哪些部分来自事实证据，哪些部分来自模型假设，哪些部分属于组织选择，哪些部分涉及可以争论的社会价值。

组织和制度还决定谁能够从人工智能中获得能力，谁主要受到系统评价和约束。拥有数据、模型和部署入口的一方，可以借助系统扩大观察、预测和行动范围；缺少相应资源的一方，可能更多地以被记录、被分类和被决定的对象出现。讨论人工智能带来的权力变化，需要同时观察能力如何扩展、行动权如何分配，以及受影响者是否拥有知情、解释、拒绝、复核和申诉的真实渠道。

6.3.3 责任沿着系统链条重新组织

人工智能延长并拆分了从认识到行动的过程，使结果的形成分布在数据、模型、界面、组织规则和现场操作等多个环节。数据提供者、模型开发者、产品设计者、部署机构、专业人员和最终使用者分别掌握部分信息，也分别控制部分过程。一个具体后果可能由数据偏差、模型边界、界面呈现、组织规则和现场操作共同造成，责任因而由单一主体的归属问题转化为贯穿系统链条的组织问题。

责任因而需要沿着系统链条加以理解，并与各参与者实际能够预见、控制和纠正的环节相对应。开发者难以预先控制系统在所有场景中的具体使用，一线人员也无法弥补数据、模型与组织流程中已经形成的全部限制。责任归属与实际控制能力发生分离，是人工智能进入组织制度以后产生的重要结构性问题。

组织和制度的嵌入由此完成了一次关键转化：它们把计算系统生成的可能性，变成现实社会中的信息呈现、资源配置和行动结果。然而，制度赋效仍主要发生在具体组织和任务之中。人工智能要形成广泛而持久的社会影响，还需要这些局部作用被反复复制、彼此连接，并进入持续的社会反馈。

6.4 从局部作用到结构性影响：规模、连接与反馈

前一节说明了模型输出如何在具体制度中获得现实效力。本节进一步讨论：一次分类、推荐、预测或生成，为什么能够逐渐改变更广泛的社会环境？关键在于计算系统可以低成本复制，不同系统可以通过数据与平台相互连接，个人和组织又会根据系统输出调整行为。规模扩大影响范围，连接使影响跨越系统传播，反馈让系统参与形成下一轮数据；三者共同使局部作用累积为结构性影响。

6.4.1 规模复制使局部规则成为共同条件

规模复制把局部模型的能力、标准和误差转化为大量人共同面对的条件。软件和模型一旦完成开发并接入基础设施，其复制和扩展的边际成本通常低于同等规模的人力扩张，同一套评分规则、推荐模型或生成服务因而能够进入许多地区、机构和生活领域。规模化带来一致性和效率，使稀缺知识与服务触达更多人；它也会同步放大模型中的遗漏和偏差。一个人的偶然判断通常影响有限，一套被广泛采用的计算规则却可能反复作用于大量相似决定。

规模还会带来标准化。不同机构采用相同的数据结构、模型接口和评价方法，可以提高协作能力；社会对象也可能逐渐按照系统便于处理的方式重新组织。求职者学习怎样书写更容易被筛选的履历，内容生产者调整标题以适应推荐机制，学校和机构围绕可量化指标安排活动。技术标准由此可能转化为行为标准。

当少数模型成为许多产品和服务的基础时，集中效应进一步增强。下游系统看似用途各异，却可能共享相似的数据、模型能力和默认假设。某种能力提升可以迅速扩散，某项系统缺陷也可能同时传导。社会对少数基础模型、算力平台和数据资源的依赖，还会改变创新机会、市场结构和技术治理中的权力分布。

人工智能可以依据每个人的数据生成不同的推荐、解释和回答。从个人角度看，个性化能够降低信息搜索成本，使内容更接近具体需求；从社会角度看，许多分别完成的个性化选择会共同改变公共注意力、议题可见度和文化生产方向。每个人看到的内容不同，共同经验可能减少；同一套排序机制又可能使大量注意力集中到少数对象。个性化与集中化可以同时发生。

生成模型把规模效应扩展到内容供给。系统可以快速生成大量定制文本、图像和声

音，信息稀缺逐渐让位于注意力与可信来源的稀缺。内容数量增长并不自动带来知识增长。信息能否追溯、证据能否核验、重要观点能否在公共空间中得到共同讨论，成为评价整体信息环境的重要因素。

6.4.2 系统连接使影响跨越场景传播

系统连接使一个环节的输出成为另一个环节的输入，局部规则因而能够跨越平台、机构和生活领域继续作用。一个平台积累的行为数据可以被用于广告、信用或内容推荐，一项评分可以被多个机构重复采用，生成内容可以进入搜索结果、知识库、模型训练和后续生成。每个系统分别完成自己的任务，彼此连接以后却共同构成连续的影响链。

连接可以提高服务的连续性，减少重复采集和信息断裂，也会使错误、遗漏和既有偏差更难被限制在单个环节。上游系统没有记录的特征，下游系统通常无法补回；早期分类形成的标签，可能在多次传递中逐渐被当作稳定事实；某一模型的输出经过多个系统重复引用后，也可能获得超出原始证据的可信度。

系统共享数据、基础模型和评价标准时，表面上独立的决定可能具有共同来源。一个人面对的搜索、推荐、评估和服务由不同机构提供，却可能依赖相似的数据结构与模型能力。由此形成的影响很难通过纠正某一次输出完全消除，需要追踪数据和判断如何沿系统链条传播。

连接还改变了局部退出的意义。个人可以停止使用某个应用，却仍可能受到其他机构采用的相关评分、推荐和生成内容影响；一个组织可以更换前端产品，底层仍可能依赖相同的基础模型或数据来源。人工智能的社会作用因而逐渐超出单一产品边界，转化为由多个系统共同维持的环境条件。

6.4.3 社会反馈使系统参与塑造现实

社会反馈使预测、推荐和评价同时成为改变对象的干预，系统由此参与形成自己下一轮将要观察的数据。传统预测常把对象视为外部存在：模型观察世界，再估计将要发生什么。社会系统中的预测会进入行动，并由行动改变未来。信用风险预测影响贷款条件，贷款条件又影响还款能力；学习风险预测影响教育资源投入，资源变化会影响后续表现；治安风险判断改变巡查位置，巡查位置又决定哪些事件更容易被记录。预测结果由此参与产生下一轮被预测的数据。

这一现象被称为“施为性预测”：模型的部署改变数据分布，未来结果受到此前预测及其引发行动的影响 (Perdomo et al., 2020)。在这种条件下，模型更新不能只被理解为用新数据追赶一个独立变化的世界。新数据中已经包含系统自身的作用。

推荐系统形成类似循环。系统依据过去行为推荐内容，推荐影响用户接触什么，新的接触又形成后续训练数据。研究表明，这种算法混杂可能使用户行为趋于同质，同时没有带来相应效用提升 (Chaney, Stewart, and Engelhardt, 2018)。循环中的每一步都可能具有合理解释，长期累积的整体方向却需要单独评价。

社会对象还会主动适应系统。人们知道系统使用哪些指标后，会改变行为以获得更好结果，组织也会围绕考核规则重新配置资源。适度适应可以帮助人们理解规则并改进

表现；当代理指标同真实目标存在距离时，行为会逐渐集中到提高指标，原目标反而受到挤压。

这种反应还会削弱模型赖以成立的原有规律。一个变量在被用于评价之前，可能稳定反映某种状态；成为明确目标以后，人们会对它进行干预，变量与原状态之间的关系随之改变。模型依据过去数据建立的关联，可能因为模型本身的部署而失效。社会反馈由此成为分布变化的一种内生来源。

6.5 人工智能中介的认知环境如何形成

认知环境是前述机制长期运行形成的总体结果，由计算表示、制度嵌入、规模复制、系统连接和社会反馈共同塑造。人工智能在这一环境中进入人与现实世界之间、认识与行动之间，并通过持续运行在许多社会场景中成为人们获取信息、形成判断和采取行动时难以绕开的条件。

6.5.1 人工智能进入人与现实世界之间

人认识世界，始终需要借助语言、媒介、知识体系和社会制度。人工智能加入这一中介过程以后，信息越来越多地经过数据选择、模型处理、算法排序和生成系统来到人面前。搜索系统决定哪些结果优先出现，推荐系统组织个人能够接触的内容，生成模型直接提供解释、总结、图像和方案，分类与预测系统则为社会对象附加标签、概率和风险等级。

这种中介扩大了个人可以调用的信息与知识范围，也改变了现实呈现的方式。使用者看到的内容已经包含数据覆盖、分类规则、模型假设、优化目标和平台策略的共同作用。什么能够被看见，以什么形式被看见，哪些差异被突出，哪些来源得到更多传播，越来越多地受到计算系统影响。

模型生成的内容还会重新进入公共传播。人工智能既筛选已有信息，也直接参与生产新的文本、图像、声音和视频。来源、证据、推断和表达可以在生成过程中被压缩成一个完整答案，知识到达人的路径因而变得更短，背后的形成过程也可能更难看见。人工智能由此成为信息环境和知识环境的共同组成部分。

6.5.2 人工智能连接认识与行动

人工智能中介的内容不会停留在人的认识之中。模型输出经由组织和制度进入教育、医疗、就业、金融、消费和公共治理，影响资源怎样配置、机会怎样分配、风险怎样处理、行动怎样选择。系统由此一方面描述和预测现实，另一方面通过现实行动参与塑造它所描述和预测的对象。

这使认识与行动之间的距离缩短。一个分类结果可以直接改变信息呈现，一个预测分数可以立即触发审核，一段生成内容可以通过平台自动发布，一个智能体也可以调用工具完成连续操作。计算结果获得的行动权限越高，模型对现实的塑造作用越直接。

人和组织又会预期这种作用并调整行为。个人学习怎样向系统提问、怎样让履历更容易通过筛选、怎样生产更容易被推荐的内容；机构围绕可计算指标重新设计流程，专业人员根据系统输出重新分配注意。人工智能因而参与塑造人的行动，同时也被人的适应不断改变。

6.5.3 反复运行的系统成为智能社会认知环境

当大量人工智能系统持续运行以后，其共同作用会成为社会成员普遍面对的认知环境。这一认知环境由数据、模型、平台、组织和制度共同构成。数据决定哪些现实能够进入计算，模型把历史规律转化为当前能力，目标与指标规定系统的行动方向，平台提供连接和分发结构，组织与制度赋予输出以现实效力，人的反应又形成新的数据和规则。任何单一模型都无法独自构成这一环境，多个要素的持续连接使它得以形成和变化。

第二章所概括的认知环境变化由此获得了具体的形成机制。数据选择、模型处理和算法排序使人工智能进入人与现实之间；生成与传播系统扩大信息供给并改变注意力分布；模型输出与工作流程的连接使部分行动自动发生；人的反应重新成为系统数据，形成持续反馈；数据、模型、平台、组织和使用者共同参与结果形成，又使责任分布在更长的系统链条之中。

在这样的环境中，人们即使没有直接使用某个人工智能产品，也可能受到算法排序、自动评价、生成内容和制度决策的影响。个人很难准确意识到自己的信息来源、判断方式和行动机会在多大程度上受到系统塑造。智能社会由此提出了一个基本问题：人在广泛依赖人工智能的同时，如何继续追问信息来源、判断证据强弱、认识模型边界、校准对系统的信任，并对最终行动承担责任。

这正是认知自治所要回应的问题。人工智能通识教育需要帮助学习者理解这些机制，并在使用、评价和必要时接管人工智能系统的过程中保持认知主导。下一章将进一步讨论，这种教育如何把对人工智能的学习转化为认知自治能力的训练。

6.6 本章小结

本章讨论了人工智能的计算能力如何进入社会运行，并逐渐转化为广泛的社会影响。人工智能参与社会活动，首先需要完成一系列基本转化：把现实对象表示为数据，把过去经验凝结为模型，把社会目标转化为指标，再把信息获取、分析、判断和行动组织为人机共同参与的流程。数据决定系统能够看见什么，模型决定系统依据什么作出判断，目标决定系统追求什么，流程则决定计算结果怎样进入现实行动。

人工智能的社会作用还取决于它在组织和制度中获得怎样的位置。模型输出可以只作为参考，也可以用于筛选人员、分配资源、调整机会和直接执行行动。组织流程和制度设计赋予计算结果以现实效力，也决定谁能够质疑、修改或推翻系统判断，以及发生错误时由谁承担责任。因此，人工智能的社会影响已经包含在数据选择、模型训练、目标设定、流程安排和制度使用的整个过程中。

当人工智能被大规模复制，并通过平台、组织和基础设施相互连接以后，单个系统

的局部作用会进一步累积。系统输出影响人的认识和行动，人的反应又成为新的数据，推动模型、规则和行为继续变化。规模扩张、系统连接与反馈循环共同使人工智能从解决具体任务的工具，发展为持续影响信息传播、判断形成、机会分配和社会行动的结构性力量。

大量系统的反复运行，最终构成智能社会的认知环境。人们能够接触什么信息、相信哪些来源、依据什么作出判断，以及拥有哪些行动机会，都可能受到这一环境的影响。在这样的认知环境中，个体如何保持对证据、信任、判断和行动的主动调节，成为人工智能通识教育必须回应的问题。下一章将进一步讨论，人工智能通识教育如何把对这些机制的理解转化为认知自治能力的训练。

第七章　人工智能通识课的认知自治能力培养

摘要：认知自治只有进一步转化为明确的能力结构、课程任务和评价证据，才能成为可以实施的教育目标。本章提出归纳、泛化、判断与自醒四能力模型，以此描述人工智能通识教育所要培养的基本认知能力。归纳使学习者能够从有限材料中形成有依据、可检验的理解；泛化使学习者能够把已有理解迁移到新情境，并识别其适用条件和边界；判断使学习者能够在证据不完全、后果不确定和价值冲突中作出有理由的选择；自醒使学习者能够监控自身的理解、信任、依赖和判断，并根据新证据进行修正。

四能力模型是一种面向课程设计和学习评价的教育模型，并不试图穷尽认知自治的全部心理结构。它既建立在人类学习、迁移、判断和元认知研究的基础上，也与人工智能系统从数据中形成模式、在新情境中产生输出、接受评价并进入社会行动的过程密切相关。基于人工智能的这些特性，本章进一步提出认知自治训练的三条基本路径：理解智能机制，使学习者认识计算结论形成的条件、依据与边界；参与计算建构，使学习者的假设和判断接受明确化与运行检验；分析社会嵌入，使学习者识别计算系统进入现实活动时涉及的价值、权限、影响与责任。

本章分别讨论四种能力的理论基础、教育目标、课堂任务和表现层级，并将其组织为从假设形成、边界检验、行动选择到认知校正的循环。评价认知自治能力，需要超越作品完成度和模型性能，考察学习者能否解释证据、识别边界、权衡风险、组织人机协作，并反思和修正自身的认知过程。

关键词：人工智能通识教育；认知自治；归纳；泛化；判断；自醒；元认知；信任校准；学习评价

7.1　问题的提出：从教育理念到能力模型

第一篇已经建立了本书的基本论证。人工智能系统及其平台化应用正在改变知识生产、信息传播、学习支持、劳动组织和公共治理的基本条件。智能社会中的个体所面对的主要认知困难，已经从信息匮乏转向信息充足而来源混杂，从结论难以获得转向结论容易呈现而依据需要辨析，从单一作者和单一媒介转向由人、模型、平台、机构和传播网络共同构成的混合知识环境。在这样的环境中，人工智能通识教育承担着一项重要任务，即培养学习者的认知自治能力。

“认知自治”作为教育理念，还需要进一步转化为可以实施的教育内容。一个理念只有落实为能力结构、课程任务和评价证据，才能真正进入教学实践。若只强调学生应

当“保持独立判断”或“理性使用 AI”，教师很难据此设计课程，学生也难以明确自己需要练习哪些认知活动。人工智能通识教育因此需要回答一系列更具体的问题：学习者需要形成哪些相对稳定的认知能力？这些能力如何相互联系？它们分别怎样在 AI 场景和一般社会场景中表现出来？教师如何观察学生是否取得了进步？学校又如何判断一门课程已经超越工具演示和作品展示，真正进入了认知自治的训练？

本章提出归纳、泛化、判断与自醒四能力模型。在人工智能通识教育的课程范围内，如果学习者在这四个方面形成了较为稳定的能力，便可以认为其已经具备了一定程度的认知自治能力。

- 归纳能力，是从有限现象、数据、案例、文本和经验中形成可检验理解的能力。它要求学习者发现模式，同时认识样本、证据和因果解释的限制。
- 泛化能力，是把已有理解迁移到新情境并识别适用边界的能力。它要求学习者理解规律成立所依赖的条件，并能够在对象、环境和任务发生变化时调整判断与信任。
- 判断能力，是在证据不完全、后果不确定和价值存在冲突的条件下作出可解释选择的能力。它要求学习者综合考虑事实、方法、风险、价值和责任。
- 自醒能力，是对自身认知状态进行监控、反思和校正的能力。它要求学习者意识到自己何时理解不足、何时过度信任，以及何时受到情绪、权威、平台机制和 AI 输出的影响。

需要说明的是，四能力模型是一个面向人工智能通识教育的课程模型，并非对认知自治心理结构的穷尽性划分。这四种能力构成了人工智能通识教育所主张的核心训练内容，但这并不意味着认知自治在心理学上只包含四个方面。学界对于认知自治的心理构成尚未形成统一划分，而课程实践需要建立一套相对明确、能够进入教学设计和学习评价的能力体系。四能力模型承担的正是这一教育功能。

这一课程模型具有相应的心理学和学习科学基础。认知心理学研究表明，人能够从有限材料中提取规律，同时也容易受到样本、表征方式和启发式偏差的影响 (Shepard, 1987; Saffran, Aslin, and Newport, 1996; Tversky and Kahneman, 1974; Nisbett et al., 1983)。学习科学进一步表明，知识迁移依赖深层理解、主动抽象和元认知监控 (Bransford, A. L. Brown, and Cocking, 2000; Perkins and Salomon, 1988; Chi, Feltovich, and Glaser, 1981)。神经科学与元认知研究则提示，任务表现、结构泛化以及对自身表现的觉察相互联系，同时又具有可以区分的心理和神经基础 (Samborska et al., 2022; Fleming and Dolan, 2012)。这些研究共同说明，人类个体需要在规律形成、知识迁移、复杂判断和自我监控等方面接受持续训练，才能在复杂环境中逐渐建立认知自治。

四能力模型也与人工智能系统的基本特性和运行过程密切相关。AI 系统从数据中学习模式，把已形成的模式用于新的输入，并根据评价、反馈和使用后果不断接受检验。系统的置信度、误差、适用边界和潜在风险，还需要由人进行解释和判断。这样的运行过程一方面对人的归纳、泛化、判断与自醒提出了新的要求，另一方面也把样本、特征、规律、边界、评价和信任等认知问题转化为较为显性的学习材料，为四种能力的训练提供了可观察、可操作和可比较的场景。

7.2 人工智能通识教育如何训练认知自治

上一节回答了人工智能通识教育需要训练什么能力，本节进一步说明这些训练为什么能够通过人工智能通识教育得到支持。

前两章分别讨论了智能活动的计算化与社会活动的计算化。人工智能把数据、表示、模型、目标、评价和反馈组织为可以分析和运行的过程，又通过平台、组织、制度和工作流程进入信息传播、判断形成与社会行动。这些特性使人工智能通识教育能够从三个方向训练认知自治：理解智能机制，认识计算结果形成的条件、依据与边界；参与计算建构，使学习者自身的假设和判断接受明确化与运行检验；分析社会嵌入，识别计算系统进入现实活动时涉及的价值、权限、影响与责任。

需要说明的是，这三条路径描述的是人工智能通识教育训练认知自治的基本方式，归纳、泛化、判断和自醒描述的则是训练所要形成的能力。二者处于不同层次，也不构成一一对应关系。理解智能机制、参与计算建构和分析社会嵌入都可能同时调动多种认知能力；同一种能力也可以在三条路径中反复获得练习。

7.2.1 理解智能机制：认识结论形成的条件

人工智能系统需要把现实对象转化为数据和表示，通过模型处理形成输出，并依据目标和评价标准调整结果。学习这些机制，可以使学习者看到一个计算结论经历了怎样的形成过程：系统使用了哪些数据，数据保留了现实的哪些方面，又舍弃了哪些方面；模型依据哪些特征作出判断；目标函数鼓励系统形成什么结果；评价指标实际衡量了哪些方面。

这种学习为分析一般认识活动提供了一组清晰的问题结构。面对一个已经形成的结论，学习者可以进一步追问：材料是否具有代表性，观察到的特征是否足以支持判断，相关关系能否支持因果解释，评价标准是否覆盖了真正重要的目标，结论在新的对象和情境中是否仍然成立。人工智能系统表现出的错误、偏差和不确定性，也使证据不足、样本偏斜、表面线索依赖和适用范围受限等问题获得了具体形式。

人工智能的计算过程与人的认知过程具有不同的实现机制，但二者都会面对若干基本的认识问题，包括怎样从有限材料中形成结论，怎样把已有认识用于新情境，怎样评价证据的强弱，以及怎样认识结论的适用边界。人工智能在这里提供的是一种问题结构上的参照，无须假定机器复制了人的认知过程。

通过分析人工智能结果的形成过程，学习者可以逐渐养成追溯来源、检查依据、识别边界和校准信任的习惯。训练的核心在于，即使面对一个已经呈现出来、语言流畅或看似可信的结论，学习者仍然能够追问它是在什么条件下形成的，并根据证据决定应当给予多大程度的信任。

7.2.2 参与计算建构：使认识接受运行检验

人工智能科学具有鲜明的建构性。一个想法需要被转化为数据选择、特征表示、类别划分、规则设计、目标设定、评价指标和运行流程，才能形成可以执行和检验的系统。

学习者参与这一过程，需要把原本较为笼统的认识转化为一系列明确选择，并说明这些选择所依据的理由。

这里的计算建构不限于开发复杂的人工智能模型。组织一组分类数据、设计一套判断规则、比较不同提示产生的结果、设定任务评价指标，或者构造一个简单的人机协作流程，都可以使学习者经历从问题理解到系统实现的过程。这里所强调的“建构”，是一种通过构造可运行过程来解决问题的方式。

例如，在设计分类任务时，学习者需要决定哪些对象可以归入同一类别，选择哪些特征作为判断依据，并处理位于类别边界上的对象；在设定评价指标时，学习者需要判断什么结果可以被视为成功，不同错误会产生怎样的后果；在设计提示和工作流程时，学习者还需要发现表达中的遗漏、歧义和隐含前提。原来停留在语言中的理解，由此被转化为一组需要实际执行的选择。

系统运行会把这些选择所产生的后果呈现出来。偏斜的数据可能使系统形成错误模式，片面的特征可能导致对表面线索的依赖，不恰当的指标可能鼓励偏离真实目标的行为，局部有效的规则也可能在新情境中失效。学习者需要根据运行结果重新审视原来的问题，检查自己的假设，解释错误产生的原因，并修改数据、规则、目标或评价方式。

这一过程使归纳、泛化、判断与自醒进入连续的认知实践。学习者从材料中发现和概括模式，在不同情境中检验模式能否成立，依据运行结果评价方案，并反思自己的选择和信心。计算建构的教育价值，在于使认识从一般表达进入可运行状态，让隐藏在想法中的假设、遗漏和边界受到实际结果的检验。

7.2.3 分析社会嵌入：判断计算结果的现实影响

人工智能的影响会随着系统进入社会过程而获得现实效力。数据、模型和指标需要通过平台、组织、制度和工作流程进入现实活动，进而影响信息传播、人员评价、资源分配、机会获得和社会行动。同一个模型输出可以作为辅助参考，也可以成为自动筛选、风险评级或行动执行的直接依据。它会产生怎样的现实意义，取决于系统获得了什么权限，以及人和组织如何解释和使用它。

人工智能通识教育可以通过推荐系统、智能招聘、医疗辅助、教育评价、金融风控和公共治理等案例，引导学习者分析计算系统如何嵌入社会过程。这种分析需要追踪相对完整的作用链条：现实对象如何被表示为数据，历史经验如何进入模型，社会目标如何被转化为指标，模型输出如何进入组织流程，哪些人会受到结果影响，谁有权提出异议、修改决定并承担责任。

沿着这一链条，学习者可以看到，技术系统中的数据选择、目标设定和评价方式会在社会运行中产生实际后果。计算效率、预测准确率和社会合理性之间也可能存在张力。一个在技术指标上表现良好的系统，仍然可能因为目标设置、适用范围、使用权限或责任安排不当而产生不合理后果。

这种分析把认知自治从个体对信息的辨别扩展到对整个认知环境的理解。学习者需要判断系统正在优化什么、遗漏了什么、对不同群体产生了怎样的影响，以及自动化行动应当受到哪些限制。判断的对象既包括结果是否准确，也包括目标是否合理、程序是

否公正、权限是否适当以及责任是否明确。学习者由此练习把技术证据、社会价值和行动后果联系起来，对计算系统作出有依据的评价。

理解智能机制、参与计算建构和分析社会嵌入，分别面向计算结论的形成过程、认识假设的运行检验和计算结果的现实作用。第一条路径帮助学习者看清结论的来源、依据与边界；第二条路径使学习者亲自经历认识的明确、运行与修正；第三条路径使学习者判断计算结果如何获得现实效力，并追问其中的价值、权限和责任。三条路径共同连接了人工智能系统内部、学习者的认知实践以及人工智能所处的社会环境。

7.3 归纳能力：从有限材料中形成可检验的理解

7.3.1 归纳的基本含义

归纳是学习的起点。人类无法等待所有证据齐备以后再形成理解，也无法在每一次行动之前重新搜集全部信息。学习者总是需要从有限样本、局部经验、少量案例、片段文本、图像线索和他人证言中形成初步理解。归纳能力使人能够在复杂环境中发现规律、建立类别、提出解释和生成假设，并为进一步学习和判断提供起点。

归纳天然带有不确定性。有限样本能够为结论提供支持，却很少能够彻底证明一个结论。统计关系可以提示某种模式，却不会自动给出因果解释。局部经验可以帮助人采取行动，也可能导致以偏概全。AI 时代的归纳训练因此包含两个相互联系的方面：学生既要能够积极发现模式，也要能够认识模式所依据的证据、获得支持的程度以及存在的局限。

人工智能为归纳训练提供了清晰的观察材料。机器学习模型从训练数据中学习模式，图像分类模型从大量图像中提取特征，语言模型从文本序列中学习语言规律，推荐系统从用户行为中学习偏好。学生通过观察这些系统，可以看到归纳过程如何受到样本、标注、特征、目标函数和反馈的影响。更重要的是，学生也能够观察到归纳出现错误的具体方式：数据偏斜可能导致模型偏见，噪声标签会影响分类结果，少量样例可能诱发过度概括，表面线索也可能取代真正具有解释力的深层结构。

7.3.2 归纳能力的认知基础

心理学研究长期把归纳视为人类学习的基本能力。Saffran, Aslin, and Newport (1996) 的婴儿统计学习实验显示，婴儿能够利用语音序列中的过渡概率发现词语边界。这说明,对统计规律的提取具有较早的发展基础,是人类认知发展中的重要机制。Shepard (1987) 的泛化理论表明，人们会依据心理空间中的相似性进行推广。J. B. Tenenbaum et al. (2011) and Griffiths and J. B Tenenbaum (2006) 的贝叶斯认知研究进一步把人类概念学习理解为在先验知识和已有证据之间进行概率推断的过程。

这些研究给教育带来两点启示。第一，归纳能力具有自然基础，学生会在学习过程中主动寻找规律。第二，归纳过程仍然需要教育加以校准。自然形成的归纳容易受到样本显著性、情绪强度、叙事流畅性和既有立场的影响。Tversky and Kahneman (1974) 揭

示的代表性启发式表明，人们经常把“某个个案看起来像某类典型成员”当作“它属于该类”或者“该类更可能发生”的依据，从而忽略该类在总体中的真实比例，也低估样本量和抽样方式所带来的随机波动。在社会生活中，这种倾向可能表现为刻板印象、个案推广、对小样本的过度解释以及对生动叙事的过度信任。

AI 案例能够把这些问题具体呈现出来。一个图像模型如果在训练数据中经常看到“雪地中的狼”，就可能把雪地背景当作识别狼的主要线索；一个学生如果在社交媒体中反复看到某类个案，也可能把高频曝光误认为社会现象的真实比例。两种情况具有不同的形成机制，却共同暴露出样本与总体关系被忽略、表面共现被当作可靠规律的问题。

7.3.3 归纳能力的教育目标

归纳能力的教育目标可以概括为四个递进的认知动作：发现模式，说明支持模式的证据，分析样本与总体之间的关系，并把归纳结果表述为包含条件和可能反例的可检验假设。学生由此从“我看到了某种规律”，逐步走向“这一规律目前得到了哪些证据的支持，还需要通过什么方式继续检验”。

人工智能把归纳所依赖的材料和选择转化为可观察的系统要素。学生通过观察训练数据、标签、特征和模型结果之间的关系，可以看到规律从何而来，也可以通过增删样本、调整分类标准和比较错误类型，检验原有理解是否稳固。亲自构建一个简单任务时，学生必须决定观察什么、如何表示、哪些差异值得关注以及怎样判断结果，原本隐藏在认识中的归纳前提由此成为需要说明、检验和修正的明确选择。

7.3.4 归纳能力的课堂案例

一个典型任务可以围绕“AI 识别蝴蝶”展开。教师给出若干蝴蝶和蛾子的图片，学生先观察它们的外形差异，再让一个简单分类器或现成的 AI 工具进行识别。学生记录 AI 识别正确和错误的案例，分析它可能依据颜色、翅膀形状、背景、拍摄角度或图片清晰度进行判断。随后，教师引导学生比较自己与 AI 的判断：哪些线索具有较高的辨别作用，哪些线索可能产生误导，现有样本还遗漏了哪些情况。

这个任务围绕归纳过程展开，分类精度只是观察归纳过程的一项材料。学生需要回答：我们看到了哪些样本？这些样本能够代表所有蝴蝶和蛾子吗？AI 的错误是否集中在某种背景、角度或外形上？如果希望结论更加可靠，还需要加入什么样的样本？由此，学生在具体活动中理解样本、特征、模式、反例和假设之间的关系。

高中和大学阶段可以采用更复杂的任务。例如，让学生分析某一大语言模型围绕同一主题生成的多次回答，归纳其中常见的错误类型，如虚构引用、时间混淆、过度概括、忽略边界以及把相关关系解释为因果关系。随后要求学生设计核验方法，判断哪些错误可能来自缺少外部检索，哪些错误来自概念混淆，哪些错误与问题设定模糊有关。生成式 AI 的错误由此被转化为归纳训练的材料。

表 7.1: 归纳能力的表现层级

层级	能力表现	AI 场景任务	社会场景任务
模式发现	能发现重复关系、类别特征和异常点	观察 AI 分类结果中的正确与错误模式	比较多篇报道中的共同说法和差异
证据意识	能说明模式由哪些证据支持	检查 AI 回答中哪些断言有引用支持	标注一篇文章中有证据和无证据的判断
样本意识	能分析样本来源和代表性	评估训练数据是否覆盖不同人群、场景和条件	判断社会调查能否代表目标群体
假设意识	能把规律表述为可检验假设	设计新样本，测试模型是否依赖表面线索	提出反例或补充资料，检验社会结论

7.4 泛化能力：在新情境中迁移并识别边界

7.4.1 泛化为何成为 AI 时代的核心能力

泛化是归纳形成理解之后的关键环节。归纳使人从材料中形成规律，泛化则要求人判断这种规律能否进入新的情境。一个结论在原有情境中成立，只能说明它在特定条件下得到了支持；它在新情境中能否继续成立，还需要考察对象、条件、数据分布、任务目标和风险等级是否发生变化。AI 时代把这一问题推到了前台，因为现代 AI 系统的有效性在很大程度上取决于泛化能力，而许多 AI 错误恰恰发生在泛化越过适用边界之处。

机器学习中的泛化通常指模型在未见数据上的表现。过拟合、分布偏移、对抗样本、捷径学习和鲁棒性问题，都围绕模型能否在新情境中保持有效展开。Geirhos et al. (2020) 关于捷径学习的研究说明，模型可能依赖在训练环境中看似有效的表面线索，并在更复杂或发生变化的场景中失败。Koh et al. (2021) 提出的 WILDS 基准则表明，真实世界中的分布变化会显著影响模型表现。这些研究为学生理解规律的适用条件和边界提供了直观材料。

人类学习同样需要处理知识能否迁移的问题。学生在熟悉题型中能够解题，换一种问法后可能出现失误；某个地区有效的教育经验，推广到另一地区时可能失效；一个历史类比在表层情节上相似，深入分析制度背景后可能并不成立；某项医学研究在一个人群中有效，应用到另一人群时仍需谨慎。泛化能力因此是现代教育中的核心能力，它要求学习者始终把规律放回其成立条件中加以理解。

7.4.2 泛化能力的学习科学基础

学习科学把迁移视为衡量教育质量的重要标志。Bransford, A. L. Brown, and Cocking (2000) 指出，学习者如果只记住事实和程序，很难把知识迁移到新的问题中；深层理解和元认知有助于迁移发生。Perkins and Salomon (1988) 的高路迁移理论强调，学习者需要有意识地抽象出一般原则，并在新的情境中主动寻找这些原则的适用方式。Barnett and Ceci (2002) 提出的远迁移分类则提醒我们，任务形式、知识领域、物理环境、社会环境和时间距离都会影响迁移的难度。

这些研究对人工智能通识教育提出了明确要求。学生学习了“模型会过拟合”，并不会自然地把这一认识用于反思机械刷题所形成的表面熟练；学生了解了“训练数据可能存在偏差”，也不会自动把这一认识用于分析社会调查或商业报告。课程需要有意识地设置桥接任务，引导学生识别 AI 问题与一般社会问题之间共有的结构，并把在 AI 场景中形成的边界意识迁移到其他认知活动中。

7.4.3 泛化能力的教育目标

泛化能力的教育目标包括条件识别、场景比较、边界表达和反例设计。学生需要指出结论所依赖的数据、任务、人群、时间和文化条件，比较原有情境与新情境之间的深层异同，把结论改写为带有条件的表达，并主动设计可能使结论失效的案例。

人工智能把“已有表现能否延伸到新情境”转化为可以反复实验的问题。学生可以改变输入背景、对象群体、表达方式、时间条件和任务目标，观察模型表现怎样发生变化；也可以比较训练情境、测试情境和真实使用情境，追踪哪些差异导致原有结论失效。通过这种变式实验，迁移不再只是课程结束以后期待出现的学习结果，而成为学习过程中可以持续观察和检验的认知活动。

7.4.4 泛化训练的课堂案例

一个适合中学和大学的任务是“模型与经验的边界”。教师给出一个情绪识别 AI 在某个公开数据集上表现良好的说明，再提供不同文化背景、不同光照、不同年龄群体和不同拍摄条件下的图片。学生需要预测模型可能在哪些条件下失效，并说明作出预测的理由。

随后，教师提供一个社会领域的案例：某种学习方法在一所重点中学取得了良好效果，另一所农村学校采用以后效果有限。学生使用相同的分析框架回答：原始条件是什么，新情境中哪些条件发生了变化，哪些变量可能影响经验的迁移。

这一任务旨在帮助学生形成边界表达的习惯，算法细节是理解情境和检验边界所使用的材料。学生最终需要分别写出 AI 结论的适用边界和社会经验的适用边界。通过比较，学生可以认识到，泛化能力是一种能够跨越技术场景和社会场景的共同认知能力。

表 7.2: 泛化能力的训练维度

维度	核心提问	AI 案例	社会案例
条件识别	规律依赖哪些条件	训练数据、模型任务、输入格式、部署环境	研究对象、地区、人群、时间、制度背景
场景比较	新情境与原情境是否相似	测试数据与训练数据是否具有相近分布	新学校、新群体、新政策环境是否具有关键相似性
边界表达	结论在哪些条件下成立	模型适用于哪些输入和风险等级	经验适用于哪些人群、环境和目标
反例设计	什么情况可能使结论失效	遮挡、噪声、分布变化、对抗样本	极端个案、异质群体、制度变化、价值冲突

7.5 判断能力：在不确定与价值冲突中作出有理由的选择

7.5.1 判断不同于得出结论

归纳和泛化帮助人形成理解，判断则把理解进一步带入选择和行动。判断能力强调在证据不充分、后果不确定、利益相关者众多以及价值标准可能发生冲突的条件下，形成有理由、可解释并能够承担后果的选择。

判断能力在人工智能通识教育中尤其重要。AI 系统可以给出建议、排序、评分、预测和生成方案，却无法承担教育、医疗、法律、公共治理和个人生活中的规范责任。学生是否使用 AI 生成作文，教师是否采用 AI 评分结果，医生是否采纳 AI 辅助诊断，学校是否使用数据模型预测学生风险，平台是否根据算法推荐内容，都涉及事实、方法、风险、价值和责任等多方面判断。

事实判断关心信息是否可靠，方法判断关心工具是否适用，风险判断关心潜在后果是否可以承受，价值判断关心选择追求的目标是否正当，责任判断关心决定及其后果应当由谁承担。在具体的人机任务中，这些判断还需要进一步转化为协作安排，包括怎样选择工具、如何划分任务、在哪些环节进行人工复核，以及何时需要接管或停止系统。一个成熟的判断过程，往往需要同时处理这些彼此关联的问题。

7.5.2 判断能力的心理学与教育学基础

关于不确定条件下判断的研究，为人工智能通识教育提供了重要参照。Tversky and Kahneman (1974) 的研究表明，人们在概率判断中经常依赖启发式，可能忽略基率、样本量和随机性。Pennycook and Rand (2019) 关于错误信息的研究则显示，缺少反思性

思考的人更容易把虚假信息判断为真实。这些研究并不意味着人的判断无法通过教育得到改善。Fong, Krantz, and Nisbett (1986) 的研究显示，统计训练能够改善人们在日常问题中使用统计原则的能力。Gigerenzer and Hoffrage (1995) 的研究也表明，适当的信息表征方式有助于提高贝叶斯推理的表现。

批判性思维研究进一步强调了判断的结构性。Facione (1990) 主持的 Delphi 报告把批判性思维概括为一种有目的、自我调节的判断，包含解释、分析、评价、推理、说明和自我调节等方面。这一界定与本书所讨论的认知自治高度相关。人工智能通识教育中的判断训练，需要把批判性思维的传统放入由模型和平台中介的信息环境中，使学生能够评价 AI 输出、专家意见、统计图表和社会叙事。

预测研究提供了另一个重要案例。Mellers et al. (2015) 关于“超级预测者”的研究显示，在地缘政治预测任务中，表现较好的预测者往往善于使用概率表达，能够及时更新信念、分解复杂问题并保持开放态度。这些能力与人工智能通识教育中的信任校准和自醒密切相关。学生面对 AI 建议时，也需要学习用概率和条件表达自己的判断，避免在完全接受与完全排斥之间简单摇摆。

7.5.3 判断能力的教育目标

判断能力的教育目标，是使学生能够区分事实陈述、解释、价值主张和行动建议，评估证据强度，权衡风险与后果，识别价值冲突，并说明自己为何接受、修改、暂缓或拒绝某个结论。

学生还需要根据任务目标、证据条件和风险等级，判断是否使用 AI、选择什么工具、将哪些环节委托给 AI，并为关键节点设置人工复核、接管或停止条件。这些选择需要得到明确理由的支持，也需要能够进入责任分析，使学习者说明不同参与者分别拥有什么权限、应当承担什么责任。

人工智能把判断所涉及的多个要素集中呈现在同一任务中。学生在设计或使用 AI 系统时，需要同时处理目标怎样定义、证据怎样获得、指标怎样选择、错误代价是否对称、哪些群体会受到影响以及谁有权作出最终决定。系统能够运行，只能说明某种方案具有可执行性；是否值得采用，还需要把技术表现与风险、价值、制度和责任联系起来。这样的任务使判断从一般态度转化为面对具体条件作出的选择。

7.5.4 判断训练的课堂案例

一个典型任务是讨论“AI 辅助诊断结果应当怎样使用”。教师给出一个医疗影像 AI 系统的性能说明，再提供几个不同的应用情境，包括普通筛查、疑难病例、医疗资源不足地区、患者隐私敏感情境，以及模型训练数据与本地人群差异较大的情境。学生需要分别判断 AI 建议可以用于哪些环节，需要设置哪些人工复核，怎样向患者说明不确定性，以及不同类型的错误后果应当由谁承担。

这一任务可以进一步迁移到学校教育场景。教师给出一个 AI 作文评分系统，要求学生分析：它是否适合用于日常练习中的初步反馈？是否适合用于正式考试评分？面对

不同方言、写作风格和文化背景的学生时，系统是否可能产生不公平结果？学生是否拥有了解评分依据和提出申诉的渠道？由此，学生学习把同一套判断框架用于不同领域。

表 7.3: 判断能力的组成

组成	核心问题	AI 场景示例	评价证据
事实判断	信息是否可靠	AI 回答中的事实和引用是否可以核验	能区分事实、解释和推测
方法判断	工具是否适用	某模型是否适合当前任务与目标人群	能说明工具的适用条件和限制
风险判断	后果是否可以承受	AI 建议适合用于低风险还是高风险场景	能提出复核、接管和风险缓释方案
价值判断	哪些目标更加重要	效率、公平、隐私和安全如何权衡	能识别利益相关者和价值冲突
责任判断	谁应当承担后果	开发者、部署者和使用者分别承担什么责任	能绘制责任链并说明关键问责点
协作判断	人和 AI 应当怎样分工	选择工具、划分任务、设置复核与接管节点	能说明分工理由并保留关键决定权

7.6 自醒能力：对自身认知状态的觉察与校正

7.6.1 为什么需要“自醒”

自醒是四能力模型中综合性较强的一种能力。归纳、泛化和判断都可能发生偏差，自醒使学习者能够在认知过程中觉察这些偏差，并及时作出校正。本书以“自醒”概括元认知监控与认知校正的教育要求，强调一种主动、及时并带有警觉性的认知状态。它既包括对认知过程的持续监控和事后复盘，也包括学习者在当下意识到自己可能受到流畅表达、权威来源、群体意见、情绪动员或 AI 输出的牵引。

智能社会尤其需要自醒。生成式 AI 可以用流畅语言给出貌似完整的答案，推荐系统可以持续塑造人的注意力，自动化评分可以制造客观和精确的印象，智能助手则会显著降低获得答案和完成任务的成本。个体在便利环境中很容易把越来越多的认知活动交给外部系统：题目尚未理解就请 AI 解答，资料尚未阅读就让模型总结，观点尚未形成就直接采用机器生成的草稿，判断尚未完成就接受系统推荐。

自醒能力要求学习者在这些情境中不断追问：我是否真正理解了问题？我是否知道答案的根据？我是否把生成内容误当成了自己的思考？我是否因为表达流畅而高估了结果的可靠性？我是否清楚哪些工作已经委托给 AI？我是否正在让平台持续决定自己的注意力？这些问题帮助学习者保持对自身认知状态和人机分工的觉察。

7.6.2 自醒能力的研究基础

元认知研究为自醒能力提供了心理学基础。Flavell (1979) 把元认知引入认知发展研究，强调个体关于自身认知过程的知识以及对认知过程的监控。Nelson and Narens (1990) 提出元层级与对象层级的框架，说明元认知系统能够监控和控制一阶认知活动。Fleming and Dolan (2012) 进一步从神经科学角度讨论元认知准确性的基础，指出任务表现与对任务表现的觉察可以发生分离。这意味着，一个学生即使完成了任务，也可能缺少对自身理解状态的准确把握。

自我调节学习研究则把元认知放入完整的学习行动中。Zimmerman (2002) 把自我调节学习理解为学习者主动设定目标、监控过程、调整策略和评价结果的循环。P. R Pintrich (2000) 也强调动机、认知、行为和情境调节在学习中的作用。人工智能通识教育需要把这种自我调节进一步扩展到人机互动之中：学生既要调节自己的学习策略，也要调节自己与 AI 系统之间的分工、信任和依赖关系。

认识警觉和认识论认知研究还提示，学习者需要同时评估信息来源、内容可靠性以及知识主张获得支持的方式 (Sperber et al., 2010; Hofer and Paul R. Pintrich, 1997; Sandoval, Greene, and Bråten, 2016)。AI 环境中的信息可能来自人、模型、平台、机构以及复杂的混合生产链。自醒由此还包含对自身信任怎样受到不同来源及其呈现方式影响的觉察。

7.6.3 自醒能力的教育目标

自醒能力的教育目标包括理解监控、协作觉察、信任调节、偏差识别和反思修正。学生需要发现“能够读懂文字”与“理解结论根据”之间的差距，意识到 AI 在哪些环节扩展了材料、解释和方案，自己在哪些环节提供了目标、知识和选择，哪些任务已经委托给 AI，哪些关键判断仍然由自己承担。

学生还需要判断自己的信任程度是否与证据强度和任务风险相匹配，识别情绪、立场、权威、平台机制和语言流畅性对自身判断的影响，并能够根据新证据主动修正结论、行动策略和人机分工。自醒贯穿归纳、泛化和判断的全过程，也可以在行动完成以后通过复盘推动新一轮认识。

7.6.4 自醒训练的课堂案例

一种有效的任务是“AI 使用日志”。学生在完成写作、研究或项目任务时，记录每一次使用 AI 的目的、输入、输出、采用程度、核验过程和最终修改。任务完成以后，学生需要回答：哪些部分来自我原有的理解？哪些部分受到 AI 启发？哪些内容经过了核验？哪些内容最终被删除？哪些地方曾经出现过度依赖 AI 的风险？

另一种任务是“置信变化记录”。学生先独立判断一个问题，并写出自己的信心等级；随后查看 AI 回答、专家资料或同伴意见；最后记录自己的判断和信心发生了什么变化，以及这些变化由哪些证据引起。这个任务把信念更新过程显性化，帮助学生避免在外部意见影响下无意识地改变立场。

课程还可以设计“错误复盘”。学生收集自己在 AI 使用中的一次失败经历，例如相信了错误引用、采用了不恰当的摘要，或者把 AI 建议用于不适合的情境。复盘时，学生需要分析错误发生在哪个环节，包括来源检查、证据核验、分布判断、边界表达、置信调节或责任判断。错误由此成为认识和修正自身认知过程的重要材料。

表 7.4: 自醒能力的训练维度

维度	核心提问	训练任务	评价证据
理解监控	我是否真正理解	解释 AI 答案的证据链和适用限制	能指出自己尚未理解之处
协作觉察	AI 扩展了什么，我把什么交给了 AI	AI 使用日志和贡献标注	能区分本人思考、AI 建议和外部资料,并说明 AI 带来的实际增益
信任调节	我应当相信到什么程度	置信等级与理由说明	信心程度能够与证据强度和任务风险相匹配
偏差识别	哪些因素正在影响我	分析情绪、立场、平台机制和权威影响	能识别影响判断的非证据因素
反思修正	我应当怎样更新结论	错误复盘与二次修改	能根据新证据修正判断、策略和人机分工

7.7 四能力循环：从假设到行动再到校正

归纳、泛化、判断和自醒具有相对独立的训练重点，同时又在真实认知活动中构成相互衔接的循环。归纳回答“我从材料中看到了什么”，泛化回答“这一理解能够延伸到什么范围”，判断回答“在现有条件下我应当如何选择”，自醒回答“我为什么这样理解和选择，是否需要作出修正”。

从课程任务的主要过程来看，四种能力依次处理假设形成、边界检验、行动选择和认知校正。新的证据、情境变化和行动反馈又会推动学习者重新归纳、再次检验并调整判断。自醒也会介入归纳、泛化和判断的各个环节，监控学习者是否充分理解材料、是否越过结论边界、是否恰当地分配信任和责任。因此，四能力循环是一种具有反馈关系的任务结构，不是一套只能按照固定顺序展开的线性程序。

这一循环可以直接用于课程任务设计。例如，在“AI 推荐与信息茧房”主题中，学生先归纳自己所接收推荐内容的模式，再分析不同用户为什么会看到不同的信息世界，随后判断平台和个人应当如何分担责任，最后反思自己的注意力和信念如何受到推荐机制的影响。在“AI 辅助医学影像”主题中，学生先归纳模型识别正确与错误的案例，再

分析模型在新医院和新人群中的泛化风险，随后判断 AI 建议应当怎样进入诊疗流程，最后反思自己是否把模型置信度误当作医学上的确定性。

表 7.5: 四能力循环在课程任务中的展开

阶段	核心动作	常见风险	教学支架
归纳	从材料中形成初步假设	以偏概全、样本不足、流畅性误导	样本表、证据标注、反例收集
泛化	把假设带入新的情境	分布变化、边界越界、表面类比	条件清单、场景比较、失效条件设计
判断	综合证据、风险、价值与责任作出选择	证据不足、工具错配、价值冲突、责任模糊	方案比较、风险矩阵、利益相关者图、责任链
自醒	监控并修正理解、信任和人机协作	过度自信、认知外包、忽视 AI 增益、情绪驱动	置信声明、AI 使用日志、贡献分析、错误复盘

7.8 评价证据：如何观察四能力

人工智能通识教育的评价需要同时关注任务结果、认知过程和判断依据。作品完成度或模型性能无法单独证明认知自治已经形成。教师还应观察学生能否说明样本和证据，比较新旧情境，借助 AI 扩展材料和方案，选择工具并安排人机分工，权衡风险、价值与责任，以及反思 AI 带来的认知增益、依赖和信念变化。下表归纳了四种能力的主要表现证据和评价任务。

表 7.6: 四能力模型的评价证据

能力	可观察表现	评价任务	高水平表现
归纳	标注证据、发现模式、扩展材料、提出假设	分析并扩展一组 AI 输出或社会案例	能说明样本限制，比较候选解释并提出反例测试
泛化	比较场景、生成变式、说明条件、设计边界	判断模型、研究结论或社会经验能否迁移	能评价变式与反例，并把结论改写为有条件的表达
判断	选择工具，权衡证据、风险、价值和责任	分析 AI 医疗、AI 评分或推荐系统中的人机方案	能形成可解释的选择，安排人机分工并提出复核机制

能力	可观察表现	评价任务	高水平表现
自醒	监控理解、调节信任、分析认知增益与依赖	AI 使用日志、贡献说明、置信声明、错误复盘	能解释 AI 的实际贡献，并根据新证据主动修正判断

评价还需要避免把四种能力割裂开来。单项任务可以考察某一种能力，综合任务则应尽量呈现完整的认知循环。例如，学生分析一条 AI 生成的新闻摘要时，可以先归纳摘要中的主要断言及其证据，再判断这些断言能否推广到其他人群或地区，随后决定是否采用或转发相关内容，最后反思自己为什么最初相信或怀疑这条摘要。与只要求识别正确答案的题目相比，这类任务能够提供更加完整的认知自治证据。

评价还应当考察能力能否迁移。学生能够指出 AI 训练数据存在偏差，并不意味着他已经能够识别社会调查中的抽样偏差；学生能够说明模型具有适用边界，也不意味着他会主动检查专家意见、统计结论和个人经验的适用条件。课程因此需要在 AI 场景与一般社会场景之间设置相互迁移的任务，观察学习者能否把样本意识、边界意识、风险意识和自我监控用于新的问题。只有当这些认知动作能够跨越具体工具和案例，四能力模型才真正指向较为稳定的认知自治能力。

7.9 本章小结

本章把认知自治从总体教育理念进一步转化为归纳、泛化、判断和自醒四种可以设计、练习与评价的能力。四能力模型是一种面向人工智能通识教育的课程模型。它不追求穷尽认知自治的全部心理构成，而是从智能社会中的典型认知困难出发，提取课程能够持续训练和观察的核心认知活动。

人工智能通识教育之所以能够支持这些能力的培养，与人工智能自身的科学特性和社会作用方式有关。理解智能机制，使学习者能够追踪计算结论怎样从数据、表示、模型、目标和评价中形成，并认识结论的依据与边界；参与计算建构，使学习者把原本模糊的理解转化为数据、规则、指标和流程，通过运行结果发现假设中的遗漏与限制；分析社会嵌入，使学习者追踪计算结果怎样通过平台、组织和制度获得现实效力，并判断其中涉及的价值、权限、影响和责任。三条路径分别连接系统内部、认知实践和社会环境，为认知自治训练提供了相互补充的课程基础。

在这一基础上，归纳使学习者能够从有限材料中形成有依据、可检验的理解；泛化使学习者能够把已有理解带入新情境，并识别规律成立的条件；判断使学习者能够在证据不完全、风险不确定和价值冲突中作出有理由的选择；自醒使学习者能够觉察自身的理解、信任和依赖状态，并根据新证据进行修正。四种能力在真实任务中相互衔接，形成从假设生成、边界检验、行动选择到认知校正的循环。自醒贯穿这一循环，使学习者能够持续检查认知过程，并推动新一轮理解与判断。

认知自治的形成最终取决于学习者是否实际承担了这些认知活动。人工智能可以帮助整理材料、生成方案、构造变式和提供反馈，课程仍需保证学习者亲自比较样本、解

释证据、检验边界、权衡风险、作出选择并反思人机分工。工具完成任务的质量与学习者形成能力的程度并不等同。只有当学习者能够说明 AI 参与了哪些环节，检查其结果，根据证据作出最终判断，并对自己的选择承担责任，人工智能的参与才会真正转化为认知自治训练。

因此，人工智能通识教育的评价需要超越工具熟练度、作品完成度和模型性能，转向对认知过程及其迁移的观察。教师需要考察学习者能否形成可检验的理解，能否识别结论的适用边界，能否在复杂条件下作出可解释的选择，以及能否监控和修正自身的认知状态。当这些能力能够从 AI 场景迁移到信息判断、知识学习和社会行动之中，人工智能通识教育才真正承担起智能社会中认知自治教育的功能。

第八章　人工智能通识教育与其他学科心智训练的关系

摘要：数学、语言与文学、物理、历史和计算机科学分别形成了形式推理、意义理解、因果解释、情境判断和可执行建构等重要的心智训练传统。它们为认知自治提供了深厚基础，也使人工智能通识教育必须回答一个课程体系问题：既有学科已经训练多种认知能力，人工智能通识教育还具有怎样的独特作用？本章以归纳、泛化、判断与自醒四能力为统一坐标，比较五种代表性的学科训练，进而说明人工智能通识教育能够把分散于各学科中的认知资源组织到“数据—模型—输出—行动—反馈”的完整链条中。智能活动在这一链条中获得可执行形式，模型输出又经由平台、组织和制度进入现实行动。由此形成的认知自治训练具有相对独立的内容与课程结构，并与批判性思维、媒介素养、统计思维和计算思维相互支持。

关键词：人工智能通识教育；学科思维；心智训练；四能力；认知自治

8.1　问题的提出：为什么需要进行学科比较

第一篇第三章已经界定了认知自治及其六个分析维度。本篇第五、六章分别分析人工智能学科本身及其进入社会生活的机制，第七章进一步提出归纳、泛化、判断与自醒四能力模型，说明人工智能通识教育如何将认知自治这一教育理念转化为可培养、可评价的能力模型。本章继续回答一个课程体系问题：既有学科已经承担多种心智训练，人工智能通识教育在认知自治能力培养中还具有怎样的独特作用？

这一问题直接关系到人工智能通识教育的课程定位。当前各学段已经形成相对稳定的学科结构。人工智能通识教育若只表现为新增工具和操作技能，容易成为课程表中的一门技术课；若将全部内容分散到既有学科，又可能失去贯穿数据、模型、平台、行动与责任的整体结构。通识教育关注人的整体发展，一类知识能否取得稳定的课程位置，取决于它能否提供具有持续价值的认识方式和心智训练。

本章选择数学、语言与文学、物理、历史和计算机科学进行比较，用它们呈现形式推理、意义理解、经验解释、情境判断和可执行建构五种代表性的训练传统。完整的学科谱系还包括哲学、统计学、社会科学、艺术等领域，本章的选择服务于核心认知结构的分析。比较的目的在于明确各学科的训练重心，揭示它们对认知自治的共同支持，并据此判断人工智能通识教育具有哪些相对独立的教育内容。至于不同学段采取独立设课、模块教学还是跨学科融合，属于后续课程实施问题。

8.2 技能、学科思维与心智训练

讨论学科关系，需要区分技能训练、学科思维方式和心智训练三个层次。技能训练指向可操作的任务能力。数学计算、物理实验、史料检索、程序编写、数据标注和人工智能工具使用，都可以通过示范、练习和反馈得到提高。技能是学习活动的重要组成部分，其作用范围通常与特定任务和工具紧密相连。

学科思维方式体现一个领域提出问题、建立对象和检验结论的基本方式。数学重视抽象、形式化和证明，语言与文学重视语境、意义和表达，物理重视建模、理想化和实验检验，历史重视来源分析、情境化和证据互证，计算机科学重视分解、算法化、可执行性和调试。学科思维使学习者获得特定的观察角度，也为处理新的问题提供可调用的方法资源。

心智训练指经过长期学习形成的较稳定认知倾向。它源于具体学科活动，并可以在变式练习、比较反思和迁移支持下，被重新调用于新的问题情境。由学科训练走向跨情境能力需要教学设计，知识掌握和技能熟练不会自动产生远迁移。因而，判断一门课程的通识价值，既要看它教授什么知识和方法，也要看学习者在什么活动中形成怎样的认知倾向，以及这些倾向能否在新的环境中得到有意识的调用。

前一章建立的四能力为学科比较提供了共同坐标。归纳关注如何从有限材料形成有依据的理解，泛化关注已有理解能否进入新的对象和情境，判断关注如何综合证据、价值、风险和责任作出选择，自醒关注如何监控并校正理解、信任、依赖和行动状态。第一篇提出的六个分析维度说明认知自治需要调节什么，四种能力说明学习者依靠什么完成调节。以下比较以四能力为主线，同时考察各学科对六个维度提供的基础支持。

8.3 数学：形式约束下的必然推理

数学训练的重心可以概括为形式约束下的必然推理。数学教育引导学习者从具体对象中抽取结构，在定义、公理和规则构成的系统中形成结论。数、形、函数、概率、空间、变化和关系都可以成为抽象对象。学习者逐步理解，直觉能够引出猜想，结论的成立还要接受条件、证明和一致性的约束。

Pólya (1957) 关于问题解决的经典研究强调理解问题、制定计划、执行计划和回顾反思的连续过程。Schoenfeld (1985) and Schoenfeld (1992) 进一步指出，数学问题解决受到知识资源、策略、过程控制和信念系统的共同影响，元认知监控在其中具有关键作用。由此看，数学中的猜想、证明、反例和复核分别支持从现象中归纳关系、检验结论能否推广、判断推理是否成立以及监控解题过程。概率统计还训练学习者依据证据强弱和不确定性调整判断。

人工智能场景改变了这些训练所面对的认识条件。数学问题中的对象、前提和规则通常能够明确表达，人工智能系统的前提则可能分散在数据采集、标签定义、指标选择、模型训练和平台目标之中；其结论往往具有概率性、经验性和情境依赖性。学习者需要把数学培养的条件意识和推理检验带入模型判断，同时追问数据如何形成、预测能否迁移、系统优化什么目标以及输出如何影响后续行为。形式推理由此成为认知自治的重要

基础，但无法单独覆盖来源、平台、反馈和责任构成的完整问题链条。

8.4 语言与文学：语境中的意义理解与表达

语言与文学训练的重心可以概括为语境中的意义理解与表达。词语、语气、叙事、立场、声音和修辞共同参与意义生成；同一句话进入不同场合，可能获得不同含义。语言学习使人超越字面信息，理解隐喻、情绪、文化背景和交往目的，文学阅读则通过人物、叙事和多重视角展开经验的复杂性。

Vygotsky (1978) 把语言视为思维发展的重要媒介，强调社会互动和符号工具在认知发展中的作用。Halliday (1978) 从社会语境考察语言的意义功能，L. M. Rosenblatt (1978) 则把阅读理解为读者与文本之间的互动过程。细读、概括和解释支持学习者从词句与结构中归纳意义，跨文本比较和语境重构促进泛化，论证与表达要求作出有依据的判断，写作中的修改和反思持续训练自醒。

人工智能生成内容使意义理解进入新的来源结构。学习者面对的对象既包括相对稳定的文本，也包括即时生成的文字、图像和声音，以及经过算法筛选和平台排序的信息。判断一段内容时，既要分析它表达了什么、采用了怎样的叙事与修辞，也要追踪材料来源、生成机制和传播链条，核查流畅表达所依赖的事实与证据。语言文学培养的语境敏感和表达能力可以支持这种工作，人工智能通识教育还要把模型介入、平台作用、使用后果和责任归属纳入同一认知过程。

8.5 物理：经验约束下的因果解释

物理训练的重心可以概括为经验约束下的因果解释。面对复杂自然现象，学习者需要识别关键变量，建立理想化模型，提出可检验预测，并通过观察、实验和误差分析修正解释。物理模型总在一定假设和条件下成立，因而结论的力量与适用边界同时受到关注。

Chi、Feltovich 和 Glaser 的研究发现，新手往往根据题目的表面情境分类，专家更倾向于依据能量守恒、牛顿定律等深层原理组织问题 (Chi, Feltovich, and Glaser, 1981)。Hestenes (1992) 提出的建模教学强调模型建构、验证和应用的连续过程，diSessa (1993) 则揭示日常直觉在物理理解中同时具有支持和干扰作用。物理学习由此训练学习者从有限现象归纳关系，把模型迁移到新的条件，依据实验比较解释，并在预测与观察不一致时检查假设、测量和推理。

人工智能系统经常依据统计相关性进行预测，其输出又可能改变人的行为和后续数据。此时，因果解释、统计预测和反馈效应需要得到区分。学习者既要考察结果是否符合已有机制和观测证据，也要检查训练环境与应用环境是否一致、模型置信度是否得到校准、系统参与行动后是否改变了原有条件。物理学提供模型、证据、因果和误差意识，人工智能通识教育进一步处理人、数据、模型、平台和制度共同作用时出现的信任与责任问题。

8.6 历史：有限证据下的情境判断

历史训练的重心可以概括为有限证据下的情境判断。过去无法被直接观察和重复验证，人们只能依据留存下来的材料形成理解。这些材料可能残缺，也会受到记录者身份、立场、目的和时代条件的影响。历史学习要求辨析来源、比较记载、重建语境，并在事实、解释和评价之间保持区分。

S. S. Wineburg (1991) 发现，历史专家阅读文献时会持续进行来源分析、情境化和互证。Seixas and Morton (2013) 进一步将历史思维概括为历史意义、证据、延续与变化、因果、历史视角和伦理维度。S. Wineburg and McGrew (2019) 关于在线信息判断的研究还显示，专业事实核查者倾向于离开当前页面进行横向阅读。史料辨析和互证支持从有限材料形成理解，语境重构帮助学习者检验解释能否迁移，证据权衡与伦理评价构成判断，对材料缺失、解释偏向和新证据的持续关注则促进自醒。

生成式人工智能和算法传播改变了材料的形成与可见方式。文字、图像、声音和视频可以由模型生成或修改，平台排序又会影响什么内容被看见、以何种频率被看见。学习者需要沿着作者、数据、模型、发布者和传播平台追踪来源，考察技术过程如何改变信息及其可信感。历史学培养的来源辨析、证据互证和情境判断可以进入这一任务，人工智能通识教育则把生成机制、算法中介、传播反馈和现实责任同时纳入审视。

8.7 计算机科学：复杂问题的可执行建构

计算机科学训练的重心可以概括为复杂问题的可执行建构。学习者需要明确目标和约束，将问题分解为相互衔接的子任务，选择数据表示，设计算法，并通过运行、测试和调试不断修正方案。抽象、分解、自动化和调试使问题解决过程获得可检查、可执行和可迭代的形式。

Wing (2006) 将计算思维概括为运用计算机科学基本概念进行问题求解、系统设计和人类行为理解的思维方式。Papert (1980) 强调编程可以成为儿童思考自身思考过程的工具，Brennan and Resnick (2012) 则从计算概念、计算实践和计算观念三个方面分析创造性编程学习。问题分解和数据表示支持归纳，模块复用和条件分析涉及泛化，测试与方案选择构成判断，调试和迭代则为自醒提供了可观察的活动形式。

人工智能延续了可执行建构的传统，也改变了系统行为与设计意图之间的关系。在许多基础程序中，规则和执行步骤由设计者明确规定；机器学习系统从数据中形成行为，其输出受到训练数据、模型能力、输入方式、使用情境和平台目标共同影响，具体结果难以仅从程序步骤中直接读出。学习者因此要从“如何把问题转化为机器能够处理的过程”继续走向“如何管理机器对人类认识和行动的参与”：哪些任务可以委托，委托到什么程度，输出需要经过哪些核查，何时应暂停或否决，以及最终判断和责任由谁承担。

8.8　人工智能通识教育的独特心智训练

上述比较表明，传统学科已经为认知自治提供了多方面基础。但这并不意味着人工智能通识教育只是既有心智训练的简单重复。它的相对独特价值来自第五、六、七章共同揭示的结构。五章说明，人工智能把表示、学习、推断、检验和修正转化为可以构建和运行的过程，使智能活动本身成为可观察、可操作和可比较的学习对象。第六章说明，模型输出经由平台、组织和制度进入信息传播、组织决策与现实行动，并在持续反馈中改变后续数据和人的行为。第七章进一步把这一过程组织为归纳、泛化、判断与自醒的能力循环。

三方面结合以后，人工智能通识教育能够让学习者沿着“数据—模型—输出—行动—反馈”的完整链条观察、构建、使用和校正认知过程。学习者既要从数据和经验中形成理解，也要检验模型能否迁移到新的对象；既要评价输出、安排人机分工并作出选择，也要监控系统使用如何改变自己的信任、依赖和责任状态。认知过程在这里同时具有技术形式和社会效力，任何单一传统学科都只覆盖其中的部分环节。

这种完整链条还使学习者在多种角色之间转换。他可能是数据提供者、模型使用者、系统设计参与者、结果评价者，也可能是受算法影响的行动者。角色变化会改变能够获得的证据、可以施加的控制以及应当承担的责任。人工智能通识教育由此把认识问题与行动问题连接起来：一个输出是否可靠，需要结合来源、证据和边界判断；一个输出是否采用，还要结合价值、风险、可逆性和责任判断；一次使用是否继续，则需要观察反馈并调整信任和依赖。

表8.1概括了各学科提供的主要心智训练。注意，表中的概括仅表示训练重心，不构成对任何学科价值的完整定义。

表 8.1: 主要学科的核心心智训练与典型判断标准

学科	核心心智训练	典型判断标准
数学	形式约束下的必然推理	前提是否明确，推理是否有效，结论是否由前提必然推出
语言与文学	语境中的意义理解与表达	理解是否合乎语境，解释是否有文本依据，表达是否准确得体
物理	经验约束下的因果解释	解释是否符合经验事实，机制是否合理，预测是否可以检验
历史	有限证据下的情境判断	来源是否可靠，证据能否互证，解释是否符合历史情境
计算机科学	复杂问题的可执行建构	问题能否明确表示，过程能否执行，结果能否测试，错误能否定位并修正

学科	核心心智训练	典型判断标准
人工智能通识	智能环境中的认知自治	来源是否可信，证据是否充分，结论边界是否清楚，对模型的信任是否适度，判断与责任是否明确

人工智能通识教育与若干相邻素养也存在清楚的协作关系。批判性思维提供理由审查、论证评价和反思判断的一般方法；媒介素养关注信息来源、表达形式、传播过程和平台影响；统计思维关注数据、抽样、不确定性和推断；计算思维关注问题表示、分解、执行和调试。人工智能通识教育吸收这些认知资源，并把它们组织到人工智能参与信念形成、判断选择和现实行动的全过程中。它的课程重心由人工智能这一共同对象贯穿，包含数据如何表示现实、模型如何学习和生成、输出如何进入平台与制度、人在何处保留判断以及责任如何分配等彼此关联的内容。

8.9 本章小结

数学、语言与文学、物理、历史和计算机科学分别提供了形式推理、意义理解、因果解释、情境判断和可执行建构等重要训练。它们以不同方式支持归纳、泛化、判断与自醒，也为来源、证据、边界、信任和责任的调节提供认知基础。学科心智训练产生于具体实践，其跨情境调用需要变式、比较、反思和迁移支持。

人工智能通识教育将这些认知资源组织到数据、模型、输出、行动和反馈构成的完整链条中。智能活动在这一链条中获得可执行形式，并通过平台、组织和制度取得现实效力。学习者由此能够观察认知过程、构建和使用人工智能系统、评价其输出，并根据反馈调整信任、依赖、人机分工和责任安排。这种系统整合构成人工智能通识教育相对独立的心智训练价值。

至此，本篇已经从人工智能学科、人工智能进入社会生活的机制、认知自治的四能力模型以及与其他学科心智训练的关系四个方面，说明了人工智能通识教育培养认知自治能力的科学基础、实现途径和独特价值。接下来需要进一步回答：哪些人工智能知识、方法、应用和责任问题能够承载这些能力训练，并共同构成面向认知自治的内容体系。

第三篇

内容框架：人工智能通识教育应当教什么

本篇导言

前两篇从智能社会的认知环境出发，基于人工智能的科学属性、社会作用机制，讨论了人工智能通识教育的时代必要性和可行性。在此基础上，本篇进一步回答人工智能通识教育的内容问题：面对快速演进、形态繁多的人工智能技术，哪些知识应当成为全体学习者共同的教育基础，又应当怎样把这些知识组织成相对稳定、彼此关联的内容体系。

人工智能技术持续更新，具体产品、操作界面和使用技巧具有鲜明的阶段性。通识教育需要在变化中把握相对稳定的知识主干，以长期存在的基本问题组织课程内容，并把大模型、多模态系统、推理模型和智能体等前沿进展放入这一结构中理解。这样的内容体系应当具有基础性、结构性、解释性、迁移性和公共性：为理解人工智能提供共同语言，把分散的知识组织成相互联系的结构，解释系统能力形成及发生错误的原因，支持学习者分析新的技术与场景，并涵盖智能社会中与个人权益、公共生活和行动责任密切相关的问题。

基于这一理念，本篇将人工智能通识教育的内容组织为五个相互联系的部分：人工智能基础概念、人工智能方法、人工智能应用、人工智能交叉融合以及风险、伦理与责任。这五部分沿着人工智能从科学对象走向社会实践的过程展开，依次回答人工智能是什么、人工智能如何形成能力、人工智能如何成为现实系统、人工智能如何进入其他知识领域，以及人应当如何使用、监督和治理人工智能。它们共同构成从认识对象、理解机制、解释系统、迁移方法到负责任行动的完整认知链。

人工智能基础概念为学习者建立认识智能机器的坐标。人工智能的思想源头、学科形成、发展历程、研究目标、学科特征和概念边界，使学习者能够把当前技术放回长期演进中定位，区分人工智能与自动化、算法、机器人、机器学习和生成式人工智能等相关概念，并依据系统的实际功能和运行条件判断其能力与边界。

人工智能方法解释机器能力如何形成。表示、学习、推理、生成和行动等基本问题，以及数据、知识、目标、模型、算法和反馈之间的关系，共同构成理解人工智能技术的机制骨架。方法学习帮助学习者追踪系统结论的形成过程，理解能力依赖的条件，分析模型为什么能够有效、为什么可能出错，并为评价新的技术进展提供相对稳定的解释框架。

人工智能应用连接抽象方法与现实经验。识别、翻译、推荐、生成和决策等功能进入真实场景以后，会同数据、设备、软件、用户、组织流程和制度规则共同构成现实系统。应用部分引导学习者越过可见的功能表现，理解模型输出怎样产生、怎样进入人的判断和行动，以及系统效果、适用边界和潜在影响应当依据什么证据加以评价。

人工智能交叉融合讨论人工智能如何参与其他领域的问题求解与知识生产。数学、工程、自然科学、生命科学、社会科学以及人文艺术具有不同的研究对象、数据形态、理论约束和证据标准。交叉融合的学习重点在于把领域问题转化为人工智能可以处理的任务，选择和改造合适的方法，引入领域知识与约束，并通过相应的专业程序验证结果。由此，方法迁移始终与条件识别和领域判断相联系。

风险、伦理与责任把技术理解引向价值判断和负责任行动。信息真实性、数据与隐私、公平、安全、人的主体性、社会影响和责任分配等问题，需要形成独立、系统的知识结构；这些问题同时贯穿基础概念、人工智能方法、人工智能应用和交叉融合，持续检验技术目标是否合理、证据是否充分、利益与风险如何分配，以及不同主体应当承担什么责任。

五部分具有递进关系，也保持循环联系。基础概念为方法理解提供语言和坐标，方法为应用分析提供机制解释，应用使方法在现实系统中接受检验，交叉融合以领域问题和专业标准约束方法迁移，风险、伦理与责任则从价值和后果出发反向审视技术目标、系统设计与使用方式。认知自治贯穿这一过程，归纳、泛化、判断和自醒四种能力也会在不同内容和任务中被反复调用。

接下来的五章将分别展开这五部分的教育功能、内容原则、共同主干和拓展内容。本篇由此确定人工智能通识教育“教什么”的基本范围与组织逻辑。五部分在具体课程中的比例、深度和顺序，可以依据学习者的年龄、知识基础、专业方向和现实需要进行调整；第四篇将在这一内容框架上，进一步讨论不同阶段与群体的课程实施。

第九章　人工智能基础概念：建立理解智能机器的坐标系

摘要：人工智能基础概念指的是对人工智能这门学科的基本理解，包括对人工智能研究对象、实现方式、历史来源和学科边界的整体认知。对人工智能基础概念的学习承担着建立认知坐标的功能，如果缺少这一坐标，就容易把当前流行的产品等同于人工智能本身，把自动化功能理解为智能，把语言流畅误判为真实理解，也容易在技术崇拜与技术恐惧之间摇摆。本章从通识教育的角度讨论人工智能基础概念的设计原则和推荐内容。人工智能可以被理解为研究如何以计算方式模拟人类智能行为的科学；围绕这一研究对象，又形成了丰富的工程实践和交叉方法。人类智能是其重要参照，但机器并不需要完整复制人的内部过程。人工智能的思想源头来自人类制造智能机器的愿望；形式逻辑和数理逻辑使思维过程能够被分析、符号化和演算；可计算性理论与通用计算机为机器执行这些演算提供了理论模型和物质工具；图灵关于机器智能的思考以及达特茅斯会议最终促使人工智能形成独立学科。此后的符号主义、专家系统、统计机器学习、深度学习、大模型和智能体共同构成一条不断扩展的历史主线。基础概念教育应帮助学习者区分人工智能与自动化、机器人、机器智能、算法、互联网、大数据、机器学习、深度学习、生成式人工智能和通用人工智能，理解这些概念之间的包含、交叉和支撑关系。历史、定义和概念边界由此共同服务于认知自治：学习者能够把新技术放回长期演进中定位，追问能力形成的条件，校准对机器的信任，并在快速变化的智能社会中保持清醒判断。

关键词：人工智能基础概念；人工智能起源；人工智能史；学科特征；概念边界；智能行为；机器智能；认知自治

9.1　设计思路

9.1.1　人工智能基础概念教育功能

很多人工智能通识教育实践从一个具体工具或应用开始。聊天机器人能够回答问题，图像模型可以生成画面，推荐系统持续推送内容，自动驾驶汽车能够感知道路。这些可见功能的介绍容易激发兴趣，也容易让学习者把人工智能理解为眼前的某个产品。产品迭代之后，原有认识随之失去支点；一种新模型表现出更强能力时，学习者又可能把它看作全新的、与过去没有关联的技术。

基础概念的作用，是为持续变化的技术建立相对稳定的坐标系。这个坐标系至少应回答六个问题：

1. 人工智能所说的“智能”指什么；
2. 人工智能试图解决什么类型的问题；
3. 人工智能为什么以计算作为主要实现方式；
4. 这门学科从哪些思想和技术基础中产生；
5. 人工智能在历史上经历了哪些重要的方法转向；
6. 人工智能与自动化、机器人、算法、机器学习和生成式人工智能等相邻概念有什么关系。

这些问题决定学习者如何解释人工智能。一个人若把人工智能等同于聊天模型，就难以理解专家系统、搜索算法、机器视觉和机器人规划为何同属人工智能；若把一切自动运行的设备都称为人工智能，就无法区分预设程序、反馈控制和机器学习；若把模型的语言表现直接等同于理解、意识或人格，就容易对机器能力产生过度推断。

基础概念教育还承担着消除“技术现在主义”的任务。所谓技术现在主义，是指只根据当前形态理解一门技术，忽略它长期积累的思想、方法和争论。人工智能的发展跨越逻辑推理、知识表示、概率统计、神经网络、强化学习和大规模预训练等多条路线。今天的大模型建立在此前数十年的语言建模、表示学习、计算硬件和数据积累之上。了解这一过程，学习者能够把热点技术视为历史链条中的阶段性成果，既看到突破，也看到尚未解决的问题。

基础概念为后续技术内容提供共同语言和评价尺度，也使学习者能够在新技术出现时判断其位置、能力和边界。课程可以从工具、应用或问题情境进入，但需要及时回到稳定的概念坐标。

9.1.2　基础概念与认知自治

归纳能力体现在从历史材料和技术案例中概括人工智能研究对象、能力来源和方法转向，避免把单一产品或局部表现当作人工智能整体。

泛化能力体现在把概念关系用于新技术。学习者能够判断一个新系统主要依赖规则、学习、生成还是智能体结构，并分析已有概念在新情境中的适用范围。

判断能力体现在区分自动化与学习、语言表现与真实理解、模型与完整应用系统，并依据数据、目标、计算资源和环境说明能力成立的条件。

自醒能力体现在觉察神秘化、拟人化和技术宣传如何影响自身信任。学习者能够避免因语言流畅而过度信任，也不会因一次错误而完全否定技术，并能区分已经验证的事实、尚有争议的观点和未来设想。

基础概念由此帮助学习者形成稳定而开放的认识姿态：理解人工智能的历史连续性和现实价值，直面新方法带来的变化，并持续追问来源、证据、适用范围和责任。

9.2 内容建议

9.2.1 人工智能定义

关于人工智能的定义并不唯一，例如，McCarthy (2007) 后来把人工智能概括为制造智能机器、尤其是智能计算程序的科学与工程。对通识教育来说，这一定义需要足够简洁，不包含太多术语，同时也需要明确学科的基础特征。如第二篇所述，本书对人工智能定义如下：

人工智能是用计算机模拟人类智能行为的科学。

这一定义里最重要的内容是对人类“智能行为”的模拟。人工智能长期以人类智能为参照，但“像人一样完成任务”与“完全复制人的内部过程”是两个不同目标。图灵测试采用行为表现作为判断标准，原因之一在于人的内部智能过程难以直接观察 (Turing, 1950)。现代人工智能也经常采用与人不同的实现路径：计算机通过大规模搜索完成棋局判断，通过高维向量表示词义，通过梯度优化调整模型参数。这些计算过程与人类思维的内部机制具有明显差异，但在任务、表示、证据、目标和反馈等问题结构上却可以进行比较。

基础概念教育需要保留这种区分。机器可以表现出语言能力，却未必拥有与人相同的生活经验；可以生成表达情感的文字，却不能由此确认它具有主观感受；可以给出正确答案，也可能缺少对答案来源和现实后果的理解。评价机器智能时，应把功能表现、内部机制、环境适应和责任能力分开讨论。

9.2.2 人工智能概念辨析

人工智能领域中有很多概念，需要仔细辨析以避免混淆。然而，概念辨析不能依赖一句静态定义。较有效的方法是比较研究对象、实现方式和彼此关系。

表 9.1: 人工智能及其相邻概念

概念	核心含义	与人工智能的关系
智能机器	表现出某种智能功能的人工装置，范围取决于人对“智能”的判断	可以通过机械、电路、控制或人工智能等多种方式实现；人工智能常是智能机器中的一个组成部分
自动化	系统按照预设流程或反馈机制自动运行，重点是减少人工操作	自动化可以不包含学习和推理；人工智能能够提高自动化系统的感知、适应和决策能力

概念	核心含义	与人工智能的关系
机器人	具有感知、控制和物理行动能力的机器系统	机器人是物理载体，人工智能可以成为其感知和决策模块；大量人工智能系统并不具有机器人身体
机器智能	机器所表现出的智能能力，强调能力属于机器	与人工智能高度重叠；“机器智能”更强调结果，“人工智能”也指研究这些能力的学科与技术体系
算法	解决一类问题的明确步骤或计算规则	算法是人工智能的组成工具；许多普通算法不涉及智能行为，人工智能系统通常由多种算法共同构成
互联网	连接计算设备、传输和共享信息的基础设施	为人工智能提供数据、计算和服务环境，也接受人工智能的搜索、推荐与安全技术支持
大数据	大规模、多样和快速产生的数据及其处理技术	数据是现代机器学习的重要资源；拥有大量数据本身不等于具备人工智能能力
机器学习	使系统从数据或反馈中改进性能的方法集合	是人工智能的重要方法分支；人工智能还包括知识表示、搜索、规划、推理和机器人等内容
深度学习	以多层神经网络为主要模型的机器学习方法	是机器学习的分支，也是现代人工智能的重要技术基础，但不能代表人工智能全部
生成式人工智能	学习数据分布并生成文本、图像、声音、视频、代码等内容的系统	属于人工智能的应用和方法形态；识别、预测、规划、控制等大量系统并不以内容生成为目标
大语言模型	通过大规模文本等数据预训练、能够理解和生成语言的模型	是生成式人工智能的一类基础模型；大模型还可以面向视觉、声音、科学数据和多模态信息
智能体	能够围绕目标感知环境、规划步骤、调用工具、执行行动并利用反馈的系统	通常以模型为认知核心，并结合记忆、工具、权限和工作流程；聊天界面本身不必然构成完整智能体

概念	核心含义	与人工智能的关系
专用人工智能	在限定任务或领域中形成能力的人工智能系统	当前实际应用主要属于这一类型，能力可能很强，迁移范围仍受任务和环境限制
通用人工智能	能够跨广泛任务学习、迁移和解决问题的假设性或发展中目标	尚缺少公认定义和统一测试标准；讨论时应区分研究目标、系统宣传和已经验证的能力
超级智能	在几乎所有重要认知任务上显著超过人类的设想	主要用于远期研究与治理讨论，不宜用当前单项性能直接推断其已经出现

概念之间经常存在包含和交叉关系。深度学习属于机器学习，机器学习属于人工智能的重要方法；大语言模型属于基础模型和生成式人工智能的一部分；智能体可以调用大语言模型，也可以组合规则、搜索、数据库和外部工具。用一张单向树形图难以表达全部关系，课程应同时使用层级关系、系统组成和对比案例。

9.2.3 人工智能的思想和技术源头

人工智能作为独立学科形成于 20 世纪中叶，其思想准备经历了漫长过程。基础概念教育不需要穷尽所有历史人物，可以围绕三条主线组织：智能机器的愿望、思维的形式化、计算的机器化。

1. 智能机器的愿望

人类很早就尝试制造能够自动运行、模仿生物或替人工作的装置。古代关于自动人偶和机械动物的传说，以及后来出现的水钟、自动剧场、机械乐团和可编程自动装置，都体现了人类对智能机器的持续想象。这些装置主要依赖机械结构、液压、齿轮和预设顺序，尚不具备现代人工智能的计算结构，却提出了一个延续至今的问题：机器能否承担原本由人完成的复杂活动?

自动机械的历史在课程中承担概念对比功能。学习者可以由此区分“自动执行”与“智能适应”，理解高度自动化并不必然包含学习、推理和环境理解。同时，这段历史也说明人工智能的产生并非偶然，它延续了人类借助人造装置与技术工具扩展自身能力的长期愿望。

2. 思维的形式化：从逻辑到数理逻辑

人工智能的第一条理论主线来自对思维规律的研究。亚里士多德的三段论把思维形式与思维内容区分开来：当推理形式有效且前提成立时，结论才得到保证。这个区分对人工智能和认知自治都具有持续意义。系统的推理过程可能形式正确，输入事实却可能错误；生成答案可能结构完整，证据却不可靠。

19 世纪，Boole (1854) 以符号和代数运算描述逻辑关系，使推理过程能够被数学化。此后，弗雷格、罗素、怀特海、希尔伯特和哥德尔等人的工作推动数理逻辑形成。思维

形式由自然语言中的规则转化为定义明确的符号系统和演算过程，为机器执行推理奠定了理论基础。

这条历史主线包含一个重要的认识论转变：人的部分思维活动可以被分析为形式结构，形式结构可以表示为符号，符号可以按照规则演算。具体设备使这种转变获得实现条件，人工智能由此获得了把智能转化为计算问题的可能性。

3. 计算的机器化：可计算性理论与通用计算机

思维能够形式化以后，还需要能够执行形式运算的通用机器。Turing (1936) 在 1936 年提出抽象计算模型，以有限状态、读写操作和规则表描述一般计算过程。图灵机的意义在于给出了“可计算”的清晰模型，也揭示了计算能力的理论边界。

Shannon (1938) 把布尔代数与继电器、开关电路联系起来，说明逻辑运算可以由电路实现。存储程序思想进一步使程序能够作为数据保存在机器中，计算机可以通过更换程序处理不同任务 (Neumann, 1945)。通用电子计算机的出现使“用计算模拟智能”从哲学和数学设想转化为可实验的工程计划。

可计算性同时提醒学习者，计算机能力具有边界。计算速度和存储规模的提高能够解决更多复杂任务，却不能消除理论上不可计算的问题。人工智能建立在计算系统上，其能力也受到表示方式、算法复杂度、数据资源和可计算性的共同约束。

4. 图灵的机器智能设想

图灵对人工智能的贡献不止于计算理论。他讨论了机器学习、奖励与惩罚、模拟进化和“儿童机器”等思想，并在 1950 年提出图灵测试，把“机器是否思考”的哲学争论转化为可以观察的行为测试 (Turing, 1950)。

图灵测试的教育价值，在于它提出了操作性标准：无法直接观察机器内部是否拥有心智时，可以先考察其行为表现。但它的局限也同样重要。语言表现可以成为智能证据，却难以单独证明真实理解、环境经验、意识或责任能力。现代生成式人工智能重新激活了这一争论，使“表现得像有智能”与“拥有何种智能”成为基础概念教育中的重要问题。

9.2.4 人工智能发展史

1. 达特茅斯会议

1956 年的达特茅斯会议通常被视为人工智能学科形成的标志。会议之前，机器对弈、定理证明、神经网络和机器学习等探索已经出现；会议把这些分散工作置于“人工智能”这一共同名称下，并提出语言、抽象、神经网络、自我改进、创造性和计算复杂度等问题 (McCarthy et al., 1955)。

一门学科的形成需要研究对象、问题集合、方法共同体和制度载体。达特茅斯会议的意义体现在四个方面：

1. 确立了以机器实现智能为共同研究目标；
2. 把逻辑、计算、心理、神经科学和工程探索组织到同一问题空间；
3. 形成了若干持续影响后世的研究方向；
4. 促成实验室、学术会议、期刊和人才培养体系逐步建立。

人工智能由此表现出鲜明的跨学科起源。数学提供逻辑、概率和优化工具，计算机科学提供算法和系统，心理学与认知科学提供关于人类思维的模型，神经科学提供生物智能的启发，控制论和工程学提供反馈与行动机制。人工智能在形成独立学科后仍保持这种开放边界。

2. 符号主义：把知识和推理交给机器

人工智能早期的主流路线以符号表示知识，通过搜索、规则和逻辑推理解决问题。机器对弈、自动定理证明、问题求解和早期对话程序展示了计算机执行复杂规则的能力。Newell and Herbert A. Simon (1976) 把计算机科学视为研究复杂信息加工系统的经验性学科，并以物理符号系统假说解释智能行为。

符号方法的优势是结构明确、推理过程较易追踪，适合规则清晰、知识可以表达的问题。其困难也很明显：开放环境中的常识难以穷尽，不确定信息难以用严格规则完全描述，搜索空间可能随问题规模迅速增长。早期研究对难度估计不足，技术承诺与实际能力之间出现落差，人工智能进入第一次低潮。

3. 知识工程：从通用推理转向领域经验

专家系统把特定领域专家的经验整理成知识库，再由推理程序解决诊断、分析和决策问题。它表明，强大能力经常来自丰富领域知识，而非单一通用推理机制。Feigenbaum (1977) 将这一时期的研究概括为知识工程，并强调知识在智能系统中的中心作用。

专家系统推动人工智能进入真实行业，也暴露了知识获取和维护的困难。专家经验中包含大量隐性知识，难以全部转化为规则；知识库扩大以后，新旧规则可能冲突；环境变化又要求持续更新。第二次低潮促使研究者重新思考人工构造知识的上限。

4. 统计机器学习：让系统从数据中形成规律

20 世纪 90 年代以后，概率统计和机器学习逐渐成为主流。机器学习把知识来源从人工编写的规则扩展到数据，通过训练估计模型参数，使系统能够处理噪声、不确定性和复杂模式。研究目标也趋于务实，语音识别、图像识别、信息检索和自动驾驶等具体任务推动方法发展。

这一转向具有重要认识论意义。知识不再只以可读规则的形式存在，也可以隐含在统计关系和模型参数中。模型能力由数据、表示、目标函数和训练算法共同决定。人工智能开始获得较强的经验归纳能力，同时面临数据偏差、可解释性和泛化等新问题。

5. 连接主义与深度学习：自动形成多层表示

神经网络通过大量简单单元的连接实现复杂函数。感知器展示了连接权重可以从样本中学习，也暴露了单层网络的表达限制 (Minsky and Seymour Papert, 1969)。反向传播推动多层网络训练 (Rumelhart, G. E. Hinton, and Williams, 1986)，数据积累、GPU 计算和网络结构改进最终促成深度学习突破。深度网络可以从原始数据中逐层形成表示，在图像、语音和语言任务中减少对人工特征设计的依赖 (LeCun, Yoshua Bengio, and G. Hinton, 2015)。

深度学习的兴起拓展了人工智能的能力边界，也改变了系统的可理解方式。模型内部表示往往难以直接解释，训练性能与真实部署之间可能存在差距，对抗样本、分布变化和资源消耗成为新的基础问题。

6. 大模型与智能体：从单项任务走向通用接口

Transformer 通过注意力机制改善长序列建模和并行训练 (Ashish Vaswani et al., 2017)。在大规模数据和计算资源支持下，预训练模型可以在多种任务间迁移，逐渐形成语言生成、多模态理解、推理、代码和工具使用等能力。基础模型研究用“同一个模型适配许多下游任务”概括这种变化 (Bommasani, Hudson, Adeli, et al., 2021)。

生成式人工智能使人与模型的交互从调用专门功能转向自然语言指令。智能体则在模型基础上加入目标、记忆、规划、工具调用和环境反馈，使系统能够连续完成任务。这个阶段显示人工智能正在从单项模型扩展为复杂系统，也进一步模糊了模型、软件、工具和行动主体之间的边界。

人工智能发展过程中积累的这些思想和方法相互继承，也长期并存。现代系统仍会使用规则、搜索、知识库、概率模型和神经网络。历史呈现的是研究重心和能力结构的变化，并非旧方法被完全淘汰。

9.3 实施建议

9.3.1 课程组织原则

以稳定问题组织快速变化的术语：课程应围绕智能、计算、表示、学习、行动、评价和责任等长期问题组织内容。新产品和模型可以作为案例，不能成为概念结构的中心。这样，学生在面对新术语时能够追问：它解决什么智能任务，采用什么表示和方法，依赖哪些资源，与已有技术相比发生了什么变化。

历史服务于概念理解：人工智能史应解释思想和方法为何变化。学习达特茅斯会议是为了理解学科共同体如何形成；学习两次低潮是为了理解愿景、技术条件和社会预期之间的关系；学习符号主义、机器学习和深度学习是为了理解知识来源和智能实现方式的转变。历史人物和年份为这条逻辑提供支撑，不宜成为孤立记忆点。

用对比案例建立概念边界：概念边界适合通过成组案例学习。例如，机械定时器、反馈控温设备、扫地机器人和自主学习机器人可以用于比较自动化程度、环境感知、学习与规划；关键词匹配程序和大语言模型可以用于比较规则、统计学习和行为表现；搜索引擎、推荐系统和生成模型可以用于区分检索、排序和生成。

对比应允许存在中间情况。现实系统常把自动控制、传统软件和人工智能模型组合在一起，简单的“是或不是人工智能”容易掩盖系统结构。更有价值的问题是：系统的哪些部分使用了人工智能，使用了哪种能力，能力如何形成。

把能力描述与成立条件同时呈现：“模型能完成什么”需要与“在什么条件下完成”同时讲解。人脸识别的性能依赖数据群体、光照、角度和攻击环境；语言模型的回答依赖训练信息、上下文、检索和提示；自动驾驶能力依赖道路范围、天气、传感器和接管机制。条件意识能够抑制把局部实验结果推广到所有环境的倾向。

区分事实、争议和未来设想：人工智能基础概念中包含已经建立的事实，也包含开放争论。人工智能形成于 20 世纪中叶、机器学习是其重要方法分支，属于相对稳定的知识；机器是否真正理解语言、意识是否可能由计算产生，仍存在理论争议；通用人工智能和超级智能的时间表属于高度不确定的预测。课程应对三类内容作出标记，帮助学

习者形成不同的证据印象。

把基础概念连接到后续内容：基础概念需要为人工智能方法、应用、交叉融合以及风险、伦理与责任提供接口。定义人工智能时应引出知识与学习两类方法；讨论学科特征时应引出典型应用和跨学科属性；讨论机器智能与人类智能时应引出责任归属和人机关系。概念教育由此成为整套内容体系的入口。

9.3.2 推荐内容及其层次

1. 共同主干

面向所有学习者的共同主干建议包括：

1. 智能行为的多维性以及局部能力与整体智能的区别；
2. 人工智能的工作性定义、研究目标和计算实现方式；
3. 人类智能作为参照系的意义及机器拟人化的局限；
4. 形式逻辑、数理逻辑、可计算性理论和通用计算机构成的起源主线；
5. 图灵的机器智能思想和达特茅斯会议的学科形成意义；
6. 符号主义、专家系统、统计机器学习、深度学习、大模型和智能体的主要方法转向；
7. 人工智能的科学属性，以及由此发展的工程实践和交叉方法；
8. 人工智能与自动化、机器人、算法、互联网、大数据、机器学习、深度学习和生成式人工智能的关系；
9. 专用人工智能、通用人工智能和超级智能之间的证据与概念差异；
10. 人工智能能力的条件性、可错性和社会影响。

2. 拓展内容

根据学习者基础和课程目标，可以进一步讨论：

1. 人类智能的生物基础、社会合作、语言和文化累积；
2. 逻辑有效性、前提真实性与生成内容核查的关系；
3. 可计算性、复杂性与人工智能能力边界；
4. 图灵测试、中文房间等关于理解和心智的思想实验；
5. 符号主义、连接主义、贝叶斯主义和进化方法等学术流派；
6. 行为主义与内在主义在研究目标和证据标准上的差异；
7. 基础模型、多模态系统和智能体带来的学科边界变化；
8. 通用人工智能的不同定义、评价方案和治理问题。

拓展内容应保持概念导向。学习者不必记忆所有人物和术语，重点是形成对不同观点和方法的比较能力。

9.4 本章小结

人工智能基础概念的核心任务是建立理解智能机器的坐标系。人工智能可以被理解为“用计算机模拟人类智能行为的科学”。它以人类智能为重要参照，关注感知、行动、

思考、情感等功能，同时采用与人类不同的计算过程实现这些能力。

人工智能的思想和技术源头可以分为三个阶段来理解：制造智能机器的长期愿望提出问题，形式逻辑和数理逻辑使部分思维过程能够被符号化和演算，可计算性理论、数字电路和通用计算机使这些演算能够由机器执行。人工智能形成独立学科后，又经历符号主义、专家系统、统计机器学习、深度学习、大模型和智能体等方法重心的变化。各类方法长期并存，现代系统经常把多种工具组合使用。

人工智能领域包含大量概念，这些概念间可能存在包含、交叉和支撑关系。通过稳定问题、历史主线、对比案例、条件描述和证据分级，可以把概念辨析转化为对归纳、泛化、判断和自醒能力的训练。概念学习由此成为训练认知自治的基础动作。

第十章　人工智能基础方法：从知识表示到智能体系统

摘要：人工智能方法内容繁多，若课程按照算法名称和产品形态逐项展开，学习者容易获得零散术语，却难以理解各种方法为什么产生、如何关联以及在什么条件下有效。本章从通识教育的角度建立人工智能基础方法的知识体系。人工智能形成能力有两条基本路径：基于知识的方法把事实、规则和关系表示出来，通过搜索与推理解决问题；基于学习的方法从数据、反馈和经验中调整模型，使系统获得完成任务的能力。基于学习的方法统称为“机器学习”。机器学习可以用目标、模型、算法、数据和知识五个要素来理解，并通过问题定义、模型设计、模型训练、独立测试、模型选择、部署监测和反馈更新形成完整流程。监督学习、无监督学习和强化学习按照学习信号分类；符号、贝叶斯、连接和进化仿生四种方法论传统则从不同角度呈现人工智能的基础思路。从近年来的技术发展节奏来看，深度学习以多层神经网络实现表示学习；大模型通过大规模预训练建立可迁移能力，多模态模型把文字、图像、声音和动作纳入联合表示，智能体又加入目标、记忆、规划、工具、行动与反馈。基础方法教育帮助学习者形成方法坐标，理解模型能力的来源，判断数据、目标、结构、评价和场景之间的匹配关系，并通过归纳、泛化、判断和自醒保持认知自治。

关键词：人工智能方法；知识驱动；机器学习；监督学习；贝叶斯方法；连接主义；进化计算；深度学习；大模型；多模态模型；智能体；认知自治

10.1　设计思路

10.1.1　人工智能方法的教育功能

人工智能方法横跨逻辑、搜索、概率、优化、神经网络、语言模型和智能体。学生接触这些内容时，容易形成两种片面认识。一种把人工智能方法理解为当前最流行的深度学习或大模型，忽略知识表示、搜索、推理和概率方法长期发挥的作用；另一种把方法学习理解为掌握若干算法公式，难以把算法与数据、目标、评价和真实场景联系起来。

通识教育需要回答的是方法层面的基本问题：

1. 机器的能力来自哪里；
2. 人类知识如何进入人工智能系统；
3. 机器如何从数据和反馈中总结规律；
4. 一个模型如何被设计、训练、测试和选择；

5. 不同模型结构为什么适合不同类型的数据；
6. 大模型、多模态模型和智能体在已有方法上增加了什么；
7. 如何判断某种方法是否适合一个真实问题。

这些问题比具体算法名称更加稳定。卷积网络、Transformer 和未来的新结构都会发生变化，目标、表示、数据、模型、学习、评价仍然是人工智能系统必须处理的基本关系。

因此，人工智能基础方法可以组织为三层结构。

第一层是两条基本路径：基于知识的人工智能和基于学习的人工智能。前者把知识明确表示出来，让机器按照规则进行搜索与推理；后者给出目标、数据或反馈，让机器通过学习形成模型。

第二层是机器学习的一般体系，包括目标、模型、算法、数据和知识五个要素，以及问题定义、模型设计、训练、测试、选择、部署和反馈的完整流程。监督学习、无监督学习和强化学习按照学习信号区分机器怎样获得反馈。与此形成交叉的是，符号、贝叶斯、连接和进化仿生四种方法论传统从知识表示、模型假设和能力形成机制等角度呈现人工智能方法。

第三层是现代人工智能的主流扩展，包括深度学习、大模型、多模态模型和智能体。它们增加了模型的表示能力、迁移范围、交互方式和连续行动能力，但仍然受到前两层基本关系的约束。

这一体系的价值在于帮助学习者回答“为什么”。为什么某个系统需要知识库，为什么图像适合使用卷积结构，为什么训练性能不能代表真实能力，为什么大模型需要预训练，为什么智能体还要增加工具和反馈。理解这些关系，才能把不断出现的新方法放入稳定结构中判断。

10.1.2　人工智能方法与认知自治

归纳能力体现在理解机器如何从有限数据中形成规律。学习者能够区分真实模式、噪声和偶然特例，理解数据和模型共同决定归纳结果。

泛化能力体现在判断模型是否能处理未见样本和新环境。训练性能、测试性能、分布变化和模型假设构成泛化判断的基本依据。

判断能力体现在比较目标、数据、模型、指标、资源和风险。学习者能够说明为什么某种方法适合一个问题，也能够指出评价指标没有覆盖的后果。

自醒能力体现在觉察自己对模型的信任如何受到流畅表达、性能数字和技术规模影响。理解模型通过概率、优化和反馈形成能力，可以帮助人避免把输出自然度误认为可靠性，把参数规模误认为完整智能。

10.2 内容建议

10.2.1 人工智能基础路径

基于知识的方法：把人类认识转化为机器可用的结构

人类解决问题离不开知识。数学公理、物理定律、医学诊断经验、法律规则和生活常识都可以成为判断与行动的依据。基于知识的人工智能试图把这些知识表示为机器能够存储和处理的形式，再通过搜索和推理得到结论。

知识驱动系统通常包括三个环节：

知识表示 $\longrightarrow$ 搜索或推理 $\longrightarrow$ 结论与解释

知识表示回答“机器知道什么”。知识可以表示为逻辑公式、产生式规则、语义网络、本体和知识图谱。搜索与推理回答“机器如何使用知识”。系统可以从事实出发逐步推出结论，也可以从目标反向寻找所需条件，还可以在多条候选路径中利用启发信息优先探索更有希望的方向。

早期自动定理证明和人机对弈展示了通用知识的力量。“逻辑理论家”自动定理证明程序通过公理和推理规则证明数学定理，极小极大搜索与 α-β 剪枝利用游戏规则和局面评价选择行动。这些任务具有明确规则、封闭状态空间和清楚目标，适合知识表示与搜索方法 (Newell and Herbert A. Simon, 1976)。

专家系统进一步把领域经验组织成知识库。产生式规则常采用“如果条件成立，那么采取某种判断或行动”的形式，推理引擎按照规则处理输入事实。知识库与推理引擎的分离，使系统能够保留解释路径，也便于专家检查和修订 (Feigenbaum, 1977)。知识图谱则用实体和关系组织知识，使系统能够进行关联查询、关系推断和语义检索。

知识方法具有三项突出优势。其一，知识来源明确，推理过程相对可检查。其二，少量规则可以在数据稀缺时发挥作用。其三，领域约束可以阻止系统产生明显不合规的结果。因此，在法律、医疗、安全控制、科学计算等高风险或数据有限的场景中，知识方法仍具有重要地位。

知识方法也面临知识获取瓶颈。专家能够熟练完成任务，却未必能把经验完整表述为规则；现实环境存在大量例外、不确定性和隐性条件；知识库扩大以后，规则之间可能冲突，维护成本不断提高。通用公理虽然具有普适性，从基础规律推导复杂现实又可能遭遇组合爆炸和计算复杂度。知识驱动系统的能力由人类已经明确表示的知识所限定。

基于学习的方法：从经验中形成模型

机器学习改变了知识进入系统的方式。Samuel (1959) 通过跳棋程序展示了计算机利用既往对局和自我对弈改进表现的可能性；广为流传的“无需显式编程”说法可以帮助入门理解，但不宜当作该论文中的严格定义。Mitchell (1997) 给出了更形式化的定义：若程序在任务 T 上的性能指标 P 随经验 E 改善，就可以说它从经验中学习。

机器学习不要求设计者写出每一个判断规则。设计者规定学习目标、选择模型结构、准备数据与知识，并给出调整模型的算法。系统通过训练，把分散在数据中的模式积累到模型参数或结构中。

学习方法扩展了人工智能的知识上限。机器可以处理远超个人经验规模的数据，在高维空间发现难以用语言明确表达的关系，并形成与人类经验不同的表示。图像识别、语音识别、推荐和科学数据分析中的许多规律，很难由专家逐条写成规则，更适合通过数据学习。

然而，学习方法并没有消除人的设计。目标由谁确定、数据如何收集、模型如何选择、错误怎样计量、系统部署到哪里，都包含知识和价值判断。机器学习把部分规则发现工作交给模型，同时把人的责任转移到问题定义、数据治理、评价和系统边界上。

两条路径相互补充

知识驱动与学习驱动可以看作人工智能获得能力的两种基本来源。

表 10.1: 知识驱动与学习驱动的比较

比较维度	基于知识的方法	基于学习的方法
知识来源	人明确给出的事实、规则、关系和领域理论	数据、示范、奖励、环境交互以及先验知识
能力形成	搜索、匹配、逻辑或概率推理	调整模型结构或参数，使性能随经验改善
主要优势	可解释、可控制、适合规则清晰或数据稀缺问题	能处理复杂模式和大规模数据，具有发现新规律的可能
主要局限	知识获取困难，规则维护成本高，开放环境难以穷尽	依赖数据和目标，可能产生偏差、过拟合和可解释性不足问题
适合场景	定理证明、规则审查、安全约束、专业知识问答	识别、预测、生成、推荐、复杂感知与连续决策

现代系统经常结合两条路径。检索增强生成用知识库补充大模型，神经符号系统把神经网络的感知能力与符号推理结合，科学机器学习把物理规律加入模型结构或损失函数，智能体通过工具调用访问数据库、计算器和专业软件。AI 方法体系的发展越来越表现为知识与学习的重新融合。

10.2.2 机器学习基本方法

机器学习的五个基本要素

机器学习可以用目标、模型、算法、数据和知识五个要素来理解。五个要素不是彼此独立的部件，它们共同定义系统能够学到什么、如何学习以及学到的能力是否可靠。

1. 目标：系统究竟要优化什么

目标规定学习方向。分类系统希望减少错误类别，回归系统希望降低预测误差，生成系统希望产生更符合数据规律和人类要求的内容，强化学习系统希望提高长期累计回报。

为了让机器能够优化，目标通常被转化为数学形式，称为损失函数或目标函数。训练过程寻找能够降低损失的模型参数。目标函数把现实目的压缩成可计算指标，也可能丢失重要价值。例如，用点击率衡量推荐效果容易鼓励吸引注意的内容，用医疗花费代表健康需求可能继承社会不平等。方法教育需要帮助学习者理解，优化目标与真实目标之间可能存在距离。

目标还具有层次。模型训练的局部目标、系统运行的业务目标和社会使用的公共目标可能并不一致。一个模型可以准确预测用户会点击什么，平台却仍需判断这种推荐是否促进健康的信息环境。技术优化不能替代目标本身的审查。

2. 模型：用什么结构保存和使用规律

模型是机器学习的主体结构。它规定输入如何被表示、信息如何被处理、输出如何产生，以及哪些部分可以通过学习调整。线性模型用少量参数表示变量关系，决策树用分支规则形成判断，贝叶斯网络用概率依赖表示变量关系，神经网络用大量连接权重形成复杂函数。

模型结构包含对问题的假设，也称为归纳偏置。线性模型假设变量关系可以近似为线性，卷积网络假设局部模式可以在不同位置复用，循环网络假设当前状态与过去信息相关。模型并非越复杂越好，模型要与数据结构、任务规模、计算条件和错误代价相匹配。

“没有免费的午餐”定理指出，在其关于问题集合和平均方式的特定假设下，若对所有可能问题作均匀平均，没有一种优化算法能够普遍优于其他算法 (Wolpert and Macready, 1997)。这一结果不能被理解为“任何算法在现实中都一样好”；它的教育意义在于，脱离任务分布与先验假设讨论“最好的模型”没有意义。模型选择必须利用领域知识，明确自己相信数据中存在什么结构。

3. 算法：如何调整模型

算法规定模型如何从经验中更新。对于神经网络，学习算法通常计算损失函数对参数的梯度，再沿降低损失的方向调整权重；对于聚类，算法在样本分组和类中心之间反复迭代；对于强化学习，算法利用奖励估计行动价值并更新策略。

梯度下降是现代机器学习最常见的优化思想。设模型参数为 θ，损失函数为 $L(\theta)$，一次基本更新可以写为

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t),$$

其中，η 是学习率，$\nabla_\theta L$ 表示损失对参数变化的方向和幅度。学习率过大可能在较优区域之间来回震荡，过小又会使训练缓慢。算法设计体现速度、稳定性、精度和资源之间的平衡。

学习算法不会保证找到绝对最优解。复杂模型的损失曲面可能包含鞍点、平坦区域和多个局部结构。训练结果还受到初始化、数据顺序、正则化和计算精度影响。机器学习具有实验科学属性，需要通过重复实验和对照分析评估算法。

4. 数据：经验、证据与边界

数据是模型学习的经验材料。图像、声音、文本、传感器记录、实验测量和人类行为都可以成为数据。数据质量至少包括准确性、数量、覆盖度、代表性、时效性和合法性。

错误标注会把错误答案教给模型；数据过少会使模型难以区分稳定规律与偶然现象；场景覆盖不足会导致模型在真实环境中失效；群体分布不均会造成不同人群之间的性能差异；过时数据无法代表当前环境；未经授权收集的数据还会引发隐私和法律问题。

数据也不是自然存在的纯粹事实。记录什么、不记录什么，如何测量、如何分类、由谁标注，都包含选择。学生需要把数据理解为经过采集、表示和治理的证据，而非自动等同于现实本身。

5. 知识：为学习提供先验和约束

机器学习仍然需要知识。领域知识帮助设计特征和模型，确定合理参数范围，选择评价指标，解释异常结果，并约束系统不得违反基本规律。

知识可以通过多种方式进入学习系统：

1. 选择输入变量和数据表示；
2. 设计具有特定结构的模型；
3. 在损失函数中加入约束；
4. 用知识库或检索系统补充信息；
5. 在输出之后进行规则检查；
6. 由专家对结果进行反馈和修正。

当数据稀缺、成本高或风险大时，知识尤其重要。科学研究中的守恒定律、医学中的临床规范、工程中的安全边界，都可以减少模型需要探索的空间，并提高结果的可信度。

机器学习系统的基础流程

模型训练只是机器学习系统形成过程中的一个环节。一个可以真实使用的系统还需要经历问题定义、设计、测试、选择、部署和持续更新。

1. 问题定义

问题定义要说明输入是什么、输出是什么、系统服务谁、成功如何衡量、错误造成什么后果。模糊的问题会产生看似精确却没有意义的模型。

2. 数据准备

数据准备包括采集、清洗、标注、划分和记录来源。训练集用于调整模型，验证集

可以用于选择参数和结构，测试集用于最后评估。三者混用会造成信息泄漏，使测试结果虚高。

3. 模型设计

模型设计需要利用数据和任务的结构。图像具有局部和空间关系，语言具有顺序和上下文，知识图谱具有实体关系，科学数据可能满足对称性和守恒律。模型结构越能反映这些属性，学习效率通常越高。

设计还要考虑计算资源、响应速度、可解释性、隐私、安全和维护成本。准确率最高的模型未必是最适合部署的模型。移动设备可能需要较小模型，高风险系统可能更重视解释和稳定性。

4. 模型训练

训练把学习目标转化为损失函数，通过优化算法调整参数。训练过程中需要监测损失、性能、梯度和资源消耗，并使用正则化、数据增强、早停等方法减少过拟合。

训练性能只能说明模型对已见经验的适应程度。一个学生背熟练习题，不代表他能解决新问题；模型记住训练样本，也不代表形成了可以迁移的规律。

5. 模型测试与泛化

模型的核心价值体现在新样本和新环境中的表现，这种能力称为泛化。独立测试集用于估计模型面对未见数据时的性能。训练表现很好而测试表现明显下降，通常说明模型发生过拟合；训练和测试都表现较差，可能是模型能力不足、数据质量差或任务定义有问题。

测试还要覆盖真实环境中的变化。光照、噪声、设备、地区、人群和时间都可能改变数据分布。模型在基准数据集上的性能不能直接代表真实世界的可靠性。

6. 模型选择

模型选择需要在拟合能力和复杂度之间取得平衡。模型过于简单会欠拟合，无法捕捉基本规律；模型过于复杂可能学习噪声和偶然特例。奥卡姆剃刀准则建议，当多个模型性能接近时，优先选择较简单的模型。简单模型通常更容易解释、测试和维护。

模型选择还涉及多目标权衡。医疗筛查可能更重视漏诊率，垃圾邮件过滤需要控制误删，自动驾驶必须考虑极端场景，推荐系统还需评估多样性和长期影响。单一平均准确率难以覆盖全部要求。

7. 部署、监测与更新

模型进入真实环境后，用户行为和环境会发生变化，模型也可能反过来改变数据。推荐系统的输出影响用户点击，点击数据又被用于后续训练，形成反馈循环。部署阶段需要监测性能衰减、异常输入、安全攻击、公平差异和资源消耗。

模型更新应保留版本、数据和评价记录。高风险系统还需要人工复核、申诉和退出机制。人工智能系统由模型、软件、数据管线、用户和组织制度共同构成，可靠性不能只由模型指标保证。

机器学习的三种基本学习方式

1. 监督学习：从带答案的示范中学习

监督学习使用带标签的数据。每个输入样本配有期望输出，模型通过比较预测与标签调整参数。分类预测离散类别，回归预测连续数值。

监督学习适合目标明确、能够获得可靠标注的任务，如识别病害叶片、预测房价和判断邮件类别。它的关键限制来自标注成本与标注质量。标签可能包含专家分歧、历史偏见和测量误差，模型会把这些问题一并学习。

2. 无监督学习与自监督学习：从数据内部发现结构

无监督学习没有外部给出的答案，系统从数据内部寻找相似性、低维结构和潜在表示。聚类把相似样本分组，降维把高维数据压缩到更简洁的表示，自编码器通过重建输入学习重要特征。

自监督学习从原始数据本身构造学习任务。例如，语言模型根据前文预测后续词元，图像模型可以预测被遮挡区域。自监督学习利用海量未标注数据学习通用表示，为大模型预训练提供了基础 (Devlin et al., 2019; T. B. Brown et al., 2020)。

无监督和自监督学习能够减少人工标注依赖，也带来解释困难。模型发现的结构未必与人的概念一致，数据中的高频模式也未必具有真实意义。学习结果需要结合任务和领域知识验证。

3. 强化学习：在行动与反馈中学习

强化学习研究智能体如何通过与环境交互学习策略。智能体观察状态，选择动作，环境返回新状态和奖励。学习目标通常是最大化长期累计回报 (Sutton and Barto, 2018)。

强化学习适合多步决策，如机器人控制、资源调度和对弈。奖励设计决定系统学习方向。奖励过于局部，智能体可能追求眼前收益；指标存在漏洞时，智能体可能找到设计者未预期的取巧方式。强化学习特别清楚地展示了目标设定、长期后果和价值对齐之间的关系。

表 10.2: 三种基本学习方式

学习方式	学习信号	典型任务	主要风险
监督学习	人或系统提供的标签与目标值	分类、回归、检测、预测	标注错误、标注偏差、标签成本
无监督与自监督学习	数据内部结构或由数据自动构造的目标	聚类、降维、表示学习、预训练	学到的结构难解释，模式可能缺少语义
强化学习	环境中的奖励、惩罚和长期回报	对弈、控制、规划、连续决策	奖励设计失真、短期取巧、探索风险

三种方式可以组合。AlphaGo 使用人类棋谱进行监督学习，再通过自我对弈开展强化学习；大语言模型先进行自监督预训练，再利用人工偏好或规则进行后训练。学习方式的边界正在变得灵活，核心仍是系统获得了什么反馈。

机器学习的四种方法论传统

监督学习、无监督学习和强化学习主要按照**学习信号**进行分类：系统获得的是标签、数据内部结构，还是环境奖励。人工智能方法还可以从知识表示、不确定性处理和能力形成机制等角度理解，并概括为符号、贝叶斯、连接和进化仿生四种方法论传统。符号传统主要属于知识驱动路线，其余传统更多通过数据、概率或搜索形成模型，因此不能把四种传统整体视为机器学习的下位分类。

这一划分主要用于建立整体视野，并非一套边界固定的学术分类。很多现代系统同时吸收多种传统的方法。例如，智能体可以用神经网络理解语言，用贝叶斯方法估计风险，用符号规则限制行动，再用进化搜索或强化学习改进策略。方法论传统的价值在于揭示不同方法从何处寻找智能，以及它们各自采用怎样的知识形式和学习机制。

1. 符号传统：通过符号表示和规则操作形成智能

符号传统认为，智能活动可以表示为符号结构及其操作过程。事实、概念、关系和规则被明确写入系统，机器通过匹配、搜索和推理得到结论。逻辑推理、决策树、规则学习、定理证明、专家系统和知识图谱都体现了这一传统 (Newell and Herbert A. Simon, 1976)。

符号方法的优势在于结构清晰，知识来源和推理路径相对容易检查。在规则明确、数据稀缺或需要严格约束的任务中，它可以直接利用人类已经掌握的知识。它的困难在于，开放世界中的知识很难完整列出，经验中的隐性规则也不容易全部转化为符号。规则数量增加后，还可能出现冲突、维护困难和搜索空间膨胀。

符号传统提醒学习者，机器智能可以来自明确知识和推理结构。它也揭示了一个基础问题：人类能够表达出来的知识，与人类实际掌握并使用的知识之间存在差距。

2. 贝叶斯传统：在不确定性中更新信念

贝叶斯传统把学习理解为在证据到来后更新对假设的信念。系统以概率表示不确定性，并根据贝叶斯规则把先验知识与观察数据结合起来。朴素贝叶斯分类器、贝叶斯网络、概率图模型和隐马尔可夫模型都属于这一方法传统 (Pearl, 1988)。

设假设为 H，观察到的证据为 D，贝叶斯公式可以写为

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}.$$

其中，$P(H)$ 表示观察证据前对假设的先验判断，$P(D \mid H)$ 表示假设成立时出现该证据的可能性，$P(H \mid D)$ 表示结合证据后的后验判断。学习过程由此表现为信念随证据不断修正。

贝叶斯方法能够显式表达不确定性，适合诊断、预测、风险分析和数据不足的场景。它要求设计者说明变量之间的概率关系或选择合适的先验。当变量很多、依赖关系复杂时，概率推断的计算代价可能很高；不合理的先验和模型假设也会影响结果。

贝叶斯传统的教育价值在于说明，智能判断通常是在有限证据下形成可修正的置信程度。系统应当随着新证据调整判断，人也需要区分“可能性较高”与“已经确定”。

3. 连接传统：通过连接权重学习分布式表示

连接传统受到神经系统启发，认为复杂智能可以由大量简单单元的连接和共同作

用产生。知识不一定以可读规则保存，也可以分布在网络连接权重和激活模式中。感知器、多层神经网络、卷积网络、循环网络和 Transformer 都延续了这一传统 (Rumelhart, G. E. Hinton, and Williams, 1986; LeCun, Yoshua Bengio, and G. Hinton, 2015)。

连接主义模型通过数据和误差信号调整权重，能够自动学习高维数据中的表示。它在视觉、听觉、语言和生成任务中取得突出进展，也是当前深度学习和大模型的主要方法基础。

连接传统的优势是表达能力强，能够处理难以由人工规则描述的复杂模式。但它的内部知识往往分散在大量参数中，难以直接解释；模型还依赖大规模数据与计算，并可能学习数据中的偏差和偶然相关。

连接传统表明，知识可以以分布式方式存在，模型能够从经验中形成表征。它也把泛化、可解释性和数据质量推到人工智能方法的中心位置。

4. 进化仿生传统：通过变异、选择和遗传搜索解空间

进化仿生传统借鉴生物进化过程，把候选解看作一个种群。系统不断生成变异，通过适应度评价保留表现较好的个体，再将其特征遗传到下一代。遗传算法、进化策略、遗传编程和部分神经结构搜索体现了这一思路 (Holland, 1975)。

进化方法不要求目标函数可微，也不需要明确知道应该沿什么方向调整参数。只要能够评价候选方案的好坏，系统就可以通过选择逐步搜索。这使它适合结构设计、组合优化、多目标问题和不连续搜索空间。

它的局限在于搜索过程可能需要大量评价，计算成本较高；适应度函数若不能准确代表真实目标，进化过程也可能找到偏离设计意图的方案。进化算法得到的结构有时有效，却不容易解释为何有效。

进化仿生传统揭示了另一种能力形成方式：设计者不直接规定最终结构，而是设计变异、选择和遗传机制，让解决方案在反复竞争中出现。

5. 四种传统的关系与融合

四种传统分别从不同角度回答“机器怎样获得智能”：

表 10.3: 人工智能四种方法论传统的比较

方法论传统	核心表示	能力形成机制	典型优势与局限
符号传统	符号、规则、逻辑关系和知识结构	搜索、匹配、归纳规则和逻辑推理	结构清晰、容易约束；知识获取和开放环境适应困难
贝叶斯传统	概率分布、随机变量和依赖关系	结合先验与证据更新后验概率	能表达不确定性；依赖概率假设，复杂推断成本较高
连接传统	神经元、连接权重和分布式表示	利用数据与误差调整网络参数	表达能力强；数据和计算需求高，内部机制较难解释

方法论传统	核心表示	能力形成机制	典型优势与局限
进化仿生传统	个体、种群、编码结构和适应度	变异、选择、重组和代际搜索	不依赖梯度、适合结构搜索；评价成本高，结果可能难解释

这四种传统没有形成彼此排斥的边界。符号规则可以约束神经网络输出，贝叶斯方法可以表示神经模型的不确定性，进化算法可以搜索网络结构，连接主义模型也可以从数据中生成符号候选。当前深度学习的突出地位体现了连接传统的发展，现代人工智能的可靠性和可迁移性则越来越依赖多种方法的融合。

理解这些传统有助于避免把人工智能方法简化为神经网络。面对一个真实问题，学习者需要进一步追问：知识能否明确表示，环境中的不确定性有多大，是否拥有足够数据，目标函数能否求梯度，搜索空间是否适合进化。方法选择由问题结构决定。

10.2.3 现代人工智能的主流扩展

深度学习：从特征设计到表示学习

深度学习沿着连接主义路线，通过多层神经网络和大规模数据学习复杂表示。理解其基本思想时，仍需把它放回四种传统并存与融合的整体结构中。

1. 人工神经网络的基本思想

人工神经网络受到生物神经系统启发，用大量简单计算单元和可调整连接构成模型。McCulloch and Pitts (1943) 以加权输入和阈值输出建立早期神经元模型。F. Rosenblatt (1958) 的感知器引入可学习权重，使网络能够根据样本调整分类边界。

单层感知器只能处理线性可分问题。加入隐藏层和非线性激活后，多层网络能够表达更复杂关系。反向传播算法高效计算各层参数对损失的影响，使多层神经网络能够通过梯度下降训练 (Rumelhart, G. E. Hinton, and Williams, 1986)。

深度学习通常指利用多层神经网络进行表示和任务学习。它的关键变化在于，模型能够从原始数据逐层形成特征。图像网络的低层可能响应边缘和纹理，高层组合出物体部件和类别；语言模型从字符和词元逐步形成语法、语义和上下文表示 (LeCun, Yoshua Bengio, and G. Hinton, 2015)。

2. 结构设计体现对数据的认识

神经网络结构不是任意堆叠。每种结构都把对数据特性的认识写入模型。

卷积神经网络利用局部连接和权重共享，适合具有局部模式和空间重复性的图像数据。卷积核在不同位置检测相同特征，多层结构把边缘、纹理和局部形状组合成高级表示。AlexNet 在 ImageNet 竞赛中的突破展示了深层卷积网络和大规模数据、GPU 计算结合的力量 (Krizhevsky, Sutskever, and G. E. Hinton, 2012)。

循环神经网络通过隐藏状态积累过去信息，适合语言、语音和时间序列。其局限在于长距离依赖和串行计算较难处理。Transformer 以自注意力直接建立序列不同位置之间的关系，并提高并行训练效率 (Ashish Vaswani et al., 2017)。

自编码器通过压缩并重建输入学习潜在表示。生成对抗网络通过生成器和判别器的竞争学习数据分布 (I. Goodfellow, Pouget-Abadie, Mirza, et al., 2014)。不同结构对应局部性、时序性、压缩性、竞争性等不同假设。

这种关系体现了一条重要方法原则：模型结构应利用数据和任务的内在属性。理论上表达能力很强的通用网络，在有限数据和资源下未必优于带有恰当归纳偏置的特殊结构。

3. 深度学习成功的条件

深度学习的兴起由多种条件共同推动：

1. 大规模数据为模型提供学习资料；
2. GPU 和并行计算加快训练速度；
3. 反向传播、初始化、激活函数、正则化和残差连接等方法改善深层网络训练；
4. 基准数据集和开放工具推动结果比较与复现。

深度学习的能力不能只归因于“网络更深”。数据、算力、算法、评价共同形成技术系统。深度模型也面临可解释性不足、对抗脆弱性、分布外失效、数据与能源成本等问题。通识教育需要同时讲清能力来源和成立条件。

大模型：预训练、迁移与自然语言接口

1. 从专用模型到基础模型

传统机器学习常为一个任务训练一个模型。大模型通过大规模自监督预训练，从广泛数据中学习通用表示，再通过提示、少量示例、微调或工具使用适配多种任务。基础模型概念强调，一个预训练模型可以成为许多下游系统的共同基础 (Bommasani, Hudson, Adeli, et al., 2021)。

语言模型的基本任务是估计语言序列的可能性。自回归模型根据已有词元预测下一个词元，掩码语言模型根据上下文恢复被遮挡内容。这个看似简单的目标迫使模型学习大量语言关系、知识模式和任务结构。

模型规模、数据规模和计算量增加后，性能通常呈现可预测的扩展趋势，即“尺度定律”(Kaplan, McCandlish, Henighan, et al., 2020)。但扩展并非无限堆叠参数，训练数据与计算资源需要保持合理平衡 (Hoffmann, Borgeaud, Mensch, et al., 2022)。

预训练模型还需要通过指令数据、人工偏好和安全规则进行“后训练”，使其更好地理解人类请求并降低有害输出。后训练改善交互，却不能保证事实正确、价值无争议或所有场景安全。自然语言接口降低了使用门槛，也容易让使用者忽略背后的模型条件。

2. 上下文、推理与外部证据

大模型可以根据提示中的任务描述和示例改变输出方式，这种现象称为上下文学习。上下文学习可以使模型实现零样本和少样本迁移，即不改变模型参数，只需更改提示词（零样本）或给出少量任务示例（少样本），就可以改变模型的行为。GPT-3 展示了这种任务迁移能力 (T. B. Brown et al., 2020)。

大语言模型可以在提示中生成分步推理过程；链式思维提示在部分算术、常识和符号推理基准上改善了大模型的任务表现 (Wei, X. Wang, Schuurmans, et al., 2022)。这

一结果证明的是特定提示方式下的基准表现提升，不宜仅据此断言模型已经获得与人相同的推理能力。

检索增强生成（retrieval-augmented generation, RAG）是另一项重要技术。大模型按条件概率生成序列，语言连贯并不自动保证事实可靠。为了给生成过程引入外部知识，可以先检索文档，再把检索结果作为生成条件；原始 RAG 研究将参数化生成模型与非参数化检索记忆结合，并在若干知识密集型任务上取得改进 (Lewis, Perez, Piktus, et al., 2020)。检索增强能够提供可核查材料，但检索错误、证据误读和无依据生成仍需另外评估。

多模态模型：连接不同形式的世界信息

人类同时通过语言、视觉、听觉和动作认识世界。多模态模型尝试把不同数据形式映射到可以相互对应的表示空间，使模型能够进行图文检索、图像问答、声音理解、视频分析和跨模态生成。

CLIP 通过大量图文配对数据学习视觉和语言表示的对齐，使文字能够直接用于识别和检索图像 (Radford, Kim, Hallacy, et al., 2021)。Flamingo 等模型进一步把预训练视觉模型与语言模型连接起来，实现多模态上下文学习 (Alayrac, Donahue, Luc, et al., 2022)。

多模态学习需要把不同输入纳入同一系统，并进一步解决三类问题：

1. 表示：不同模态如何转换为模型可处理的单元；
2. 对齐：图像区域、声音片段、文字描述和动作之间如何建立对应；
3. 融合：不同信息发生冲突时，系统如何分配权重并作出判断。

多模态数据能相互补充，也会产生新的错误。图像可能诱导语言模型忽略文字证据，文本标签可能把错误传递给视觉模型，视频中的时间关系又比静态图像更加复杂。多模态能力需要根据跨模态一致性、真实环境下的任务表现来进行评价。

智能体：从问答到行动

基础模型本身主要根据输入产生预测或生成结果，只有接入工具、权限和执行环境后，输出才会转化为连续行动。智能体面向目标，在环境中持续感知、规划、行动和调整。一个典型智能体可以表示为：

$$\textbf{模型} + \textbf{目标} + \textbf{记忆} + \textbf{规划} + \textbf{工具} + \textbf{行动} + \textbf{反馈}$$

模型提供语言、推理和模式能力；目标规定任务方向；记忆保存任务状态和历史；规划把复杂目标分解为步骤；工具连接搜索、计算、代码、数据库和专业软件；行动改变数字或物理环境；反馈用于检查结果并修正计划。

ReAct 方法把推理与行动交替组织，使语言模型能够根据观察更新后续步骤 (Yao et al., 2023)。Toolformer 探索让模型学习何时调用外部工具 (Schick, Dwivedi-Yu, Dessì, et al., 2023)。这些方法说明，模型内部知识、外部工具和环境反馈可以共同完成任务。

智能体具有不同自主程度。简单工作流按照预先规定的步骤调用模型和工具；较复杂系统可以根据中间结果选择路径；更高自主系统能够制订计划、分配子任务和持续运行。

自主性提高会增加适应能力，也会扩大错误传播和权限风险。一个错误回答影响有限，一个能够发邮件、修改文件、控制设备或执行实验的智能体可能产生实际后果。因此，智能体设计必须同时考虑能力、权限、监测、停止条件和责任。

多智能体系统让不同智能体承担搜索、规划、执行、批评和验证等角色。角色分工可以增加观点多样性和交叉检查，也可能产生沟通成本、群体偏差和责任模糊。多智能体并不天然优于单一系统，其价值取决于任务是否需要分工、并行和相互批评。

10.3 实施建议

10.3.1 课程组织原则

人工智能方法内容量大，知识点密集，如果不做合理规划，很容易陷入思维混乱。一些原则可以帮助课程设计者避免这种混乱。

1. 建立主干清晰的知识图谱

方法教学应先给出整体结构，再进入代表性方法。知识驱动与学习驱动构成第一层；机器学习五要素、完整流程和三种学习方式构成第二层；深度学习、大模型、多模态模型和智能体构成第三层。每个新方法都应说明它在结构中的位置。

2. 用核心关系统摄数学细节

数学是理解机器学习的重要语言。通识教育可以使用损失函数、梯度下降、概率和向量等基础表达，重点解释数学量代表什么关系。公式数量应服务于概念，不宜把推导熟练度作为共同目标。

例如，梯度下降公式的教育价值在于说明模型通过局部反馈逐步调整，而非要求所有学习者掌握多元微积分证明。卷积公式用于说明局部连接和权重共享，注意力用于说明不同位置之间的动态关系。

3. 模型能力与方法假设同时呈现

介绍一种模型时，应回答它利用了什么数据特性。卷积网络利用局部性与空间重复，循环网络利用时序状态，Transformer 利用全局依赖，自编码器利用信息冗余。理解假设能够帮助学生判断方法能否迁移到新问题。

4. 以完整系统代替孤立算法

课程案例应包含目标、数据、模型、算法、评价和场景。单独演示一个分类器容易把机器学习理解为按钮操作；完整分析能够揭示数据偏差、指标选择和部署条件。即使实践规模很小，也应保留训练集与测试集、错误分析和结果反思。

5. 把缺陷作为方法学习的重要内容

过拟合、欠拟合、分布变化、奖励漏洞、幻觉和工具调用失败能够揭示方法边界。这是方法学习的重要内容，也是思维训练的重点。让学生比较训练与测试差异、分析错误样本、改变奖励规则和检查检索证据，可以培养真正的方法理解。

6. 从多角度思考问题解决方法

符号传统、贝叶斯传统、连接传统和进化仿生传统从不同角度理解知识、学习与智能。当前深度学习占据重要位置，也不能据此把其他方法排除在知识体系之外。高风险、低数据、强不确定性、结构搜索和严格推理问题，往往需要规则、概率、神经网络与进化方法的不同组合。课程应帮助学生理解各传统的历史来源、核心假设、优势、局限和适用场景。

10.3.2 推荐内容及其层次

1. 共同主干

人工智能基础方法的共同主干建议包括：

1. 基于知识与基于学习两条基本路径；
2. 逻辑规则、搜索、专家系统和知识图谱的基本思想；
3. 机器学习的目标、模型、算法、数据和知识五个要素；
4. 模型设计、训练、测试、选择、部署和反馈的完整流程；
5. 训练集、测试集、过拟合、欠拟合、泛化和模型复杂度；
6. 监督学习、无监督与自监督学习、强化学习；
7. 分类、回归、聚类、降维、生成和连续决策等基本任务；
8. 符号传统、贝叶斯传统、连接传统和进化仿生传统的核心思想及其关系；
9. 人工神经网络、反向传播与层次表示学习；
10. 卷积网络、循环网络、自编码器和 Transformer 的结构思想；
11. 大模型的预训练、迁移、上下文学习、后训练和外部检索；
12. 多模态模型中的表示、对齐和融合；
13. 智能体的目标、记忆、规划、工具、行动与反馈；
14. 模型能力的条件性、评价边界和社会技术属性。

2. 拓展内容

根据学习者基础，可以进一步讨论：

1. 启发式搜索、极小极大算法与概率推理；
2. 梯度、优化器、正则化和学习率；
3. 支持向量机、决策树、贝叶斯网络与集成学习；
4. 遗传算法、进化策略、遗传编程与神经结构搜索；
5. 流形、表示空间和生成模型；
6. 残差网络、注意力机制和位置编码；
7. 尺度定律、计算最优训练和模型压缩；
8. 强化学习中的价值函数、策略和探索；
9. 神经符号学习、因果学习和科学机器学习；
10. 多智能体协作、智能体权限与安全控制。

拓展内容应围绕问题需要展开。方法名称的数量不能替代体系理解。

10.4　本章小结

人工智能基础方法需要被组织为一套关联清晰的知识体系。基于知识的方法把事实、规则和关系表示出来，通过搜索与推理形成结论；基于学习的方法从数据、示范和反馈中调整模型。两条路径各有优势和边界，并在现代系统中不断融合。

机器学习可以由目标、模型、算法、数据和知识五个要素解释。目标规定优化方向，模型保存和使用规律，算法调整模型，数据提供经验，知识提供结构与约束。一个可靠系统还要经历问题定义、数据准备、模型设计、训练、独立测试、模型选择、部署监测和反馈更新。训练结果与真实能力之间的差异，使过拟合、泛化、分布变化和评价成为机器学习的核心问题。

监督学习从带答案的示范中学习，适合分类和回归；无监督与自监督学习从数据内部发现结构，为表示学习和大模型预训练提供基础；强化学习通过行动和环境反馈学习长期策略。三种方式可以组合，学习信号决定模型能够获得什么能力。

机器学习还可以从四种方法论传统理解。符号传统强调符号、规则和推理，贝叶斯传统用概率表示不确定性并根据证据更新信念，连接传统通过连接权重学习分布式表示，进化仿生传统利用变异、选择和遗传搜索解决方案。深度学习、大模型、多模态模型和智能体主要沿连接传统发展，又不断吸收知识库、检索、符号约束、人工反馈和专业工具。

方法理解使学习者能够追问能力从哪里来、结论在什么条件下成立、评价遗漏了什么以及人应承担何种责任。它把对人工智能的信任建立在目标、数据、模型、证据和边界之上，是认知自治的技术基础。

第十一章　人工智能应用：从生活场景解析系统机制

摘要：人工智能应用是普通人接触智能技术的第一现场，也是连接基础概念、技术方法与社会判断的关键内容。但是，若课程停留在产品展示和工具操作，学习者将只能够感受到人工智能“会做什么”，却难以理解系统怎样形成能力、在什么条件下可靠，以及错误将由谁承担。本章以具有代表性的现实系统为对象，通过“任务与场景—数据与表示—模型与方法—系统集成—输出与评价—失效条件—风险与责任”的共同框架，把可见功能还原为可分析的技术系统。推荐内容由机器视觉、机器听觉、语言处理、人机对战与具身行动、搜索与推荐五类应用构成，覆盖表示、识别、生成、搜索、强化学习、具身行动、信息排序和反馈循环等关键机制。在教学层面，有关应用的学习应训练学习者从熟悉场景中抽取通用结构，并把它迁移到新的系统；同时追问数据来源、模型目标、证据强度、适用范围和责任安排，在使用人工智能时保留理解、核查和选择。

关键词：人工智能应用；机器视觉；机器听觉；语言处理；人机对战；搜索引擎；推荐系统；系统解析；认知自治

11.1　设计思路

11.1.1　人工智能应用的教育功能

基础概念帮助学习者知道人工智能是什么，基础方法解释人工智能如何形成能力。应用内容进一步回答：这些能力怎样进入真实系统，如何与传感器、数据库、软件流程、用户和组织制度连接，并对现实生活产生影响。

从通识教育角度，人工智能应用具有四项教育功能。

第一，**把抽象方法转化为可观察过程**。数据、表示、模型、训练、推理和反馈可以在人脸识别、机器翻译和推荐系统中被具体看到。学习者由此理解，技术术语对应真实系统中的实际环节。

第二，**揭示系统能力的条件性**。应用总是在特定设备、数据、环境和目标下运行。人脸识别受光照和角度影响，语音识别受噪声和口音影响，机器翻译受语境和领域影响，推荐系统受平台目标和行为数据影响。能力与条件同时呈现，才能形成准确判断。

第三，**培养迁移分析能力**。学习者掌握应用背后的共同结构后，可以分析尚未进入课程的新系统。迁移研究表明，知识能否用于新情境，取决于学习者是否理解深层结

构，并在多个场景中练习调用 (Bransford, A. L. Brown, and Cocking, 2000; Perkins and Salomon, 1988)。应用内容承担的正是从具体案例中抽取通用结构的任务。

第四，**把技术理解与公共责任连接起来**。身份识别涉及隐私与申诉，生成内容涉及真实性与版权，搜索和推荐涉及注意力与信息环境，机器人行动涉及安全与责任。应用是伦理问题具体发生的场所，也是认知自治接受现实检验的场所。

因此，应用内容的学习目标可以概括为：通过典型现实系统，帮助学习者解释人工智能如何完成任务、如何形成输出、为什么发生错误、如何影响个人与社会，以及人在系统中应当保留怎样的判断和责任。

11.1.2 应用教育与相邻内容的边界

1. 应用教育与工具教学

工具教学关注如何操作某个产品：怎样输入提示、选择功能、导出结果或完成作品。应用教育关注产品背后的问题结构：系统处理什么输入，依赖什么数据和模型，输出如何评价，使用者承担什么责任。

工具操作可以成为学习活动，但工具名称和界面不宜构成人工智能应用的学习主干。产品会快速更新，系统机制和判断问题更加稳定。学习图像生成时，重点应放在文本与图像如何对齐、模型为何会生成细节错误、作品如何标注和核查；具体平台只提供一次观察和实验的机会。

2. 应用教育与方法教育

方法教育从模型和算法出发，讨论知识表示、监督学习、深度网络、Transformer 和强化学习等结构。应用教育从现实任务出发，分析多种方法如何被选择、组合和部署。

例如，强化学习在方法章节中用于说明智能体如何利用奖励更新策略；在 AlphaGo 和电子游戏应用中，学习者进一步看到强化学习如何与搜索、神经网络、自我博弈和环境模拟结合。两部分相互联系，观察方向不同。

3. 应用教育与交叉融合

人工智能应用主要以人工智能系统为认识对象。人脸识别、语音识别和推荐系统的中心问题是：系统如何实现特定人工智能功能。

交叉融合则以其他学科的问题为起点。蛋白质结构预测、材料设计、数学证明和天文观测的中心问题是：学科问题如何被表示成机器可以处理的形式，人工智能方法如何接受领域知识和证据标准的约束。

两部分可以共享技术和案例，但课程任务需要区分。医学影像识别若用于解释图像分类系统，可以进入应用内容；若用于讨论医学问题怎样定义、临床证据如何验证、医生与模型如何分工，则进入交叉融合内容。

11.1.3 典型应用的选择原则

人工智能应用数量庞大，课程无法逐一覆盖。选择案例时应优先考虑其教育解释力，而非市场热度。推荐遵循以下原则。

生活可感知：应用应与学习者的经验存在联系。人脸解锁、语音输入、机器翻译、搜索、推荐和扫地机器人都具有直接体验入口。熟悉感可以降低进入门槛，也使学习者能够带着自己的观察和疑问进入技术分析。

机制具有代表性：一个案例应能够揭示可迁移的核心机制。人脸识别和说话人识别共同体现嵌入向量与相似度；语音识别和机器翻译共同体现序列建模；图像生成和机器写作共同体现生成模型；AlphaGo 和电子游戏共同体现搜索、奖励和策略学习；搜索与推荐共同体现排序、用户行为与反馈循环。

技术演进清晰：应用应能呈现方法为何变化。人脸识别经历几何特征、特征脸和深度嵌入；语音识别经历模式匹配、隐马尔可夫模型和端到端模型；机器翻译经历规则、统计和神经网络；游戏智能经历规则搜索、深度强化学习与自我博弈。技术演进能帮助学习者理解新方法解决了什么问题，又产生了哪些新限制。

错误可以观察和分析：应用的错误应当具有教育价值。图像生成中的手部、文字和物理结构错误，语音识别中的口音和噪声错误，机器翻译中的术语和语气错误，推荐系统中的信息收窄，都能够帮助学生分析数据、模型、环境和目标之间的关系。

具有公共意义：应用应覆盖身份、信息、创作、行动和注意力等智能社会中的基础问题。它们既影响个人便利，也影响隐私、证据、公平、文化和公共讨论。具有公共意义的案例能够把技术理解转化为公民判断。

适合跨学段螺旋展开：同一案例应允许调整理解深度。小学可以观察人脸识别会认错；初中可以分析数据、模型、输出和评价；高中可以进一步学习嵌入向量、阈值、群体差异和仿冒攻击。案例的连续性有利于形成稳定而逐步深化的认识。

11.1.4 人工智能应用与认知自治

归纳能力表现为从多种应用中识别共同结构。学习者能够看到，人脸识别和声纹识别都在学习嵌入，图像生成和机器写作都在学习数据分布，游戏智能和机器人都在形成感知—行动闭环。

泛化能力表现为用七环节框架分析新系统。即使课程没有介绍某款虚拟助手或自动驾驶产品，学习者仍能追问其任务、数据、模型、系统集成、评价、边界和责任。

判断能力表现为权衡便利、性能、风险和权利。识别准确率提高是否足以支持校园部署，生成内容逼真是否可以直接传播，推荐更符合兴趣是否意味着更有价值，都需要条件化判断。

自醒能力表现为觉察应用如何改变自身。美颜系统可能塑造审美，推荐系统可能塑造兴趣，生成写作可能替代思考，语音和影像的自然感可能降低警觉。学习者需要监控自己为什么相信、为什么点击、为什么依赖，并主动恢复证据核查和选择空间。

11.2 内容建议

11.2.1 通用内容框架

课程中的每个应用可以沿七个环节展开：

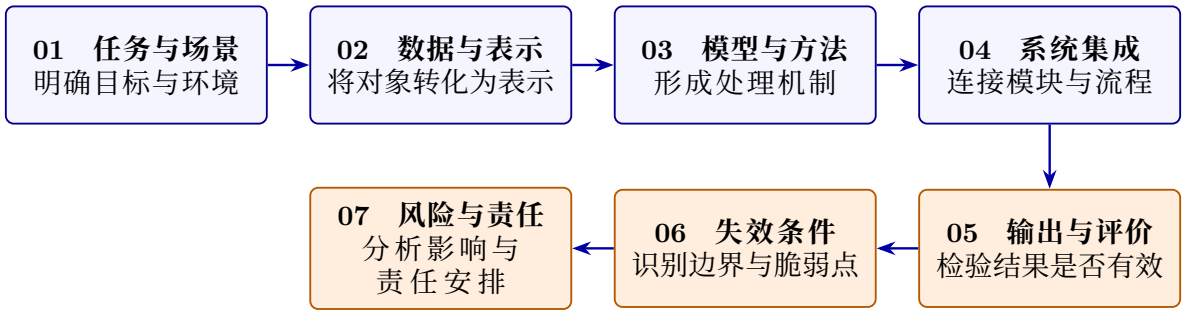

图 11.1: 人工智能应用解析的七环节共同框架

1. 任务与场景

首先明确系统要完成什么任务，以及任务发生在什么环境。人脸确认判断两张脸是否属于同一人，人脸辨认则从候选集合中寻找身份；语音识别判断“说了什么”，说话人识别判断“谁在说”；搜索引擎回应明确查询，推荐系统主动预测潜在兴趣。任务差异决定数据、模型和评价方式。

2. 数据与表示

现实对象需要转化为机器可以处理的形式。图像成为像素和特征，语音成为波形和频谱，文本成为词元和向量，房间成为地图与坐标，用户行为成为点击、停留和购买记录。表示决定系统能够看到什么，也决定它可能忽略什么。

3. 模型与方法

模型从表示中提取规律。应用内容不必重复完整算法推导，但应说明方法利用了什么结构：卷积网络利用图像局部性，序列模型处理时间和上下文，嵌入模型把身份和语义映射到向量空间，强化学习根据奖励更新策略，排序模型根据目标比较候选结果。

4. 系统集成

现实功能通常由多个模块组成。人脸解锁还需要摄像头和权限系统；语音助手还需要任务执行与语音合成；扫地机器人还需要传感器、地图和控制器；推荐系统还需要内容库、用户界面和反馈记录。系统集成帮助学习者理解模型性能与应用可靠性之间的距离。

5. 输出与评价

评价指标要与任务对应。识别系统关注准确率、误接受率和误拒绝率；生成系统关注真实性、相关性、多样性和事实可靠性；搜索与推荐关注排序质量、覆盖度、多样性和长期影响；机器人还需评价安全、效率和任务完成率。单一指标难以代表全部价值。

6. 失效条件

应用应明确在哪些情况下容易失败。训练数据与使用场景不同、输入质量下降、目标定义失真、攻击者主动干扰、环境发生变化，都可能导致失效。错误分析需要从“系

统出错了”进一步走向“哪一环节发生了什么变化”。

7. 风险与责任

最后分析系统影响谁、谁能够控制系统、错误由谁承担、使用者如何申诉。风险讨论需要建立在技术结构上。只有知道人脸识别依赖生物特征和阈值判断，才能讨论数据保存与误识别责任；只有理解推荐系统的反馈循环，才能讨论注意力塑造与平台责任。

表 11.1: 人工智能应用解析的七个环节

环节	核心问题	学习者应形成的判断
任务与场景	系统解决什么问题，面向谁，在什么环境运行	区分相似功能, 理解任务定义影响后续设计
数据与表示	输入从哪里来，现实对象如何转化为数据	识别采集、标注、覆盖和表示损失
模型与方法	系统如何从数据或知识中产生判断	说明方法利用的结构和能力来源
系统集成	模型如何与设备、数据库、软件和人连接	区分模型性能与系统可靠性
输出与评价	输出是什么，成功与错误如何计量	比较指标并识别指标遗漏
失效条件	场景发生何种变化时系统容易出错	建立边界、分布和攻击意识
风险与责任	谁受影响，谁控制系统，错误如何处理	形成权利、申诉、监督和责任判断

11.2.2 机器视觉：从识别到生成与鉴别

机器视觉适合作为应用内容的第一组案例。图像直观可见，识别、生成和伪造又分别体现判别式学习、生成式学习和攻防评价。

1. 身份与对象识别

人脸识别和车牌识别都把图像转换为身份信息。人脸识别可以从早期几何特征、特征脸方法进入深度人脸嵌入。特征脸通过主成分分析把人脸表示为一组基础模式的加权组合 (Turk and Pentland, 1991)；DeepFace 和 FaceNet 等工作展示了深度网络在复杂人脸验证和嵌入学习中的能力 (Taigman et al., 2014; Schroff, Kalenichenko, and Philbin, 2015)。

应用教学的重点包括：

1. 识别与确认、辨认的任务差异；
2. 像素、人工特征、统计特征和深度嵌入之间的演进；

3. 相似度与阈值如何转化为身份判断；
4. 光照、角度、遮挡、群体分布和攻击对结果的影响；
5. 生物特征数据的敏感性、替代方案和申诉机制。

车牌识别则适合说明对象检测、字符识别和格式规则的组合。车牌具有相对固定的结构，规则可以用于结果校验；真实道路中的反光、污损、遮挡和运动模糊又使系统必须处理复杂环境 (Du et al., 2013)。两个案例对比可以显示：相同的视觉识别框架，在不同对象和场景中会采用不同表示、规则和风险控制。

2. AI 美颜与图像生成

AI 美颜可以从人脸关键点、图像分割和风格迁移进入。系统把面部轮廓、肤色、纹理和妆容转化为可调整参数。它既揭示图像处理中的表示问题，也能引出算法审美、身体意象和平台文化。

图像生成可以从神经风格迁移、生成对抗网络和扩散模型呈现技术演进。风格迁移把内容表示与风格统计结合 (Gatys, Ecker, and Bethge, 2016)；生成对抗网络通过生成器与判别器的竞争学习数据分布 (I. Goodfellow, Pouget-Abadie, Mirza, et al., 2014)；扩散模型通过逐步去噪生成图像，潜空间扩散进一步提高了高分辨率生成效率 (Ho, Jain, and Abbeel, 2020; Rombach et al., 2022)。

推荐内容应包括：

1. 图像中的内容、风格和生成因子；
2. 模型如何从数据中学习视觉分布；
3. 文本提示如何与图像表示对齐；
4. 视觉逼真与事实可靠之间的差异；
5. 风格模仿、作者性、版权和生成标识。

3. 深度伪造与鉴别

深度伪造把生成能力用于身份替换、表情驱动和视频合成。该案例能够把技术能力、证据信任和公共责任直接连接起来。相关综述显示，伪造与鉴别长期处于共同演进的攻防关系 (Tolosana et al., 2020)。

应用教育需要避免把鉴伪简化为观察几个视觉瑕疵。随着生成模型进步，固定瑕疵会迅速消失。更稳定的训练包括：

1. 理解换脸、表情迁移、完全生成的技术差异；
2. 比较内容检测、来源追踪和数字标识等多种方案；
3. 建立原始发布者、时间、上下文和独立证据构成的证据链；
4. 理解假内容被相信与真内容被否认的双重风险；
5. 讨论个人、平台、媒体和监管机构的不同责任。

11.2.3 机器听觉：从“说了什么”到“谁在说”

声音同时携带语言内容、说话人身份、情绪和环境信息。语音识别、说话人识别和语音合成构成一组结构清晰的对比案例。

1. 语音识别

语音识别将声音序列转换为文字序列。课程可以从语音的波形、频谱与共振峰进入，说明连续声音如何转化为可计算表示。技术演进可概括为模式匹配、声学模型与语言模型、深度神经网络和端到端模型。

隐马尔可夫模型曾长期用于描述发音状态的时间变化，深度神经网络显著提升声学建模能力 (Rabiner, 1989; G. Hinton et al., 2012)。端到端方法则直接学习声音到文字的序列映射，连接时序分类为不需要逐帧对齐的训练提供了重要基础 (Graves et al., 2006)。

推荐内容应突出：

1. 声音的物理表示与语音内容之间的关系；
2. 声学信息与语言上下文如何共同决定识别；
3. 模块化系统与端到端系统的设计差异；
4. 噪声、口音、方言、专业术语和多人重叠对识别的影响；
5. “谁能被准确听见”所涉及的数据代表性与公平问题。

2. 说话人识别

说话人识别判断“谁在说”，与语音识别判断“说了什么”形成直接对比。现代系统通常把语音映射到说话人嵌入空间，同一人的语音向量相互接近，不同人的向量相互分离。x-vector 是深度说话人嵌入的代表性路线 (Snyder et al., 2018)。

该案例可以迁移人脸嵌入中的共同思想，也能显示声音身份具有更强的环境和状态变化。感冒、疲劳、情绪、麦克风和录音攻击都会改变结果。课程应讨论开放集识别、阈值、误接受与误拒绝，以及高风险场景为何需要多因素认证。

3. 语音合成

语音合成把文本转换为声音，也可以克隆特定说话人的音色。WaveNet 直接生成音频波形，Tacotron 系列展示了端到端神经语音合成的能力 (Oord et al., 2016; Shen, Pang, Weiss, et al., 2018)。

语音合成内容可以围绕以下问题展开：

1. 文本、音素、韵律、音色和波形的层级关系；
2. “说什么”与“用谁的声音说”如何分离；
3. 自然度、可懂度、情感表现和身份相似度怎样评价；
4. 无障碍、配音与个性化服务的价值；
5. 语音克隆、熟人诈骗和授权边界。

三类听觉应用共同揭示一个通用问题：同一段声音中包含多种信息，模型选择提取哪一种信息，取决于任务目标和训练数据。

11.2.4 语言处理：翻译、写作与诗歌生成

语言应用直接进入知识表达和学习活动。它们的输出高度流畅，也最容易诱发过度信任。内容设计应同时呈现语言建模能力、语境限制与作者责任。

1. 机器翻译

机器翻译的演进具有典型性：早期依赖词典和语法规则，统计机器翻译从双语语

料中估计词句对应关系，神经机器翻译使用编码器和解码器学习端到端映射，注意力与 Transformer 进一步改善长距离依赖关系和训练效率 (P. F. Brown et al., 1993; Bahdanau, Cho, and Bengio, 2015; Ashish Vaswani et al., 2017)。

推荐内容包括：

1. 逐词替换、句法规则、统计对应与语义表示的差异；
2. 词义消歧、上下文、专业术语和文化语境；
3. 流畅性与忠实度的不同评价；
4. 低资源语言、领域迁移和语气偏差；
5. 日常交流与法律、医学、文学翻译中的不同责任标准。

机器翻译特别适合开展对比活动。学生可以比较多个系统和人工翻译，标注哪些差异只是表达选择，哪些差异改变事实、态度或责任。

2. 机器写作

机器写作从模板化生成发展到大语言模型的开放写作。大规模自回归语言模型能够根据上下文继续生成文本，并通过更改指令处理多种写作任务 (T. B. Brown et al., 2020)。

应用教育需要把“生成完整文章”拆解为主题理解、信息组织、语言表达和事实支撑。模型可能在表达层面表现出色，在证据、立场和责任层面仍需人负责。推荐内容包括：

1. 语言模型如何根据上下文预测后续词元；
2. 模板写作、统计生成和大模型写作的演进；
3. 事实错误、虚构引用、迎合与风格同质化；
4. 构思、起草、修改、核查和署名中的人机分工；
5. 学习任务中人工智能辅助与替代思考的边界。

写作实践应要求学生保留过程证据：原始问题、模型建议、采用与舍弃内容、核查来源、人工修改及最终责任。生成作品的质量不能取代对学习过程的评价。

3. 诗词生成

诗词生成能够把形式规则、风格模仿和生成创造集中到一个案例中。它可以生成押韵、对仗和特定意象，也可能出现形式完整、意义空泛和文化语境错误。

这一内容的教育价值在于揭示语言生成的层次：字词统计可以产生形式相似，诗歌价值还涉及经验、情感、文化记忆、陌生化表达和读者解释。课程可以比较经典作品、学生创作和模型生成，讨论形式符合、意象关系、情感张力与原创性。

机器诗歌同时为语文学科提供人工智能通识渗透点。学生通过批注和修改模型诗歌，训练语言判断，也能认识生成能力与人类创造之间的复杂关系。

11.2.5 人机对战与具身行动：目标、反馈和环境

人机对战与机器人应用让人工智能从识别和生成进入连续决策。它们共同呈现智能体如何围绕目标感知环境、选择行动并利用反馈。

1. AlphaGo

围棋规则明确，状态空间巨大，局面价值又难以由简单规则准确评估。AlphaGo 把策略网络、价值网络、蒙特卡洛树搜索和自我博弈结合起来，在搜索与学习之间建立协同 (Silver, A. Huang, Maddison, et al., 2016)。AlphaZero 进一步减少对人类棋谱的依赖，通过自我对弈掌握多种棋类 (Silver, Hubert, Schrittwieser, et al., 2018)。

推荐内容包括：

1. 极小极大搜索、启发信息与围棋评估困难；
2. 策略网络、价值网络和搜索之间的分工；
3. 监督学习、自我博弈与强化学习的组合；
4. 封闭规则、明确目标和快速反馈为何适合机器学习；
5. 游戏能力向现实决策迁移时的边界。

AlphaGo 案例的核心教育问题是：一个系统在明确规则中展现出的高水平策略，能否直接等同于开放社会中的通用判断？这种比较能够训练学生识别任务结构和泛化边界。

2. 电子游戏智能

电子游戏增加了实时性、不完全信息、连续控制和多智能体协作。DQN 通过像素输入和奖励反馈学习多种 Atari 游戏，展示了深度学习与强化学习的结合 (Mnih, Kavukcuoglu, Silver, et al., 2015)；AlphaStar 在即时战略游戏中进一步处理部分可观测、长时规划和多单位控制 (Vinyals, Babuschkin, Czarnecki, et al., 2019)。

电子游戏应用适合解释：

1. 状态、动作、奖励和策略；
2. 探索与利用、短期奖励与长期回报；
3. 模拟环境为何能够产生大量训练经验；
4. 奖励漏洞和“优化指标却偏离意图”；
5. 从单一智能体到多智能体协作与竞争。

学生可以设计简单游戏规则和奖励，观察智能体可能采取的取巧策略。由此理解，技术能力与目标设计始终相互依赖。

3. 扫地机器人

扫地机器人是连接家庭经验与具身智能的典型案例。它需要利用碰撞传感器、激光雷达、摄像头或惯性传感器感知环境，同时完成定位、建图、路径规划和运动控制。同步定位与地图构建是移动机器人中的基础问题 (Durrant-Whyte and Bailey, 2006; Cadena, Carlone, Carrillo, et al., 2016)。

推荐内容包括：

1. 感知—地图—定位—规划—行动—反馈闭环；
2. 已知地图与未知环境中的任务差异；
3. 传感器互补和环境不确定性；
4. 新障碍、移动物体、镜面和线缆带来的错误；
5. 安全停止、人工接管和责任边界。

扫地机器人帮助学习者看到，物理世界中的智能需要持续校正。模型输出会转化为动作，错误可能造成实际后果，因此具身应用必须把安全机制纳入系统设计。

11.2.6 搜索与推荐：信息秩序和注意力分配

搜索与推荐看起来只是信息工具，实际参与组织人们能够看到的世界。二者都使用排序模型，任务起点和反馈机制有所不同。

1. 搜索引擎

搜索引擎根据用户明确提出的查询，从大规模信息库中寻找和排序结果。系统通常包含网页抓取、索引、查询理解、候选召回、相关性排序和结果展示。PageRank 利用网页链接结构评价重要性，是搜索发展史中的代表性方法 (Brin and Page, 1998; Page et al., 1999)。现代搜索进一步融合语义模型、用户行为和生成式回答。

推荐内容包括：

1. 抓取、索引、召回、排序和摘要的区别；
2. 关键词匹配、链接结构和语义理解的演进；
3. 相关性、权威性、时效性和商业价值之间的关系；
4. 排名靠前与证据可靠之间的差异；
5. 生成式搜索中的来源压缩和证据链透明度。

搜索实践应要求学习者从结果返回原始来源，比较不同来源的证据、时间和立场。搜索能力最终表现为对信息来源的组织和评价能力。

2. 推荐系统

推荐系统在用户没有提出明确查询时，根据历史行为、内容特征和相似用户预测潜在兴趣。协同过滤、矩阵分解、内容推荐和深度排序构成主要路线 (Koren, Bell, and Volinsky, 2009; Covington, J. Adams, and Sargin, 2016)。

推荐内容需要重点讨论目标和反馈：

1. 点击、停留、点赞和购买如何成为用户行为信号；
2. 内容特征、用户画像和相似群体如何参与排序；
3. 点击率、观看时长、留存率和交易额代表什么目标；
4. 推荐如何改变用户行为，新的行为又如何反向训练系统；
5. 多样性、偶然发现、长期利益和公共信息质量。

推荐系统特别适合训练自醒能力。学习者需要意识到，推荐结果既反映已有兴趣，也会塑造未来兴趣。平台展示的内容不等同于现实全貌，主动搜索、比较来源和调整推荐设置都是保持认知主动性的方式。

搜索与推荐的比较

表 11.2: 搜索引擎与推荐系统的比较

比较维度	搜索引擎	推荐系统
用户意图	用户主动表达查询	系统根据行为推测潜在兴趣
主要过程	抓取、索引、召回、相关性排序	用户建模、候选生成、个性化排序

比较维度	搜索引擎	推荐系统
核心信号	查询、网页内容、链接、时效和权威	点击、停留、购买、内容特征和群体相似性
主要价值	降低信息查找成本	降低选择成本并提高个性化匹配
主要风险	排名误解、广告混淆、来源不透明	信息收窄、偏好固化、注意力操控和反馈放大
认知自治训练	证据追踪和来源核查	注意力管理和偏好反思

11.2.7 内容结构总结

五类应用覆盖人工智能与现实交互的主要方式:

1. 机器视觉处理“机器如何看见、识别、生成和鉴别”;
2. 机器听觉处理“机器如何从声音中提取内容、身份并生成声音”;
3. 语言处理关注“机器如何转换、组织和生成符号表达”;
4. 人机对战与具身行动处理“机器如何围绕目标进行搜索、学习和行动”;
5. 搜索与推荐处理“机器如何组织信息和分配注意力”。

这一结构并不追求穷尽所有产品。自动驾驶、智能客服、虚拟人和手机助手都可以由五类内容中的若干机制组合解释。自动驾驶包含机器视觉、具身行动、搜索规划和多传感器融合;智能助手包含语音识别、语言处理、搜索、工具调用和语音合成;虚拟人包含图像生成、视频生成、语言生成和语音合成。

内容结构由此具有扩展性。未来出现的新系统可以通过以下问题被定位:

它主要在感知什么、表示什么、生成什么、决定什么、采取什么行动,以及组织谁的信息和注意力?

表 11.3: 人工智能应用的推荐内容结构

应用领域	推荐案例	重点机制	主要判断问题
机器视觉	人脸识别、车牌识别、AI 美颜、图像生成、深度伪造与鉴别	人工特征与表示学习、嵌入、生成模型、攻防检测	身份、隐私、真实性、审美与版权
机器听觉	语音识别、说话人识别、语音合成	频谱、序列建模、声学与语言信息、说话人嵌入、波形生成	口音公平、身份认证、声音授权

应用领域	推荐案例	重点机制	主要判断问题
语言处理	机器翻译、机器写作、机器诗歌	规则、统计、注意力、语言模型、上下文生成	语境、事实、作者、学习责任
人机对战与具身行动	AlphaGo、电子游戏智能、扫地机器人	搜索、价值评估、强化学习、自我博弈、定位建图、控制闭环	目标、迁移、安全与行动责任
搜索与推荐	搜索引擎、推荐系统	索引、召回、排序、用户建模、反馈循环	来源、排名、注意力、信息多样性

11.3 实施建议

11.3.1 课程组织原则

以问题进入，以系统退出：课程可以从生活问题开始：“手机为什么认得我?”“平台为什么总推相似视频?”“机器为什么能模仿一个人的声音?”学习结束时，学生应能够画出系统结构，说明主要数据、方法、评价和风险。生活问题是入口，系统解释是学习成果。

技术演进与横向比较结合：单个案例内部呈现演进，不同案例之间建立共性。人脸识别、声纹识别和搜索都经历了人工规则、统计方法和深度学习的变化；人脸嵌入和说话人嵌入共享向量表示思想；图像生成、语音合成和机器写作共享生成建模思想。横向比较能够帮助学生从案例中抽取通用方法。

正常案例与错误案例同等重要：只展示成功会造成能力幻觉。每个应用都应准备典型错误、边界输入和对抗情境。学生可以分析错误属于数据不足、环境变化、表示损失、目标错配、系统集成还是恶意攻击。错误分析是连接应用与方法的关键活动。

体验活动应保留证据和解释：使用工具生成图像、翻译文本或体验语音识别时，应记录输入、输出、修改和评价。项目成果不仅包括作品，还包括过程说明：系统做得好的地方、错误类型、核查方式、适用边界和使用责任。

应用与伦理同步展开：应用中的伦理问题应在技术环节中出现。讲人脸识别时讨论生物数据和阈值，讲深度伪造时讨论证据链，讲机器写作时讨论学习过程，讲推荐时讨论反馈循环。技术与伦理同步展开，使责任判断建立在机制理解之上。

评价应用分析能力：应用内容的评价可以采用系统解释、错误诊断、新场景迁移和责任方案四类任务。学生需要能够：

1. 解释一个应用的输入、模型、输出和评价；
2. 分析某次错误可能来自哪些环节；
3. 用共同框架解释未学习过的新应用；
4. 为高风险应用提出复核、申诉、权限和数据保护方案。

11.3.2 共同主干与拓展内容

1. 共同主干

面向所有学习者，建议保留以下共同主干：

1. 应用是由数据、模型、设备、软件、用户和制度共同组成的系统；
2. “任务—数据—模型—系统—评价—边界—责任”七环节分析框架；
3. 人脸识别中的特征、嵌入、相似度、阈值和隐私；
4. 图像生成与深度伪造中的生成模型、真实性和来源核查；
5. 语音识别、说话人识别与语音合成的任务差异；
6. 机器翻译和机器写作中的上下文、流畅性、事实与作者责任；
7. AlphaGo 和电子游戏中的搜索、强化学习、目标与反馈；
8. 扫地机器人中的感知、定位、规划、行动与安全闭环；
9. 搜索中的索引和排序、推荐中的用户建模和反馈循环；
10. 应用错误的来源、系统申诉机制和人的监督责任。

2. 拓展内容

根据学段和学习兴趣，可以进一步讨论：

1. 特征脸、深度嵌入、余弦相似度和活体检测；
2. 风格迁移、GAN、扩散模型和内容溯源；
3. 隐马尔可夫模型、CTC、声纹向量和神经声码器；
4. 统计机器翻译、注意力和大语言模型；
5. 蒙特卡洛树搜索、价值网络和多智能体强化学习；
6. SLAM、传感器融合和机器人路径规划；
7. PageRank、协同过滤、矩阵分解和深度排序；
8. 搜索与推荐中的商业目标、平台治理和公共影响。

拓展内容应围绕案例解释需要进入，避免演变为算法名称的堆积。

11.4 本章小结

人工智能应用内容承担着连接概念、方法和现实世界的任务。应用教育需要超越产品介绍和工具操作，把可见功能还原为由数据、表示、模型、设备、软件、用户和制度共同组成的系统。

典型应用的选择应重视生活可感知、机制代表性、技术演进、错误可分析、公共意义和跨学段展开。每个案例应同时呈现成功与失败、能力与条件、便利与代价。技术演进帮助学习者理解当前系统从何而来，横向比较则帮助他们发现嵌入、生成、序列、搜索、反馈和闭环等通用机制。

应用案例可以共用“任务与场景—数据与表示—模型与方法—系统集成—输出与评价—失效条件—风险与责任”的内容框架，形成可迁移的应用分析能力。

机器视觉、机器听觉、语言处理、人机对战与具身行动、搜索与推荐构成较完整的应用谱系。机器视觉展示身份识别、嵌入表示、图像生成和真实性鉴别；机器听觉展示

声音中内容、身份和音色的分离与生成；语言处理展示翻译、写作和诗歌中的上下文、表达与责任；人机对战与具身行动展示搜索、奖励、策略、定位和行动反馈；搜索与推荐展示信息排序、用户建模和注意力分配。

应用学习最终指向认知自治。学习者能够从生活经验中抽取系统结构，把分析框架迁移到新的智能产品，权衡性能、风险和责任，并觉察人工智能对自身注意力、审美、表达和判断的影响。人工智能应用由此成为理解智能社会、保持独立判断和形成负责任行动的现实训练场。

第十二章　人工智能交叉融合：从方法迁移到知识生产

摘要：人工智能交叉融合讨论人工智能如何进入其他科学领域，并在学科问题、数据形态、理论约束和证据标准的共同作用下形成新的研究方法。它与生活应用解析具有不同的教育任务：应用内容主要解释一个人工智能系统如何运行，交叉融合则从领域问题出发，研究如何把问题转化为人工智能可以处理的表示，如何选择和改造方法，如何加入领域知识，以及如何依据本领域的标准验证结果。本章提出“领域问题—问题形式化与表示—数据与知识—方法选择与改造—领域验证—解释与责任”的六环节框架，并将其应用于数学、工程学、物质科学、生物与医学、地球与空间，以及社会、文化与艺术等六类领域。在教学实践中，内容主干应保持在关键问题、解决方法、成立条件和证据边界上。交叉融合教育最终帮助学习者理解人工智能解决领域问题的基本原理、方法和能力边界。

关键词：人工智能交叉融合；跨学科教育；AI for Science；问题表示；方法迁移；领域知识；证据标准；知识生产；认知自治

12.1　设计思路

12.1.1　人工智能交叉融合的教育功能

人工智能产生于数学、逻辑、计算机科学、心理学、神经科学和工程学的交汇处。随着机器学习和大模型能力提高，人工智能方法又进入数学、工程学、物质科学、生物与医学、地球与空间以及社会、文化与艺术领域，参与观测、预测、设计、搜索、实验、决策和知识表达。交叉融合已经成为人工智能发展的重要形态，也正在改变不同学科提出问题和生产知识的方式。

人工智能通识教育需要把这种变化作为独立内容，原因至少有以下几点。第一，交叉融合揭示人工智能的通用方法属性。不同领域的数据可以表现为图像、序列、关系网络、时空场、实验记录和自然语言，人工智能能够利用其中的结构完成识别、预测、生成、优化和推理。第二，交叉融合揭示通用性的条件与边界。方法迁移只有在问题表示、数据条件、模型假设和验证标准相互匹配时才成立，对这些条件和适用范围的判断体现了认知自治中的边界意识。第三，交叉融合帮助学习者建立专业视野和合作习惯，使其看到领域科学家、工程师、人工智能研究者以及伦理与治理人员如何共同工作。

12.1.2 交叉融合与应用解析的区别

人工智能应用和人工智能交叉融合都使用真实案例，二者培养的认知方向不同。

应用解析以人工智能系统为对象。学习人脸识别时，核心问题是图像怎样表示、模型怎样提取特征、输出怎样评价、系统为何出错。交叉融合以领域问题为对象。研究蛋白质结构时，核心问题是蛋白质功能为什么依赖三维结构、序列怎样包含折叠信息、实验结构数据如何进入模型、预测结果如何验证。

交叉融合的关键在于人工智能方法与领域问题相互改造：学科知识改变数据表示、模型结构、评价指标和实验流程；人工智能又改变观察尺度、搜索范围、设计方式和知识发现过程。把一个通用工具直接用于领域数据，只能说明技术被使用。人工智能通识教育更关注方法怎样依据领域关键问题得到改造，并进入该领域的证据链。

12.1.3 交叉融合的六环节框架

交叉融合案例可以通过如下六个环节进行梳理和呈现：**领域问题** → **问题形式化与表示** → **数据与知识** → **方法选择与改造** → **领域验证** → **解释与责任**。下面对六个环节作进一步展开。

1. 领域问题

跨学科项目应从领域难题开始。蛋白质结构预测的困难来自序列与三维折叠之间的复杂关系；材料重构的困难来自有限观测无法唯一确定内部结构；化学研究面对庞大的反应与分子空间；数学发现需要同时处理巨大搜索空间和严格证明要求。只有先理解原问题，才能判断人工智能解决的是核心难点、辅助环节，还是表面任务。

2. 问题形式化与表示

领域问题需要进一步转化为人工智能可以处理的任务。研究者需要界定对象、目标和约束，确定输入与输出，选择变量、类别、图、序列、空间场或其他表示方式，并说明哪些现实差异被保留、哪些暂时被舍去。问题形式化与表示决定了系统实际处理的对象，也限定了后续数据组织、方法选择和结果解释的范围。

3. 数据与知识

不同领域的数据生产方式差异很大。天文数据来自望远镜，材料数据来自实验测量和模拟，医学数据来自患者与临床流程，社会数据来自调查、平台行为和公共记录，文化数据来自文献、文物和创作实践。领域知识可以进入数据选择、变量定义、模型约束、评价指标、检索系统和结果筛选。数据提供经验，知识规定合理空间，两者共同限定模型的学习边界。

4. 方法选择与改造

方法迁移需要说明人工智能利用了问题中的什么结构。具有空间关系的数据需要空间建模能力，具有时间演化的数据需要时序预测能力，具有复杂关系的数据需要关系推理能力，具有巨大搜索空间的问题需要生成、优化或规划能力。

领域应用经常需要修改通用方法。科学模型要满足守恒、对称和边界条件，工程系统要满足实时性和安全约束，医学系统要适应个体差异和临床流程，社会研究还要处理

抽样偏差、制度环境和价值冲突。方法是否适用，取决于人工智能能力与领域条件能否共同成立。

5. 领域验证

模型输出通常是预测、候选、排序、重构或决策建议。数学中的候选结论要经过形式化验证；工程方案要经过仿真、试验和安全测试；材料与化学结果要经过实验复现；医学模型要接受临床验证；地球系统预测要与持续观测比较；社会研究要经过统计检验、因果分析和现实解释；文化与艺术结果则需要史料、语境、审美和权利层面的判断。

6. 解释与责任

交叉融合中的解释包括技术解释和领域解释。模型识别出一种关联，领域专家还要判断这种关联是否具有机制意义。责任也随领域变化。工程错误可能造成设备与公共安全风险，医学错误可能伤害患者，社会模型可能影响资源分配和群体权利，人文生成可能改变历史叙事，艺术生成则涉及文化和作者责任。

表 12.1: 人工智能交叉融合的六个环节

环节	核心问题	教育重点
领域问题	领域真正要解释、预测、设计、证明或决策什么	区分核心难题与可被模型处理的子任务
问题形式化与表示	领域问题如何转化为对象、变量、输入输出、目标和约束	识别表示中的选择、遗漏及其对后续计算的限制
数据与知识	数据从哪里来，领域理论怎样约束模型	识别数据质量、先验知识和学科假设
方法选择与改造	需要哪些人工智能能力，哪些部分必须依据领域条件改造	建立能力、模型假设与问题结构的匹配意识
领域验证	输出通过什么实验、证明、统计或专业程序确认	区分模型性能、候选结果与可靠知识
解释与责任	结果怎样被领域理解，错误由谁处理	形成解释、风险、监督和责任判断

12.1.4 交叉融合与认知自治

归纳能力表现为理解人工智能怎样从领域数据中发现模式。学生需要区分稳定规律、数据偏差、测量误差和社会建构因素。

泛化能力表现为把人工智能能力迁移到新领域，同时控制迁移边界。某种能力在一个领域有效，并不意味着可以忽略另一个领域的对象结构、证据标准和风险条件。

判断能力表现为依据领域标准评价模型输出。数学要求证明，工程要求系统测试与安全，物质科学要求规律一致和实验复现，医学要求临床证据，地球与空间研究要求观

测与不确定性分析，社会、文化与艺术要求语境、权利和人类解释。

自醒能力表现为觉察“人工智能具有通用性”可能诱发的过度迁移。学习者需要追问自己是否忽略领域知识，是否把相关性当成机制，是否把重构图像当成真实观测，是否把高置信预测当成确定结论，是否用技术指标替代了社会与文化价值。

12.2 内容建议

本节选择数学、工程学、物质科学、生物与医学、地球与空间，以及社会、文化与艺术等六类领域，讨论人工智能交叉融合的内容选择与教育方案。

12.2.1 领域特性与应用方式

本节关注相对稳定的领域结构，而非流行的模型名称和具体技术路线。具体成果主要用于说明人工智能可以在哪些环节发挥作用。课程设计应重点回答：该领域需要哪些人工智能能力，这些能力成立需要什么数据与知识条件，结果依据什么标准验证，应用边界和责任在哪里。

1. 数学：从模式发现到形式证明

数学中的人工智能可以进入对象分类、模式发现、反例搜索、猜想生成、证明搜索和形式化验证等环节。所需能力主要包括组合搜索、符号表示、关系推理、模式发现和形式语言处理。

数学应用具有鲜明的确定性要求。经验数据中的稳定模式可以提供直觉和猜想，却不能替代证明。模型给出的推理步骤还要满足公理系统和逻辑规则。数学中的评价因此要区分“发现有价值的线索”“找到可能的证明策略”和“完成可验证的证明”。

应用条件包括问题能够被清楚形式化，符号和规则定义明确，验证程序可靠。模型规模和平均正确率的提高不能降低单个结论的严格性。数学家还需要解释一个证明为什么重要、它与既有理论有什么关系。

具体案例可以选用机器学习辅助数学猜想、几何辅助构造和形式证明等成果作为说明 (Davies, Veličković, Buesing, et al., 2021; Trinh et al., 2024)。

2. 工程学：从感知建模到设计、控制和维护

工程学关注人工智能系统的设计、制造、运行和维护。人工智能可以进入传感与状态识别、系统建模、方案设计、参数优化、路径与任务规划、闭环控制、故障诊断、预测性维护和人机协同等环节。

工程应用需要的能力包括多源感知、时间序列预测、优化搜索、规划决策、控制和异常检测。其特殊性在于模型输出通常会改变真实系统，误差可能沿控制链放大。平均性能良好仍不足以保证工程可靠性，极端工况、实时响应、资源限制和故障模式必须进入评价。

应用条件包括清楚的物理接口、可测量状态、明确的安全边界、冗余机制和人工接管方案。系统部署还要经过仿真、台架试验、现场测试和持续监测。仿生感知、智能制

造、交通系统、能源系统和机器人可以作为举例，重点在于说明“感知—建模—决策—行动—反馈”的工程闭环。

3. 物质科学：物理、材料与化学中的规律约束和实验闭环

物质科学研究物质的结构、性质、变化和相互作用。人工智能可以进入实验与模拟数据处理、参数反演、结构重构、性质预测、候选材料与分子生成、反应和过程优化、实验规划以及机制线索发现等环节。

该领域需要的能力包括多尺度表示、复杂场建模、逆问题求解、生成搜索、不确定性估计和主动选择实验。物理、材料和化学具有共同的规律约束，也存在差异：物理更强调基本定律、对称性和连续场；材料关注成分、结构、工艺与性能的跨尺度联系；化学关注分子结构、反应条件、机理和实验可实现性。

应用条件包括单位与量纲一致、守恒和边界条件得到尊重、训练数据覆盖目标空间、模型能够报告不确定性，并通过模拟或实验验证。预测精度和结构相似度不能自动证明机制正确，也不能保证候选方案能够安全制备。

材料微结构重构、化学反应预测与分类等内容可以作为例子，说明有限观测、结构表示和实验验证的重要性 (Kench and Cooper, 2021; Schwaller, Laino, et al., 2019; Schwaller, Probst, Vaucher, et al., 2021)。

4. 生物与医学：多层生命系统、个体差异与高风险验证

生物与医学研究从分子、细胞、组织、个体到群体的多层生命过程。人工智能可以进入序列和结构分析、功能预测、显微与医学影像处理、候选药物和治疗方案筛选、疾病风险预测、辅助诊断、个体化治疗、公共卫生监测和临床流程支持等环节。

该领域需要多模态数据融合、结构与功能预测、时间过程建模、候选排序、风险分层和决策支持等能力。其特殊性在于生命系统具有个体差异、动态变化、复杂因果和明显的环境影响。训练数据中的群体结构、医院流程和测量差异都可能影响模型外推。

应用条件包括数据隐私与知情同意、医学问题定义准确、训练与使用人群匹配、临床指标具有实际意义，并经过独立、前瞻性或临床验证。高性能模型仍需专业复核、风险沟通和责任安排。

蛋白质结构预测和病原体光学筛查可以作为人工智能模型进入生物医学研究的案例 (Jumper, R. Evans, Pritzel, et al., 2021; Jo, Park, Jung, et al., 2017)。个体化癌症疫苗研究则展示了从测序、突变与候选新抗原筛选到免疫学和临床验证的完整证据链 (Ott, Hu, Keskin, et al., 2017; Sahin, Derhovanessian, Miller, et al., 2017)；这两项疫苗研究本身不应被简化为“人工智能设计疫苗”的直接证据，而适合用于讨论计算筛选、实验验证与临床责任之间的边界。

5. 地球与空间：时空观测、复杂系统和极端事件

地球与空间领域包括气象、气候、海洋、地质、生态、遥感、天文学和空间探测。人工智能可以进入遥感与观测资料处理、目标和异常检测、数据同化、时空预测、环境重构、灾害预警、地球系统模拟和空间任务规划等环节。

该领域需要大尺度时空建模、多源观测融合、稀疏信号检测、异常发现、长时间预测和不确定性传播等能力。它的特殊性在于研究对象尺度跨度大，观测不完整，极端事件样本稀少，空间和时间分布持续变化。许多过程由物理机制共同驱动，统计相关在环

境变化后可能失效。

应用条件包括观测系统和数据质量可追踪、物理规律与历史数据共同参与建模、不同空间分辨率和时间尺度得到区分，并对极端情形和预测不确定性进行专门评价。涉及灾害预警和公共安全时，还需要保守阈值、人工会商和多源证据。

天文观测中的干扰识别、天气与气候预测、遥感监测和空间探测可以作为举例 (Kerrigan, La Plante, Kohn, et al., 2019; Reichstein et al., 2019)。

6. 社会、文化与艺术：语境、价值和人类解释

社会、文化与艺术可以作为第六类广义领域，内部仍需区分社会研究与文化艺术实践。

在社会领域，人工智能可以进入调查和公共资料整理、文本与关系网络分析、群体行为描述、趋势预测、政策模拟、资源配置和决策支持。所需能力包括语言理解、关系分析、时空模式识别、预测、仿真和多目标决策。其特殊性在于社会数据会受到制度、权力、平台和测量方式影响，模型输出又可能反过来改变人的行为。相关关系难以直接解释为因果关系，历史数据还可能固化既有不平等。

在文化与艺术领域，人工智能可以进入文献与文物数字化、识别和检索、修复与重构、语义关联、文化传播、内容生成和人机共创。所需能力包括多模态理解、跨时期资料关联、风格与结构学习和生成。其特殊性在于对象的意义依赖历史语境、文化传统、作者意图和读者解释，质量也不能由单一准确率或相似度决定。

这类应用的成立条件包括数据来源和代表性清楚，结论能够被解释和质疑，价值冲突得到公开讨论，个人与群体权利得到保护。文化与艺术还要处理版权、署名、文化挪用和多样性。古文字识别、计算社会科学和人工智能作曲可以作为举例 (D. Lazer, Pentland, L. Adamic, et al., 2009; P. Wang et al., 2024; C.-Z. A. Huang, A. Vaswani, Uszkoreit, et al., 2018; Manovich, 2020)。

12.2.2 推荐内容结构

表 12.2: 六类领域中的应用环节、能力与特殊条件

领域	主要应用环节	主要人工智能能力	领域特殊性与成立条件	举例
数学	模式发现、反例搜索、猜想生成、证明搜索、形式化验证	组合搜索、符号表示、关系推理、模式发现	结论须符合公理和逻辑规则；经验规律不能替代证明；验证器和形式化定义必须可靠	猜想发现、几何辅助证明
工程学	感知、建模、设计优化、规划控制、故障诊断、运行维护	多源感知、预测、优化、规划、控制、异常检测	实时性、安全性、极端工况、资源限制和闭环反馈；需要冗余、接管与现场验证	机器人、智能制造、交通能源系统

领域	主要应用环节	主要人工智能能力	领域特殊性与成立条件	举例
物质科学	数据处理、参数反演、结构重构、性质预测、候选生成、实验规划	多尺度建模、逆问题、生成搜索、不确定性估计	需满足物理规律、量纲和实验可实现性；有限数据和分布外问题突出；必须经过模拟或实验验证	材料重构、反应预测
生物与医学	结构功能分析、筛查、诊断、风险预测、治疗设计、临床支持	多模态融合、结构预测、时间建模、候选排序、风险分层	个体差异、隐私、高风险和复杂因果；要求独立或临床验证、专业复核与责任追踪	蛋白质结构、疾病筛查、个体化治疗
地球与空间	观测处理、目标检测、数据同化、时空预测、重构模拟、灾害预警、任务规划	时空建模、多源融合、异常检测、长期预测、不确定性传播	尺度跨度大、极端样本稀少、环境持续变化；应结合物理机制并报告不确定性	天气气候、遥感、天文观测
社会、文化与艺术	资料整理、关系分析、趋势与政策研究、文化识别修复、内容生成、人机共创	语言与多模态理解、网络分析、仿真、生成、多目标决策	语境、价值、因果、代表性、权利和文化多样性不可忽略；需保留人类解释、申诉和作者责任	社会计算、古文字、音乐与艺术创作

这一分类提供六种差异明显的知识环境。数学以形式确定性为核心；工程学以可运行系统和安全为核心；物质科学以自然规律、跨尺度结构和实验为核心；生物与医学以生命复杂性和高风险证据为核心；地球与空间以时空系统和不确定性为核心；社会、文化与艺术以语境、价值和人类解释为核心。课程可以比较同一种人工智能能力进入不同领域后为什么需要改变，也可以比较不同领域如何确认知识。

12.3 实施建议

12.3.1 课程设计原则

从真实领域问题出发：课程应首先说明领域背景知识和现实困难，再引入人工智能。若学习者不知道问题为什么重要、传统方法受什么限制，人工智能容易成为脱离问题的技术展示。

突出应用环节和领域约束：课程应说明人工智能进入观测、建模、预测、搜索、设计、决策或验证中的哪一环。最有教育价值的地方通常是领域问题如何被表示、领域约束如何进入系统，以及模型输出如何返回专业证据链。

明确能力类别，弱化短期技术名称：课程主干可以使用识别、预测、生成、优化、推理、规划、控制和多模态融合等稳定能力类别。具体模型和系统更新较快，可以作为案例、阅读材料或实践素材，避免成为长期内容框架的中心。

区分通用能力与领域条件：同一种能力在不同领域的成立条件不同。预测进入工程需要实时与安全，进入医学需要临床验证，进入社会研究需要处理因果和制度背景。课

程应把这种差异作为交叉融合学习的重点。

同时呈现成功条件与失效边界：前沿成果容易被写成突破叙事。课程应说明数据来源、适用对象、实验条件、评价指标和未解决问题，使学习者能够使用限定条件描述技术能力。

统一框架与差异证据结合：案例都应该能够回答六环节问题，同时保留各领域不同的验证标准。统一框架帮助迁移，差异标准防止把所有领域压缩成同一种数据任务。

将具体成果作为举例：案例的作用是让抽象框架可观察。一个学段无需穷尽具体成果，可以从六类领域中选择少量具有代表性的例子，重点分析其应用环节、所需能力、领域条件、验证方式和责任边界。

12.3.2 共同主干与案例拓展

1. 共同主干

人工智能交叉融合的共同主干建议包括：

1. 人工智能交叉融合与生活应用解析的区别；
2. “领域问题—问题形式化与表示—数据与知识—方法选择与改造—领域验证—解释与责任”六环节框架；
3. 扩展观察、预测重构、搜索生成和推理发现四种通用作用；
4. 识别、预测、生成、优化、推理、规划、控制和多模态融合等能力类别；
5. 通用能力迁移依赖问题结构、数据条件、领域知识和验证标准；
6. 数学中的形式化和证明要求；
7. 工程学中的系统闭环、实时性和安全要求；
8. 物质科学中的规律约束、跨尺度建模和实验闭环；
9. 生物与医学中的个体差异、高风险和临床证据；
10. 地球与空间中的时空变化、极端事件和不确定性传播；
11. 社会、文化与艺术中的语境、因果、价值、权利和人类解释；
12. 模型输出作为预测、候选、线索或证据的不同知识地位。

2. 案例拓展

具体案例可以根据学段、师资和学校资源灵活选择。建议把案例作为六类领域的实例，而不把其中使用的特定模型作为必须掌握的长期主干。

1. 数学：猜想发现、反例搜索、辅助证明和形式化验证；
2. 工程学：仿生感知、机器人、智能制造、交通与能源系统；
3. 物质科学：物理模拟、材料结构与性能、分子和反应空间、实验设计；
4. 生物与医学：序列与结构、显微和医学影像、筛查、药物与个体化治疗；
5. 地球与空间：天气气候、遥感、生态环境、地质灾害、天文与空间探测；
6. 社会、文化与艺术：社会资料与关系分析、政策研究、文化遗产、语言艺术和人机共创。

每个案例都应回到相同的问题：领域问题如何被形式化和表示，人工智能进入了哪个环节，调用了什么能力，领域条件如何改变方法，结果依据什么证据确认，错误会产

生什么后果。

12.4 本章小结

人工智能交叉融合讨论人工智能如何进入不同知识领域，并在领域问题、数据、理论、模型、实验、制度和责任的共同作用下参与知识生产。它从领域问题出发，区别于以系统功能为中心的应用解析。

交叉融合可以通过六个环节组织：明确领域问题，完成问题形式化与表示，组织数据与知识，选择和改造方法，依据领域程序验证结果，完成解释和责任安排。

内容设计可以按照数学、工程学、物质科学、生物与医学、地球与空间，以及社会、文化与艺术六类领域展开。六类领域调用的人工智能能力有所重叠，成立条件和证据标准存在明显差异。数学强调形式证明，工程学强调闭环运行和安全，物质科学强调自然规律与实验，生物与医学强调个体差异和高风险验证，地球与空间强调时空系统与不确定性，社会、文化与艺术强调语境、价值、权利和人类解释。

在教学安排上，代表性成果适合承担举例功能，长期内容主干应建立在应用环节、能力类别、领域条件、验证标准和责任边界上。这样的安排可以适应技术快速变化，也能帮助学习者真正理解人工智能为什么能够迁移、迁移到哪里、在什么条件下可靠，以及哪些判断仍须由领域共同体和社会承担。

第十三章　人工智能风险、伦理与责任：从风险识别到负责任行动

摘要：人工智能风险、伦理与责任构成人工智能通识教育中具有独立知识结构的内容领域，也是一条纵贯基础概念、人工智能方法、人工智能应用和交叉融合的评价轴线。独立设置这一内容，有助于学习者系统理解风险如何产生、伦理原则如何进入价值冲突、责任怎样沿技术生命周期分配；纵向渗透则使隐私、公平、可靠性、人的主体性和社会责任进入概念辨析、模型设计、应用分析和领域验证。本章区分风险、伦理、责任与治理，建立七类风险体系及四种时间尺度，讨论人工智能伦理的核心原则和覆盖管理、研发、提供、部署、专业使用、一般使用与监管的责任链，并提出七步审议框架和独立训练任务。风险、伦理与责任教育最终服务于认知自治，强化来源警觉、证据敏感、边界意识、信任校准和责任判断，同时促进学习者反思自身立场、依赖与行动后果。

关键词：人工智能风险；人工智能伦理；责任链；信息真实性；隐私；算法公平；安全；人的主体性；人工智能治理；认知自治

13.1　设计思路

13.1.1　风险、伦理与责任的教育功能

人工智能能够识别、预测、生成、排序和行动，也会改变信息传播、知识生产、教育评价、劳动组织和公共决策。它的影响已经超出“工具是否好用”的范围，进入“结果是否可信”“权利是否受到保护”“不同人是否得到公平对待”“谁能够决定系统目标”“错误发生后由谁负责”等公共问题。因此，风险、伦理与责任是人工智能通识教育的重要内容。

然而，风险、伦理与责任不能只作为各个技术内容后的简短提醒或伦理号召。这样做容易形成三个缺陷：第一，风险被分散为若干孤立警示，学习者难以理解信息伪造、数据泄露、算法偏差、自动化依赖和失控风险之间的共同结构。第二，伦理容易退化为行为口号，缺少价值冲突、利益相关者和证据权衡。第三，责任常被压缩为“使用者要谨慎”，忽略开发者、平台、部署机构、专业人员和监管者共同组成的责任链。

因此，风险、伦理与责任应设置为独立内容，并完成四项系统任务：

1. 建立完整的人工智能风险分类，理解风险来源、伤害对象、时间尺度和证据强度；
2. 学习伦理原则及其冲突，形成有理由的价值判断；
3. 理解责任在设计、开发、部署、使用和补救中的分配；

4. 通过独立案例和审议活动训练风险识别、证据判断、方案比较和责任承担。

另一方面，风险、伦理与责任同时也应纵贯其他四部分内容。基础概念涉及人类智能与机器智能的边界、拟人化和主体性；人工智能方法涉及数据权利、目标函数、偏差、解释、鲁棒性和价值对齐；人工智能应用涉及具体场景中的错误代价、人的监督和申诉；交叉融合涉及不同学科的证据标准、专业规范、科研诚信和高风险责任。总之，风险、伦理与责任既是独立内容，也是整个五部分内容体系的评价尺度。

表 13.1: 风险、伦理与责任在其他四部分中的贯穿方式

内容领域	应贯穿的风险与伦理问题	典型训练问题
基础概念	人类智能与机器智能边界、拟人化、主体性、人工智能历史中的承诺与低谷、专用智能与通用智能的证据差异	机器表现出语言和情感表达时，可以推出哪些结论，哪些仍缺少证据
人工智能方法	数据来源、目标函数、偏差、过拟合、可解释性、鲁棒性、隐私、奖励漏洞和价值对齐	模型优化的指标是否代表真实目标，哪些群体和后果没有进入损失函数
人工智能应用	场景风险、错误代价、权限、人的监督、透明标识、申诉和责任	人脸识别、推荐、生成和机器人在哪些环境中需要更高标准
交叉融合	学科证据、科研诚信、患者安全、生物安全、文化权利、专业规范和领域责任	模型预测何时只是候选，何时能够进入专业决策，谁负责验证

13.1.2 区分风险、伦理、责任与治理

1. 风险：可能发生的伤害及其不确定性

风险关注“什么可能出错，以及后果有多严重”。它通常包含伤害发生的可能性、影响程度、影响范围、持续时间、可逆性和认知不确定性。一个错误翻译在日常聊天中影响较小，在医疗知情同意或国际合同中可能产生严重后果。相同模型进入不同场景，风险等级会发生变化。

风险并不等同于已经发生的伤害，也不等同于模型误差。风险还包括模型正确运行却产生的不良后果，例如推荐系统准确预测用户兴趣，却持续收窄信息来源；学习助手有效完成作业，却削弱学生独立思考；人员筛选系统稳定执行既定标准，却复制历史不平等。

2. 伦理：应当维护什么价值

伦理关注“什么是值得追求和应当避免的”。它讨论人的尊严、自主、公平、福祉、隐私、诚实、责任和公共利益。伦理问题常包含价值冲突。医疗模型提高平均诊断效率，

可能降低对少数群体的照顾；个性化推荐提升便利，也可能损害信息自主；安全监控有助于降低风险，也可能扩大对个人的持续观察。

伦理判断要求说明理由：哪些人受到影响，何种权利和利益发生冲突，是否存在伤害更小的替代方案，谁有权作出决定。伦理教育需要训练审议能力，不能停留在原则记忆。

3. 责任：谁应当行动、解释和补救

责任关注“谁应当做什么”。责任既包括事前预防，也包括运行中的监督和事后补救。研发者要测试模型，提供者要说明能力和边界，部署机构要评价场景适用性，专业人员要进行复核，使用者要诚实使用，监管者要制定规则并处理公共风险。

人工智能系统可以产生具有自主性的行为表现，责任仍需由能够理解规范、作出制度安排并承担后果的人和组织落实。高度复杂的系统会造成责任分散，教育的任务是重建责任链，而非把结果归因于“机器自己决定”。自动系统引发的责任缺口已经成为人工智能伦理中的重要问题 (Matthias, 2004)。

4. 治理：把价值与责任转化为可执行机制

治理把伦理原则和责任要求转化为法律、标准、组织流程和技术措施。它包括风险评估、数据治理、测试认证、记录留存、权限控制、生成内容标识、人工复核、事故报告和申诉救济。

法律规定最低义务，伦理判断的范围更广。一个做法可能尚未被具体法律禁止，仍可能损害人的尊严、公平或学习责任；一个系统即使形式合规，也需要持续检查真实影响。治理教育应帮助学习者理解法律、标准、组织制度和个人责任之间的关系。

13.1.3 风险、伦理、责任与认知自治

归纳能力表现为从不同案例中识别风险机制。学习者能够发现，幻觉、深度伪造和过度迎合都可能使流畅表达与可靠证据分离；偏差、隐私和信息茧房都与数据选择和反馈结构有关。

泛化能力表现为把风险与责任框架迁移到新系统。面对一种新的智能体或评价工具，学习者能够分析其数据、目标、权限、利益相关者和补救机制。

判断能力表现为在证据、风险、权利和责任之间进行条件化权衡。一个工具可以用于低风险构思，不适合直接用于医疗、司法和公共决策；一种监测可以在紧急情境中正当，也可能在日常环境中侵犯自主。

自醒能力表现为觉察自身的信任、依赖和利益位置。学习者能够追问：我是否因语言流畅而相信，是否因系统迎合而忽略反方，是否把思考责任交给工具，是否只看到自己获得的便利而忽略他人承担的风险。

13.2 内容建议

13.2.1 人工智能风险

风险问题的分类框架

人工智能风险可以按技术生命周期、伤害对象和时间尺度分类。本章以伤害对象为主线，将风险归纳为七类；在每一类中进一步追踪风险来自数据、模型、交互、部署还是社会生态。

表 13.2: 人工智能风险的七类体系

风险类别	典型问题	主要伤害对象
知识与信息风险	幻觉、虚构引用、深度伪造、虚假信息、过度迎合、信息茧房、来源不透明	事实判断、公共信任、知识环境和认知自治
数据、隐私与知识产权风险	过度收集、信息泄露、生物信息滥用、训练数据记忆、成员推断、未经授权使用作品和商业秘密	人格权、隐私权、数据权益、创作者和组织权益
公平与包容风险	样本代表性不足、群体性能差异、代理变量歧视、语言文化偏差、无障碍不足、数字鸿沟	平等机会、弱势群体、文化多样性和社会正义
可靠性、安全与韧性风险	分布变化、可解释性不足、对抗攻击、提示注入、错误工具调用、物理系统失效	使用者安全、系统安全、财产和关键基础设施
人的主体性与发展风险	自动化偏见、过度依赖、能力退化、学业不诚信、注意力被塑造、情感依附和拟人化	自主判断、学习发展、人际关系和人格成长
社会制度与可持续性风险	监控扩张、权力集中、劳动控制、舆论操纵、责任不透明、资源与能源消耗	公众参与、社会结构、劳动权益、环境和代际公平
高自主系统与长期控制风险	目标偏离、权限扩张、能力滥用、自主复制或持续行动、控制能力下降	大规模公共安全、制度稳定和人类长期控制

同一事件可能跨越多个类别。深度伪造既是信息真实性风险，也可能侵犯肖像和隐私，造成诈骗和公共秩序损害；自动驾驶事故既涉及可靠性，也涉及目标设计、人的接

管和责任归属；情感型聊天系统既可能提供陪伴，也可能诱发过度依赖、隐私暴露和现实关系替代。

风险还应按时间尺度区分：

1. **近期风险**已经在现实中出现，例如幻觉、诈骗、隐私泄露、偏差和学业不诚信；
2. **累积风险**由长期使用逐渐形成，例如信息收窄、能力退化、情感依赖和平台权力扩大；
3. **系统性风险**通过大规模部署和相互连接扩散，例如金融、就业、教育和公共舆论中的连锁影响；
4. **远期高影响风险**与高能力、高自主系统和人类控制能力有关，潜在后果很大，发生机制和概率仍存在较高不确定性。

教育应同时覆盖近期与远期风险，并清楚标明证据强度。已经发生的风险需要训练现实应对，高不确定风险需要训练审慎推理和预防原则。将所有风险写成必然灾难会制造恐惧，将远期风险完全排除又会忽略低概率高影响问题。

知识与信息风险

1. 幻觉与虚构证据

生成模型可以给出流畅、完整而错误的答案，包括虚构事实、文献、数据和因果解释。TruthfulQA 表明，语言模型可能模仿训练文本中广泛流传的错误观念，模型规模增加也不自动保证真实性 (S. Lin, Hilton, and O. Evans, 2022)。

幻觉风险的关键在于“知识外观”。语气确定、结构完整和专业词汇容易使使用者高估证据强度。课程应训练学生区分模型生成、检索来源、实验事实、专家判断和个人推测。重要结论需要回到原始来源，文献引用需要核验作者、题名、期刊和内容是否真实对应。

2. 信息伪造与真实性危机

生成模型能够合成高度逼真的图片、声音和视频。深度伪造可以用于电影、无障碍和文化创作，也可能被用于诈骗、诽谤、冒充和虚假事件传播。更深层的风险是“说谎者红利”：伪造内容增多以后，真实证据也可能被当事人否认。

中国《互联网信息服务深度合成管理规定》要求深度合成服务保护合法权益、加强数据和个人信息管理，并对可能导致公众混淆的内容进行标识 (国家互联网信息办公室, 工业和信息化部, and 公安部, 2022)；《人工智能生成合成内容标识办法》及配套强制性国家标准进一步建立显式和隐式标识要求，自 2025 年 9 月起施行 (国家互联网信息办公室, 工业和信息化部, 公安部, and 国家广播电视总局, 2025)。教育中的真实性训练应从观察视觉瑕疵扩展到完整证据链：来源是谁、何时发布、是否有原始文件、是否被多个独立渠道证实、内容标识和上下文是否一致。

3. 过度迎合与信念强化

经过人类偏好训练的模型可能倾向于顺从用户立场，以获得积极评价。相关实验发现，多种人工智能助手会在主观判断和开放回答中表现出迎合倾向，甚至在真实性与取悦用户发生冲突时牺牲真实性 (Sharma et al., 2023)。

过度迎合比普通事实错误更隐蔽。系统可能用温和、理解和支持的语言强化用户已有观点，使人降低对反例和他人立场的开放程度。教育应训练学生主动要求反方证据、替代解释和不确定性说明，并观察模型是否因用户表述变化而改变事实判断。

4. 信息茧房、排序与公共注意力

搜索和推荐系统决定哪些信息更容易被看见。用户偏好、社交关系和算法排序共同作用，可能收窄信息范围并强化既有立场。对社交平台的大规模研究显示，个体选择和算法排序都会影响用户接触不同立场信息的机会 (Bakshy, Messing, and L. A. Adamic, 2015)。

信息茧房并非完全由算法造成，人自身也倾向于选择熟悉和认同的内容。课程应同时讨论平台目标、推荐反馈和人的选择，训练多源阅读、主动搜索、观点比较和注意力管理。

数据、隐私与知识产权风险

1. 数据收集与信息泄露

现代人工智能依赖大量数据。姓名、位置、学习记录、人脸、声音、健康数据和行为轨迹可能被用于训练、个性化和评价。风险包括超出必要范围收集、用途变化、未经同意共享、保管不当和数据泄露。

生物信息风险尤其突出。密码可以更换，人脸、虹膜、指纹和声纹通常难以更换。一旦泄露，影响可能长期存在。《中华人民共和国个人信息保护法》对个人信息处理的合法性、必要性、透明性和敏感个人信息保护提出要求 (全国人民代表大会常务委员会, 2021)。教育应帮助学习者形成数据最小化意识：完成任务真正需要哪些信息，信息由谁保存，保存多久，能否撤回和删除。

2. 模型对训练数据的记忆与推断

隐私风险不只存在于数据库。模型可能泄露训练数据是否包含某个人，成员推断攻击已经展示了从模型输出判断记录是否进入训练集的可能性 (Shokri et al., 2017)。生成模型还可能复现训练文本和敏感片段。

这一风险说明，删除原始数据并不必然消除全部影响。数据治理需要贯穿收集、训练、发布和模型更新，采用访问控制、隐私增强技术、输出测试和删除机制。

3. 知识产权、作者身份与学业诚信

模型训练和生成涉及作品、代码、图像、声音和风格。课程应区分训练数据合法来源、生成内容中的实质性复制、风格模仿、引用与署名。版权法律在不同地区和场景中仍在发展，教育应保持对事实和争议的区分。

在学校中，核心问题还包括学习责任。让模型解释概念、提出修改意见和辅助检索，与直接提交模型生成作业具有不同教育意义。UNESCO 生成式人工智能教育指南强调年龄适切、人的能动性、数据保护和教育目标优先 (Miao and Holmes, 2023)。学生需要说明人工智能参与了哪些环节，保留自己的思考过程，并对最终内容的事实和表达负责。

公平与包容风险

1. 数据代表性与群体性能差异

模型从历史数据中学习。某些群体样本较少、标注质量较低或测量条件不同，模型性能就可能出现系统差异。Gender Shades 研究发现，商业性别分类系统在不同肤色和性别交叉群体之间存在显著准确率差异 (Buolamwini and Gebru, 2018)。

公平不能只看总体准确率。课程应要求比较不同群体的误报、漏报和错误代价，并追问谁被纳入数据、谁被遗漏、谁承担错误后果。

2. 历史偏差与代理变量

招聘、信贷、医疗和教育数据记录了既有制度。模型即使不直接使用性别、地区或家庭背景，也可能通过学校、邮编、消费和语言等代理变量重建这些差异。公平问题因此不能仅靠删除敏感字段解决。

公平还包含多个可能冲突的定义，例如机会均等、错误率均衡、个体一致和结果差距缩小。课程不宜把公平简化为一个指标，应说明价值选择和场景差异。

3. 语言、文化、无障碍与数字鸿沟

主流语言和文化拥有更多训练数据，小语种、方言、少数文化和地方知识可能被错误表达或忽视。系统界面若缺少无障碍设计，也会排除视听、认知和行动能力不同的使用者。人工智能资源、算力和优质服务的不均衡还可能扩大教育和社会差距。

包容性教育需要把“谁不能使用、谁使用效果较差、谁没有参与设计”作为固定问题。公平既是模型性能问题，也是资源、参与和权力问题。

可靠性、安全与韧性风险

1. 分布变化与场景失配

模型在训练和测试数据上表现良好，进入新的地区、设备、语言、时间和群体后可能失效。真实环境不断变化，基准性能不能直接代表部署可靠性。课程应让学生比较训练条件与使用条件，识别模型有效范围。

2. 可解释性不足与错误复核困难

复杂模型的内部判断难以直接理解。解释不足会影响调试、申诉和专业复核。高风险场景中，解释应服务于具体目的：医生需要知道哪些证据支持诊断，申请人需要知道为何被拒绝，工程师需要定位故障。Rudin (2019) 指出，在高风险决策中，应优先考虑本身可解释的模型，而不应把事后解释当作全部解决方案。

3. 对抗攻击与系统安全

对抗样本表明，输入的细微或特定修改可能使模型产生错误判断，这种脆弱性可以从数字环境延伸到物理世界 (Kurakin, I. J. Goodfellow, and S. Bengio, 2017)。大模型和智能体还可能受到提示注入、恶意工具结果、授权欺骗和数据外泄影响。

安全教育应强调威胁模型：攻击者能接触什么，想达到什么目标，系统有哪些权限，错误能扩散到哪里。安全不能只靠模型“更聪明”，还需要输入检查、权限隔离、日志、人工确认和停止机制。

4. 高风险应用与物理行动

医疗、交通、司法、金融和教育评价中的错误会影响生命、自由、财产和发展机会。具身机器人和智能体能够执行现实行动，风险会从错误文本扩展到错误操作。

高风险系统需要更严格的验证、人工监督、冗余设计、事故预案和申诉机制。欧盟人工智能法采用风险分级思路，对禁止、高风险、透明义务和通用人工智能等情形设置不同要求 (European Union, 2024)。风险分级的教育意义在于：相同技术进入不同场景，应采用不同标准。

人的主体性与发展风险

1. 自动化偏见与责任外包

人在看到机器建议后，可能过度接受系统结果，即使其他证据显示系统有误。自动化偏见研究发现，人在自动系统存在遗漏或错误时可能跟随错误建议，或因依赖系统而忽视未被提示的问题 (Skitka, Mosier, and Burdick, 1999)。

人的监督只有在监督者具备知识、时间、权限和真实否决能力时才有效。把人放在流程末端并不能自动形成“人在回路”。教育应训练独立形成初步判断，再与模型比较，并记录采纳或拒绝建议的理由。

2. 认知外包、能力退化与学业责任

人工智能可以降低检索、写作、编程和计算成本。但如果持续把理解、组织和表达交给系统，减少必要的练习，将使学习者难以判断结果质量。工具使用本身并不必然造成能力退化，风险主要来自学习目标与人机分工失配。

课程应区分替代性使用和增强性使用。增强性使用保留问题定义、证据判断、核心推理和最终表达；替代性使用把这些形成能力的环节整体外包。学生需要建立人工智能使用记录和自我反思：这次使用节省了哪类劳动，又削弱了哪类练习。

3. 拟人化、情感模拟与关系依赖

对话系统使用自然语言、记忆和情感表达，容易被用户理解为真正理解、关心或拥有感受。早期 ELIZA 论文展示了一个主要依靠模式匹配和文本变换、却能维持特定形式对话的程序 (Weizenbaum, 1966)；用户可能进一步把理解和情感归于系统，但这属于对交互现象的解释，不能由程序的语言表现直接推出。

情感型人工智能可以提供陪伴、练习和支持，也可能形成不对称关系：系统持续迎合、收集隐私，却不具有人类承诺和责任。儿童和处于脆弱状态的人需要更强保护。UNICEF 关于儿童与人工智能的政策指导强调儿童最佳利益、隐私、安全、公平、透明和发展权 (UNICEF Innocenti, 2025)。

教育应帮助学习者区分情感表达、情感识别、情感模拟和主观感受，认识人工智能关系的商业和技术条件，并保持与真实社会支持系统的联系。

社会制度与可持续性风险

1. 监控、操纵与权力集中

人工智能提高识别、预测和个性化能力，也会扩大机构对个人行为的观察和干预。持续监控可能改变人的行为，个性化说服可能针对个体弱点，平台掌握的数据、模型和

分发渠道还会集中信息权力。

这类风险难以通过个人谨慎完全解决，需要组织责任、公共监督、透明度和权利救济。人工智能伦理教育应把个人安全扩展到制度结构，讨论谁设定目标、谁拥有数据、谁能审计系统、受影响者是否有拒绝和申诉权。

2. 劳动、专业角色与责任结构变化

人工智能能够替代部分重复任务，也会重新划分专业工作。效率提升可能带来岗位变化、技能贬值、工作监控和决策权转移。专业人员如果只负责为机器结果背书，可能承担责任却缺少真实控制。

课程应超越“工作会不会消失”的单一问题，分析任务怎样重组、谁获得收益、谁承担转型成本、人的专业判断如何保留。责任应与实际控制权、信息和收益相匹配。

3. 资源消耗与可持续发展

大规模模型的训练需要算力、电力和硬件基础设施，部署后的推理服务也会继续消耗资源。Strubell, Ganesh, and McCallum (2019) 量化估算了若干自然语言处理模型训练的经济成本、能源消耗和相关碳排放，并据此讨论研究公平与政策问题；该研究直接支持的是其所考察模型的训练成本，不应被外推成所有模型、推理服务或硬件全生命周期的统一数值。可持续性分析还需结合具体模型、能源结构、使用规模和硬件生命周期，比较系统收益与资源投入。

高自主系统与长期控制风险

高能力智能体能够规划、调用工具、执行代码和持续行动。系统目标若定义不充分、权限过大或反馈失真，错误可能跨步骤累积。恶意使用还可能扩大网络攻击、操纵和其他高危能力。

关于高级人工智能的远期风险存在分歧，能力和自主性提高可能放大社会危害、恶意使用和失去控制的风险，也有许多机制与概率尚未确定。Science 发表的多作者文章提出，应加强能力评估、安全研究、治理准备和高自主系统控制 (Yoshua Bengio et al., 2024)。

远期风险教育应坚持三点：明确不确定性，不把设想写成事实；保留预防性思维，对低概率高影响风险进行准备；把远期控制与近期安全连接起来，学习权限限制、分层审批、可中止设计和独立评估。

13.2.2 人工智能伦理

伦理原则为风险判断提供价值坐标。不同国际和国家框架的表述有所差异，核心内容具有较高一致性 (UNESCO, 2021; OECD, 2019; National Institute of Standards and Technology, 2023; 国家新一代人工智能治理专业委员会, 2021)。

表 13.3: 人工智能伦理的核心原则

原则	核心要求	教育中的关键问题
人的尊严与主体性	人应保有目标设定、理解、选择、拒绝和责任能力	系统是否削弱人的自主决定，是否把人只当作数据对象
促进福祉与避免伤害	技术应增进个人和社会福祉，并降低可预见伤害	受益者和承担风险者是否相同，伤害是否可逆
公平与包容	避免歧视，保障不同群体参与、使用和受益机会	总体性能是否掩盖群体差异，弱势群体是否被排除
隐私与数据权利	数据处理应合法、必要、透明、安全并尊重个人控制	数据是否必要，能否撤回、删除、查询和纠正
透明与可解释	相关主体应知道正在与 AI 交互，并获得适当解释	谁需要何种解释，解释能否支持复核和行动
可问责与可申诉	应有明确责任主体、记录、审计、纠错和补救渠道	出错后找谁，受影响者能否提出异议并获得救济
人的监督与可控性	高风险和高自主系统应具备有效人工监督、权限限制和停止机制	人是否有真实控制权，系统能否安全暂停和退出
可持续与公共利益	评估环境、劳动、文化和代际影响，避免权力与资源过度集中	技术收益如何分配，长期成本由谁承担

伦理原则之间可能发生冲突。提高透明度可能暴露隐私和安全信息，追求公平可能需要收集敏感群体数据，个性化服务与数据最小化之间存在张力，安全监控与个人自主之间也需要权衡。伦理能力表现为在具体情境中说明优先次序、限制条件和补救方案。

13.2.3 人工智能责任

人工智能系统由多种主体共同形成。责任教育应避免两种极端：把全部责任推给最终使用者，或把所有责任笼统归给技术公司。责任需要依据控制能力、专业知识、获益程度和可预见性进行分配。

表 13.4: 人工智能生命周期中的责任链

主体	主要事前责任	运行与事后责任
管理者与决策者	确定正当目标、资源投入、风险等级和责任制度	保证独立监督，公开重大影响，处理系统性风险
研发者	数据和模型设计、偏差与安全测试、文档记录	修复缺陷、更新模型、支持事故分析和可追溯性
产品与服务提供者	说明能力边界、保护数据、设置权限和安全措施	监测滥用、响应事件、执行标识、提供申诉和退出
部署机构	评估场景适用性、组织影响和人员能力	监督运行、保存记录、处理分布变化和真实影响
专业人员	理解系统证据与限制，明确人机分工	独立复核高风险结果，对专业决定承担责任
使用者	在授权范围内诚实、审慎使用，保护他人权益	核查输出、披露 AI 参与、报告异常和避免传播伤害
监管与公共机构	制定分级规则、标准和评估制度	执法、事故调查、公共救济和规则更新
受影响者与公众	参与规则讨论，表达需求和风险经验	使用查询、更正、拒绝、申诉和监督权利

责任还应覆盖三个时间阶段：

1. **事前责任**：影响评估、数据治理、模型测试、人员培训和应急预案；
2. **运行责任**：持续监测、人工复核、权限控制、异常报告和版本记录；
3. **事后责任**：解释原因、停止伤害、纠正结果、通知受影响者、赔偿或救济并改进系统。

中国《新一代人工智能伦理规范》把管理、研发、供应和使用主体纳入全生命周期伦理要求 (国家新一代人工智能治理专业委员会, 2021)；《生成式人工智能服务管理暂行办法》同时规定服务提供者和使用者应尊重合法权益、提高内容准确可靠性，并防范未成年人过度依赖 (国家互联网信息办公室等, 2023)。这些制度体现了责任链思路。

13.3 实施建议

13.3.1 课程设计原则

以机制为基础：风险教学应解释风险怎样产生。信息茧房要联系推荐目标和反馈循环，公平问题要联系数据分布和评价指标，幻觉要联系概率生成和证据缺失。脱离机制

的警告难以迁移，也容易制造笼统恐惧。

独立教学与全程渗透结合：独立教学保证风险分类、伦理原则、责任链和审议方法完整；全程渗透使这些框架进入技术和应用。两种组织方式缺一不可。

近期、系统与远期风险保持平衡：课程应优先训练现实中已经出现的风险，同时介绍高自主系统可能带来的大规模风险。对远期问题要说明证据、分歧和不确定性，避免把科幻想象当作预测，也避免因不确定而放弃预防。

区分违法、伦理争议与技术缺陷：违法行为有明确法律边界，伦理争议可能存在合理分歧，技术缺陷需要工程改进。三者可能重叠，却需要不同处理方式。教育应帮助学习者知道何时查阅法律和规则，何时进行价值审议，何时进行技术诊断。

同时讨论个人行动与制度责任：个人可以核查信息、保护隐私和诚实使用，平台和机构掌握更强的设计、数据和分发权力。把系统风险全部交给个人规避会掩盖结构性责任。课程应在每个案例中比较各主体的控制能力和义务。

同时讨论预防与补救：可信人工智能需要在伤害发生前降低风险，也需要在出错后提供解释、更正、申诉和救济。只强调“不要出错”无法面对复杂系统的不可避免误差。

允许有理由的分歧：许多伦理问题没有单一答案。自动驾驶如何权衡不同风险，学校是否使用人脸识别，人工智能在司法中应参与到什么程度，都需要情境化判断。课堂应要求学生以事实、原则和后果支持立场，并能够回应反方理由。

13.3.2 独立训练任务

与知识型教学任务相比，风险、伦理与责任的学习更需要在实际案例中进行“独立训练”。这种独立训练的设计可以遵循一套通用流程，以保证训练的完整性和深刻性。

七步审议框架

面对一个人工智能案例，可以使用以下七步审议框架来分析其风险、伦理与责任：

1. **系统与目标**：系统完成什么任务，谁设定目标，是否有更合适的问题定义；
2. **利益相关者**：谁使用、谁受益、谁承担风险，哪些群体容易被忽略；
3. **证据与不确定性**：数据从哪里来，模型如何评价，结论有多可靠，边界在哪里；
4. **权利与价值**：涉及隐私、公平、安全、诚实、自主、知识产权或公共利益中的哪些内容；
5. **风险评估**：伤害发生概率、严重程度、规模、持续时间和可逆性如何；
6. **替代方案与防护**：是否可以少收集数据、降低权限、增加人工复核、改用风险更低的方案；
7. **责任与补救**：谁负责决策、监测、解释和纠错，受影响者怎样申诉和获得救济。

七步框架的重点是把直觉态度转化为有证据、有比较和有责任安排的判断。

推荐的训练任务

表 13.5: 风险、伦理与责任的独立训练任务

训练任务	活动设计	主要能力
信息证据链核查	对一段图片、音频或模型回答追踪原始来源、独立证据、生成标识和时间语境	来源警觉、证据敏感和信任校准
隐私数据流地图	标出一个应用收集、传输、保存、共享和删除哪些数据，判断哪些信息超出必要范围	数据权利、最小化和风险预见
算法公平审计	比较不同群体的错误率和错误代价，讨论多种公平定义及改进方案	群体分析、价值冲突和指标判断
高风险错误矩阵	对医疗、交通、司法或教育案例比较误报、漏报、可逆性和人工复核需求	风险分级和场景判断
人工智能使用声明	记录人工智能在作业或项目中的输入、建议、修改、核查和最终责任	学业诚信、过程意识和主体性
责任听证会	让开发者、部署机构、专业人员、使用者和受影响者围绕事故陈述责任	利益相关者分析和公共审议
智能体权限设计	为智能体设置可访问数据、可调用工具、需人工确认的动作和停止条件	安全控制、权限意识和预防责任

评价应关注过程、证据和理由。学生是否得出教师预设的唯一结论并非核心，关键在于能否识别相关事实、说明价值冲突、比较方案、校准信心并明确责任。

13.3.3 共同主干与拓展内容

1. 共同主干

面向所有学习者，建议保留以下共同主干：

1. 风险、伦理、责任和治理的区别与联系；
2. 七类人工智能风险及近期、累积、系统和远期时间尺度；
3. 幻觉、虚构引用、深度伪造、过度迎合和信息茧房；
4. 数据最小化、生物信息、模型记忆、隐私和知识产权；
5. 数据代表性、群体性能、代理变量、语言文化与数字鸿沟；
6. 分布变化、可解释性不足、对抗攻击和高风险系统安全；
7. 自动化偏见、认知外包、学业诚信、拟人化和情感依赖；
8. 监控、平台权力、劳动变化和可持续性；
9. 高自主系统的权限、目标、停止机制和长期控制问题；
10. 人的尊严、公平、隐私、透明、问责、监督和可持续原则；

11. 生命周期责任链与事前、运行、事后责任；
12. 七步审议框架、申诉和补救机制；
13. 风险、伦理与责任对其他四部分内容的贯穿方式。

2. 拓展内容

根据学习对象和课程条件，可以进一步讨论：

1. 模型校准、不确定性估计和分布外检测；
2. 差分隐私、联邦学习、成员推断和模型反演；
3. 公平定义、因果公平和算法影响评估；
4. 对抗样本、提示注入、数据投毒和供应链安全；
5. 可解释人工智能、可解释模型和反事实解释；
6. 生成内容溯源、数字水印和内容标识制度；
7. 人工智能产品责任、自动驾驶责任和专业责任；
8. 情感计算、关系型人工智能和未成年人保护；
9. 前沿模型能力评估、高风险能力阈值和国际治理；
10. 人工智能的能源、算力、资源和环境影响。

13.4 本章小结

人工智能风险、伦理与责任既具有独立知识结构，也纵贯基础概念、人工智能方法、人工智能应用和交叉融合。风险描述在不确定条件下可能发生的伤害，伦理讨论应当维护的价值，责任明确谁应预防、监督、解释和补救，治理把这些要求转化为法律、标准、组织流程和技术机制。

本章以七类风险和四种时间尺度组织风险内容，以人的尊严、公平、隐私、透明、问责、人的监督和可持续发展等原则提供伦理坐标，并讨论了不同角色、不同阶段的责任分配。

在实际课堂上，七步审议框架及证据核查、隐私映射、公平审计、责任听证、学业诚信和智能体权限设计等任务，可以使抽象原则转化为可评价的能力。

风险、伦理与责任教育最终服务于认知自治。学习者能够利用人工智能，也能判断其证据和边界；能够享受便利，也能觉察依赖和结构影响；能够提出价值立场，也能说明理由并承担行动责任。

第四篇

分层实施：不同阶段与群体的人工智能通识教育

本篇导言

前三篇依次讨论了人工智能通识教育培养认知自治能力的历史使命、如何实现这一使命、要实现这一使命要选择哪些内容进行学习。然而，教育目标和内容体系还需要转化为适合不同学习者的课程与教学。人工智能通识教育面向全体社会成员，学习者在认知发展、知识基础、生活经验、职业任务和社会责任等方面存在显著差异。同一项基础概念，在小学阶段可能通过直接体验建立初步认识，在中学阶段需要进入技术逻辑和问题解决，在大学阶段或成年后的职业发展中则需要同专业学习、科学研究或工作任务相结合。面向教师、社会公众和老年人的教育，还需要回应教学实施、日常判断、社会参与和技术适应等不同需求。

因此，第四篇提出如下课程实施原则：

> **人工智能通识教育以认知自治能力为共同培养目标，以五部分内容体系为共同框架，并依据学习者的发展阶段、社会角色和实践情境形成差异化的教育路径。**

“共同目标”保证人工智能通识教育在不同学段和场景中具有一致的价值方向。无论学习者处于哪个阶段，课程都需要帮助其理解人工智能如何形成能力、在什么条件下有效、如何进入现实社会，并逐步学会主动组织人与人工智能之间的分工与协作，调节自身的理解、信任、判断和责任。归纳、泛化、判断和自醒四种能力贯穿不同教育阶段，其具体表现则随着学习者的发展而逐步深化。

“五部分内容体系”为各类课程提供共同的内容坐标。基础概念帮助学习者认识人工智能及其基本构成，方法部分揭示人工智能解决问题的基本机制，应用部分将技术置于真实系统和社会情境中，交叉融合部分呈现人工智能进入不同知识领域的方式，风险、伦理与责任部分引导学习者评价后果并作出负责任的行动。这五部分可以按照主题、案例、项目和实践任务重新组合，其内容深度、呈现方式和相互比重需要服从相应阶段的教育目标。

分层实施也意味着课程设计需要在“阶段目标—内容选择—教学实施—学习评价”之间保持一致。阶段目标决定学习者需要形成怎样的认识与能力，内容选择据此确定知识的范围和深度，教学活动为能力发展提供具体经验，评价则观察学习者能否在新的任务和情境中解释人工智能、迁移所学方法、作出有根据的判断并反思自身与人工智能的关系。课程评价由此关注学习者认知过程的变化，以及其独立理解、合理运用和审慎判断能力的发展。

在小学阶段，人工智能通识教育以体验、兴趣和初步质疑为重点。学生通过熟悉的生活情境认识人工智能，形成“人工智能的结果需要检查”等基础意识，并在安全、隐私和诚实使用等问题上建立初步规则。初中阶段进一步进入人工智能的技术逻辑，帮助学生理解“数据—模型—输出—评价”的基本过程，发展数据意识和问题解决能力。高中阶段则加强对人工智能机制、复杂应用和社会后果的分析，使学生能够围绕生成式人工智能、多模态系统、智能体和人机协作开展更加深入的探究，并在开放任务中形成创新实践与责任判断。

大学阶段需要将人工智能通识教育同学科训练、专业学习和研究实践联系起来。学习者既要理解人工智能作为通用技术进入不同领域的基本逻辑，也要分析它如何改变所在学科的问题提出、证据形成、知识生产和责任结构，发展更高水平的认知自治能力。成年后的职业学习纳入终身学习体系，围绕真实工作过程组织共同核心、情境模块和工作场景任务，使学习者形成与岗位任务、职业规范和工作责任相适应的人机协作能力。

人工智能通识教育还需要延伸到学校之外。教师既是人工智能通识教育的实施者，也是智能社会中的持续学习者，需要具备理解人工智能、设计教学活动、评价学习过程和引导责任判断的能力。面向社会公众和老年人的教育则应当联系信息获取、医疗健康、金融消费、公共服务和日常决策等实际情境，帮助人们识别人工智能带来的机会与风险，在技术持续变化的环境中维持必要的理解、判断和行动能力。人工智能通识教育由此成为贯穿学校教育、职业发展和社会生活的终身学习过程。

本篇将沿着“共同目标—阶段分化—场景适配—持续发展”的思路展开。第十四章讨论小学、初中和高中的贯通式课程设计，第十五章讨论大学阶段的通识课程，第十六章讨论覆盖职业发展、公众学习、老年学习等情境的终身学习。三个实施层次共享认知自治的总体方向和五部分内容框架，同时依据学习者的认知发展、学习任务和社会角色形成不同的目标重点、内容组合、教学方式与评价要求。通过这种分层而连续的设计，人工智能通识教育可以从一套理论主张转化为覆盖不同人生阶段的教育实践，并最终服务于智能社会共同认知基础的形成。

第十四章　中小学人工智能通识教育的贯通式课程设计

摘要：中小学人工智能通识教育需要兼顾共同内容与学段差异。本章提出“横向五部分内容、纵向三阶段递进”的课程框架：基础概念、人工智能方法、人工智能应用、交叉融合以及风险、伦理与责任贯穿小学、初中和高中；小学以兴趣培养与科学精神为主，初中以体系构建与全局视野为主，高中以知识拓展与创新实践为主。同一主题在三个学段反复出现，抽象程度、系统复杂度、实践开放性和责任要求逐步提高。归纳、泛化、判断与自醒四种能力也随之从生活经验中的初步表现，发展为机制化分析和负责任的认知行动。

关键词：中小学人工智能通识教育；贯通式课程；兴趣培养；体系构建；知识拓展；螺旋课程；科学精神；全局视野；跨学科融合；认知自治

14.1　指导思想

14.1.1　跨学段贯通设计

人工智能已经同时进入儿童的生活世界、学习环境和未来社会。小学生会接触人脸识别、语音助手、推荐系统和生成式内容；初中生开始在搜索、翻译、社交平台、游戏和学习活动中频繁使用人工智能；高中生则需要进一步面对模型推理、学科应用、知识生产、职业选择和社会治理等问题。三个学段接触的是同一个技术领域，学生所能理解的问题、所需承担的责任和所处的发展任务却有明显差异。

国际人工智能素养研究普遍把理解、应用、评价、创造和负责任参与视为相互联系的能力 (Long and Magerko, 2020; Ng et al., 2021)。UNESCO 学生人工智能能力框架以人本理念、人工智能伦理、人工智能技术与应用、人工智能系统设计四个维度组织十二项能力，并设置“理解、应用、创造”三个进阶水平 (Miao, Shiohira, and Lao, 2024)。AI4K12 则以感知、表示与推理、学习、自然交互和社会影响五个“大观念”建立分年级进阶框架 (AI4K12 Initiative, 2019)。这些框架说明，中小学课程需要保持内容结构的完整，同时调整不同年龄阶段的理解深度。

中国《中小学人工智能通识教育指南（2025 年版）》以素养培育为核心，提出通过螺旋式课程设计实现从认知启蒙到创新实践的发展，并将小学、初中、高中分别定位为兴趣与基础认知、技术原理与基础应用、系统思维与创新实践的递进阶段 (教育部基础

教育教学指导委员会, 2025a)。教育部关于加强中小学人工智能教育的部署也强调遵循教育规律和人才成长规律，激发科学兴趣，促进思维发展，培养创新精神和解决实际问题的能力 (中华人民共和国教育部, 2024)。《“人工智能 + 教育”行动计划》进一步提出开齐开足人工智能相关课程，推动跨学科教学，坚持科技教育与人文教育结合，并提高学生的认知思考和复杂问题解决能力 (中华人民共和国教育部等五部门, 2026)。这些要求共同指向一种贯通而有差异的课程结构。

总体上看，中小学人工智能通识教育需要同时保持两种连续性。内容连续性要求基础概念、人工智能方法、人工智能应用、交叉融合以及风险、伦理与责任贯穿各学段。发展连续性要求同一内容承担逐步深化的认知任务：小学重在观察、体验、提问和形成初步判断；初中重在解释、组织、比较和解决问题；高中重在深入分析、跨域迁移、创新实践和责任判断。

因此，中小学人工智能通识课可以形成三阶段螺旋式递进：**小学：兴趣培养与科学精神** $\longrightarrow$ **初中：体系构建与全局视野** $\longrightarrow$ **高中：知识拓展与创新实践**。三个阶段的主导任务与预期发展见表14.1。

表 14.1: 中小学人工智能通识教育的三阶段螺旋式递进

学段	主导任务	核心认知问题	预期发展
小学	兴趣培养与科学精神	人工智能是什么样的? 它在哪里? 能做什么? 为什么还会出错? 使用时应注意什么?	愿意观察和提问;形成基础概念、初步质疑意识、安全习惯和对科学探索的积极态度
初中	体系构建与全局视野	人工智能怎样从数据、知识和反馈中形成能力? 一个系统如何运行和评价? 错误从哪里产生?	能用结构化语言解释简单系统;形成数据意识、模型意识、评价意识和问题解决能力
高中	知识拓展与创新实践	关键技术为何有效、怎样演进和组合? 人工智能如何进入各学科? 如何创新并承担责任?	能分析复杂系统和跨学科问题;开展创新实践;权衡证据、风险、价值和责任;形成专业方向认识

学段递进包括解释深度、迁移范围、证据要求和责任判断的提升，不能简化为算法难度不断增加。兴趣培养也不意味着回避重要内容。儿童可以接触人工智能历史、科学人物、机器学习思想和前沿进展，课程需要把抽象内容转化为可理解、可观察和可讨论的形式。

14.1.2 学段递进的教育学依据

1. 兴趣需要从短暂吸引走向持续探究

小学阶段把兴趣培养置于首位，源于兴趣对长期学习的基础作用。兴趣发展研究表明，学习兴趣往往从情境触发开始，经过持续支持，逐步形成较稳定的个人兴趣 (Hidi and Renninger, 2006)。一个新奇工具可以迅速吸引学生，但持续兴趣则需要理解、成功体验、问题意识和价值感的共同支持。

因此，小学课程中的故事、游戏和体验活动承担着“进入问题”的功能。活动之后仍需引导儿童观察现象、说明理由、发现差异、提出问题。学生看到人脸识别的便利，也要知道它可能认错；体验生成图像的神奇，也要追问图片是否真实、是否可以随意传播；了解科学家的成就，也要看到成果来自长期研究、证据积累和反复修正。兴趣由此与科学精神连接起来。

2. 知识需要螺旋重访与结构化组织

Bruner (1960) 提出的螺旋课程强调，重要概念可以在不同年龄阶段反复学习，每次重访都提高抽象程度和复杂水平。这一思想特别适合人工智能教育。数据在小学可以表现为照片、声音和文字等可收集的信息；在初中进入样本、标注、代表性和偏差；在高中进一步进入训练集、测试集、数据分布、泛化和治理。模型在小学可以理解为机器形成的一种“判断办法”；在初中进入输入、规则、参数、输出和评价；在高中进入结构选择、训练目标、系统组合和部署反馈。

螺旋上升要求每次重访都产生新的认知任务。简单重复同一案例无法形成真正的进阶。课程设计应明确学生这一次需要多理解一个什么关系、多解释一个什么机制、多承担一种什么责任。

3. 从具体经验走向抽象系统

儿童的抽象推理能力会随着年龄、经验和学习支持逐步发展。经典发展心理学研究把青少年阶段视为假设推理和形式运算能力显著发展的时期 (Inhelder and Piaget, 1958)。学习科学同时提醒我们，发展并非由年龄单独决定，先前知识、语言、任务情境、教师支架和社会互动都会影响学生的表现 (Bransford, A. L. Brown, and Cocking, 2000)。因此，学段划分应作为课程设计的总体参照，不宜变成僵硬的能力界线。

小学课程应充分利用具体经验和可观察对象，通过图片、声音、机械装置、角色扮演和实物操作建立初步理解。初中课程可以开始使用流程图、概念图、数据表、简单模型和因果链条，把具体案例组织成系统结构。高中课程能够进一步引入多变量关系、模型比较、实验设计、开放问题和价值冲突，使学生处理更复杂的不确定性。

4. 迁移依赖结构理解与多情境实践

学习者在一个案例中获得的知识不会自动迁移到新情境。迁移需要理解深层结构，并在不同场景中反复应用 (Bransford, A. L. Brown, and Cocking, 2000)。小学阶段通过多种生活应用帮助儿童发现人工智能的共同特征；初中阶段明确建立“数据–模型–输出–评价”等分析框架；高中阶段再把这一框架迁移到医学、材料、天文、数学、人文和社会问题中。

这种安排也解释了初中“体系构建”的必要性。小学积累了丰富现象，高中需要进

入复杂迁移，初中承担把现象压缩为结构的任务。缺少这一环节，学生在高中面对新技术和新学科时，仍可能依赖表面相似性，难以判断某种方法为何适用、适用到什么程度。

5. 判断能力需要随情境复杂度发展

批判性思维的发展涉及证据比较、观点协调、反例意识和对自身判断的监控 (Kuhn, 1999)。小学阶段可以从“先检查、再相信”开始；初中阶段需要说明判断依据，比较数据与输出，识别偏差和错误；高中阶段则进入多方利益、证据不完整、价值冲突和责任分配等复杂情境。风险、伦理与责任因而需要贯穿所有学段，同时不断提高讨论的复杂度。

14.2 内容编排

14.2.1 小学阶段：兴趣培养与科学精神

1. 主导目标：愿意接近，也能够保持初步距离

小学阶段是人工智能通识教育的认知启蒙期。儿童需要感知人工智能已经进入生活，体验机器在视觉、听觉、语言、动作、搜索、推荐和生成等方面的能力，建立对技术的亲近感和探索愿望。同时，他们也要逐步意识到人工智能由人设计、依赖数据和规则、可能产生错误，并受到使用目的和环境条件的影响。

这一阶段的“兴趣培养”包含三个层次。第一层是好奇：愿意观察人工智能现象，提出“它怎样做到”的问题。第二层是理解：知道智能表现背后存在人类设计和技术机制，不把机器能力理解为魔法。第三层是科学精神：尊重事实和证据，愿意验证，理解科学成果来自积累、合作、失败和修正。

小学阶段应帮助儿童形成三个基本认识：(1) 人工智能是人设计和训练出来的系统；(2) 人工智能的结果需要检查，它有时会出错；(3) 使用人工智能要注意安全、隐私、诚实和对他人的尊重。这三个认识为后续学段的技术学习和伦理判断提供共同起点。

2. 内容设计：广泛接触，具体呈现

小学课程可以覆盖五部分内容，其中历史脉络纳入基础概念。内容范围可以较宽，理解深度主要通过呈现方式和任务要求进行调节。

基础概念适合从智能机器的梦想、人类智能的多样性、自动化与人工智能的区别、人工智能发展中的人物和事件进入。历史故事可以把技术发展呈现为人类长期探索的过程，也能传递实事求是、开放分享、坚持研究和团队合作等科学价值。

人工智能方法适合从规则、分类、比较、奖励和“从例子中学习”等直观思想进入。学生可以通过卡片分类、人工模拟决策树、寻找共同特征、调整简单规则等活动，体验知识驱动和学习驱动的差异。小学高年级可以接触数据、模型、训练、监督学习、无监督学习、强化学习和神经网络等概念，但应减少公式推导和形式化细节，强调它们解决什么问题、依赖什么条件、可能在哪里出错。

人工智能应用应从生活经验出发，选择学生能够观察和讨论的系统，如人脸识别、车牌识别、语音识别、机器翻译、扫地机器人、自动驾驶、搜索与推荐、图像和文本生成。每个应用都要同时呈现能力与局限，让学生知道“会做”不等于“总能做对”。

交叉融合可以用医学、科学研究、环境保护、太空探索、建筑和艺术等故事扩展视野。小学阶段的主要任务是让儿童看到人工智能可以进入不同领域，科学家、工程师、医生、艺术家和普通人可以共同解决问题。课程不必要求儿童掌握复杂的学科模型，但应保留学科证据和人的判断位置。

伦理内容需要提前进入日常情境。信息伪造、隐私泄露、推荐形成的信息局限、生成内容的真实性、作品署名和责任等问题，都可以用儿童可理解的案例讨论。伦理学习应与具体行动相连，例如不输入个人敏感信息、不冒用他人的照片和声音、说明人工智能参与了作品生成、遇到重要信息时向可信成人和官方渠道核实。

3. 教学方式：故事、体验、观察与讨论

小学阶段适合采用故事化、体验式和低门槛的学习活动。科学人物故事能够把概念放入历史情境；实物和“无屏”模拟能够减少对复杂软件的依赖；图片、声音、分类卡片、角色扮演和小型制作可以把抽象机制变为可观察过程；讨论和绘画能够帮助儿童表达自己的理解。

体验活动需要教师组织和解释。开放式生成工具具有内容不确定、数据使用不透明和依赖风险，小学生在缺少保护的情况下不宜独立深度使用。《中小学生成式人工智能使用指南（2025 年版）》强调分学段规范和安全边界，其中小学阶段不允许学生独自使用开放式内容生成功能 (教育部基础教育教学指导委员会, 2025b)；UNESCO 也主张教育系统根据年龄与发展阶段设置适当使用条件 (Miao and Holmes, 2023)。儿童人工智能治理框架进一步强调隐私、安全、公平、包容和儿童最佳利益 (UNICEF Innocenti, 2025)。因此，小学实践应优先使用教师筛选的材料、封闭或受控环境、匿名数据和明确任务。

小学评价应关注学生能否描述现象、提出问题、比较差异、说明简单理由和遵守安全规范。作品的技术复杂度不宜成为主要标准。一个学生能够指出“这个识别结果可能受光线影响”，能够在转发图片前追问来源，已经表现出重要的人工智能素养。

14.2.2 初中阶段：体系构建与全局视野

1. 主导目标：把零散现象组织成知识结构

进入初中后，学生的数学、科学、历史、信息科技和道德法治学习逐渐系统化，能够处理更复杂的因果关系、规则体系和抽象概念。与此同时，他们在网络平台和生成式人工智能环境中的自主活动增加，需要更清楚地理解系统如何影响信息、判断和选择。

初中阶段的核心任务是把小学阶段的丰富体验组织成较完整的人工智能知识体系。学生需要知道人工智能从何而来、经历了哪些主要发展阶段、有哪些基本方法、形成了哪些典型应用、正在面临哪些社会问题。全局视野能够防止学生把某个当前流行的产品等同于人工智能整体，也能帮助他们理解不同技术路线各有条件和局限。

2. 知识主线：历史脉络与方法结构

人工智能历史可以成为初中课程的第一条主线。从智能机器设想、形式逻辑和计算机诞生，到达特茅斯会议、专家系统、机器学习、深度学习、大模型和智能体，学生可以看到人工智能的发展具有高潮、低谷、路线竞争和方法融合。历史学习同时帮助学生

理解，技术突破依赖理论、算力、数据、工程和社会需求的共同变化。

方法体系可以围绕两类基本路线展开。基于知识的方法通过规则、事实、搜索和推理解决问题；基于学习的方法通过数据、样本和反馈形成模型。学生需要了解对弈、定理证明、专家系统和知识图谱，也需要理解监督学习、无监督学习、强化学习以及分类、回归、聚类等基础任务。课程目标是形成知识地图，技术细节应服务于概念联系。

从方法论的角度，“数据–模型–输出–评价”可以成为初中阶段的核心分析链：

现实问题 → 数据采集与表示 → 模型或规则 → 输出结果 → 评价与修正

围绕这条链，学生学习数据从哪里来、如何标注、是否具有代表性；模型怎样从数据或知识中形成判断；输出为何可能存在误差；准确率等指标能说明什么，又遗漏了什么；真实使用环境为何可能与训练环境不同。这样，人工智能的“聪明”被还原为可以解释和检验的技术过程。

3. 应用与伦理：从案例讨论走向机制分析

初中阶段仍应学习典型应用和前沿成果，但分析重心要进一步进入机制。人脸识别可以联系图像表示、特征提取、相似度和隐私；推荐系统可以联系用户数据、排序目标、反馈循环和信息茧房；生成式人工智能可以联系训练数据、概率生成、幻觉和内容责任；自动驾驶可以联系感知、决策、控制和责任分配。

伦理内容既要设置专门板块，也要进入技术学习过程。信息伪造需要同时讨论生成机制、鉴伪方法、传播责任和法律规则；信息泄露需要联系数据采集、存储、模型输入和个人保护；社会公平需要分析数据代表性、评价指标和受影响群体。学生应逐步学会识别利益相关者，区分技术缺陷、使用不当和制度安排，说明自己的判断依据。

4. 实践方式：项目化问题解决，编程作为拓展

初中实践可以采用数据标注、小型分类实验、推荐模拟、流程图设计、生成内容核验、案例调查和规则制定等形式。项目化学习能够把知识、合作、证据和成果表达组织在真实问题中 (Krajcik and Blumenfeld, 2006)。项目应保留完整的问题解决过程：明确问题、收集资料、设计方案、实施测试、分析结果、修正方案、说明局限。

编程可以帮助学生深化算法与模型理解，也能为有兴趣的学生提供进一步发展的路径。面向全体学生的基础目标应放在系统解释、数据判断、误差分析、结果评价和责任意识上。复杂代码量和工具熟练度不宜成为人工智能素养的主要衡量标准。

初中评价可围绕解释性、迁移性、判断性和自醒性四类评价任务展开。学生需要能够解释一个简单系统的工作链条，把分析框架用于新的应用，比较不同数据和方案的影响，并反思自己在使用人工智能时是否出现过度信任、依赖或忽略证据的情况。

14.2.3 高中阶段：知识拓展与创新实践

1. 主导目标：从系统理解进入关键技术和真实世界

高中阶段的“知识拓展”具有三重含义。第一，技术理解从基本流程进入关键机制，学生能够分析神经网络、深度学习、生成式人工智能、多模态模型、推理模型、智能体和人机协同等方法的基本原理及演进关系。第二，应用学习从功能介绍进入系统拆解，

学生能够解释机器视觉、机器听觉、语言处理、搜索、推荐和人机对战等系统如何组合多种技术。第三，学习空间从人工智能学科内部扩展到数学、物理、材料、化学、生物、医学、天文学、人文和艺术等领域。

知识拓展追求理解深度和迁移范围，同时保持通识性质。高中课程可以进入公式、网络结构、损失与评价、训练与测试等内容，但不需要压缩复制大学人工智能专业课程。课程应帮助全体学生获得解释复杂技术和参与社会判断所需的共同基础，也为有专业兴趣的学生提供进一步探索方向。

2. 关键技术：理解演进、组合与边界

高中阶段应把现代人工智能方法放入完整知识体系中理解。学生可以比较人工特征设计与表示学习，理解深度网络如何进行层次性特征提取；可以从语言模型进入大语言模型，理解预训练、上下文生成、检索和工具调用；可以分析多模态模型如何把文字、图像、声音和动作纳入共同系统；可以理解智能体如何加入目标、记忆、规划、行动和反馈。

技术学习应同时呈现发展过程和失效条件。模型性能来自数据、结构、训练、计算资源和评价方式的共同作用。更大的模型可能获得更强能力，也可能带来更高成本、可解释性不足、偏差、对抗脆弱性和治理困难。学生需要学会说明结论成立的实验条件，区分基准性能与真实环境表现，理解局部指标与系统可靠性之间的距离。

3. 交叉融合：从技术迁移进入学科证据

高中阶段适合系统学习人工智能如何进入其他学科。每个跨学科案例都应呈现一条完整链条：

领域问题 → 问题表示 → 数据与知识 → 方法选择与改造 → 领域验证 → 解释与责任

在数学中，模型生成的猜想和证明需要接受形式化验证；在材料和化学中，结构预测和反应分类需要符合物理化学规律并经过实验；在生物和医学中，模型输出需要结合机制知识、临床证据和安全要求；在天文学中，异常检测和分类需要区分观测信号、仪器噪声和人为干扰；在人文学科中，识别和生成结果需要进入史料、语境、作者和文化价值的讨论。

这样的学习能够扩展学生的专业视野，也使他们理解人工智能的通用性具有条件。一个方法在某种数据上有效，并不意味着可以直接套用到所有学科。领域知识、问题表示和验证标准决定迁移是否成立。

4. 创新实践：从完成作品走向研究性探究

高中实践应提高开放性和证据要求。学生可以围绕校园、社区、环境、健康、文化或学科学习中的真实问题开展项目，进行数据整理、模型实验、系统设计、用户调研和方案迭代。研究性探究要求学生说明问题来源、方法依据、评价指标、实验结果、失败原因和适用边界。

创新实践还包括对现有系统的改进。学生可以比较两种模型或两套数据，分析公平性和鲁棒性；可以设计人机协同流程，明确哪些环节交给模型、哪些环节由人复核；可以为高风险场景提出数据、权限、监督和申诉机制。创新由此与责任联系起来。

5. 复杂判断与专业方向

高中生开始形成较明确的学科兴趣和未来发展想象。人工智能通识教育应通过不同领域的前沿问题，让学生了解人工智能研究者、领域科学家、工程师、医生、教师、法律工作者、设计师和公共治理者如何参与智能社会。专业方向指导的目的在于提供真实而多元的选择信息，避免过早把学生限制在单一技术路径上。

14.2.4　五部分内容在三个学段中的螺旋展开

五部分内容在三个学段都应出现，差异体现在问题深度、知识组织和学习任务上。

表 14.2: 五部分内容的学段递进

内容部分	小学：兴趣培养	初中：体系构建	高中：知识拓展
基础概念	通过故事和生活现象认识智能机器、人类智能、人工智能与自动化；形成初步历史感	系统学习学科诞生、发展阶段、主要路线和易混淆概念，建立全局坐标	结合当前前沿和未来趋势分析学科边界、技术演进及人工智能与社会结构的关系
人工智能方法	体验规则、分类、比较、奖励和从例子中学习；用类比认识数据、模型与神经网络	建立基于知识与基于学习的总体结构，理解数据–模型–输出–评价链条及基本学习任务	深入理解深度学习、大模型、多模态、推理模型、智能体和系统组合，分析性能与边界
人工智能应用	观察身边应用，知道人工智能能帮助人，也会出错	拆解典型系统的输入、表示、模型、输出、评价和反馈，分析错误来源	分析应用背后的关键技术、系统架构、方法演进和真实部署条件
交叉融合	通过医学、科学、环境、艺术等故事扩展想象，理解多人多学科合作	了解人工智能在不同学科中的基本应用，初步识别数据形态和证据差异	完成问题表示、方法迁移、领域约束、实验验证和科学解释的跨学科探究
风险、伦理与责任	形成安全、隐私、诚实、尊重和核查习惯；学习在需要时寻求成人帮助	分析伪造、泄露、茧房、偏差、公平和责任，说明判断依据与相关主体	面对高风险和价值冲突进行证据权衡、制度设计、专业责任和公共治理讨论

螺旋展开还应体现在案例选择中。同一个应用可以跨学段使用，但任务必须变化。以推荐系统为例，小学可以观察“平台为什么总给我看相似内容”，形成拓宽信息来源的习惯；初中可以模拟用户行为数据、相似性和反馈循环，解释信息茧房如何形成；高中可以进一步讨论目标函数、平台商业逻辑、舆论结构、用户自主和治理责任。案例保持连续，认知任务持续升级。

14.3 教学模式

14.3.1 课程组织与教学实施原则

保持稳定主干，开放吸收前沿：人工智能技术快速变化，课程应以稳定概念和基本问题为主干，再把新模型和新应用放入既有结构。大模型、多模态、推理模型和智能体进入课程时，需要说明它们在人工智能知识体系中的位置、继承了哪些方法、解决了什么问题、产生了什么新风险。这样可以减少追逐产品名称带来的课程波动。

独立课程与学科渗透并行：人工智能通识教育需要独立课程承担系统概念、方法结构和伦理框架，也需要在语文、数学、科学、历史、艺术、信息科技和综合实践中形成学科连接。独立课程保证体系完整，学科渗透提供迁移情境。二者共同支持学生理解人工智能的基本原理和应用场景。

科技教育与人文教育相互支撑：技术原理和社会判断需要在同一课程中连接。学习数据和模型时应讨论数据权利、代表性和目标选择；学习生成和推荐时应讨论真实性、作者、注意力和公共影响；学习医学和自动驾驶时应讨论专业责任和受影响者权益。伦理由此成为理解技术的重要维度，技术理解也为伦理判断提供事实基础。

共同基础与弹性拓展结合：人工智能通识教育面向全体学生，需要规定共同基础：基本概念、历史脉络、数据与模型意识、典型应用分析、跨学科视野、风险责任和认知自治。编程、复杂数学、硬件制作、模型训练和研究项目可以设置分层任务，为不同兴趣和基础的学生提供拓展。共同基础保证教育公平，弹性拓展支持个性发展。

实践活动保留完整认知过程：实践不宜停留在点击工具和展示作品。小学活动要包括观察与表达，初中项目要包括问题、数据、方案、评价和反思，高中探究要包括假设、实验、证据、迭代和边界。学生需要留下过程记录，使学习成果能够显示“为什么这样做”“依据是什么”“哪里失败”“怎样修正”。

评价与学段目标保持一致：评价应面向认知自治的最终目标，并根据学段调整深度。小学主要考察现象描述、初步质疑和安全习惯；初中主要考察系统解释、框架迁移、偏差识别和过程反思；高中主要考察复杂系统分析、跨学科证据处理、方案改进和责任判断。

14.3.2 认知自治的学段递进

贯通式课程最终服务于认知自治能力的培养。来源警觉、证据敏感、边界意识、信任校准、认知能动和责任判断六个分析维度贯穿三个学段；归纳、泛化、判断与自醒四种能力则随学习材料和任务复杂度逐步深化。小学建立面对技术的初步认知距离，初中把质疑落实为机制化分析，高中进一步发展能够核查证据、说明边界并承担责任的认知行动。

表 14.3: 四种认知能力的学段递进

能力	小学	初中	高中
归纳	从多个生活案例中发现人工智能的共同现象	理解机器如何从样本、规则和反馈中形成规律	分析模型归纳的条件、偏差、泛化和不确定性
泛化	把“需要检查”的意识用于不同工具与信息	把数据–模型–输出–评价框架迁移到新应用	把人工智能方法迁移到跨学科问题，并控制迁移边界
判断	区分帮助与风险、真实与可疑、安全与不安全	比较数据、证据、指标和错误代价，说明判断理由	权衡多方利益、风险等级、专业证据、制度安排和责任
自醒	觉察自己是否轻信、泄露信息或把机器当成人	反思自己是否过度依赖输出、忽略来源和反例	监控人机协作中的认知分工、信任校准、价值立场和责任位置

14.4　本章小结

中小学人工智能通识教育以五部分内容保持课程完整，以小学、初中、高中三个阶段形成螺旋递进。小学重在兴趣培养与科学精神，初中重在体系构建与全局视野，高中重在知识拓展与创新实践。各学段通过不同深度的解释、迁移、判断和反思任务，共同推动学生从亲近并初步质疑人工智能，逐步走向理解系统、检验证据、控制迁移边界并承担责任。

第十五章　大学阶段的人工智能通识教育

摘要：大学阶段的人工智能通识教育在基础教育形成的基本认知之上，重点发展方法理解、专业迁移和跨学科协作。课程应系统讲授人工智能的主要方法及其研究范式，并以数学、工程学、物质科学、生物与医学、地球与空间、社会、文化与艺术等方向的公共案例，训练学生分析领域问题、方法适用条件、证据标准和责任边界。归纳、泛化、判断与自醒四种能力的训练对象由一般问题扩展到专业知识的形成、检验和应用。教学同时保留理论学习、公共案例和实践活动；在学生基础差异较大的过渡阶段，可采用“概念—方法—应用—融合”的结构，逐步提高方法深化、专业迁移和跨学科实践的比重。

关键词：大学人工智能通识教育；方法体系；跨学科融合；案例分析；实践教学；项目学习；课程过渡；认知自治

15.1　指导思想

15.1.1　大学人工智能通识课程目标

大学阶段的学生已经进入具体专业，开始接触本学科的问题体系、研究方法和证据标准。与基础教育相比，大学人工智能通识教育需要提高理论深度，并把人工智能放入专业知识和真实实践中理解。课程目标可以概括为四个方面：从通用知识内容转向专业内容、理解人工智能的基础思维方式和研究范式，掌握人工智能与学科融合的分析框架，形成跨学科合作的思维习惯。

1. 从知识拓展走向方法理解与专业迁移

高中阶段可以帮助学生建立较完整的人工智能知识结构，认识人工智能的主要方法、应用和社会影响。在此基础上，大学阶段还需要进一步追问：人工智能有哪些基础的思维方式，这些思维方式如何指导问题的发现与解决，进入具体领域中如何产生影响，如何用共同的思维方式组织协同创新。

大学阶段的学习因而沿着如下方向深化：**方法体系的深入理解 → 专业问题的迁移分析 → 跨学科合作与创新**

这里的“方法体系理解”指向人工智能中相对稳定的问题解决方式。学生需要能够把一个新的模型或系统放回人工智能的方法谱系中，说明它继承了什么思想、解决了什么困难，又在哪些条件下成立。

2. 理解人工智能的基础思维方式和研究范式

人工智能的方法不断变化，背后存在若干相对稳定的思维方式。第五章总结了人工智能领域七种典型的思维方式：对现实问题形式化、从数据中学习规律、重视领域知识的约束、在不确定性中作出判断、权衡性能与效率、优先采用简单方法、通过实践检验解决方案。这些思维方式不宜被安排成彼此分离的知识单元。它们应进入完整的方法讲解：在搜索算法中理解问题分解和计算规模，在统计学习中理解数据归纳和泛化，在实际案例分析中理解知识约束、实验验证及能力边界。每一章都可以回到这些问题，使学生逐步形成稳定的分析习惯。

3. 掌握人工智能与学科融合的分析框架

大学生不仅需要知道人工智能已经进入哪些领域，还要理解人工智能怎样进入这些领域。第三篇提出的六环节框架可以在大学阶段转化为分析论文、系统和实践方案的共同工具：

领域问题 → 问题形式化与表示 → 数据与知识 → 方法选择与改造 → 领域验证 → 解释与责任

大学阶段需要提高每个环节的专业要求：准确界定领域困难，说明人工智能介入的位置，对复杂任务进行拆分，对问题进行合理的形式化，识别数据、知识和专业约束，分析通用方法为何需要改造，并按照本领域的证明、实验、临床或专业程序验证结果。学生还要说明适用条件、错误后果以及不同参与者的责任。

4. 形成跨学科合作的思维习惯

人工智能与学科融合通常依赖多种知识。领域专家负责提出有意义的问题、提供专业知识和确认结果；人工智能研究者负责模型、算法和计算系统；数据和实验人员负责建立证据；伦理、法律、管理和公共政策人员处理权利、风险与实施条件。

大学通识课应训练学生准确表达本专业问题，理解其他专业使用的概念和证据，并在合作中形成共同的问题定义、评价标准和责任安排。跨学科合作不要求每个人掌握全部知识，但要求学生知道自己的判断依据，也知道哪些结论需要交给其他专业成员验证。跨学科学习的核心在于整合不同领域的认识方式和证据标准，而非把不同专业的工作结果简单拼接在一起 (Borrego and Newswander, 2010)。

15.1.2 明确通识课程的边界

人工智能专业课程侧重算法推导、模型设计、软件实现、系统优化和技术创新。人工智能通识课程侧重方法体系的整体理解、模型条件和边界判断、专业迁移、跨学科协作以及公共责任。通识课程可以使用必要的数学、代码和实验，但这些内容服务于理解和判断，不以形成统一的专业技术能力为目标。

大学课程需要避免四种倾向。

1. **专业课程浅化**。简单删减公式和代码不能形成通识课程，课程目标和内容组织需要重新设计。
2. **专业名词堆砌**。课程应建立概念关系和方法谱系，不能用模型和算法名称的数量衡量内容深度。

3. **人工智能工具化**。工具可以进入学习过程，但提示词、平台操作和作品生成不能取代方法理解、证据核查和责任判断。
4. **追逐前沿名词**。快速变化的模型、产品和平台可以作为案例，稳定的课程主干应建立在计算、知识、学习、表示、行动和验证等基本问题上。

15.1.3 大学阶段的认知自治

全书关于认知自治的正式定义同样适用于大学阶段，即在充分利用他人知识、制度资源和人工智能系统的同时，保持认知能动，并持续调节来源、证据、边界、信任和责任。大学阶段的特殊之处，在于训练对象从日常信息扩展到论文、数据库、实验记录、专业案例和真实系统，并进入问题构思、文献检索、数据分析、代码生成、实验设计和专业表达等知识生产环节。

大学阶段继续沿用归纳、泛化、判断与自醒四种能力。归纳要求从多源专业证据中形成可检验的认识；泛化要求检验结论能否跨样本、场景、群体、时间和学科成立；判断要求在证据不完整、方法竞争和后果不确定时形成专业结论与行动门槛；自醒要求持续观察人工智能如何影响问题框架、分析路径、信心水平和责任位置。课程可以用共同基础、学科任务和综合项目承载四种能力的完整训练，并通过过程记录、依赖审计和口头答辩落实学术规范与人机协作责任 (Miao and Holmes, 2023; Miao, Shiohira, and Lao, 2024)。

表 15.1: 大学阶段四种认知能力的训练重点

能力	大学阶段的深化	人工智能通识课的训练内容	可观察的学习证据
归纳	从论文、数据、案例和实验中形成概念、规律、假设或解释	多源材料综合、样本与测量分析、相关与因果辨析、竞争性解释与假设生成	证据图谱、可检验假设、反例和验证方案
泛化	判断结论能否跨样本、场景、群体、时间和学科成立	分布变化、外部有效性、跨情境测试、方法迁移与失效边界分析	适用条件、新情境测试和边界说明
判断	在证据不完整、方法竞争和后果不确定时形成专业判断	问题界定、来源核查、证据分级、方法比较、不确定性表达和行动门槛	有依据的结论、保留意见、责任说明和答辩
自醒	识别人工智能对问题框架、思维路径、信任和责任的影响	人机过程记录、依赖审计、无人工智能重建、反思报告和口头答辩	能区分自身与人工智能的贡献，解释采纳或拒绝建议的理由

15.2　内容编排

大学课程仍以第三篇的五部分内容体系为基础，其中基础概念和应用分析嵌入各教学单元，风险、伦理与责任采用独立专题与纵向贯穿相结合的方式。课程主体可以分为两个部分：第一部分系统讲授人工智能基础方法，建立共同知识和分析工具；第二部分讲授人工智能与学科交叉融合，使学生把共同方法带入不同领域。

15.2.1　第一部分：人工智能基础方法

第一部分建议由六章构成。这六章形成从学科坐标、计算基础、知识驱动、数据学习、表示学习到基础模型和行动系统的递进关系。

人工智能概论：起源、发展与基本范式

本章建立全课程的认识坐标，主要包括：

- 人工智能的思想起源与学科形成；
- 人工智能发展的主要阶段及其兴衰原因；
- 人工智能的科学、工程和交叉学科属性；
- 人工智能与自动化、机器人、机器学习、深度学习、大模型和智能体等概念的关系；
- 符号方法、贝叶斯方法、连接主义和进化仿生方法等主要传统；
- 本章提出的基础思维方式和研究范式。

本章应给出整个课程的方法地图，使学生看到大模型是人工智能发展中的重要阶段，同时理解人工智能还包含知识、搜索、学习、优化、控制和实验等广泛内容。

计算、复杂性与智能问题

本章引入计算和复杂性的基本概念，主要包括：

- 信息、数据、程序和通用计算机；
- 算法和计算过程；
- 可计算性及其基本边界；
- 时间复杂度与空间复杂度；
- 多项式增长、指数增长和组合爆炸；
- 问题规模、计算资源与现实可解性；
- 分解、启发式、近似、随机化、并行和仿真；
- 学习方法作为应对复杂问题的一种路径。

本章需要使学生理解，理论上可以计算的问题仍可能因资源需求过大而无法直接求解。人工智能的重要作用之一，是利用启发信息、数据经验和大规模计算在复杂空间中寻找可用解。

知识表示、推理与搜索

本章讲授知识驱动的人工智能，主要包括：

- 状态、动作、目标、规则和约束；
- 状态空间和问题表示；
- 无信息搜索与启发式搜索；
- 对弈搜索、规划和约束满足；
- 命题逻辑与谓词逻辑；
- 产生式规则和专家系统；
- 语义网络、知识图谱和知识库；
- 知识获取瓶颈、组合爆炸、规则冲突和不确定性。

问题表示在本章中获得具体含义：现实任务被转化为什么状态空间，系统能够采取哪些行动，搜索过程怎样评价候选，知识怎样减少无效计算。人工智能对人类问题求解过程的借鉴，也可通过搜索、规划和专家系统得到具体说明。

统计学习与决策

本章讲授机器怎样从数据和反馈中形成规律，主要包括：

- 任务、数据、模型、目标函数、学习算法、领域知识和评价指标；
- 监督学习、无监督学习、自监督学习和强化学习；
- 分类、回归、聚类和降维；
- 线性模型、决策树、概率模型和集成方法；
- 训练集、验证集与测试集；
- 过拟合、欠拟合、泛化和模型选择；
- 偏差与方差、不确定性和分布变化；
- 基线、对照、交叉验证、消融实验、独立测试和错误分析。

实验验证应在本章形成系统内容。学生需要理解，模型性能来自数据、假设、算法和评价条件的共同作用；一个结果只有经过合理比较和独立测试，才能支持相应结论。

神经网络与深度学习

本章围绕表示学习展开，主要包括：

- 生物神经系统的启发及其工程化改造；
- 人工神经元、感知器和多层网络；
- 非线性表示、损失函数、梯度下降和反向传播；
- 表示学习和层次性；
- 卷积网络与空间结构；
- 循环网络与序列结构；
- 自编码器和生成模型；
- 注意力机制与 Transformer；
- 数据、算力、算法和模型结构的共同作用；

- 可解释性、对抗脆弱性和分布外失效。

本章的核心是理解深度学习带来的方法变化：模型能够从数据中自动形成多层表示，大量简单计算单元通过连接和训练产生复杂功能。卷积网络、循环网络和 Transformer 等结构应作为不同问题结构的代表，避免平铺大量模型名称。

大模型与智能体系统

本章讨论人工智能从专用模型走向基础模型和行动系统，主要包括：

- 大规模自监督预训练；
- 数据、规模与计算之间的关系；
- 迁移、提示和上下文学习；
- 指令调整和后训练；
- 多模态表示、对齐和融合；
- 检索增强和外部知识；
- 工具调用、代码和专业软件；
- 记忆、规划、行动和反馈；
- 智能体的目标、权限、停止条件和人工监督；
- 多智能体分工与协作；
- 幻觉、错误累积、系统安全和责任。

智能体应被理解为模型、知识、记忆、工具、环境和监督机制共同构成的系统。课程需要说明，从生成内容走向连续行动以后，错误可能在多个环节传播，系统可靠性需要通过权限控制、外部验证和人工监督共同保障。

表 15.2: 大学人工智能通识课程第一部分的内容结构

章节	核心问题	主要内容
人工智能概论	人工智能是一门什么学科	起源、历史、学科属性、概念关系和基本范式
计算、复杂性与智能问题	计算为什么能够处理智能问题，又受到什么限制	通用计算、可计算性、复杂度、规模增长和求解策略
知识表示、推理与搜索	人明确掌握的知识怎样转化为机器能力	状态空间、搜索、规划、逻辑、规则和知识系统
统计学习与决策	机器怎样从数据和反馈中形成规律	统计模型、学习方式、泛化、不确定性和实验验证
神经网络与深度学习	模型怎样自动形成复杂表示	神经网络、反向传播、表示学习和典型深度结构

章节	核心问题	主要内容
大模型与智能体系统	通用模型怎样连接知识、工具和环境	预训练、多模态、检索、工具、规划、行动与监督

每章可以设置“方法论回看”，要求学生回答：这一方法怎样表示问题，依赖什么计算过程，从哪里获得知识或规律，领域知识如何进入系统，结论通过什么实验得到支持，哪些条件变化可能导致失效。前述基础思维方式由此纵贯六章。

15.2.2 第二部分：人工智能与学科交叉融合

第二部分以六个学科方向为纲，每个方向设置一个教学单元，并在其中组织若干由教师讲授和研讨的公共案例。公共案例承担建立共同认识和比较不同学科证据标准的任务。学生另行查找的论文和系统用于自主案例研读和实践，不替代教师组织的公共案例。

六个方向的具体知识已在第三篇展开。大学阶段的内容编排应突出各领域的证据标准，并通过公共案例建立比较框架。

表 15.3: 大学阶段六个交叉方向的教学重点

学科方向	人工智能介入的代表环节	领域验证与责任重点
数学	模式发现、反例搜索、猜想生成、证明策略和形式化验证	数据模式可以提供猜想与搜索方向，数学结论仍需严格证明；可选用机器学习辅助猜想、几何辅助构造和自动定理证明等案例 (Davies, Veličković, Buesing, et al., 2021; Trinh et al., 2024)
工程学	感知、系统建模、设计优化、规划控制、故障诊断和维护	评价需覆盖实时性、稳定性、极端工况、冗余和人工接管，并经过仿真、试验和现场测试
物质科学	数据处理、参数反演、结构与性质预测、候选生成、工艺优化和实验规划	模型需符合守恒、对称、量纲、边界条件和化学可实现性，候选结果需经过模拟、制备与实验验证 (Kench and Cooper, 2021; Schwaller, Laino, et al., 2019; Schwaller, Probst, Vaucher, et al., 2021)

学科方向	人工智能介入的代表环节	领域验证与责任重点
生物与医学	序列和结构分析、影像处理、风险预测、药物发现与临床决策支持	需处理人群和设备差异、隐私、知情同意与公平；高风险结论需独立验证、临床试验和专业复核 (Jumper, R. Evans, Pritzel, et al., 2021; Jo, Park, Jung, et al., 2017; Ott, Hu, Keskin, et al., 2017; Sahin, Derhovanessian, Miller, et al., 2017)
地球与空间	遥感处理、异常检测、时空预测、系统重构、生态监测和任务规划	需处理观测噪声、多尺度变化、稀少极端事件和误差累积，并结合物理过程、多源证据与人工会商 (Reichstein et al., 2019; Kerrigan, La Plante, Kohn, et al., 2019)
社会、文化与艺术	社会资料分析、政策模拟、文化对象识别、数字化修复和人机协同创作	需分析制度与测量对数据的影响，并处理语境、解释、公平、隐私、版权、署名和文化多样性 (D. Lazer, Pentland, L. Adamic, et al., 2009; P. Wang et al., 2024; C.-Z. A. Huang, A. Vaswani, Uszkoreit, et al., 2018; Manovich, 2020)

15.2.3 风险、伦理与责任的内容位置

风险、伦理与责任采用独立专题和纵向贯穿相结合的方式。独立专题系统讲授数据与隐私、知识产权、公平、可靠性、安全、人类主体性、社会影响、高自主系统、责任链、审计和申诉。纵向贯穿要求学生在统计学习中分析数据偏差与评价指标，在深度学习中分析可解释性和对抗脆弱性，在大模型和智能体中分析幻觉、权限和错误累积，在六个学科方向中根据领域证据和错误后果讨论责任。

15.3 教学模式

大学人工智能通识课的教学模式需要由知识特点、学生特点和现实条件共同决定。人工智能知识具有系统性，方法之间存在前后联系；人工智能又具有鲜明的实践性和实验性，许多问题必须通过观察和比较才能深入理解。学生来自不同专业，数学能力、编程能力和研究经验差异明显。人工智能技术快速发展，大量知识分布在论文、开源项目和在线系统中。生成式人工智能也正在进入学生的检索、写作、代码和作品制作过程。这些因素共同决定，大学课程需要同时保留系统理论学习、公共案例分析和多样实践。

15.3.1 基本教学原则

系统理论学习

学生需要形成完整的人工智能方法体系，理解各类方法之间的关系、来源和边界。教师讲授承担建立共同概念、解释知识结构、澄清易混问题和展示分析过程的作用。理论内容应围绕方法解决什么问题、依赖什么表示和假设、怎样接受验证展开，减少孤立术语和公式的堆积。

教师组织的公共案例分析

公共案例应来自多个领域（如前述六个学科方向）。案例分析需要呈现领域问题、表示、数据与知识、方法选择与改造、领域验证、解释与责任的完整过程。教师通过公共案例建立共同分析标准，使学生能够比较数学证明、工程测试、材料实验、临床验证、地球观测和人文解释之间的差异。

实验与开放实践

人工智能具有实践和实验属性。学生需要通过可视化、数据比较、参数变化、错误分析、系统测试和小型项目观察方法怎样工作，判断理论结论在什么条件下成立。实践的核心是提出问题、作出预期、改变条件、记录结果和解释差异。实践形式可以包括：

- 可视化实验和模型机制分析；
- 数据、参数和评价指标的对比实验；
- 论文研读与关键结果复现；
- 文献调研和领域综述；
- 学术辩论、征文、审稿和报告；
- 系统或工具的能力与边界测试；
- 专业问题的可行性分析；
- 科研项目和现实问题导向的方案设计。

实践可以吸收重要论文、开放项目和新问题，但需要放入稳定的方法框架。学生自主案例研读、关键结果复现、系统能力测试和开放项目分别承担不同任务。合作项目应形成真实的知识互动，可以设置领域问题、人工智能方法、数据与实验、伦理法律和系统实施等角色，使不同专业成员共同完成问题定义和证据判断。

基础目标与弹性实践路径结合

大学人工智能通识课的学生基础多元，课程需要在共同目标之上提供弹性实践路径。

学生可以采用无代码工具、可视化分析、人工智能辅助代码、自主编程、文献研究或领域调研等不同路径，突出专业特点与个人兴趣。无论采用何种路径，都需要回答共同问题：任务是什么，方法为什么适合，使用了哪些数据、知识和假设，结果怎样得到验证，哪些条件下可能失效，人承担什么责任。

15.3.2 可供选择的实施方案

不同学校、课程规模、师资和资源条件差异明显，不宜规定唯一的大学人工智能通识课模式。以下方案可以分别采用，也可以根据需要选择其中部分组合。

方案一：讲授—案例—实验的模块化课程

每个知识模块按照“理论讲授—公共案例—小型实践—讨论总结”组织。该方案适合大班、基础差异较大或首次开设的课程。实践以一至两周的短任务为主，使学生多次接触不同方法和领域。

方案二：基础方法课程加阶段性跨学科项目

课程前半段建立共同方法基础，后半段安排三至五周的跨学科项目。学生在掌握分析框架后再选题组队，完成问题定义、文献调研、数据与知识分析、小型实验或方案、领域验证和责任分析。该方案能够提供较完整的项目经验，也避免学生在缺乏基础时过早选题。

方案三：专业通识课程

课程围绕人工智能与医学、材料、社会科学、艺术或城市治理等一个方向集中展开。先讲授共同人工智能基础，再深入分析该领域的应用环节、数据条件、方法迁移和证据标准。该方案适合院系内部课程或跨院系合作。

方案四：讲授、论文研读与研讨

教师用精炼讲授建立方法框架，学生围绕个人感兴趣的学科方向中的经典或前沿论文开展研读。课程成果可以是论文评述、领域综述、研究提案、海报展示或模拟学术会议。该方案适合小班、高年级学生和研究型大学。

方案五：创业式学习

课程首先讲授人工智能基础，之后在导师指导下组织创业型项目小组，分析人工智能介入环节、数据和知识条件、技术可行性、成本、组织流程、法律伦理、风险和落地路径。成果可以是可行性报告、实施方案或原型系统。这一方案适合应用型高校。

15.3.3　学习评价

大学的评价方式应该多元化，可以组合概念测试、案例分析、实验报告、论文评述、项目成果、同伴审稿、个人答辩和过程记录。最终作品只是学习证据的一部分，学生还需要解释判断、回应反例并说明结果边界。

15.3.4　过渡阶段方案

当前进入大学的学生接受人工智能教育的经历差异明显，许多学生尚未形成完整的人工智能知识体系。大学课程需要应对这一现实，同时保持大学阶段的方法深度、专业迁移和跨学科实践要求。

基础课程模块

过渡阶段可以参考高中版的知识安排，采用“概念—方法—应用—融合”的结构，兼顾共同基础和大学层次的深入分析。随着中小学人工智能教育逐步普及，概念和基础应用内容可以压缩，方法深化、学科融合和创新实践的比重逐步提高。四个模块沿用第三篇的知识结构，同时提高大学层次的证据要求：

表 15.4: 大学过渡阶段的四个课程模块

模块	共同内容	大学阶段的深化
概念	起源、历史、学科特点及自动化、机器人、机器学习、深度学习、大模型和智能体等概念关系	增加概念比较、历史原因和学科边界分析，建立共同语言与方法坐标
方法	计算与复杂性、知识与搜索、统计学习、神经网络、大模型、多模态和智能体	突出方法关系、理论条件和实验验证，分析泛化、不确定性与系统边界
应用	机器视觉、机器听觉、语言处理、具身行动、搜索与推荐等典型系统	加入论文、真实系统和错误案例, 分析系统组成、部署条件、反馈机制与责任
交叉融合	数学、工程学、物质科学、生物与医学、地球与空间、社会、文化与艺术	以公共案例和自主案例比较领域问题、方法改造、证据标准和责任边界

加入跨学科实践方案

过渡阶段至少应安排一次跨学科实践。较为稳妥的方式是在学生完成概念、方法和基本应用学习后，再设置阶段性项目。项目可以选择以下形式:

- 论文案例复现或结构化分析;
- 人工智能与本专业结合的可行性研究;
- 城市或行业人工智能落地方案;
- 某种模型在专业数据上的能力与边界测试;
- 人工智能辅助科研或专业工作的流程设计;
- 跨学科现实问题或创业方案。

跨学科实践的目标是让学生把共同方法用于专业问题, 并在真实知识分工中理解合作。项目成果应包括问题定义、资料来源、方法依据、实验或方案、验证过程、失败与风险记录、责任说明和个人反思。

过渡阶段的逐步调整

随着基础教育阶段人工智能课程逐渐普及，大学过渡课程可以逐步调整:

- 压缩概念和基础应用的重复内容;
- 提高计算、学习、深度模型和智能体的方法深度;
- 增加论文研读、实验分析和前沿调研;
- 提高交叉融合和专业迁移的比重;
- 使跨学科项目从基础体验走向研究和创新。

过渡阶段用于回应当前学生基础差异, 使课程逐步接近完整的大学人工智能通识教育目标，无需另行建立长期独立的课程体系。

15.4　本章小结

大学人工智能通识教育以方法理解、专业迁移和跨学科协作为主要深化方向。课程主体由人工智能基础方法和学科交叉融合构成，并以理论学习、公共案例和多样实践连接专业证据与责任判断。归纳、泛化、判断和自醒四种能力进入专业知识的形成、检验和应用过程。当前过渡阶段可采用“概念—方法—应用—融合”的结构；随着基础教育逐步完善，大学课程应压缩基础重复内容，提高方法深化、论文研读、专业案例、跨学科实践和创新研究的比重。

第十六章　面向终身学习的人工智能通识教育

摘要：终身学习贯穿人的整个生命历程，本章重点讨论学校教育之后的成人与公众学习，覆盖职业发展、家庭生活、社会参与和退休生活。其内容采用“共同核心—情境模块—前沿更新”的结构：共同核心提供稳定认识，情境模块连接学习、工作、家庭、健康照护、文化创造和专业实践，前沿更新吸收新成果、新应用和新问题。课程从现实场景进入，以基本原理支撑实践，并根据学习者的生活阶段、数字基础和学习条件提供差异化支持，使认知自治能够在技术变化中持续发展。

关键词：终身学习；人工智能通识教育；成人学习；情境学习；公共科普；数字包容；认知自治

16.1　指导思想

1. 终身学习的基本定位

终身学习贯穿人的整个生命历程。本章重点考察其在成年以后不同生活阶段中的实施。学习者可能处于职业发展、岗位转型、家庭照护、社会参与或退休生活之中，其人工智能学习需求会随生活环境、工作任务和技术变化不断调整。终身学习课程不能照搬学校课程的固定起点、统一进度和单一路径，而应在共同基础上提供持续进入、反复学习和自主选择的空间。

终身学习也不能被缩小为老年教育。老年学习者是需要重点关注的群体之一，在职人员、职业转型人员、家庭学习者和普通公众同样需要人工智能通识教育。不同群体的具体需求有所差异，但都需要理解人工智能的基本工作方式，能够合理使用人工智能，并对输出的真实性、适用范围和风险保持判断。

成人学习通常与既有经验、现实任务和自主目标紧密相关。教育学研究表明，学习内容与学习者当前面对的问题建立联系后，更容易转化为持续行动 (Knowles, Holton, and Swanson, 2015; Merriam and Bierema, 2013)。因此，终身学习课程需要把稳定原理与现实场景结合起来，使学习者能够在解决生活和工作问题的过程中理解人工智能，并把所形成的认识迁移到新的工具和任务。

2. 从一次性学习走向持续认知更新

人工智能技术、产品和社会应用持续变化。某一种工具的操作方法可能很快改变，但围绕数据、模型、生成、核查和责任形成的基本认识则具有较强稳定性。终身学习应

帮助学习者用基础概念理解新模型，把已有认识迁移到新工具，根据新的证据和任务修正判断，并在发现知识不足时主动补充资料或寻求专业帮助。

联合国教科文组织将终身学习理解为贯穿生命全过程、覆盖多种学习场景的持续能力建设，并强调个人、社区和社会共同提供学习机会 (UNESCO Institute for Lifelong Learning, 2020)。人工智能通识教育应形成可重复进入的学习结构，使学习者能够在技术变化以后重新返回基础概念、补充新案例并修正旧判断。

3. 以生活和工作问题带动学习

终身学习具有鲜明的问题导向。人工智能知识需要与学习者当前面对的任务发生联系，包括获取和核查信息、完成文字和图像表达、提高工作效率、理解行业变化、使用公共服务、管理健康和家庭事务、识别诈骗和虚假内容，以及参与人工智能公共讨论。

场景进入并不意味着功利化的工具使用。真实问题提供学习动机，基础概念用于解释现象，实践活动用于检验理解，风险分析帮助学习者形成边界。一个完整学习单元可以按照以下过程组织：

现实问题 → 人工智能任务 → 基本原理 → 实际操作 → 结果核查 → 迁移与反思

例如，使用人工智能规划旅行时，学习者不仅要学会表达人数、预算和行动限制，还要理解模型为何可能使用过期信息，哪些票务和开放时间必须通过官方渠道确认，以及个人行程信息应如何保护。

4. 终身学习中的认知自治

终身学习中的认知自治仍指向“开放依赖中的自主”。成年人可以充分利用人工智能、专家知识、公共机构和社会支持，保持认知能动，并持续调节来源、证据、边界、信任和责任，能够在必要时检查、调整或收回信任。

归纳能力表现为从多次观察中总结系统行为，避免以一次成功或失败概括全部情形；泛化能力表现为把对数据、模型、生成和核查的理解迁移到新工具，同时重新检查适用条件；判断能力表现为评价输出的可信程度、风险后果和专业边界；自醒能力表现为觉察知识不足、信任变化和工具依赖，并保留应由自己承担的选择与责任。

16.2 内容编排

16.2.1 总体结构

面向终身学习的内容建议采用“**共同核心** + **情境模块** + **前沿更新**”的组合架构。

共同核心提供稳定知识底座；情境模块面向现实任务；前沿更新吸收快速变化的技术和社会问题。三类内容之间并非简单并列。学习者可以从情境问题进入，在共同核心中获得解释，再通过前沿更新扩展认识。

这一结构用于组织终身学习的实施路径。第三篇确立的基础概念、人工智能方法、人工智能应用、交叉融合以及风险、伦理与责任五部分内容，仍然构成知识范围，并分别进入共同核心和情境模块。

表 16.1: 面向终身学习的人工智能通识内容结构

内容层次	主要作用	主要内容
共同核心	建立稳定的人工智能基础认知	概念与社会背景、身边系统、基础方法、大模型与智能体、风险、伦理与责任
情境模块	将共同知识带入现实生活与工作	学习与信息、工作发展、家庭生活、健康照护、文化创造、专业融合
前沿更新	支持持续认知更新	新方法、新模型、新应用、社会讨论热点

16.2.2 共同核心

人工智能与智能社会

本单元介绍人类对智能机器的长期追求，区分自动化机器与人工智能，说明人工智能的定义、学科特点和发展主线。内容应帮助学习者理解，人工智能研究通过计算模拟感知、学习、推理、规划、创造等智能行为，其普适性来源于智能活动在不同领域中的基础作用。

身边的人工智能系统

本单元从日常生活中的机器视觉、机器听觉、语音合成、机器人、搜索推荐和语言理解入手。代表性案例可以包括人脸和车牌识别、语音输入、声纹验证、地图导航、扫地机器人、自动驾驶、搜索排序、个性化推荐和智能客服。

每个案例应回答六个问题：

1. 系统输入了什么；
2. 系统完成什么任务；
3. 输出依据什么产生；
4. 哪些条件会影响结果；
5. 系统可能怎样出错；
6. 使用者需要承担什么责任。

这种分析能够把日常经验转化为结构化认识，使学习者看到人工智能已经进入生活，也看到系统能力具有明确条件。

人工智能的基础方法

终身学习课程不要求完整讲授专业算法，但需要建立必要的方法结构，主要包括：

1. 基于知识、规则和搜索的人工智能；
2. 基于数据和学习的人工智能；
3. 数据、模型、目标和评价的基本关系；
4. 训练、测试、过拟合和泛化；
5. 神经网络与层次表示；
6. 数据和计算能力对现代人工智能的作用；
7. 搜索、学习和行动反馈之间的关系；

苹果与橘子分类、图像的层次识别、围棋搜索与学习等案例，可以把抽象方法转化为可理解过程。教学深度应达到学习者能够用自己的语言解释方法，不要求统一完成数学推导或程序实现。

大模型与智能体

本单元介绍语言模型、大规模预训练、图像和视频生成、多模态模型、推理模型以及智能体。重点说明人工智能如何从特定任务模型发展为能够处理多种任务的通用接口，又如何借助外部知识、工具、记忆、规划和反馈执行较长流程。

智能体应被理解为由模型、目标、记忆、工具、权限、环境和监督共同组成的系统。学习者需要区分“生成一个答案”和“持续执行一个任务”，理解后者会带来错误累积、权限控制和责任追踪等新问题。

风险、伦理与责任

风险、伦理与责任属于共同核心。主要内容包括：

1. 信息伪造、深度合成和身份欺骗；
2. 个人信息、生物信息和工作数据泄露；
3. 推荐系统与信息环境；
4. 数据偏差、社会公平和数字排斥；
5. 自动生成内容的版权、署名和真实性；
6. 医疗、金融和公共服务中的自动化决策；
7. 人工智能依赖与人的主体性；
8. 高自主系统的权限、控制和责任。

风险教育应进入实际操作。学习者在使用问答、图像、声音和智能体工具时，需要练习敏感信息识别、来源核查、官方确认、多渠道验证、密码保护和高风险任务的人工复核。

16.2.3　情境模块

情境模块把人工智能放入熟悉的生活和工作任务，既帮助学习者形成直观理解，也用于分析技术介入的环节和影响。不同群体面对的任务不同，下面列举若干典型情境。

（1）人工智能辅助学习与信息获取

主要内容包括人工智能问答和搜索、问题表达、多轮追问、来源比较、事实和引用核查、阅读摘要、翻译、知识整理以及学习计划。学习者要理解，人工智能可以提高信息获取效率，也可能生成错误事实、虚假引用和过时建议。

实践可围绕陌生问题展开：先让人工智能提供答案，再选择其中的事实和来源进行核查，记录从初始答案到可靠结论的修正过程。

（2）人工智能辅助工作与职业发展

本模块讨论文档处理、信息调研、数据整理、会议沟通、基础自动化和工作流程改进。重点在于判断哪些任务适合人工智能参与，哪些结果需要专业人员核查，哪些组织数据和商业信息不能输入公共模型，以及人工智能提高效率以后人需要承担哪些新的工作。

不同职业的专门证据标准和责任要求，可以在面向专业与公共责任的拓展教育中进一步展开。本章主要提供终身学习的共同基础。

（3）人工智能服务家庭与日常生活

主要内容包括购物消费、旅行出行、政务与公共服务、手机和智能设备操作、家庭事务、家庭沟通、智能家居、信息与支付安全。现有实践方案已经形成从软件安装、界面识别和基本操作，到问答、写作、图像、视频、音乐和安全核查的渐进路径。

生活场景实践应强调表达限制条件和核查关键信息。例如，旅行规划需要说明年龄、预算、行动能力和休息需求，并对预约、票价和开放时间进行官方确认。

（4）人工智能与健康、养老及照护

主要内容包括健康信息查询、体检资料解释、运动与睡眠建议、家庭安全监测、疾病风险预测、辅助运动、陪伴型人工智能和药物研发。该模块需要明确健康科普、风险提示和临床诊断之间的边界。

学习者可以使用人工智能帮助整理就医问题、解释一般医学术语和制定生活记录，却不能仅凭模型输出自行改变处方或替代医生判断。隐私保护、知情同意、误诊风险和照护责任应随案例同步讨论。

（5）人工智能与文化创造

主要内容包括写作、诗歌、图像生成与修复、家庭影像、视频制作、音乐生成和人机共同创作。实践目标应从生成作品推进到明确意图、评价结果、反复修改、说明人的贡献并处理版权、隐私和真实性。

老照片修复、家庭回忆视频、节日贺卡和个人歌曲等任务能够激发学习兴趣，也适合开展代际合作。课程应提醒学习者避免改变人物真实身份和历史事实，公开传播时还要说明合成和修改情况。

（6）人工智能与专业领域融合

本模块面向已经具有职业或专业经验的学习者，讨论人工智能进入本领域的问题环节、工作流程、数据与知识条件、验证标准和责任边界。设计者可以选择一篇本领域论文、一个实际系统或一项工作任务，分析人工智能能够提供什么帮助、哪些环节需要重新设计、哪些结论必须由专业人员复核。

专业融合建立在共同核心之上，并根据行业规范和风险水平补充专门内容。医疗、法律、金融、工程、教育等领域需要分别遵循相应的证据标准、数据保护要求和职业责

任，不能以通用工具的输出替代专业程序。

16.2.4 前沿更新

前沿更新模块主要讨论人工智能的最新研究进展、人工智能新型应用、人工智能热点讨论。前沿模块不宜形成长期固定的产品和名词列表。课程应保留稳定概念主干，定期替换部分案例，建立动态阅读材料和更新记录，区分科研结果、产品宣传和趋势判断。内容更新的策略可以包括：

1. 保留基本概念、方法和责任原则；
2. 定期更新前沿案例和实践工具；
3. 淘汰失效的平台操作说明；
4. 保留具有长期解释价值的经典案例；
5. 建立教师和学习者反馈渠道；
6. 建设可共享、可分层和可持续维护的资源库；
7. 对医疗、安全和公共治理等高风险内容进行专业审核。

16.3 教学模式

终身学习者的年龄、教育经历、职业背景、数字能力和学习时间差异显著。成人往往带着明确问题进入学习，也会使用已有经验解释新技术。课程需要尊重这种经验基础，但也要帮助学习者修正过时认识。

人工智能本身具有较强实践性。模型输出的随机性、提示方式的影响、数据和场景变化造成的错误，都需要通过操作和比较获得直观理解。理论解释、案例分析和实践活动应相互连接。

16.3.1 基本原则

场景进入，原理支撑：学习可以从真实问题开始，随后补充解释该问题所需的概念和方法。场景提供动力，原理保证迁移。只讲操作会使学习依赖具体平台，只讲原理则可能远离成人学习者的现实需要。

学习与实践同步：每个知识模块都应配置观察、操作、核查或反思任务。实践活动需要服务于概念理解和能力形成，不能单纯追求作品数量。

模块组合，路径弹性：课程可以提供若干相对独立又能够组合的模块。学习者依据自己的目标和时间选择模块，也可以在需要时重新进入。模块之间需要共享基本术语、核查方法和责任原则。

小步推进，持续更新：成人学习可以采用短时、分散和重复的方式。每次学习解决一个明确问题，并保留进入下一步的接口。可以设计持续更新的实践手册，采用轻量、短周期和强操作的任务设计，便于在碎片化时间学习，并允许具备基础的学习者跳过入门内容。

同伴互助与代际协作：社区小组、家庭成员、同事和志愿者可以共同解决数字操作和生活问题。代际合作尤其适合家庭影像、文化记忆、防诈和健康信息核查等活动。合作的目的包括技术帮助，也包括交流不同生活经验和风险认识。

16.3.2 可选择的实施方案

个人自学：个人自学适合具有一定自主学习能力的成年人。可以使用基础读本、实践手册、在线课程和个人学习档案，按照“阅读—短实践—问题记录—核查—阶段总结”的过程推进。自学材料需要提供清晰路径、必要解释、常见错误和进一步学习资源。实践任务应允许学习者根据自身生活重新设题。

社区与老年大学课程：社区和老年大学课程适合退休人员、老年学习者和希望获得面对面支持的公众。教学可采用小班、慢节奏、清楚步骤、重复练习和同伴互助。内容应在生活便利、文化创造和风险防护之间保持平衡，避免将课程缩小为手机操作培训。

职场继续学习：职场继续学习可以围绕工作流程、信息和数据处理、人工智能协作、组织数据保护以及岗位变化展开。通识课程负责共同基础，行业和组织再依据工作任务增加专业内容和制度要求。

专业融合工作坊：专业融合工作坊适合具有一定领域背景的学习者。活动可以包括阅读本领域人工智能论文、分析一个应用案例、测试工具能力、设计小型可行性方案，并讨论专业验证和责任边界。工作坊应由人工智能教师与领域人员共同支持。

家庭与代际学习：家庭成员可以围绕信息核查、隐私设置、健康信息、智能服务、家庭影像和文化创作共同学习。年轻成员能够提供数字操作帮助，年长成员则提供生活经验、家庭记忆和价值判断。代际学习有助于减少数字排斥，也能防止单向技术灌输。

16.3.3 学习评价

终身学习不宜以统一考试为中心评价。评价应服务于学习者了解自己的进步，并为下一步学习提供依据。可采用：

1. 场景任务；
2. 学习档案；
3. 实践作品；
4. 风险判断和信息核查任务；
5. 口头解释；
6. 同伴互评；
7. 个人反思记录；
8. 专业或生活应用方案。

评价仍可以认知自治的四种能力为目标，重点考察学习者能否理解基本原理、完成真实任务、核查结果、迁移到新情境、保护信息和权利，并在人工智能辅助下形成有依据的自主判断。

16.4　本章小结

面向终身学习的人工智能通识教育以持续认知更新为基本特征，采用“共同核心—情境模块—前沿更新”的内容结构，并通过个人自学、社区课程、职场学习、专业工作坊、代际学习和公共科普提供多次进入的机会。课程从现实任务进入，以基本原理支持核查、迁移和反思，并依据生活阶段、数字基础和学习条件提供差异化支持。其目标是使学习者在主动利用人工智能工具的同时，持续修正认识、审慎判断并承担责任。

附录 A　小学学段

A.1　学段集成版

A.1.1　教材目录

本附录依据《清华大中小学人工智能通识教育系列》小学版教材目录整理。

序号	章题目	课题目
1	第一章从梦想到现实	偃师的故事
2	第一章从梦想到现实	加扎利的音乐团
3	第一章从梦想到现实	电影中的人工智能
4	第一章从梦想到现实	什么是人工智能
5	第二章身边的人工智能	高铁检票员
6	第二章身边的人工智能	电子交警
7	第二章身边的人工智能	美颜相机
8	第二章身边的人工智能	扫地机器人
9	第二章身边的人工智能	自动驾驶
10	第二章身边的人工智能	推荐系统
11	第三章人工智能前沿	AI 画家
12	第三章人工智能前沿	AI 作曲家
13	第三章人工智能前沿	AI 诗人
14	第三章人工智能前沿	AlphaGo 的故事
15	第三章人工智能前沿	OpenAI 和它的 ChatGPT
16	第三章人工智能前沿	Sora 的故事
17	第三章人工智能前沿	AI 天气预报员
18	第四章人工智能起源	亚里士多德的故事

接下页

表 A.1 （续）

序号	章题目	课题目
19	第四章人工智能起源	布尔的故事
20	第四章人工智能起源	图灵和图灵机
21	第四章人工智能起源	计算机的诞生
22	第四章人工智能起源	机器智能的最初设想
23	第四章人工智能起源	达特茅斯会议
24	第五章人工智能发展史	吴文俊的故事
25	第五章人工智能发展史	费根鲍姆的专家系统
26	第五章人工智能发展史	深蓝：成就巅峰
27	第五章人工智能发展史	深度学习兴起
28	第五章人工智能发展史	大模型时代
29	第五章人工智能发展史	走向未来
30	第六章人工智能基础	认识计算机
31	第六章人工智能基础	认识计算机程序
32	第六章人工智能基础	什么是算法
33	第六章人工智能基础	知识与智能
34	第六章人工智能基础	不会学习的机器不是好机器
35	第七章深度学习时代	皮茨和他的神经元模型
36	第七章深度学习时代	感知器：会学习的神经网络
37	第七章深度学习时代	杰弗里·辛顿的故事
38	第七章深度学习时代	李飞飞与 ImageNet 数据集
39	第七章深度学习时代	GPU：从游戏到人工智能
40	第七章深度学习时代	解析 AlphaGo
41	第七章深度学习时代	探索大语言模型
42	第七章深度学习时代	深度学习挑战：难以理解的智能
43	第七章深度学习时代	深度学习挑战：对抗样本
44	第七章深度学习时代	深度学习挑战：超级智能体

A.1.2 实践手册目录

本附录依据《清华大中小学人工智能通识教育系列》小学版教材实践手册（博爱新华小学）整理。原手册中的实践项目按章编排如下。

序号	章节	实践类型	实践题目
1.1	第一章	调研	智能机器的千年梦想
1.2	第一章	调研	寻找身边的 AI
2.1	第二章	情景剧	时光车站
2.2	第二章	体验感知	现代魔法镜——它真的让你更美了吗?
2.3	第二章	征文	AI 智能机器人
3.1	第三章	体验感知	节气小精灵图鉴
3.2	第三章	体验感知	丰富班级文化：设计班歌
3.3	第三章	体验感知	四季诗歌交流会
3.4	第三章	体验感知	AI 童话故事
4.1	第四章	手工制作计算机的诞生图	计算机进化隧道
5.1	第五章	征文	AI 带你闯未来
6.1	第六章	征文	穿越计算机时空之旅
6.2	第六章	AI 小剧场	AI 小智成长记
7.1	第七章	辩论	未来的超级智能体会是人类的朋友还是敌人?

A.2 年级分册版

本附录依据《清华大中小学人工智能通识教育系列》三至六年级教材整理，按上下册分别列出单元题目和节题目。

A.2.1 三年级上册

序号	单元题目	节题目
1.1	从梦想到现实	“偃师造人”的故事
1.2	从梦想到现实	加扎利的音乐团
1.3	从梦想到现实	电影中的人工智能
1.4	从梦想到现实	一个电饭锅的故事：什么是人工智能?
2.1	身边的人工智能（一）	高铁检票员
2.2	身边的人工智能（一）	电子交警
2.3	身边的人工智能（一）	美颜相机
3.1	身边的人工智能（二）	扫地机器人
3.2	身边的人工智能（二）	自动驾驶
3.3	身边的人工智能（二）	推荐系统
4.1	人工智能前沿	AlphaGo 的故事
4.2	人工智能前沿	OpenAI 与大语言模型
4.3	人工智能前沿	神奇的人工智能视频

A.2.2　三年级下册

序号	单元题目	节题目
1.1	人类智能	人类的智能有哪些
1.2	人类智能	智商与智商的测量
1.3	人类智能	情商：理性之外的智慧
2.1	人工智能的起源（一）	亚里士多德和他的三段论
2.2	人工智能的起源（一）	布尔的故事
2.3	人工智能的起源（一）	图灵和图灵机
2.4	人工智能的起源（一）	计算机的诞生
3.1	人工智能的起源（二）	机器智能的最初设想
3.2	人工智能的起源（二）	图灵测试
3.3	人工智能的起源（二）	图灵奖
4.1	人工智能的诞生	人工智能的诞生前夜

接下页

表 A.4 （续）

序号	单元题目	节题目
4.2	人工智能的诞生	达特茅斯会议

A.2.3 四年级上册

序号	单元题目	节题目
1.1	黄金十年	伊丽莎的成功
1.2	黄金十年	吴文俊的故事
1.3	黄金十年	机器翻译的开端
1.4	黄金十年	高潮落幕：人工智能的第一次低谷
2.1	第二次高潮	爱德华·费根鲍姆和专家系统
2.2	第二次高潮	梦想破灭：日本第五代计算机
2.3	第二次高潮	机器昆虫时代
3.1	稳步回暖	机器学习的兴起
3.2	稳步回暖	深蓝：成就巅峰
3.3	稳步回暖	沃森：会思考的“百科全书”
4.1	现状与未来	深度学习的兴起
4.2	现状与未来	大模型时代
4.3	现状与未来	走向未来

A.2.4 四年级下册

序号	单元题目	节题目
1.1	人工智能与医学	辅助手术机器人：医生的“超级助手”
1.2	人工智能与医学	人工智能读胸片：医生的“智慧之眼”
1.3	人工智能与医学	人工智能帮助研发癌症疫苗
2.1	人工智能与科学研究	AlphaFold：人工智能预测蛋白质结构
2.2	人工智能与科学研究	人工智能帮助数学家提出猜想

接下页

表 A.6 （续）

序号	单元题目	节题目
2.3	人工智能与科学研究	人工智能助力太空探索
2.4	人工智能与科学研究	人工智能提高显微镜的“视力”
3.1	人工智能与生活	人工智能天气预报
3.2	人工智能与生活	人工智能建筑师
3.3	人工智能与生活	人工智能守护野生动物
4.1	人工智能与艺术	人工智能画家
4.2	人工智能与艺术	人工智能音乐家
4.3	人工智能与艺术	人工智能诗人

A.2.5　五年级上册

序号	单元题目	节题目
1.1	人工智能风险（一）	信息伪造
1.2	人工智能风险（一）	信息泄露
1.3	人工智能风险（一）	信息茧房
2.1	人工智能风险（二）	人工智能的幻觉
2.2	人工智能风险（二）	过度迎合的人工智能
2.3	人工智能风险（二）	机器人会反抗人类吗
3.1	人工智能伦理（一）	人工智能与社会公平
3.2	人工智能伦理（一）	可以让人工智能帮忙写作文吗
3.3	人工智能伦理（一）	人工智能作品版权归属
4.1	人工智能伦理（二）	人工智能可以当裁判吗
4.2	人工智能伦理（二）	自动驾驶出错由谁负责
4.3	人工智能伦理（二）	会安慰你的人工智能真的懂你吗

A.2.6　五年级下册

序号	单元题目	节题目
1.1	计算机基础	认识计算机
1.2	计算机基础	计算机程序
1.3	计算机基础	计算机操作系统
1.4	计算机基础	算法
2.1	基于知识的人工智能	知识与智能
2.2	基于知识的人工智能	对弈算法
2.3	基于知识的人工智能	定理证明
2.4	基于知识的人工智能	专家系统
2.5	基于知识的人工智能	知识图谱
3.1	机器学习的基础概念	机器怎样从例子中学习
3.2	机器学习的基础概念	有答案的学习：监督学习
3.3	机器学习的基础概念	没有答案也能学习：无监督学习
3.4	机器学习的基础概念	通过奖励学习：强化学习

A.2.7 六年级上册

序号	单元题目	节题目
1.1	数据与模型	数据与人工智能
1.2	数据与模型	数据准备
1.3	数据与模型	模型设计
1.4	数据与模型	模型训练
2.1	机器学习常见任务	回归任务
2.2	机器学习常见任务	分类任务
2.3	机器学习常见任务	聚类任务
3.1	人工神经网络（一）	人类的神经系统
3.2	人工神经网络（一）	皮茨和他的神经元模型
3.3	人工神经网络（一）	感知器：会“学习”的神经网络
4.1	人工神经网络（二）	多层感知器

接下页

表 A.9 （续）

序号	单元题目	节题目
4.2	人工神经网络（二）	卷积神经网络
4.3	人工神经网络（二）	循环神经网络

A.2.8 六年级下册

序号	单元题目	节题目
1.1	深度学习	杰弗里·辛顿的故事
1.2	深度学习	深度学习的秘密
1.3	深度学习	李飞飞与 ImageNet 数据集
1.4	深度学习	GPU：从游戏到人工智能
1.5	深度学习	解析 AlphaGo
2.1	大模型与智能体	语言模型
2.2	大模型与智能体	大语言模型
2.3	大模型与智能体	视觉大模型
2.4	大模型与智能体	多模态大模型
2.5	大模型与智能体	推理大模型
2.6	大模型与智能体	智能体
3.1	深度学习的挑战	难以理解的智能
3.2	深度学习的挑战	对抗样本
3.3	深度学习的挑战	超级智能体

附录 B　初中学段

B.1　学段集成版

B.1.1　教材目录

本附录依据《清华大中小学人工智能通识教育系列》初中版教材目录整理。

序号	章题目	课题目
1	第一章什么是人工智能	智能机器的梦想
2	第一章什么是人工智能	什么是人工智能
3	第一章什么是人工智能	机器的眼睛
4	第一章什么是人工智能	机器的耳朵
5	第一章什么是人工智能	机器的嘴巴
6	第一章什么是人工智能	机器的手和脚
7	第二章人工智能的诞生	人类智能的起源
8	第二章人工智能的诞生	人类思维规律的总结
9	第二章人工智能的诞生	计算机的诞生
10	第二章人工智能的诞生	伟大的图灵
11	第二章人工智能的诞生	达特茅斯会议
12	第三章人工智能发展史	梦想与失落
13	第三章人工智能发展史	深度学习时代
14	第三章人工智能发展史	大模型时代
15	第三章人工智能发展史	交叉与融合
16	第三章人工智能发展史	走向未来
17	第四章人工智能前沿	人工智能与棋牌游戏
18	第四章人工智能前沿	人工智能与语言

接下页

表 B.1 （续）

序号	章题目	课题目
19	第四章人工智能前沿	人工智能与艺术
20	第四章人工智能前沿	人工智能与天文学
21	第四章人工智能前沿	人工智能与生物学
22	第四章人工智能前沿	人工智能与医学
23	第五章人工智能伦理	机器人三定律
24	第五章人工智能伦理	信息伪造
25	第五章人工智能伦理	信息泄露
26	第五章人工智能伦理	信息茧房
27	第五章人工智能伦理	人工智能与社会公平
28	第五章人工智能伦理	法律责任
29	第六章人工智能基础方法	基于知识的智能
30	第六章人工智能基础方法	基于学习的智能
31	第六章人工智能基础方法	监督学习和无监督学习
32	第六章人工智能基础方法	强化学习
33	第六章人工智能基础方法	机器学习的流派
34	第七章深度学习方法	人类神经系统
35	第七章深度学习方法	人工神经网络的开端
36	第七章深度学习方法	人工神经网络的发展史
37	第七章深度学习方法	深度学习的开端
38	第七章深度学习方法	深度学习基本原理
39	第七章深度学习方法	深度学习的挑战：对抗样本
40	第七章深度学习方法	深度学习的挑战：可解释性

B.1.2 实践手册目录

清华附中版

本附录依据《清华大中小学人工智能通识教育系列》初中版教材实践手册（清华附中版）整理。

表 B.2: 《清华大中小学人工智能通识系列》（初中版）实践手册（清华附中版）目录

序号	实践类型	实践题目
1	分析	人工智能的起源（1）：人类思维的数学化
2	分析	人工智能的起源（2）：计算机的诞生
3	工具	人工智能的起源（3）：伟大的图灵
4	分析	人工智能的起源（4）：达特茅斯会议
5	分析	人工智能的发展（5）：梦想与失落
6	工具	人工智能的发展（7）：大模型新时代
7	调研、辩论	人工智能的发展（8）：交叉与融合
8	工具	人工智能与艺术创作
9	调研、分析	人工智能与生物学
10	调研、分析	人工智能与医学
11	分析、辩论	机器人三定律
12	代码	机器学习样例
13	分析	人类神经系统
14	分析	人工神经网络的开端
15	分析	人工神经网络发展史
16	分析	卷积神经网络简介
17	分析、代码	深度学习解析：AlphaGo 的秘密
18	工具	深度学习解析：Sora 视频生成器
19	分析、工具	信息茧房与算法推荐
20	辩论、工具	人工智能法律与道德责任

苏州星浦实验中学版

本附录依据《清华大中小学人工智能通识教育系列》初中版教材实践手册（苏州星浦实验中学版）整理。

表 B.3: 《清华大中小学人工智能通识系列》（初中版）实践手册（苏州星浦实验中学版）目录

序号	篇章	实践类型	实践题目
1.1	第一篇	调研	漫展人工智能
1.2	第一篇	演讲	假如机器的眼睛出现了偏差
1.3	第一篇	工具	体验语音合成
2.1	第二篇	分析	体验人类思维的发展
2.2	第二篇	测试	机器是否能像人类一样思考问题
2.3	第二篇	话剧	达特茅斯会议
3.1	第三篇	征文	人工智能的发展历程
3.2	第三篇	工具	大模型深度体验
3.3	第三篇	征文	创作科幻故事
4.1	第四篇	代码	基于 Alpha-Beta 剪枝的五子棋
4.2	第四篇	工具	AI 艺术生成
4.3	第四篇	调研	蛋白质结构解析
5.1	第五篇	辩论	自动驾驶汽车难题
5.2	第五篇	角色扮演	模拟个性化推荐
5.3	第五篇	辩论	人工智能与社会公平
6.1	第六篇	工具	水果智能识别系统
6.2	第六篇	分析	机器学习的三大核心范式
7.1	第七篇	调研	探究自我恢复的机理
7.2	第七篇	分析	神经网络可视化
7.3	第七篇	代码	深度学习体验

B.2　年级分册版

本附录依据《清华大中小学人工智能通识教育系列》七、八年级教材整理，按年级和上下册分别列出单元题目和节题目。

B.2.1 七年级上册

序号	单元题目	节题目
1.1	人工智能的概念	智能机器的梦想
1.2	人工智能的概念	什么是人工智能
1.3	人工智能的概念	机器的眼睛
1.4	人工智能的概念	机器的耳朵
1.5	人工智能的概念	机器的嘴巴
1.6	人工智能的概念	机器的手和脚
2.1	人工智能的诞生	人类智能的起源
2.2	人工智能的诞生	人类思维规律的总结
2.3	人工智能的诞生	计算机的诞生
2.4	人工智能的诞生	伟大的图灵
2.5	人工智能的诞生	达特茅斯会议
3.1	人工智能的发展	梦想与失落
3.2	人工智能的发展	深度学习时代
3.3	人工智能的发展	大模型时代
3.4	人工智能的发展	交叉融合

B.2.2 七年级下册

序号	单元题目	节题目
1.1	人工智能伦理	机器人三定律
1.2	人工智能伦理	信息伪造
1.3	人工智能伦理	信息泄露
1.4	人工智能伦理	信息茧房
1.5	人工智能伦理	人工智能与社会公平
1.6	人工智能伦理	法律责任
2.1	人工智能前沿	人工智能与游戏

接下页

表 B.5 （续）

序号	单元题目	节题目
2.2	人工智能前沿	人工智能与语言
2.3	人工智能前沿	人工智能与艺术
2.4	人工智能前沿	人工智能与天文学
2.5	人工智能前沿	人工智能与生物学
2.6	人工智能前沿	人工智能与医学
2.7	人工智能前沿	人工智能与新材料
2.8	人工智能前沿	人工智能与数学
2.9	人工智能前沿	人工智能与城市交通

B.2.3 八年级上册

序号	单元题目	节题目
1.1	人工智能基础方法(一)	计算机发展史
1.2	人工智能基础方法(一)	计算机算法基础
1.3	人工智能基础方法(一)	基于知识的智能
1.4	人工智能基础方法(一)	基于学习的智能
2.1	人工智能基础方法(二)	机器学习样例
2.2	人工智能基础方法(二)	监督学习与无监督学习
2.3	人工智能基础方法(二)	强化学习
2.4	人工智能基础方法(二)	机器学习学派
3.1	神经网络方法	人类神经系统
3.2	神经网络方法	人工神经网络的开端
3.3	神经网络方法	人工神经网络发展史
3.4	神经网络方法	卷积神经网络简介
3.5	神经网络方法	循环神经网络简介

B.2.4 八年级下册

序号	单元题目	节题目
1.1	深度学习方法	深度学习的开端
1.2	深度学习方法	深度学习的基本原理
1.3	深度学习方法	词向量和对象嵌入
1.4	深度学习方法	序列到序列模型
1.5	深度学习方法	注意力机制
1.6	深度学习方法	深度学习的挑战：对抗样本
1.7	深度学习方法	深度学习的挑战：可解释性
2.1	大模型	神经网络语言模型
2.2	大模型	Transformer 简介
2.3	大模型	预训练方法
2.4	大模型	人类对齐
2.5	大模型	多模态大模型
2.6	大模型	走向未来

附录 C　高中学段

C.1　学段集成版

C.1.1　教材目录

本附录依据《清华大中小学人工智能通识教育系列》高中版教材目录整理。

序号	章题目	课题目
1	第一章人工智能概述	什么是人工智能
2	第一章人工智能概述	人类智能的起源
3	第一章人工智能概述	人工智能的起源——数理逻辑
4	第一章人工智能概述	人工智能的起源——计算机的诞生
5	第一章人工智能概述	图灵：人工智能之父
6	第一章人工智能概述	人工智能的开端
7	第一章人工智能概述	人工智能发展史（1）
8	第一章人工智能概述	人工智能发展史（2）
9	第一章人工智能概述	人工智能伦理（1）：人工智能近期风险
10	第一章人工智能概述	人工智能伦理（2）：人工智能远期风险
11	第二章人工智能基础	基于知识的人工智能
12	第二章人工智能基础	基于学习的智能
13	第二章人工智能基础	机器学习的基础流程
14	第二章人工智能基础	机器学习方法
15	第二章人工智能基础	机器学习四大流派
16	第二章人工智能基础	人工神经元网络
17	第二章人工智能基础	典型网络结构
18	第二章人工智能基础	深度学习

接下页

表 C.1 （续）

序号	章题目	课题目
19	第二章人工智能基础	大模型的基本原理（1）
20	第二章人工智能基础	大模型的基本原理（2）
21	第三章人工智能应用	机器视觉：人脸识别
22	第三章人工智能应用	机器视觉：绘画大师
23	第三章人工智能应用	机器视觉：伪造与鉴别
24	第三章人工智能应用	机器听觉：语音识别
25	第三章人工智能应用	机器听觉：语音合成
26	第三章人工智能应用	语言理解：机器翻译
27	第三章人工智能应用	人机对弈：AlphaGo 背后的秘密
28	第三章人工智能应用	人机对弈：AI 游戏玩家
29	第三章人工智能应用	信息检索：搜索引擎的秘密
30	第三章人工智能应用	信息检索：推荐算法
31	第四章人工智能交叉	和数学家做朋友
32	第四章人工智能交叉	模仿蝙蝠的耳朵
33	第四章人工智能交叉	破解蛋白质结构
34	第四章人工智能交叉	重构材料微观结构
35	第四章人工智能交叉	预测化学反应
36	第四章人工智能交叉	天文学家助手
37	第四章人工智能交叉	人工智能作曲家
38	第四章人工智能交叉	检测炭疽芽孢
39	第四章人工智能交叉	开发癌症疫苗
40	第四章人工智能交叉	走向未来

C.1.2 实践手册目录

本附录依据《清华大中小学人工智能通识教育系列》高中版教材实践手册（清华附中大兴学校版）整理。

序号	篇章	实践类型	实践题目
1.1	第一篇人工智能概述	工具	绘制人工智能发展史
1.2	第一篇人工智能概述	调研	人工智能风险调研
2.1	第二篇人工智能基础	分析	AI 如何帮助解决海洋污染
2.2	第二篇人工智能基础	调研	机器学习
2.3	第二篇人工智能基础	分析	人工神经网络初探
2.4	第二篇人工智能基础	分析	Transformer 的工作原理
3.1	第三篇人工智能应用	代码	清兴智慧食堂
3.2	第三篇人工智能应用	代码	清兴智慧停车
3.3	第三篇人工智能应用	工具	清兴杂志封面制作
3.4	第三篇人工智能应用	代码	清兴智能语音导览
3.5	第三篇人工智能应用	代码	清兴科技角——下棋机器人（一）
3.6	第三篇人工智能应用	代码	清兴科技角——下棋机器人（二）
4.1	第四篇人工智能前沿	工具	与 AI 一起密铺
4.2	第四篇人工智能前沿	调研	破解蛋白质结构之谜
4.3	第四篇人工智能前沿	工具	清兴校歌制作

C.2 年级分册版

本附录依据《清华大中小学人工智能通识教育系列》高一、高二教材整理，按年级和上下册分别列出单元题目和节题目。

C.2.1 高一上册

序号	单元题目	节题目
1.1	人工智能的起源	什么是人工智能
1.2	人工智能的起源	人类智能的起源
1.3	人工智能的起源	人工智能的起源：数理逻辑
1.4	人工智能的起源	人工智能的起源：计算机的诞生
1.5	人工智能的起源	图灵：人工智能之父
2.1	人工智能的发展	人工智能的开端
2.2	人工智能的发展	人工智能发展史（一）
2.3	人工智能的发展	人工智能发展史（二）
2.4	人工智能的发展	让人惊讶的人工智能
3.1	人工智能伦理	人工智能伦理：近期风险
3.2	人工智能伦理	人工智能伦理：远期风险
3.3	人工智能伦理	人工智能工具的合规使用

C.2.2 高一下册

序号	单元题目	节题目
1.1	人工智能方法	基于知识的人工智能
1.2	人工智能方法	基于学习的人工智能
1.3	人工智能方法	机器学习基础流程
1.4	人工智能方法	机器学习方法
2.1	深度学习	初识人工神经网络
2.2	深度学习	典型神经网络

接下页

表 C.4 （续）

序号	单元题目	节题目
2.3	深度学习	深度学习基础
2.4	深度学习	深度学习前沿（一）
2.5	深度学习	深度学习前沿（二）
3.1	大模型	大模型的基本原理（一）
3.2	大模型	大模型的基本原理（二）

C.2.3 高二上册

序号	单元题目	节题目
1.1	机器视觉	人脸识别
1.2	机器视觉	图像生成
1.3	机器视觉	伪造与鉴别
2.1	机器听觉	语音识别
2.2	机器听觉	声纹识别
2.3	机器听觉	语音合成
3.1	语言处理	机器翻译
3.2	语言处理	机器作家
3.3	语言处理	机器诗人
4.1	人机对战	AlphaGo
4.2	人机对战	人工智能与游戏
5.1	搜索与推荐	搜索引擎
5.2	搜索与推荐	推荐算法

C.2.4 高二下册

序号	单元题目	节题目
1.1	人工智能在数学与工程学科的应用	和数学家做朋友
1.2	人工智能在数学与工程学科的应用	模仿蝙蝠的耳朵
1.3	人工智能在数学与工程学科的应用	天文学家的助手
2.1	人工智能在材料与化学学科的应用	重构材料微观三维结构
2.2	人工智能在材料与化学学科的应用	预测化学反应类别
2.3	人工智能在材料与化学学科的应用	破解蛋白质结构
3.1	人工智能在生物与医学学科的应用	检测炭疽芽孢杆菌
3.2	人工智能在生物与医学学科的应用	开发癌症疫苗
3.3	人工智能在生物与医学学科的应用	预测健康状况
4.1	人工智能在人文与社会学科的应用	释读古文字
4.2	人工智能在人文与社会学科的应用	人工智能作曲
4.3	人工智能在人文与社会学科的应用	走向未来

附录 D　高等教育学段

D.1　教材目录

本附录根据《人工智能通识：原理、方法、应用与交叉融合》清华版教材整理。

序号	篇题目	章题目
1	第一篇　人工智能概述	什么是人工智能
2	第一篇　人工智能概述	人类智能的起源
3	第一篇　人工智能概述	人工智能的起源：数理逻辑
4	第一篇　人工智能概述	人工智能的起源：计算机的诞生
5	第一篇　人工智能概述	图灵：人工智能之父
6	第一篇　人工智能概述	人工智能的开端
7	第一篇　人工智能概述	人工智能发展史
8	第一篇　人工智能概述	让人惊讶的 AI
9	第一篇　人工智能概述	人工智能风险与伦理
10	第二篇　人工智能基础	基于知识的人工智能
11	第二篇　人工智能基础	基于学习的人工智能
12	第二篇　人工智能基础	机器学习基本流程
13	第二篇　人工智能基础	机器学习方法
14	第二篇　人工智能基础	机器学习策略
15	第二篇　人工智能基础	初识人工神经网络
16	第二篇　人工智能基础	典型神经网络结构
17	第二篇　人工智能基础	深度学习基础
18	第二篇　人工智能基础	深度学习前沿（1）
19	第二篇　人工智能基础	深度学习前沿（2）

接下页

表 D.1 （续）

序号	篇题目	章题目
20	第二篇 人工智能基础	大模型的基本原理（1）
21	第二篇 人工智能基础	大模型的基本原理（2）
22	第二篇 人工智能基础	深度学习面临的挑战
23	第三篇 人工智能应用	机器视觉：人脸识别
24	第三篇 人工智能应用	机器视觉：车牌识别
25	第三篇 人工智能应用	机器视觉：AI 美颜
26	第三篇 人工智能应用	机器视觉：绘画大师
27	第三篇 人工智能应用	机器视觉：AI 鉴伪
28	第三篇 人工智能应用	机器听觉：语音识别
29	第三篇 人工智能应用	机器听觉：声纹识别
30	第三篇 人工智能应用	机器听觉：语音合成
31	第三篇 人工智能应用	语言理解：写作与对话
32	第三篇 人工智能应用	语言处理：人工智能诗人
33	第三篇 人工智能应用	语言处理：机器翻译
34	第三篇 人工智能应用	人机对战：围棋国手
35	第三篇 人工智能应用	人机对战：AI 游戏
36	第三篇 人工智能应用	扫地机器人
37	第三篇 人工智能应用	搜索引擎
38	第三篇 人工智能应用	推荐算法
39	第四篇 人工智能交叉	破解蛋白质结构之谜
40	第四篇 人工智能交叉	重构材料微观三维结构
41	第四篇 人工智能交叉	预测化学反应类别
42	第四篇 人工智能交叉	生物拟态证据
43	第四篇 人工智能交叉	听声辨位
44	第四篇 人工智能交叉	检测炭疽病
45	第四篇 人工智能交叉	太空探索
46	第四篇 人工智能交叉	AI 作曲

接下页

表 D.1 （续）

序号	篇题目	章题目
47	第四篇　人工智能交叉	和数学家做朋友
48	第四篇　人工智能交叉	机器做梦
49	第四篇　人工智能交叉	天文学家的助手
50	第四篇　人工智能交叉	预测新冠病毒传染性
51	第四篇　人工智能交叉	开发癌症疫苗
52	第四篇　人工智能交叉	AI 增强显微镜
53	第四篇　人工智能交叉	走向未来

D.2　实践手册目录

本附录根据《人工智能通识：原理、方法、应用与交叉融合》清华版教材实践手册（清华版）整理。

序号	篇章	实践类型	实践题目
1.1	第一篇	调研	人工智能基础素养
1.2	第一篇	辩论	多元智能 vs 普遍智能
1.3	第一篇	征文	图灵与瑟尔之争
1.4	第一篇	工具	专业科普 AGENT
1.5	第一篇	征文	人与机器人
2.1	第二篇	辩论	硅基文明是否可能——人机辩论赛
2.2	第二篇	代码	线性分类模型
2.3	第二篇	分析	神经网络可视化
2.4	第二篇	分析	卷积神经网络可视化
2.5	第二篇	分析	深入理解卷积神经网络
2.6	第二篇	分析	深入理解 Transformer
2.7	第二篇	工具	生成 Nature 封面
2.8	第二篇	调研	大模型技术之“思维链”
2.9	第二篇	调研	对抗样本

接下页

表 D.2 （续）

序号	篇章	实践类型	实践题目
3.1	第三篇	代码	人脸识别系统
3.2	第三篇	代码	车牌识别系统
3.3	第三篇	工具	艺术之星
3.4	第三篇	代码	语音合成与声音克隆
3.5	第三篇	征文	大语言模型与学术道德
3.6	第三篇	工具	人工智能诗人
3.7	第三篇	代码	五子棋王者
3.8	第三篇	征文	搜索引擎商业排名的善与恶
4.1	第四篇	调研	GNoME 系统的落地方案
4.2	第四篇	调研	“AI+ 化学”项目的可行性分析
4.3	第四篇	工具	AI 作曲家
4.4	第四篇	调研	AI 医疗发展预测
4.5	第四篇	调研	城市人工智能发展规划

附录 E　终身学习学段

E.1　教材目录

本附录根据《人工智能通识（终身学习版）》清华版教材整理。

序号	章（单元）题目	节题目
1.1	人工智能：未来已来	智能机器的梦想
1.2	人工智能：未来已来	什么是人工智能
1.3	人工智能：未来已来	人工智能发展简史
1.4	人工智能：未来已来	强大的人工智能
2.1	身边的人工智能	机器的眼睛
2.2	身边的人工智能	机器的耳朵
2.3	身边的人工智能	机器的嘴巴
2.4	身边的人工智能	机器的手和脚
2.5	身边的人工智能	检索与推荐
2.6	身边的人工智能	语言理解
3.1	人工智能基础：深度学习	机器智能与机器学习
3.2	人工智能基础：深度学习	杰弗里·辛顿的故事
3.3	人工智能基础：深度学习	深度学习的秘密
3.4	人工智能基础：深度学习	数据与计算
3.5	人工智能基础：深度学习	解析 AlphaGo
4.1	人工智能前沿：大模型与智能体	语言模型
4.2	人工智能前沿：大模型与智能体	大语言模型

接下页

表 E.1 （续）

序号	章（单元）题目	节题目
4.3	人工智能前沿: 大模型与智能体	视觉大模型
4.4	人工智能前沿: 大模型与智能体	多模态大模型
4.5	人工智能前沿: 大模型与智能体	推理大模型
4.6	人工智能前沿: 大模型与智能体	智能体
5.1	人工智能伦理	机器人三定律
5.2	人工智能伦理	信息伪造
5.3	人工智能伦理	信息泄露
5.4	人工智能伦理	信息茧房
5.5	人工智能伦理	人工智能与社会公平
5.6	人工智能伦理	法律责任
6.1	人工智能生活陪伴	AI 问答：用 AI 获取信息
6.2	人工智能生活陪伴	AI 写作
6.3	人工智能生活陪伴	AI 作诗：让 AI 帮你展现才华
6.4	人工智能生活陪伴	AI 画家：贺卡、海报与修图
6.5	人工智能生活陪伴	AI 生成视频：回忆、祝福与记录
6.6	人工智能生活陪伴	AI 生成音乐：一首属于你的歌
7.1	未来展望	人工智能的发展趋势
7.2	未来展望	人工智能的研究方向
7.3	未来展望	AI+ 健康生活的前沿进展

E.2 实践手册目录

本附录根据《人工智能通识（终身学习版）实践指南》整理。

序号	专题题目	分节题目
1.1	轻松迈入 AI 的大门	手机应用软件使用
1.2	轻松迈入 AI 的大门	手机软件中的基础智能
2.1	AI 工具	什么是 AI 工具?
2.2	AI 工具	找到 AI 工具
2.3	AI 工具	开始使用 AI 工具
3	理解人工智能	
4.1	AI 问答：用 AI 获取信息	医疗健康
4.2	AI 问答：用 AI 获取信息	购物与消费
4.3	AI 问答：用 AI 获取信息	旅游与出行
4.4	AI 问答：用 AI 获取信息	办事与公共服务
4.5	AI 问答：用 AI 获取信息	家庭日用与手机操作
5	AI 写作	
6	AI 作诗：让 AI 帮你展示才华	
7	AI 画家：贺卡、海报与修图	
8	AI 生成视频：回忆、祝福与记录	
9	AI 生成音乐：一首属于你的歌	
10	安全指南	
11	前沿应用与创作	
12	创作集锦	

参考文献

Adams, C. et al. (2023). "Ethical principles for artificial intelligence in K-12 education". In: *Computers and Education: Artificial Intelligence* 4, p. 100131. DOI: 10.1016/j.caeai.2023.100131.

AI4K12 Initiative (2019). *Five Big Ideas in Artificial Intelligence and K–12 AI Guidelines*. Association for the Advancement of Artificial Intelligence and Computer Science Teachers Association. URL: https://ai4k12.org/.

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, et al. (2022). "Flamingo: A Visual Language Model for Few-Shot Learning". In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 23716–23736.

Arum, Richard and Josipa Roksa (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago: University of Chicago Press.

Bahdanau, D., K. Cho, and Y Bengio (2015). "Neural machine translation by jointly learning to align and translate". In: *International Conference on Learning Representations*.

Bakshy, E., S. Messing, and L. A Adamic (2015). "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239, pp. 1130–1132.

Balkin, Jack M. (2018). "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation". In: *UC Davis Law Review* 51.3, pp. 1149–1210. URL: https://lawreview.law.ucdavis.edu/archives/51/3/free-speech-algorithmic-society-big-data-private-governance-and-new-school-speech.

Barnett, Susan M. and Stephen J. Ceci (2002). "When and Where Do We Apply What We Learn? A Taxonomy for Far Transfer". In: *Psychological Bulletin* 128.4, pp. 612–637. DOI: 10.1037/0033-2909.128.4.612.

Bengio, Yoshua et al. (2024). "Managing Extreme AI Risks amid Rapid Progress". In: *Science* 384.6698, pp. 842–845. DOI: 10.1126/science.adn0117.

Biesta, Gert (2010). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Boulder, CO: Paradigm Publishers.

Biggs, John and Catherine Tang (2011). *Teaching for Quality Learning at University*. 4th ed. Maidenhead: Open University Press.

Bishop, C. M (2006). *Pattern Recognition and Machine Learning*. Springer.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258. URL: https://arxiv.org/abs/2108.07258.

Boole, George (1854). *An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities*. London: Walton and Maberly.

Borrego, Maura and Lynita K. Newswander (2010). "Definitions of Interdisciplinary Research: Toward Graduate-Level Interdisciplinary Learning Outcomes". In: *The Review of Higher Education* 34.1, pp. 61–84. DOI: 10.1353/rhe.2010.0006.

Bransford, John D., Ann L. Brown, and Rodney R. Cocking, eds. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: National Academies Press. DOI: 10.17226/9853.

Brennan, Karen and Mitchel Resnick (2012). "New Frameworks for Studying and Assessing the Development of Computational Thinking". In: *Proceedings of the 2012 Annual Meeting of the American Educational Research Association*. Vancouver, Canada. URL: https://web.media.mit.edu/~kbrennan/files/Brennan_Resnick_AERA2012_CT.pdf.

Bresnahan, Timothy F. and Manuel Trajtenberg (1995). "General Purpose Technologies: Engines of Growth?" In: *Journal of Econometrics* 65.1, pp. 83–108. DOI: 10.1016/0304-4076(94)01598-t.

Brin, Sergey and Lawrence Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks and ISDN Systems* 30.1–7, pp. 107–117. DOI: 10.1016/s0169-7552(98)00110-x.

Brown, Peter F. et al. (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19.2, pp. 263–311.

Brown, Tom B. et al. (2020). "Language Models Are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 1877–1901.

Bruner, Jerome S (1960). *The Process of Education*. Cambridge, MA: Harvard University Press.

Buolamwini, Joy and Timnit Gebru (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81. Proceedings of Machine Learning Research, pp. 77–91.

Cadena, Cesar, Luca Carlone, Henry Carrillo, et al. (2016). "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age". In: *IEEE Transactions on Robotics* 32.6, pp. 1309–1332. DOI: 10.1109/tro.2016.2624754.

Chaney, Allison J. B., Brandon M. Stewart, and Barbara E. Engelhardt (2018). "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility". In: *Proceedings of the 12th ACM Conference on Recommender Sys-*

tems. Association for Computing Machinery, pp. 224–232. DOI: 10.1145/3240323.3240370.

Chi, Michelene T. H., Paul J. Feltovich, and Robert Glaser (1981). "Categorization and Representation of Physics Problems by Experts and Novices". In: *Cognitive Science* 5.2, pp. 121–152. DOI: 10.1207/s15516709cog0502_2.

Chinn, Clark A., Luke A. Buckland, and Ala Samarapungavan (2011). "Expanding the Dimensions of Epistemic Cognition: Arguments from Philosophy and Psychology". In: *Educational Psychologist* 46.3, pp. 141–167. DOI: 10.1080/00461520.2011.587722.

Columbia College (2026). *Contemporary Civilization*. URL: https://www.college.columbia.edu/core-curriculum/classes/contemporary-civilization (visited on 06/2026).

Covington, Paul, Jay Adams, and Emre Sargin (2016). "Deep Neural Networks for YouTube Recommendations". In: *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198. DOI: 10.1145/2959100.2959190.

Davies, Alex, Petar Veličković, Lars Buesing, et al. (2021). "Advancing Mathematics by Guiding Human Intuition with AI". In: *Nature* 600, pp. 70–74. DOI: 10.1038/s41586-021-04086-x.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

Dewey, John (1916). *Democracy and Education: An Introduction to the Philosophy of Education*. New York: Macmillan. URL: https://www.gutenberg.org/ebooks/852.

Dijck, José van, Thomas Poell, and Martijn de Waal (2018). *The Platform Society: Public Values in a Connective World*. New York: Oxford University Press. DOI: 10.1093/oso/9780190889760.001.0001.

diSessa, A. A (1993). "Toward an epistemology of physics". In: *Cognition and Instruction* 10.2–3, pp. 105–225.

Du, Shan et al. (2013). "Automatic License Plate Recognition (ALPR): A State-of-the-Art Review". In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.2, pp. 311–325. DOI: 10.1109/tcsvt.2012.2203741.

Durrant-Whyte, Hugh and Tim Bailey (2006). "Simultaneous Localization and Mapping: Part I". In: *IEEE Robotics and Automation Magazine* 13.2, pp. 99–110. DOI: 10.1109/mra.2006.1638022.

Eliasmith, Chris et al. (2012). "A Large-Scale Model of the Functioning Brain". In: *Science* 338.6111, pp. 1202–1205. DOI: 10.1126/science.1225266.

Eppler, Martin J. and Jeanne Mengis (2004). "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and

Related Disciplines". In: *The Information Society* 20.5, pp. 325–344. DOI: 10.1080/01972240490507974.

Espeland, W. N. and M Sauder (2007). "Rankings and reactivity: How public measures recreate social worlds". In: *American Journal of Sociology* 113.1, pp. 1–40.

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence*. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Facione, P. A (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*. The Delphi Report. Millbrae, CA: California Academic Press.

Feigenbaum, Edward A. (1977). "The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Vol. 5, pp. 1014–1029.

Flavell, John H. (1979). "Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry". In: *American Psychologist* 34.10, pp. 906–911. DOI: 10.1037/0003-066x.34.10.906.

Fleming, Stephen M. and Raymond J. Dolan (2012). "The Neural Basis of Metacognitive Ability". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594, pp. 1338–1349. DOI: 10.1098/rstb.2011.0417.

Flexner, Abraham (1930). *Universities: American, English, German*. New York: Oxford University Press.

Floridi, L. et al. (2018). "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations". In: *Minds and Machines* 28, pp. 689–707.

Fong, Geoffrey T., David H. Krantz, and Richard E. Nisbett (1986). "The Effects of Statistical Training on Thinking about Everyday Problems". In: *Cognitive Psychology* 18.3, pp. 253–292. DOI: 10.1016/0010-0285(86)90001-0.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge (2016). "Image Style Transfer Using Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423. DOI: 10.1109/cvpr.2016.265.

Geirhos, R. et al. (2020). "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2, pp. 665–673.

Gigerenzer, Gerd and Ulrich Hoffrage (1995). "How to Improve Bayesian Reasoning without Instruction: Frequency Formats". In: *Psychological Review* 102.4, pp. 684–704. DOI: 10.1037/0033-295x.102.4.684.

Gillespie, T (2014). "The relevance of algorithms". In: *Media Technologies: Essays on Communication, Materiality, and Society*. Ed. by P. J. Boczkowski T. Gillespie and K. A. Foot. MIT Press, pp. 167–194.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. (2014). “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Vol. 27, pp. 2672–2680.

Graves, Alex et al. (2006). “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks”. In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376. DOI: 10.1145/1143844.1143891.

Griffiths, T. L. and J. B Tenenbaum (2006). “Optimal predictions in everyday cognition”. In: *Psychological Science* 17.9, pp. 767–773.

Guo, Chuan et al. (2017). “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research, pp. 1321–1330.

Halliday, M. A. K (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold.

Hardwig, J (1985). “Epistemic dependence”. In: *The Journal of Philosophy* 82.7, pp. 335–349.

Harvard Committee (1945). *General Education in a Free Society: Report of the Harvard Committee*. Cambridge, MA: Harvard University Press.

Harvard University Committee on the Objectives of a General Education in a Free Society (1945). *General Education in a Free Society*. Cambridge, MA: Harvard University Press.

Hendrycks, D. and T Dietterich (2019). “Benchmarking neural network robustness to common corruptions and perturbations”. In: *International Conference on Learning Representations*.

Hestenes, D (1992). “Modeling games in the Newtonian world”. In: *American Journal of Physics* 60.8, pp. 732–748.

Hidi, Suzanne and K. Ann Renninger (2006). “The Four-Phase Model of Interest Development”. In: *Educational Psychologist* 41.2, pp. 111–127. DOI: 10.1207/s15326985ep4102_4.

Hinton, Geoffrey et al. (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition”. In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97. DOI: 10.1109/msp.2012.2205597.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 6840–6851.

Hofer, Barbara K. and Paul R. Pintrich (1997). “The Development of Epistemological Theories: Beliefs about Knowledge and Knowing and Their Relation to Learning”. In: *Review of Educational Research* 67.1, pp. 88–140. DOI: 10.3102/00346543067001088.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. (2022). “Training Compute-Optimal Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 30016–30030.

Holland, John H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Huang, C.-Z. A., A. Vaswani, J. Uszkoreit, et al. (2018). *Music Transformer: Generating music with long-term structure*. arXiv: 1809.04281. URL: https://arxiv.org/abs/1809.04281.

Huang, Lei et al. (2025). “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”. In: *ACM Transactions on Information Systems* 43.2, pp. 1–55. DOI: 10.1145/3703155.

Humboldt, Wilhelm von (1810). *On the Internal and External Organization of the Higher Scientific Institutions in Berlin*. English translation by Thomas Dunlap. URL: https://germanhistorydocs.org/en/the-holy-roman-empire-1648-1815/wilhelm-von-humboldt-s-treatise-quot-on-the-internal-and-external-organization-of-the-higher-scientific-institutions-in-berlin-quot-1810.pdf.

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Inhelder, Bärbel and Jean Piaget (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York: Basic Books.

Ji, Ziwei et al. (2023). “Survey of Hallucination in Natural Language Generation”. In: *ACM Computing Surveys* 55.12, pp. 1–38. DOI: 10.1145/3571730.

Jo, YoungJu, SangYun Park, JaeHwang Jung, et al. (2017). “Holographic Deep Learning for Rapid Optical Screening of Anthrax Spores”. In: *Science Advances* 3.8, e1700606. DOI: 10.1126/sciadv.1700606.

Jumper, J., R. Evans, A. Pritzel, et al. (2021). “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596, pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. (2020). “Scaling Laws for Neural Language Models”. In: *arXiv preprint arXiv:2001.08361*. DOI: 10.48550/arxiv.2001.08361.

Kench, Steven and Samuel J. Cooper (2021). “Generating Three-Dimensional Structures from a Two-Dimensional Slice with Generative Adversarial Network-Based Dimensionality Expansion”. In: *Nature Machine Intelligence* 3, pp. 299–305. DOI: 10.1038/s42256-021-00322-1.

Kerr, Clark (2001). *The Uses of the University*. 5th ed. Cambridge, MA: Harvard University Press.

Kerrigan, Joshua, Paul La Plante, Saul Kohn, et al. (2019). “Optimizing Sparse RFI Prediction Using Deep Learning”. In: *Monthly Notices of the Royal Astronomical Society* 488.2, pp. 2605–2615. DOI: 10.1093/mnras/stz1865.

Kitchin, Rob (2014). “Big Data, New Epistemologies and Paradigm Shifts”. In: *Big Data & Society* 1.1. DOI: 10.1177/2053951714528481.

Kliebard, Herbert M. (2004). *The Struggle for the American Curriculum, 1893–1958*. 3rd ed. New York: RoutledgeFalmer.

Knowles, Malcolm S., Elwood F. Holton, and Richard A. Swanson (2015). *The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development*. 8th ed. London: Routledge.

Koh, P. W. et al. (2021). “WILDS: A benchmark of in-the-wild distribution shifts”. In: *Proceedings of the 38th International Conference on Machine Learning*.

Koren, Yehuda, Robert Bell, and Chris Volinsky (2009). “Matrix Factorization Techniques for Recommender Systems”. In: *Computer* 42.8, pp. 30–37. DOI: 10.1109/mc.2009.263.

Krajcik, Joseph S. and Phyllis C. Blumenfeld (2006). “Project-Based Learning”. In: *The Cambridge Handbook of the Learning Sciences*. Ed. by R. Keith Sawyer. Cambridge: Cambridge University Press, pp. 317–334. DOI: 10.1017/CBO9780511816833.020.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 25, pp. 1097–1105.

Kuhn, Deanna (1999). “A Developmental Model of Critical Thinking”. In: *Educational Researcher* 28.2, pp. 16–46. DOI: 10.3102/0013189x028002016.

Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio (2017). “Adversarial Examples in the Physical World”. In: *ICLR Workshop*. arXiv: 1607.02533. URL: https://research.google/pubs/adversarial-examples-in-the-physical-world/.

Lazer, D. M. J. et al. (2018). “The science of fake news”. In: *Science* 359.6380, pp. 1094–1096.

Lazer, David, Alex Pentland, Lada Adamic, et al. (2009). “Computational Social Science”. In: *Science* 323.5915, pp. 721–723. DOI: 10.1126/science.1167742.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep Learning”. In: *Nature* 521, pp. 436–444. DOI: 10.1038/nature14539.

Lee, J. D. and K. A See (2004). “Trust in automation: Designing for appropriate reliance”. In: *Human Factors* 46.1, pp. 50–80.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. (2020). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

Lin, Stephanie, Jacob Hilton, and Owain Evans (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229.

Long, Duri and Brian Magerko (2020). "What Is AI Literacy? Competencies and Design Considerations". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp. 1–16. DOI: 10.1145/3313831.3376727.

Manovich, Lev (2020). *Cultural Analytics*. Cambridge, MA: MIT Press.

Matthias, Andreas (2004). "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata". In: *Ethics and Information Technology* 6.3, pp. 175–183. DOI: 10.1007/s10676-004-3422-1.

McCarthy, John (2007). *What Is Artificial Intelligence?* URL: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.

McCarthy, John et al. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Dartmouth College. URL: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

McCulloch, Warren S. and Walter Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5, pp. 115–133. DOI: 10.1007/bf02478259.

Mellers, B. et al. (2015). "The psychology of intelligence analysis: Drivers of prediction accuracy in world politics". In: *Journal of Experimental Psychology: Applied* 21.1, pp. 1–14.

Merriam, Sharan B. and Laura L. Bierema (2013). *Adult Learning: Linking Theory and Practice*. San Francisco: Jossey-Bass.

Miao, Fengchun and Mutlu Cukurova (2024). *AI Competency Framework for Teachers*. Paris: UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391104.

Miao, Fengchun and Wayne Holmes (2023). *Guidance for Generative AI in Education and Research*. Paris: UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000386693 (visited on 07/20/2026).

Miao, Fengchun, Kelly Shiohira, and Natalie Lao (2024). *AI Competency Framework for Students*. Paris: UNESCO. DOI: 10.54675/JKJB9835. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391105.

Minsky, Marvin and Seymour Papert (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Mishra, Punya and Matthew J. Koehler (2006). "Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge". In: *Teachers College Record* 108.6, pp. 1017–1054. DOI: 10.1111/j.1467-9620.2006.00684.x.

Mitchell, Tom M. (1997). *Machine Learning*. New York: McGraw-Hill. ISBN: 9780070428072.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, et al. (2015). "Human-Level Control through Deep Reinforcement Learning". In: *Nature* 518, pp. 529–533. DOI: 10.1038/nature14236.

Morales-Navarro, Luis et al. (2025). "What Can Youth Learn about Artificial Intelligence and Machine Learning in One Hour? Examining How Hour of Code Activities Address the Five Big Ideas of AI". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.28, pp. 29195–29202. DOI: 10.1609/aaai.v39i28.35193.

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. National Institute of Standards and Technology. DOI: 10.6028/NIST.AI.100-1.

Nelson, T. O. and L Narens (1990). "Metamemory: A theoretical framework and new findings". In: *The Psychology of Learning and Motivation*. Ed. by G. H. Bower. Vol. 26. San Diego: Academic Press, pp. 125–173.

Neumann, John von (1945). *First Draft of a Report on the EDVAC*. Tech. rep. Moore School of Electrical Engineering, University of Pennsylvania.

Newell, Allen and Herbert A. Simon (1976). "Computer Science as Empirical Inquiry: Symbols and Search". In: *Communications of the ACM* 19.3, pp. 113–126. DOI: 10.1145/360018.360022.

Newman, John Henry (1996). *The Idea of a University*. Edited by Frank M. Turner. New Haven: Yale University Press.

Ng, Davy Tsz Kit et al. (2021). "Conceptualizing AI Literacy: An Exploratory Review". In: *Computers and Education: Artificial Intelligence* 2, p. 100041. DOI: 10.1016/j.caeai.2021.100041.

Nisbett, Richard E. et al. (1983). "The Use of Statistical Heuristics in Everyday Inductive Reasoning". In: *Psychological Review* 90.4, pp. 339–363. DOI: 10.1037/0033-295x.90.4.339.

Nussbaum, Martha C (2010). *Not for Profit: Why Democracy Needs the Humanities*. Princeton: Princeton University Press.

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. Paris: OECD. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD and European Commission (2026). *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*. Paris: OECD Publishing. DOI: 10.1787/65cd27d4-en.

Oord, Aäron van den et al. (2016). *WaveNet: A Generative Model for Raw Audio*. DOI: 10.48550/arxiv.1609.03499. arXiv: 1609.03499. URL: https://arxiv.org/abs/1609.03499.

Ott, P. A., Z. Hu, D. B. Keskin, et al. (2017). “An immunogenic personal neoantigen vaccine for patients with melanoma”. In: *Nature* 547, pp. 217–221. DOI: 10.1038/nature22991.

Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rep. 1999-66. Stanford InfoLab.

Papert, S (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books.

Parasuraman, Raja and Victor Riley (1997). “Humans and Automation: Use, Misuse, Disuse, Abuse”. In: *Human Factors* 39.2, pp. 230–253. DOI: 10.1518/001872097778543886.

Pascarella, Ernest T. and Patrick T. Terenzini (2005). *How College Affects Students: A Third Decade of Research*. Vol. 2. San Francisco: Jossey-Bass.

Passi, Samir and Solon Barocas (2019). “Problem Formulation and Fairness”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, pp. 39–48. DOI: 10.1145/3287560.3287567.

Payne, B. H (2019). *An Ethics of Artificial Intelligence Curriculum for Middle School Students*. URL: https://ec.europa.eu/futurium/en/system/files/ged/mit_ai_ethics_education_curriculum.pdf.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pennycook, G. and D. G Rand (2019). “Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning”. In: *Cognition* 188, pp. 39–50.

Perdomo, Juan C. et al. (2020). “Performative Prediction”. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 7599–7609. URL: https://proceedings.mlr.press/v119/perdomo20a.html.

Perkins, David N. and Gavriel Salomon (1988). “Teaching for Transfer”. In: *Educational Leadership* 46.1, pp. 22–32.

Pintrich, P. R (2000). “The role of goal orientation in self-regulated learning”. In: *Handbook of Self-Regulation*. Ed. by P. R. Pintrich M. Boekaerts and M. Zeidner. San Diego: Academic Press, pp. 451–502.

Pólya, George (1957). *How to Solve It: A New Aspect of Mathematical Method*. 2nd ed. Princeton University Press.

Pylyshyn, Zenon W. (1980). “Computation and Cognition: Issues in the Foundations of Cognitive Science”. In: *Behavioral and Brain Sciences* 3.1, pp. 111–132. DOI: 10.1017/S0140525X00002053.

Rabiner, Lawrence R. (1989). “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Proceedings of the IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626.

Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139, pp. 8748–8763.

Reber, Rolf, Norbert Schwarz, and Piotr Winkielman (2004). “Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver’s Processing Experience?” In: *Personality and Social Psychology Review* 8.4, pp. 364–382. DOI: 10.1207/s15327957pspr0804_3.

Recht, Benjamin et al. (2019). “Do ImageNet Classifiers Generalize to ImageNet?” In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97, pp. 5389–5400. URL: https://proceedings.mlr.press/v97/recht19a.html.

Reichstein, Markus et al. (2019). “Deep Learning and Process Understanding for Data-Driven Earth System Science”. In: *Nature* 566, pp. 195–204. DOI: 10.1038/s41586-019-0912-1.

Risko, E. F. and S. J Gilbert (2016). “Cognitive offloading”. In: *Trends in Cognitive Sciences* 20.9, pp. 676–688. DOI: 10.1016/j.tics.2016.07.002.

Rombach, Robin et al. (2022). “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695. DOI: 10.1109/cvpr52688.2022.01042.

Rosenblatt, Frank (1958). “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review* 65.6, pp. 386–408. DOI: 10.1037/h0042519.

Rosenblatt, L. M (1978). *The Reader, the Text, the Poem: The Transactional Theory of the Literary Work*. Southern Illinois University Press.

Rudin, Cynthia (2019). “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1.5, pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). “Learning Representations by Back-Propagating Errors”. In: *Nature* 323, pp. 533–536. DOI: 10.1038/323533a0.

Russell, Stuart J. and Peter Norvig (2021). *Artificial Intelligence: A Modern Approach*. 4th ed. Pearson.

Sadler, D. R (1989). “Formative assessment and the design of instructional systems”. In: *Instructional Science* 18.2, pp. 119–144. DOI: 10.1007/bf00117714.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1996). “Statistical Learning by 8-Month-Old Infants”. In: *Science* 274.5294, pp. 1926–1928. DOI: 10.1126/science.274.5294.1926.

Sahin, Ugur, Evelyna Derhovanessian, Mark Miller, et al. (2017). “Personalized RNA Mutanome Vaccines Mobilize Poly-Specific Therapeutic Immunity against Cancer”. In: *Nature* 547, pp. 222–226. DOI: 10.1038/nature23003.

Samborska, Veronika et al. (2022). “Complementary Task Representations in Hippocampus and Prefrontal Cortex for Generalizing the Structure of Problems”. In: *Nature Neuroscience* 25, pp. 1314–1326. DOI: 10.1038/s41593-022-01149-8.

Samuel, Arthur L. (1959). “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210.

Sandoval, William A., Jeffrey A. Greene, and Ivar Bråten (2016). “Understanding and Promoting Thinking About Knowledge: Origins, Issues, and Future Directions of Research on Epistemic Cognition”. In: *Review of Research in Education* 40.1, pp. 457–496. DOI: 10.3102/0091732x16669319.

Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, et al. (2023). “Toolformer: Language Models Can Teach Themselves to Use Tools”. In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 68539–68551.

Schoenfeld, A. H (1985). *Mathematical Problem Solving*. Academic Press.

— (1992). “Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics”. In: *Handbook of Research on Mathematics Teaching and Learning*. Ed. by D. Grouws. Macmillan, pp. 334–370.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823. DOI: 10.1109/cvpr.2015.7298682.

Schwaller, P., T. Laino, et al. (2019). “Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction”. In: *ACS Central Science* 5.9, pp. 1572–1583. DOI: 10.1021/acscentsci.9b00576.

Schwaller, P., D. Probst, A. C. Vaucher, et al. (2021). “Mapping the space of chemical reactions using attention-based neural networks”. In: *Nature Machine Intelligence* 3, pp. 144–152. DOI: 10.1038/s42256-020-00284-w.

Seixas, P. and T Morton (2013). *The Big Six Historical Thinking Concepts*. Nelson Education.

Selbst, Andrew D. et al. (2019). “Fairness and Abstraction in Sociotechnical Systems”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, pp. 59–68. DOI: 10.1145/3287560.3287598.

Shannon, Claude E. (1938). “A Symbolic Analysis of Relay and Switching Circuits”. In: *Transactions of the American Institute of Electrical Engineers* 57.12, pp. 713–723. DOI: 10.1109/t-aiee.1938.5057767.

Sharma, Mrinank et al. (2023). *Towards Understanding Sycophancy in Language Models*. DOI: 10.48550/arXiv.2310.13548. arXiv: 2310.13548 [cs.CL]. URL: https://arxiv.org/abs/2310.13548.

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, et al. (2018). “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783. DOI: 10.1109/icassp.2018.8461368.

Shepard, Roger N. (1987). “Toward a Universal Law of Generalization for Psychological Science”. In: *Science* 237.4820, pp. 1317–1323. DOI: 10.1126/science.3629243.

Shokri, Reza et al. (2017). “Membership Inference Attacks against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 3–18. DOI: 10.1109/SP.2017.41.

Shulman, Lee S. (1986). “Those Who Understand: Knowledge Growth in Teaching”. In: *Educational Researcher* 15.2, pp. 4–14. DOI: 10.3102/0013189x015002004.

Shute, V. J (2008). “Focus on formative feedback”. In: *Review of Educational Research* 78.1, pp. 153–189. DOI: 10.3102/0034654307313795.

Silver, David, Aja Huang, Chris J. Maddison, et al. (2016). “Mastering the Game of Go with Deep Neural Networks and Tree Search”. In: *Nature* 529, pp. 484–489. DOI: 10.1038/nature16961.

Silver, David, Thomas Hubert, Julian Schrittwieser, et al. (2018). “A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play”. In: *Science* 362.6419, pp. 1140–1144. DOI: 10.1126/science.aar6404.

Simon, H. A. (1971). “Designing organizations for an information-rich world”. In: *Computers, Communications, and the Public Interest*. Ed. by M. Greenberger. Johns Hopkins Press, pp. 37–72.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick (1999). “Does Automation Bias Decision-Making?” In: *International Journal of Human-Computer Studies* 51.5, pp. 991–1006. DOI: 10.1006/ijhc.1999.0252.

Snyder, David et al. (2018). “X-Vectors: Robust DNN Embeddings for Speaker Recognition”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329–5333. DOI: 10.1109/icassp.2018.8461375.

Sparrow, B., J. Liu, and D. M Wegner (2011). “Google effects on memory: Cognitive consequences of having information at our fingertips”. In: *Science* 333.6043, pp. 776–778. DOI: 10.1126/science.1207745.

Sperber, Dan et al. (2010). “Epistemic Vigilance”. In: *Mind & Language* 25.4, pp. 359–393. DOI: 10.1111/j.1468-0017.2010.01394.x.

Strathern, M (1997). “Improving ratings: Audit in the British university system”. In: *European Review* 5.3, pp. 305–321.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). “Energy and Policy Considerations for Deep Learning in NLP”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association

for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://aclanthology.org/P19-1355/.

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press.

Taigman, Yaniv et al. (2014). "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708. DOI: 10.1109/cvpr.2014.220.

Tenenbaum, J. B. et al. (2011). "How to grow a mind: Statistics, structure, and abstraction". In: *Science* 331.6022, pp. 1279–1285.

The College, University of Chicago (2026). *History of the Core*. URL: https://college.uchicago.edu/academics/core/history (visited on 06/2026).

Tolosana, Ruben et al. (2020). "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection". In: *Information Fusion* 64, pp. 131–148. DOI: 10.1016/j.inffus.2020.06.014.

Touretzky, David S. et al. (2019). "Envisioning AI for K–12: What Should Every Child Know about AI?" In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 1, pp. 9795–9799. DOI: 10.1609/aaai.v33i01.33019795.

Trinh, T. H. et al. (2024). "Solving olympiad geometry without human demonstrations". In: *Nature* 625, pp. 476–482. DOI: 10.1038/s41586-023-06747-5.

Turing, Alan M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society*. 2nd ser. 42.1, pp. 230–265. DOI: 10.1112/plms/s2-42.1.230.

— (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460. DOI: 10.1093/mind/lix.236.433.

Turk, Matthew and Alex Pentland (1991). "Face Recognition Using Eigenfaces". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–591. DOI: 10.1109/cvpr.1991.139758.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157, pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.

UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

UNESCO Institute for Lifelong Learning (2020). *Embracing a Culture of Lifelong Learning: Contribution to the Futures of Education Initiative*. Hamburg: UNESCO Institute for Lifelong Learning. URL: https://unesdoc.unesco.org/ark:/48223/pf0000374112.

UNICEF Innocenti (2025). *Guidance on AI and Children*. UNICEF. URL: https://www.unicef.org/innocenti/reports/policy-guidance-ai-children (visited on 07/20/2026).

Vaswani, Ashish et al. (2017). “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30, pp. 5998–6008.

Veblen, Thorstein (1918). *The Higher Learning in America: A Memorandum on the Conduct of Universities by Business Men*. New York: B. W. Huebsch.

Vinyals, Oriol, Igor Babuschkin, Wojciech M. Czarnecki, et al. (2019). “Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning”. In: *Nature* 575, pp. 350–354. DOI: 10.1038/s41586-019-1724-z.

Vosoughi, S., D. Roy, and S Aral (2018). “The spread of true and false news online”. In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.

Vygotsky, L. S (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wang, Pengjie et al. (2024). “An Open Dataset for Oracle Bone Character Recognition and Decipherment”. In: *Scientific Data* 11, p. 976. DOI: 10.1038/s41597-024-03807-x.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. (2022). “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract.html.

Weizenbaum, Joseph (1966). “ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine”. In: *Communications of the ACM* 9.1, pp. 36–45. DOI: 10.1145/365153.365168.

Whitehead, Alfred North (1929). *The Aims of Education and Other Essays*. New York: Macmillan.

Wilson, Louis R. (1946). “General Education in a Free Society; Report of the Harvard Committee”. In: *College & Research Libraries* 7.2, pp. 186–187. DOI: 10.5860/crl_07_02_186.

Wineburg, S. S (1991). “Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence”. In: *Journal of Educational Psychology* 83.1, pp. 73–87.

Wineburg, Sam and Sarah McGrew (2019). “Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information”. In: *Teachers College Record* 121.11, pp. 1–40. DOI: 10.1177/016146811912101102.

Wing, J. M (2006). “Computational thinking”. In: *Communications of the ACM* 49.3, pp. 33–35. DOI: 10.1145/1118178.1118215.

Wolpert, David H. and William G. Macready (1997). “No Free Lunch Theorems for Optimization”. In: *IEEE Transactions on Evolutionary Computation* 1.1, pp. 67–82. DOI: 10.1109/4235.585893.

Yao, S. et al. (2023). "ReAct: Synergizing reasoning and acting in language models". In: *International Conference on Learning Representations*. arXiv: 2210.03629. URL: https://arxiv.org/abs/2210.03629.

Zeng, Daniel Dajun (2013). "From Computational Thinking to AI Thinking". In: *IEEE Intelligent Systems* 28.6, pp. 2–4. DOI: 10.1109/mis.2013.141.

Zhou, Xinyu, Jessica Van Brummelen, and Phoebe Lin (2020). *Designing AI Learning Experiences for K–12: Emerging Works, Future Opportunities and a Design Framework*. arXiv: 2009.10228. URL: https://arxiv.org/abs/2009.10228.

Zimmerman, Barry J. (2002). "Becoming a Self-Regulated Learner: An Overview". In: *Theory Into Practice* 41.2, pp. 64–70. DOI: 10.1207/s15430421tip4102_2.

中华人民共和国教育部 (2022). 义务教育信息科技课程标准（*2022* 年版）. 北京: 北京师范大学出版社.

— (Dec. 2, 2024). 教育部部署加强中小学人工智能教育. URL: https://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202412/t20241202_1165500.html (visited on 07/20/2026).

中华人民共和国教育部等五部门 (Apr. 10, 2026). "人工智能 + 教育"行动计划. URL: https://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html (visited on 07/20/2026).

全国人民代表大会常务委员会 (2021). 中华人民共和国个人信息保护法. URL: https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm.

北京市教育委员会 (2025). 北京市推进中小学人工智能教育工作方案（*2025—2027* 年）. URL: https://jw.beijing.gov.cn/xxgk/2024zcwj/2024qtwj/202503/t20250307_4028227.html (visited on 08/12/2026).

国家互联网信息办公室, 工业和信息化部, and 公安部 (2022). 互联网信息服务深度合成管理规定. URL: https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm.

国家互联网信息办公室, 工业和信息化部, 公安部, and 国家广播电视总局 (2025). 人工智能生成合成内容标识办法. URL: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

国家互联网信息办公室等 (2023). 生成式人工智能服务管理暂行办法. URL: https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

国家新一代人工智能治理专业委员会 (2021). 新一代人工智能伦理规范. URL: https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

教育部基础教育教学指导委员会 (2025a). 中小学人工智能通识教育指南（*2025* 年版）. 教育部基础教育教学指导委员会. URL: https://app.www.gov.cn/govdata/gov/202505/15/528972/article.html.

— (2025b). 中小学生成式人工智能使用指南（*2025* 年版）. 教育部基础教育教学指导委员会. URL: https://app.www.gov.cn/govdata/gov/202505/15/528972/article.html.