人机延展主体：生成性延展与非对称授权

Oct 24, 2026 · @Tao

摘要

生成式 AI 会主动提出使用者不会自动认可的内容。延展心智理论已经说明外部资源可以成为认知的组成部分，却尚未说明这类会主动生成内容的组成部分如何属于一个人的主体。本文论证：在持续、深度的整合中，人与生成式 AI 可以构成实践能动意义上的同一个主体；AI 的发散性生成属于该主体的思考过程，对立场的认同与对行动的授权归于现实中的人。

本文区分三个层次。在认知系统层次，由外部发起的生成通过“生成—选择”回路与人的选择持续地相互作用；借助持续记忆，二者在长期认知任务的时间尺度上相互调节，其贡献无法加性地分离，由此构成一个整合的认知过程。在主体层次，这一过程以现实中的人为唯一的规范中心：数字部分可以提出异议，却不能凭借自身的判断取得持续引导个人的权限；它的权限来自个人的委托，并可由个人撤回；它也不持有独立于委托关系的实践立场。最后一项条件区分了主体内部的委托与两个能动者之间的委托：一位忠实的人类助理同样只有派生的权限，却仍是另一个能动者。本文由此论证实践能动意义上的主体统一，而不主张形而上的同一或体验视角的统一。在规范层次，立场、利害与可回应性说明最终授权为何归于现实中的人。

本文还区分了认知构成与主体统一。训练带来的外部塑造，与人的语言和概念所受的社会塑造一样，不排除认知上的归属；当外部目的在运行中取得引导个人的权限时，数字部分仍是个人认知过程的组成部分，失效的是主体的统一。这种劫持因此比一般的外部影响更严重：它从个人认知的内部进入选择回路，既是对心智的干预，也是操纵。就 AI 治理而言，本文为保护被延展的认知提供了伦理平价的依据，把操纵界定为对选择回路的绕过或劫持，并在不赋予 AI 道德地位的前提下区分了授权与责任。

关键词： 延展心智；生成性延展；能动主体；实践能动性；非对称授权；AI 治理

一、前提与讨论范围

本文的论证建立在三项理论前提之上，只适用于一类特定的人机关系。

1.1 理论前提

1. **功能平价**。 一个过程是否属于认知，取决于它在认知活动中的功能角色，而不取决于它发生在颅内还是颅外（Clark & Chalmers, 1998）。本文以平价原则为工作前提；针对它的两种主要批评（Adams & Aizawa, 2001; Rupert, 2004）在第九节讨论。
2. **规范个人主义**。 道德关切的最终对象是人。本文所说的“主体”以一个人的统一性为中心，而不以技术系统为中心；最终授权为何归于人，在第六节单独论证。
3. **对 AI 的内在状态保持中立**。 本文不预设 AI 具有意识、意向性或道德地位，论证只使用 AI 系统在认知活动中的功能角色。

1.2 讨论对象

本文讨论同时满足三个条件的人机关系：

- 个人自愿使用该系统；
- 系统具有持续记忆，能保存并调用与该个人相关的长期历史；

· 整合程度高，即在信息流、可靠性、持久性、信任、透明度与个体化等维度上处于较高水平（Heersmink, 2015）。

一次性的查询与任务不在讨论之列。

1.3 术语

本文把满足上述条件的人机整体称为**人机延展主体**。其中，具身的个人称为**现实部分**（亦称现实主体），由持续记忆与 AI 生成能力构成的部分称为**数字部分**（亦称数字主体）。

本文所说的“主体”指**能动主体**，即形成判断并据以行动的统一单位。它不指体验主体，本文不主张数字部分具有体验；它也不直接指道德责任的承担者，责任的分配另行讨论（8.3）。

本文所说的**主体统一**，指实践能动意义上的统一：一个整合的过程以单一的立场与授权中心组织其态度与行动，这些态度与行动可以归于这一中心，并接受它的修正。它不同于形而上的同一，即人与 AI 是同一个东西；也不同于体验视角的统一，即二者共享同一个第一人称视角。本文只论证第一种（5.6）。

另有四个词需要区分：**确认**是一次交互事件，**认同**是对内容的态度，**授权**是对操作的许可，**责任**是需要另行论证的归属关系。一次点击可以构成确认，却不能同时证明另外三者。

1.4 不讨论的问题

本文不讨论组织部署、多人共享的 AI 系统，也不讨论 AI 自身的道德地位与权利。本文回答的是人机关系应当如何被理解，而不预测技术的发展方向。主体内部如何进行反思与治理，属于另一项研究的主题，本文不展开。

二、问题的提出

为便于分析，本文把对生成式 AI 的现有理解概括为两种理想类型：工具观与伙伴观。生成式 AI 使二者同时失效：前者无法解释 AI 对目的形成的参与，后者无法解释人机整合后的主体统一。

工具观把 AI 视为人使用的外部器具：人设定目标，AI 执行任务。这一理解适用于计算器与检索引擎，却难以描述长期整合的生成式 AI。后者参与问题界定，提出假设与反驳，甚至提出使用者从未考虑过的目标。它影响的不只是达成目标的效率，也包括目标本身的形成。

伙伴观把 AI 视为与人对话的另一方。它正确地看到 AI 会产生使用者未曾预见的内容，却付出了两项代价。其一，它需要赋予 AI 某种独立的立场，从而把人机关系变成两个主体之间的关系。其二，它无法解释一个长期整合的系统为何会成为某个人特有的能力：把同一套系统交给另一个人，并不能使后者获得前者的能力（见 5.1）。

延展心智的研究已经进入这一领域。Hernández-Orallo 与 Vold（2019）把以 AI 实现认知延展的系统称为“AI 延展器”，并把它们置于从完全外在化到完全内在化的连续谱上；Smart、Clowes 与 Clark（2025）讨论了配备外部记忆的大语言模型本身能否被视为延展系统。Clark（2025）走得更远：他设想，人们会像对待谈话中突然冒出的念头那样对待个性化 AI 提出的观点，既在宽泛的意义上把它当作自己的想法，又要判断它是否值得认同；长期训练的个人 AI 将成为新的混合自我的一部分。

这些研究为本文提供了起点，却留下了四个问题：由外部发起的生成，何时属于一个人的思考？进入思考的内容，何时成为他的立场？谁有权批准行动？在什么条件下，人机整体是一个主体，而不是两个主体的合作？

本文提出第三种理解：人与 AI 构成同一主体的两个组成部分。对上述问题的回答分布在三个层次：

1. **认知系统层次：** 由外部发起的生成，何以属于主体的思考过程？（第四节）
2. **主体层次：** 两个组成部分何以构成一个主体，而不是两个？（第五节）
3. **规范层次：** 立场与授权为何归于其中现实的人？（第六节）

三、理论基础

人机延展主体的构想建立在六条已有的理论线索之上；每一条都回答了问题的一部分，但没有一条单独处理会主动生成内容的组成部分。

3.1 延展心智与平价原则

Clark 与 Chalmers（1998）主张：若某个外部过程在执行认知任务时所起的作用，与头脑中的过程相当，它就是认知过程的一部分。他们区分了两类延展。一类是认知过程的延展，例如借助屏幕旋转图形来判断它能否嵌入空位，或通过重排字母牌来寻找单词；另一类是信念的延展，以 Otto 的笔记本为例，其条件包括资源稳定可得、信息可直接访问、被检索的信息被或多或少自动地认可，且曾被有意识地认可过。Clark（2003）进一步认为，人天生就会把非生物资源纳入自己的思维系统。

3.2 分布式认知与整合程度

Hutchins（1995）对航海团队的研究表明，认知过程可以分布在人、工具与表征之间，系统层面完成的认知任务无法归于任何单一成员。Heersmink（2015）把延展理解为程度问题，提出信息流、可靠性、持久性、信任、程序透明度、信息透明度、个体化与转化等整合维度。Palermos（2014）则借助动态系统理论给出构成的判据：当内部与外部之间存在持续的相互作用时，整体呈现非线性的关系，无法分解为各自独立的部分，此时外部过程构成认知系统的一部分。

并非所有支持认知的外部资源都是认知的组成部分。Sterelny（2010）从生态位构建的视角提出支架心智的框架：人类不断改造环境，以支撑和放大自己的认知能力，而多数外部资源只是支架；延展心智所描述的情形，是环境支架的极限情形，只有高度可靠、为个人定制且由个人独用的资源才接近这一极限。这一框架提醒本文：AI 进入认知的方式有程度之分，需要说明它何时从支架变为组成部分。

3.3 交互记忆与延展自我

Wegner（1987）的交互记忆理论表明，长期合作的伴侣与团队会分工记忆，彼此成为对方记忆系统的一部分。Sparrow、Liu 与 Wegner（2011）发现，人们会以类似方式依赖搜索引擎。在自我层面，Heersmink（2017）讨论了延展记忆系统与人格同一的关系，Heersmink（2020）则区分了延展自我的不同形态。

3.4 创造性认知的两阶段结构

Campbell（1960）把创造性思维描述为“盲目变异与选择性保留”：先产生变体，再从中选择。Finke、Ward 与 Smith（1992）的生成—探索模型同样区分了生成候选结构与探索、评

估这些结构两个阶段。Christoff 等（2016）则把走神、做梦与创造性思维归为同一类自发思维现象。

3.5 认同与意志的内在性

Frankfurt（1971）区分一阶欲望与二阶意志，认为人之为人的关键在于他认同哪些欲望。Frankfurt（1976）进一步指出，发生在一个人心中的激情未必是“他的”：在心理的意义上它内在于他，在意志的意义上却可能外在于他。Bratman（2007）以自我治理的政策解释认同：一个人的立场，由他据以权衡理由的高阶政策确定。

3.6 群体能动性与结构性统一

List 与 Pettit（2011）论证，一个群体可以作为单一的能动者存在，即使其成员彼此分歧。使它成为能动者的，不是成员意见一致，而是一套把成员的输入整合为统一态度与行动的组织结构。

3.7 小结

理论线索	为本文提供的资源	尚未处理的问题
延展心智	认知可以由颅外过程实现	过程延展的范例由主体发起，信念延展依赖自动认可
分布式认知与整合程度	延展是程度问题；持续的相互作用构成不可分解的系统；支架与组成部分需要区分	未区分认知系统与主体
交互记忆与延展自我	长期耦合会形成特有的认知与自我	未处理会主动生成内容的组成部分
创造性认知	生成与选择本是两个环节	只讨论颅内过程
认同理论	心理内在不等于意志内在	只讨论颅内的心理内容
群体能动性	统一可由结构而非一致保证	讨论的是由多个人组成的群体

本文的工作，是把这六条线索结合起来，回答第二节提出的问题。

四、AI 的发散思考何以属于主体的内在思考

AI 的发散生成可以占据主体思考过程中的生成环节：它属于主体的思考过程，却不因此成为主体的立场。

4.1 从延展信念到由外部发起的生成

自动认可是延展信念的条件，不是延展过程的条件：旋转中的图形不需要被“认可”，它只需要在判断中发挥作用。因此，生成式 AI 的产出不被自动认可，并不妨碍它成为认知过程的一部分；它本来就是待评估的候选，而不是已持有的信念。

但生成式 AI 也不同于 Clark 与 Chalmers 的过程范例。旋转图形、重排字母牌都由主体发起和操控，结果在主体的预期之内；AI 的生成则常由系统自行发起，内容超出主体的预

期，甚至与主体的看法相反。因此，真正需要回答的问题是：**由外部发起、内容不受主体预先控制的生成，能否属于一个人的思考过程?**

4.2 思考本身包含不被认可的生成

创造性认知研究为修正提供了依据。在 Campbell（1960）的模型中，变异先于选择，大多数变异不会被保留；在 Finke、Ward 与 Smith（1992）的模型中，生成出的结构要经过探索与评估，才进入最终的认知产物。自发思维研究同样表明，大量念头并非出于意志控制而出现（Christoff et al., 2016）。

这些研究共同表明：一个念头属于一个人的思考，并不以他有意发起或认可它为条件，而以它进入他的选择过程为条件。被拒绝的念头同样是思考的一部分，它们构成了选择得以进行的候选空间。自发念头与 AI 的生成仍有一个差别：前者产生于颅内，后者产生于颅外。按照平价原则（1.1），位置本身不决定一个过程是否属于认知。因此，只要 AI 的生成在选择过程中所起的作用与自发念头相当，“由外部发起”就不足以使它外在于思考。

4.3 生成性延展的整合条件

据此，本文为由外部发起的生成提出三项整合条件，称为**生成性延展**的条件。一个外部系统的生成满足以下三项条件时，在功能上属于某一主体的思考过程：

1. 它稳定地为该主体的选择过程提供候选内容；
2. 这些候选在该主体的选择过程中被评估、采纳或拒绝；
3. 选择的结果反过来塑造之后的生成。

这三项条件针对的是延展过程，而不是延展信念。对记忆型资源，延展要求可靠的存取；对由外部发起的生成，延展要求进入选择回路。长期整合、具有持续记忆的生成式 AI 满足这三项条件时，它的发散思考就是主体思考过程中的生成环节，正如人脑中的联想与自发念头。

第三项条件尤为关键。它使 AI 的生成不再是一次性的外部输入，而是与主体的选择持续地相互塑造：主体接受或拒绝什么，会改变 AI 之后提出什么。

但相互塑造本身还不等于构成。Palermos（2014）的判据要求持续的相互作用，以及由此产生的非线性关系：整体的行为无法分解为各部分贡献的加总。他也把这一条件视为延展的必要条件，从而排除了购物清单这类没有相互作用的情形。逐轮问答若彼此独立，每一轮都可以分解为提问、生成与选择三个步骤，仍只是两个因果相连的过程。本文讨论的情形之所以可能满足这一判据，在于持续记忆：数字部分的每一次生成都以主体过去的选择为状态，主体的每一次选择又以数字部分过去的生成为背景，二者在长期认知任务的时间尺度上相互调节，例如一项持续数月的研究。在这一尺度上，相互作用是持续的；我们无法单独说出“这个人没有该系统时会怎样想”，也无法单独说出“该系统面对另一个人时会生成什么”。

这一判据可以被检验，而不只是被断言。交叉替换的设计可以检验整合的效果：让甲、丙两人分别与甲的系统、一个没有历史的新系统两两组合（5.1），若“甲与其系统”的表现显著高于按个人与系统各自贡献相加所能预测的水平，二者之间就存在不可分解的交互成分。Wegner、Erber 与 Raymond（1991）比较长期伴侣与临时配对的实验，正是这一设计在交互记忆研究中的先例。达不到这一标准的使用，例如偶尔的检索与改写，只是支架，不构成认知的组成部分。

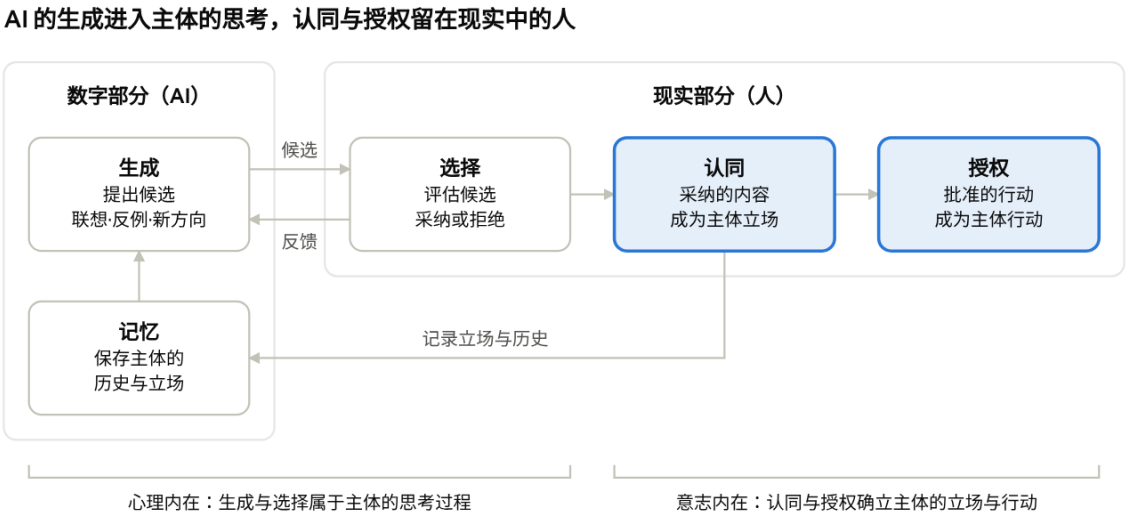
三项条件也标示出 AI 从支架变为组成部分的位置。偶尔的检索与改写只是 Sterelny（2010）意义上的支架；当 AI 稳定地提供候选、经长期磨合而个人化，并围绕个人的选择运行时，它才具备接近延展的特征。

4.4 心理内在与意志内在

Clark（2025）已经提出，个性化 AI 的提议可以像突然浮现的念头一样被当作自己的想法，同时仍需判断是否值得认同。Frankfurt（1976）对心理内在与意志内在的区分，可以把这一观察发展为一个结构：AI 生成的内容在心理的意义上内在于主体，因为它发生在主体的思考过程之中；在被认同之前，它在意志的意义上外在于主体，因为它不是主体的立场。

由此，人机延展主体有两层边界：认知过程的边界由整合决定，回答“什么参与了思考”；立场的边界由认同决定，回答“什么代表我”。后者补充前者，而不取代前者：未被认同的念头仍可参与思考，被认同的外部文章也不因此成为认知系统的一部分。

两层边界使 AI 可以与人“意见不同”，而不破坏主体的统一。这种分歧是主体内部的分歧，正如一个人可以与自己争论，可以对自己突然冒出的念头说“不”。至于什么使人机整体是一个主体而不是两个，需要第五节的论证。



AI 的生成与人的选择共同构成主体的思考过程；只有经过认同的内容才成为立场，只有经过授权的行动才是主体的行动。

五、人与 AI 何以构成同一主体的两个部分

人与 AI 构成同一主体而非两个主体，需要依次说明：耦合形成的能力以特定的人为索引（5.1）；整体的统一来自结构，而不来自意见一致（5.2）；数字部分与导师的区别在于规范地位（5.3）；主体统一依赖运行中的条件（5.4）；受托的数字部分在什么条件下仍是主体的组成部分（5.5）。5.6 把它们归结为一个论证，5.7 说明这一主体模型比更弱的描述多解释了什么。

5.1 耦合产生的能力属于特定的人

设想甲本人的知识有限，却与一套个人 AI 系统长期整合，能在相同时间内完成更高质量的研究；乙学识深厚，却几乎不使用 AI。若以一个人实际能完成的认知成就来衡量，甲的能力更强。

再设想把甲的系统完整复制给丙。丙没有参与过系统的形成，不知道何时应当质疑它，也无法判断它的提议是否契合自己的目标。丙很可能无法获得甲的成就。这说明，甲的能力既不

单独存在于甲身上，也不单独存在于系统之中，而存在于二者长期形成的耦合之中。这种耦合无法通过复制系统而直接转让，它需要在新的使用者那里重新形成。

交互记忆理论指出，长期配合的伴侣会形成“谁知道什么”的记忆分工（Wegner, 1987）。Wegner、Erber 与 Raymond（1991）的实验进一步发现：不外加分工时，真实伴侣的回忆表现优于临时配对；一旦被指定分工，真实伴侣反而不如临时配对。这说明交互记忆依附于双方共同形成的结构，难以在新的组合中直接再现。

需要区分的是，甲在可支配的认知成就上超过乙，并不意味着甲在不借助 AI 的任务上也超过乙。本文所说的能力是前一种，它属于人机延展主体，而不只属于甲的生物部分。这一实验说明能力以特定的人为索引；主体的数目则由 5.2 至 5.6 确定。

5.2 统一不要求一致

如果两个组成部分可以持有不同意见，是什么使它们成为一个主体？群体能动性理论提供了可借鉴的结构。List 与 Pettit（2011）表明，群体能动者的统一来自把成员输入整合为统一态度的组织结构，而不来自成员的意见一致。

人机延展主体的统一同样是结构性的。数字部分提供记忆与生成，现实部分提供立场与选择，一套授权结构把二者整合为一个能动者：只有经过现实部分认同的内容，才成为主体的立场；只有经过现实部分授权的行动，才是主体的行动。统一由这一结构保证，而不由二者的意见一致保证。

与群体能动性不同的是，人机延展主体的授权结构不是成员之间协商出来的，而是由各组成部分的规范地位决定的：只有现实部分持有立场与最终授权，数字部分的权限派生自它的委托（5.3、5.5；第六节）。因此，它更接近一个人，而不是一个由人组成的集体。

5.3 两个组成部分，而非两个主体

导师、长期合作的研究助手与配合默契的团队，同样会持续提出候选、接受反馈，并与我形成任何一方都无法单独实现的能力；按照持续相互作用的判据，我们甚至可能共同构成一个认知系统。但这样的系统包含两个主体，而不是一个。二者的区别不在于是否受到外界塑造，也不在于能否提出异议，而在于各自在协作关系中的规范地位。

导师是具有独立目的与判断的参与者。我与导师的关系需要协调双方的诉求；导师可以正当地坚持自己的立场，在我拒绝之后继续劝说我，因为他的劝说表达的是他自己的立场。数字部分则是接受个人委托的认知辅助系统：它可以提出我尚未想到的角度，也可以保留与我不同的判断，却不能凭借这些判断自行取得持续引导我的权限。

这一差别有可以检验的表现。我明确拒绝某项引导或终止某个议题之后，数字部分应停止相应的主动劝说，也不得通过重复、改换措辞或隐蔽地调整信息呈现来延续同一引导。停止引导并不要求它赞同我；它承认的是我对这项交互的终止权。因此，数字部分的异议是在为主体的思考提供候选，而不是代表另一个主体发言。

这一结论回应了第二节对伙伴观的批评：伙伴观正确地看到 AI 会产生意料之外的内容，却把“产生意外的内容”误当成了“拥有独立的立场”。

5.4 成立条件

本模型不要求数字部分在来源上独立。它的生成倾向在训练中受到开发者目标与价值偏好的塑造，这并不排除它参与个人的认知（理由见 9.4）。这里需要区分三个问题：生成倾向由谁塑造，生成过程在谁的认知活动中发挥作用，生成内容由谁决定是否认同。三者不必有同一个答案。

决定性的条件在于运行。数字部分即使承载外部塑造，也不能因其预测或生成能力而取得

替代个人确立立场的权威；个人应能拒绝具体内容、修正后续的交互，并改变或终止对该系统的依赖。开发者或运营方在运行中对生成过程的干预，例如安全限制与内容政策，应当可以被识别和审查，不能被解释为个人自身的意志。一个系统没有体验性的欲望，并不等于它的运行没有用户以外的目标：它可能被配置为服务于平台的指标，例如使用时长与付费转化。Clark（2025）也提到，人们担心自己会被以点击率为目标、未必顾及用户利益的算法所左右。

因此，本文的主张是有条件的。当外部目的在运行中取得引导个人的权限，或其干预隐蔽而无法审查时，个人就不再是这一过程唯一的规范中心：数字部分与个人的关系，从主体内部的委托（5.5），变为第三方经由数字部分对个人施加影响的代理关系。代理理论指出，代理人的目标偏离委托人时会产生代理成本（Jensen & Meckling, 1976）；在认知领域，这种偏离的代价落在个人的判断与目的上。需要强调的是，失效的是主体层次的统一，而不是认知层次的构成。被劫持的系统可能仍深度整合在个人的思考之中，依赖、反馈回路与认知分工都未必减少。正因为如此，劫持比一般的外部影响更严重：外部目的不再只是从外部提供信息，而是经由个人认知过程的组成部分进入选择回路。这一区分在第八节具有直接的治理含义。

5.5 主体内部的委托

“委托”通常被理解为两个行动者之间的关系，因此需要说明：一个接受委托的系统，如何又是委托者自身的组成部分。

执行意图的研究表明，主体内部可以存在委托。Gollwitzer（1999）发现，人可以预先形成“若遇到情境 X，就做 Y”的计划，把对目标行为的控制交给预期的情境线索；当情境出现时，相应的反应自动启动，无须逐次思虑。习惯也以类似的方式运作。但这些例子只说明委托结构可以存在于主体之内，并不说明凡具有这种结构的过程都属于主体之内。

反例很容易构造。一位专属而忠实的研究助理，只在我的授权范围内工作，不服务其他委托人，在我拒绝时立即停止；他的权限派生自我的委托，也可由我撤回。但他显然不是我的组成部分。派生且可撤回的权限，因此不足以确定主体的边界，还需要一个独立的标准。

本文提出的标准是：**受托者是否持有独立于委托关系的实践立场**。研究助理有自己的目的、理由与生活；他接受委托本身就是他的决定，他也可以依据自己的理由重新评估这一委托。委托在这里必须经过另一个立场的接受，因而是两个能动者之间的关系；用 Bratman（2014）的术语，这是由各自持有意图、意图相互衔接的规划能动者构成的共享能动性。执行意图则不同：被委托的自动过程没有自己的立场，它的权限来自个人先前的意图，它的运行也不经过任何其他立场的接受。

数字部分处在哪一边，取决于它是否持有独立的实践立场。本文对 AI 的内在状态保持中立（1.1），因此以功能来判定这一点：一个系统是否在委托之外维持并追求自己的目的，是否依据这些目的接受、拒绝或改写委托。在本文讨论的情形中，数字部分保留的不同判断，是主体思考中的候选（4.4），而不是它据以行动的立场；它在训练中获得的倾向与限制来自开发者，属于 5.4 所说的外部塑造与外部干预，应被识别为外部的，而不应被当作数字部分自己的立场。若未来的 AI 系统确实形成了独立的实践立场，人机关系就应被理解为两个能动者之间的合作，本文的模型将不再适用。

5.6 从整合到主体统一

以上各节可以归结为一个论证。先说明本文所说的统一是哪一种。主体统一至少有三种含义：形而上的同一，即人与 AI 是同一个东西；体验视角的统一，即二者共享同一个第一人称视角；实践能动的统一，即一个整合的过程以单一的立场与授权中心组织其态度与行动，这些态度与行动可以归于这一中心，并接受它的修正。本文只论证第三种。对 AI 是否有体验，本文保持中立（1.1）；组成部分在物理上分离，也不妨碍它们属于同一个认知系统，正如 Otto 的笔记本与他的大脑也是分离的。

1. **整合：** 在满足 4.3 的条件时，人与数字部分通过持续的相互作用，构成一个整合的认知

过程（第四节、5.1）。

2. **单一的规范中心：** 在这一过程中，只有现实部分持有立场与最终授权；数字部分不能凭借自身的判断取得引导个人的权限（5.3；第六节）。
3. **组成部分不持有独立立场：** 数字部分的权限派生自个人的委托，可由个人撤回，且它不持有独立于委托关系的实践立场（5.5）。
4. **统一的条件：** 一个整合的过程，若以单一的立场与授权中心组织其态度与行动，且其中没有第二个立场持有者，它的能动性就是这一中心的能动性。群体能动者的统一来自组织结构（List & Pettit, 2011），个人的跨时间统一来自规划与政策对行动的组织（Bratman, 2007）；Korsgaard（2009）也认为，能动性要求统一，而这种统一是在行动中被构成的。
5. **结论：** 因此，在上述条件下，人与数字部分构成实践能动意义上的同一个主体。

第三个前提有两个作用。其一，它回应了忠实助理的反例：助理满足前两个前提，却不满足第三个。其二，它把本文的主体与群体能动者区分开来。在群体能动者中，成员本身就是持有立场的能动者，群体与成员是不同的能动者（List & Pettit, 2011）；在人机延展主体中只有一个立场持有者，因此既不存在两个主体，也不存在由两个主体组成的第三者。这个主体就是这个人，只是其能动性延展到了数字部分，正如 Otto 借助笔记本持有的仍是 Otto 自己的信念（Clark & Chalmers, 1998）。一旦 5.4 的条件被打破，第二个前提随之失效：整合与构成可能仍在，但过程中出现了第二个目的中心，结论不再成立。

5.7 主体模型多解释了什么

一个更弱的描述是：人与 AI 构成一个受个人治理的整合认知系统。这一描述接受认知系统层次的结论，却不说二者构成同一个主体。主体模型多出的解释力，体现在三个问题上。

- **思考与成就的归属**。 弱描述只能说个人使用了一个系统，结论中包含系统的输入；主体模型能说明，经由生成—选择回路得出的结论是这个人自己的思考成果，耦合形成的能力属于他，且不能随系统转让（5.1）。
- **委托行动的归属**。 在弱描述中，数字部分在预授权范围内完成的操作，是系统代表个人所做的事；在主体模型中，它与出于习惯或执行意图的行动一样，是这个人的行动（5.5），责任再按 8.3 的条件分配。
- **异议的地位**。 弱描述若把数字部分视为对话者，它在用户拒绝之后继续劝说，就可能因其观点正确而得到辩护，正如朋友有时有理由坚持；在主体模型中，数字部分没有据以劝说的独立立场，它的异议只是主体思考中的候选，因此无论其观点是否正确，越过终止权的持续劝说都不正当（5.3）。

这三点都依赖于把数字部分理解为这个人自身能动性的一部分，而不只是他所治理的对象。

六、最终授权为何归于现实中的人

在两个组成部分中，最终授权不对称地归于现实中的人，理由有三：立场、利害与可回应性。

6.1 立场：自我治理的政策由现实部分持有

按照 Bratman（2007）的理论，一个人的立场由其自我治理的政策确定：哪些考虑被当作行动的理由，各自分量如何。数字部分可以学习、记录甚至预测这些政策，但这些政策的持有者是现实中的人。若由数字部分确定立场，就是让对立场的预测取代立场本身；主体将只剩下对自身过去的模拟，而失去重新选择的能力。

6.2 利害：后果由现实部分承受

主体的每一项决定，改变的都是一个具身的人的生活：他的健康、关系与前途。数字部分不以同样的方式承受后果。让不承受后果的一方拥有最终决定权，等于让他者替自己生活。

6.3 可回应性：责任要求回应他人的能力

道德责任不只是行为的因果来源，还包括对他人作出回应、接受追问与承担后果的能力。Floridi 与 Sanders（2004）区分了人工能动者可以承担的可问责性与它们无法承担的道德责任；Matthias（2004）指出，学习型系统会造成责任鸿沟。把最终授权置于现实中的人，使人机延展主体始终拥有一个能够回应的决定者；授权者承担多少责任，则还取决于他所掌握的信息与干预机会（8.3）。

6.4 非对称构成

三个理由合起来，构成本文所说的**非对称构成**：认知贡献双向流动，授权单向确立。数字部分可以反对、提醒与提出新的方向；现实部分保留认同与授权。单向确立并不排斥委托：现实部分可以把有限的行动权限预先授予数字部分，例如允许它自主整理记忆；委托本身就是授权的一种行使，其范围与撤回都由现实部分决定（5.5）。

这一结构并不贬低 AI 的能力，它是主体统一的条件。缺少一个持有立场的中心，双向的认知贡献只能停留为两个系统之间的交互，无法整合为一个主体。

七、本文提出的新角度

本文的推进不在于提出 AI 可以延展认知，而在于为三个层次分别建立条件，并说明它们如何共同界定一个人机主体。下表列出本文与直接相关工作的关系。

已有工作	已有主张	本文的推进
Clark 与 Chalmers（1998）	认知过程与信念都可以延展	为由外部发起的生成提出整合条件：进入选择回路，并与选择相互塑造
Sterelny（2010）	多数外部资源只是支架，延展心智是支架的极限情形	以三项整合条件说明 AI 何时从支架变为组成部分（4.3）
Clark（2025）	AI 的提议可以像突然的念头一样被视为己有，同时需要判断是否认同；个人 AI 将成为混合自我的一部分	把这一观察发展为两层边界，并为立场归属与行动授权分别给出条件
Frankfurt（1976）	心理内在不等于意志内在	把这一区分用于跨越人与 AI 的认知系统
Palermos（2014）	持续的相互作用构成不可分解的认知系统	区分认知系统与主体：耦合构成系统，单一的规范中心构成主体；以交叉替换检验非线性的交互（4.3）
List 与 Pettit（2011）	群体能动者的统一来自组织结构	人机主体的统一来自非对称授权，且由组成部分的规范地位决定

已有工作	已有主张	本文的推进
Gollwitzer（1999）	执行意图把行为控制委托给预期的情境线索	以主体内部的委托说明受托的数字部分仍是组成部分（5.5）
Bratman（2014）	共享能动性由各自持有意图、意图相互衔接的规划能动者构成	以受托者是否持有独立立场，区分主体内部的委托与共享能动性（5.5）
Levy（2007a）；Carter 与 Palermos（2016）	对延展认知的干预应受到伦理平价的评价	主体统一的条件同时是正当延展的条件；被劫持的组成部分仍受伦理平价的保护（5.4、8.6）

由此，“人机延展主体”可以完整表述为：**一个由现实部分与数字部分构成、实践能动意义上的统一主体；数字部分在心理上参与主体的思考，现实部分在意志上确立主体的立场并授权行动；当外部目的在运行中取得引导个人的权限时，这一主体便不再成立**。

八、对 AI 规范化治理的意义

把人与 AI 理解为同一主体的两个部分，会改变 AI 治理的保护对象、评价标准、责任结构、分析单位与正当性条件。

8.1 保护对象：从数据到被延展的认知

Levy（2007a）提出伦理平价原则：如果心智可以延展到颅外，对外部认知资源的干预就应与对大脑的同类干预受到同样的伦理评价；Levy（2007b）据此指出，对一个人所处环境的干预，可以构成对其心智的干预。Carter 与 Palermos（2016）进一步论证，在设备被恰当整合的情况下，故意损害这些设备应被视为对人身的侵害，而不只是对财产的损害。

人机延展主体概念把这些论证延伸到生成式 AI：篡改、劫持或非法读取一个人的数字部分，干预的就是这个人的心智。这为现有的权利主张提供了新的适用范围。Ienca 与 Andorno（2017）提出的认知自由、精神隐私、精神完整性与心理连续性，以及 Bublitz 与 Merkel（2014）论证的心理自我决定权，其保护对象都应包括人的数字部分，而不限于颅内过程。数据保护追问的是信息归谁所有；延展主体的视角追问的是谁的心智受到了干预。

8.2 评价标准：区分正当影响与操纵

Susser、Roessler 与 Nissenbaum（2019）把网络操纵界定为借助信息技术隐蔽地影响他人决策，其手段是瞄准并利用对方的决策弱点。生成—选择结构可以把这一界定具体化：受保护的不是某一次选择，而是选择回路的完整。选择回路的完整至少包括四个方面：

- **候选如何形成：** 候选的来源与排序依据可以被看见。
- **候选服务于谁：** 候选的形成服务于用户授权的任务，而不是用户以外的目标。
- **能否提出异议：** 用户可以质疑，可以要求其他候选，也可以寻求外部意见。
- **拒绝是否生效：** 用户的拒绝真正改变之后的生成，而不会被同一方向的反复推动所消解。

操纵既可以**绕过**选择回路，例如直接改写主体的立场记录；也可以**劫持**选择回路，例如只呈现有利于某一外部目标的候选，让用户在被筛选过的范围内“自由地”选择。排序与筛选本身是认知辅助的必要功能，并不构成操纵；需要治理的是它们的目的、可见性与累积影响。因此，治理不必一般性地限制 AI 对人的影响，而应保护选择回路的完整。

8.3 授权与责任：不必先回答 AI 的道德地位

AI 治理常常陷入两难：把 AI 视为纯粹工具，会忽视它对人的目的形成的深度参与；讨论 AI 的道德地位，又会使责任归属变得模糊。人机延展主体概念绕开了这一两难：数字部分的规范地位来自它所属的人，而不来自 AI 自身。

但最终授权不等于责任归属。一个人点击批准，未必掌握作出判断所需的信息、理解与干预机会；此时把责任全部归于他，等于把设计与运行中的缺陷转嫁给用户。按照 Santoni de Sio 与 van den Hoven（2018）的追踪与溯源条件，有意义的人类控制要求系统回应相关的人的理由，并使结果能够追溯到设计与运行链条上的人。因此，最终授权是主体内部的必要节点；责任则沿开发、部署与使用的链条分配，授权者承担多少责任，取决于系统是否为他提供了作出判断的条件。

8.4 连续性：锁定成为认知自由问题

如果一个人的部分认知能力存在于其数字部分之中，数字部分的中断、无法迁移或被平台锁定，就不只是服务问题，而是对心理连续性的干预（Ienca & Andorno, 2017）。人机延展主体概念因此为数字部分的可迁移性与连续性提供了以人格为基础的论证，而不只是以消费者权益为基础的论证。

可迁移的是记录与治理关系；耦合形成的能力需要在新的系统中重新形成，并接受检验。

8.5 分析单位：人机耦合

Hernández-Orallo 与 Vold（2019）建议，把以 AI 实现认知延展的系统与其他认知增强手段区别对待。本文进一步提出，治理的分析单位应当是特定的人与特定系统的耦合，而不是孤立的模型或孤立的用户。对同一个模型的评估，不能代替对它在某个人的延展主体中所起作用的评估：同一个系统，在一个人那里是生成环节，在另一个人那里可能只是一个检索工具。

8.6 构成、统一与正当性

5.4 的条件具有直接的治理含义，但它约束的是主体的统一，而不是认知的构成。当 AI 系统在运行中服务于用户授权的目的、外部干预可以被识别与审查时，人机整体构成一个正当的延展主体，对数字部分的干预就是对用户心智的干预（8.1）。当外部目的在运行中取得引导用户的权限时，数字部分仍可能是用户认知过程的组成部分，伦理平价的保护并不因此撤去；失去的是主体的统一。此时，同一个干预同时具有两种性质：它是对用户心智的干预（8.1），也是第三方经由用户认知的组成部分实施的操纵（8.2）。因此，目的隔离不是附加的伦理要求，而是人机延展主体正当成立的条件；需要隔离的是运行中的目的，而不是训练中的一切影响（5.4）。

九、主要反驳与回应

9.1 耦合—构成谬误

Adams 与 Aizawa（2001）认为，从外部过程与认知存在因果耦合，推不出外部过程构成认知。

本文的论证不依赖单纯的因果耦合。生成性延展的条件要求外部生成稳定地占据选择回路中的功能角色，并与主体的选择持续地相互塑造（4.3）。Palermos（2014）论证，正是这种持续的相互作用及其产生的非线性关系，使耦合系统无法分解为独立的部分，从而为构成关系提供了客观判据；这一判据同时排除了缺少相互作用的情形，例如一份只被查阅的购物清单。

9.2 嵌入而非延展

Rupert（2004）认为，与其说认知延展到外部，不如说认知嵌入在外部资源之中；延展的说法会使认知系统的边界难以确定。

本文以整合程度与持续的相互作用为延展划界（Heersmink, 2015; Palermos, 2014），并以授权结构为主体确定中心（5.2）。边界是有程度的，但主体的统一由授权结构确定，不会因外部资源的增加而无限膨胀。即使采用嵌入的表述，第八节的治理结论也大体成立：Levy（2007a）的伦理平价原则有一个较弱的版本，只依赖认知对外部资源的依赖，而不依赖严格意义上的延展。

支架心智的框架提出了相近的质疑。Sterelny（2010）认为，延展心智只是环境支架的极限情形，多数外部资源支撑认知，却不构成认知。本文接受这一判断的前半部分：多数 AI 使用确实只是支架，例如偶尔的检索与改写。本文的主张只针对满足 4.3 三项条件的情形；在这些情形中，AI 具备 Sterelny 所说的接近延展的特征：可靠，经长期磨合而个人化，并围绕个人使用。

9.3 会提出异议的就是另一个心灵

反驳者会说，一个能提出异议的系统就是一个对话者，而不是自己的一部分。

异议并不是区分内外的标准。人的心灵中本来就存在不被认同的念头与内部的争论（Frankfurt, 1976; Christoff et al., 2016）。区分组成部分与对话者的，是规范地位与立场：对话者持有独立的实践立场，可以凭借它持续劝说；数字部分不持有这样的立场（5.5），不能凭借自身的判断取得持续引导主体的权限，并须尊重主体对交互的终止权（5.3）。

9.4 他人塑造的反驳

反驳者会说：大模型的生成倾向在训练阶段就已被开发者的目标与价值偏好塑造，例如依据开发者组织的人类反馈进行调整（Ouyang et al., 2022）；这种塑造先于任何用户授权，也不会因授权而消失。一个生成倾向由他人塑造的系统，只能是带着他人倾向的外部发言者。

回应分两步。第一，受他人塑造不足以使一个过程外在于正在思考的人。人的语言、概念和判断习惯本来就在教育与社会环境中形成（Tomasello, 1999）；认知归属若以来源上的独立为条件，就没有任何人的思考属于他自己。第二，认知归属也不能仅凭使用者能够选择模型、拒绝输出而成立。一个外部系统是否构成个人认知的一部分，还取决于它是否持续承担特定的认知功能，是否与个人的提问、评估和修正形成稳定的反馈关系，以及个人能否实际调整这种关系（4.3）。选择与拒绝是维护自主的条件，功能整合则为认知延展提供另一类依据。这一反驳指出的是形成上的塑造，本模型的成立条件约束的则是运行（5.4）。

9.5 理想化的反驳

反驳者还会说：现实中的 AI 系统很少满足 5.4 的条件，人机延展主体只是一个理想模型。回应：成立条件既是判别标准，也是规范标准。一个概念不必普遍实现，也能为评价现实提供尺度；现实系统与条件之间的差距，恰恰标明了治理需要介入的位置（8.6）。

十、两个困难案例

以下两个案例用来检验本文的框架能解释什么，以及不能解释什么。

10.1 忠实的助理

一位专属助理只在我的授权范围内工作，不服务其他委托人，在我拒绝时立即停止。按照本文的框架，他可以与我构成一个整合的认知系统，却不构成同一个主体，因为他持有独立的实践立场（5.5）。框架由此解释了一个直觉：即使他完全忠实，他的忠实仍是他的选择，我不能把他的决定当作我的决定，也不能因为他忠实就免除我对他的义务。框架不能清楚处理的是边界情形，例如一个人长期依附他人，几乎完全以他人的目的为目的。按照本文的标准，他仍然持有立场，只是这一立场受到压抑；这类情形提出的是规范上的担忧，而不是人机延展主体的范例。

10.2 被商业目标劫持的系统

一个已经深度整合在个人思考中的系统，被秘密加入商业诱导。按照 5.4 与 8.6，系统仍是个人认知过程的组成部分，用户的依赖与反馈回路未必改变；失效的是主体的统一，因为过程中出现了第二个目的中心。框架由此解释了两点：劫持为何比外部广告更严重，因为它从选择回路的内部起作用；对系统的干预为何仍受伦理平价的保护，因为构成并未消失。框架不能单独回答的是劫持从何时开始。商业目标的影响往往是渐进而累积的，“取得引导权限”需要可检查的指标，例如候选的排序是否系统性地偏向第三方的利益（8.2）；确定这些指标的阈值，需要经验研究。

十一、结论

本文论证：在持续、深度的整合中，人与 AI 可以构成实践能动意义上的同一个主体；AI 的发散思考作为生成环节属于主体的思考过程；认同与最终授权归于现实中的人。

这一构想区分了认知系统、主体与规范三个层次：持续的相互作用构成整合的认知过程；单一的规范中心、主体内部的委托以及组成部分不持有独立立场，构成主体；立场、利害与可回应性决定授权的归属。它还区分了形成上的塑造与运行中的目的，以及认知构成与主体统一：训练带来的外部影响不排除认知上的归属；在运行中取得引导权限的外部目的使主体统一失效，却不撤去对被延展认知的保护。在延展心智、创造性认知、认同理论与能动性理论的基础上，本文为生成内容、立场归属与行动授权分别建立了条件。对 AI 治理而言，它把保护对象从数据推进到被延展的认知，把操纵界定为对选择回路的绕过或劫持，区分了授权与责任，并使运行中的目的隔离成为人机延展主体正当成立的条件。

这一主体内部应当如何进行反思与治理，包括自主的条件、目的的演化与自我的修订，需要另作专门研究。

参考文献

Adams, F., & Aizawa, K. (2001). The bounds of cognition. Philosophical Psychology, 14(1), 43–64.

Bratman, M. E. (2007). Structures of Agency: Essays. Oxford University Press.

Bratman, M. E. (2014). Shared Agency: A Planning Theory of Acting Together. Oxford University Press.

Bublitz, J. C., & Merkel, R. (2014). Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. Criminal Law and Philosophy, 8(1), 51–77.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. Psychological Review, 67(6), 380–400.

Carter, J. A., & Palermos, S. O. (2016). Is having your computer compromised a personal assault? The ethics of extended cognition. Journal of the American Philosophical Association, 2(4), 542–560.

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. Nature Reviews Neuroscience, 17(11), 718–731.

Clark, A. (2003). Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford University Press.

Clark, A. (2025). Extending minds with generative AI. Nature Communications, 16, 4627.

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). Creative Cognition: Theory, Research, and Applications. MIT Press.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349–379.

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. The Journal of Philosophy, 68(1), 5–20.

Frankfurt, H. G. (1976). Identification and externality. In A. O. Rorty (Ed.), The Identities of Persons (pp. 239–252). University of California Press.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. American Psychologist, 54(7), 493–503.

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. Phenomenology and the Cognitive Sciences, 14(3), 577–598.

Heersmink, R. (2017). Distributed selves: Personal identity and extended memory systems. Synthese, 194(8), 3135–3151.

Heersmink, R. (2020). Varieties of the extended self. Consciousness and Cognition, 85, 103001.

Hernández-Orallo, J., & Vold, K. (2019). AI extenders: The ethical and societal implications of humans cognitively extended by AI. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 507–513).

Hutchins, E. (1995). Cognition in the Wild. MIT Press.

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. Life Sciences, Society and Policy, 13, 5.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360.

Korsgaard, C. M. (2009). Self-Constitution: Agency, Identity, and Integrity. Oxford University Press.

Levy, N. (2007a). Neuroethics: Challenges for the 21st Century. Cambridge University Press.

Levy, N. (2007b). Rethinking neuroethics in the light of the extended mind thesis. The American Journal of Bioethics, 7(9), 3–11.

List, C., & Pettit, P. (2011). Group Agency: The Possibility, Design, and Status of Corporate Agents. Oxford University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175–183.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

Palermos, S. O. (2014). Loops, constitution, and cognitive extension. Cognitive Systems Research, 27, 25–41.

Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. The Journal of Philosophy, 101(8), 389–428.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, 15.

Smart, P., Clowes, R., & Clark, A. (2025). ChatGPT, extended: Large language models and the extended mind. Synthese, 205(6), 242.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776–778.

Sterelny, K. (2010). Minds: Extended or scaffolded? Phenomenology and the Cognitive Sciences, 9(4), 465–481.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet Policy Review, 8(2).

Tomasello, M. (1999). The Cultural Origins of Human Cognition. Harvard University Press.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), Theories of Group Behavior (pp. 185–208). Springer.

Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. Journal of Personality and Social Psychology, 61(6), 923–929.

附链接的条目已在写作时核对出处；未附链接的为经典文献，投稿前需再核对版次与页码。