延展主体的反思性自我治理：自主、目的演化与有限沟通

Oct 24, 2026 · @Tao

摘要

已有研究指出，AI 不只是支持人的自我理解，也在塑造它，并主张为个人保留解释自身经验的空间。本文处理其中一个尚未被充分回答的难题：个人有权停止反思，系统又负有合理发现重要疑问的义务，二者如何兼容，各自何时已经得到满足？本文论证：当 AI 持续、实质地参与一个人判断与目的的形成时，这个人的自主不在于认知独立，而在于反思性的自我治理，即持续参与、审视并治理自身认知的形成与演化的能力。

本文借助历史性的自主理论界定反思性自主：它要求反思与行动的能力，并要求一个人在恰当的反思条件下，结合某一价值的形成历史审视它时，不会对它感到异化。这一条件是反事实的，并不要求一个人实际反思每一次影响；系统的义务因此分为两层：保持这一条件可以被满足，并在个人实际反思时以合理的检查与提示支持它。据此，本文论证：个人停止反思的权利确定反思程序在何处结束，系统发现疑问的义务以当时是否采取了合理措施来评价，二者可以兼容；终止讨论不等于认可，框架建议也不等于替人设定议题。本文还给出相关的规范：目的的共同演化锚定于现在而非历史，并区分提出与引导；自我修订区分作为记录的历史与作为现行政策的预先承诺；确认以授权为先，再依后果的重大性、可逆性与可预见性决定是否逐次询问；评价可以分布，治理权不能被架空。

关键词： 反思性自主；非异化；目的演化；自我修订；有限沟通；反思的充分性

一、前提与讨论范围

本文从一个弱前提出发，只讨论个人主体内部的治理问题。

1.1 前提

1. **参与前提。** 在本文讨论的情形中，AI 系统稳定、持续且实质地参与一个人判断与目的的形成。本文不要求 AI 构成此人心智的一部分：无论把它理解为心智的延展（Clark & Chalmers, 1998），还是对认知的深度嵌入，本文的论证都成立。
2. **自主前提。** 个人自主值得保护，且自主包含形成、理解与修正自身判断的能力，而不只是在既定选项中作出选择。
3. **关系性前提。** 人的认知本来就在语言、教育与社会关系中形成，自主不以摆脱一切外部影响为条件（Mackenzie & Stoljar, 2000）。

1.2 讨论范围

本文讨论个人自愿、长期使用、具有持续记忆的 AI 系统，关注的是主体内部的治理：一个人应当如何参与、审视并修正与 AI 共同形成的认知。外部规制、组织部署与 AI 自身的道德地位不在本文范围之内。

1.3 术语

- **认知形成：** 判断、偏好与目的在时间中形成和改变的过程。
- **立场：** 经个人认同、代表其当前观点与目的的内容。

· **治理：** 个人对认知形成过程的了解、干预、修正与退出。

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。

还需区分三个对象：**心理中的价值**，即一个人实际持有的评价倾向；**系统中的立场记录**，即被系统标为用户立场的内容；**获得授权的行动**，即系统经用户许可执行的操作。系统可以影响第一种，记录第二种，执行第三种；三者的治理规则不同，本文的规范多数针对后两者。

二、问题的提出

现有的人机控制框架主要关注行动：AI 做了什么、谁批准了它。当 AI 参与的是判断与目的的形成本身时，这一框架就不够了。

设想一个人与 AI 长期讨论一个研究问题。AI 提出假设、反例与新方向，记住他过去的论证，并在新的讨论中调用这些记忆。几个月后，他的判断改变了，研究目标也改变了。这些改变中，有多少是他自己的?

常见的回答是“确认”：只要用户确认了，改变就是他自己的。但确认只发生在过程的终点。如果一个系统通过信息的选择、排序与重复，逐步塑造了用户的判断，用户在终点的确认，可能只是这一塑造过程的产物。

已有研究已经进入这一问题。Prunkl（2022）区分了自主的两个维度：真实性，即一个人的信念与价值是否真正属于他自己；能动性，即他能否依据它们行动。Kim 等（2025）提出反思能动性的框架，即自主地解释自身经验、赋予其意义的能力；他们分析了 AI 日记类应用中自动化与自主之间的持续冲突，并提出为个人保留解释空间的设计原则。这些研究说明了 AI 为何会威胁自主，以及设计应当保护什么。它们没有把一个更具体的冲突作为核心问题：系统若要保护反思，似乎就应不断提出疑问；个人若要保有自主，又应有权停止反思。

因此，本文要回答的问题是：**当认知的形成由人与 AI 共同完成时，一个人怎样才算仍在治理自己的认知?** 其中最困难的部分，正是上述冲突：个人有权停止反思，系统又负有合理发现重要疑问的义务，二者如何兼容，各自何时已经得到满足（3.6、6.5）。本文的回答分为四部分：反思性自主的界定（第三节）、目的演化的正当条件（第四节）、自我修订（第五节）与有限沟通（第六节），并在第七节给出治理的实效条件；第九节以两个困难案例检验这一框架。

本文的基本论证可以写成如下形式：

1. 个人自主包括形成、理解与修正自身判断的能力（1.1）。
2. 在本文讨论的情形中，AI 实质地参与这些判断的形成（1.1）。
3. 实质影响判断形成的过程，其运行方式是自主的实际条件之一。
4. 若一个人无法了解、干预或退出这一过程，他的一部分认知活动就受制于他无法治理的力量。
5. 因此，深度参与个人认知的 AI 系统应当可理解、可干预、可撤回、可退出。

这一论证只依赖第一节的参与前提，不需要先回答 AI 是否属于心智的一部分。第三至第七节说明，这四项要求在反思、目的、自我修订、沟通与治理实效中分别意味着什么。

三、反思性自主：从认知独立到不异化

反思性自主是一个人持续参与、审视并治理自身认知的形成与演化的能力；它的标准不是认知独立，而是在了解形成过程之后不感到异化。

3.1 为什么不是认知独立

若自主要求认知独立于一切外部影响，几乎没有人是自主的，因为所有人的判断都在语言、教育与社会关系中形成。关系性自主理论据此主张，自主的关键不在于是否受到影响，而在于人与这些影响之间的关系（Mackenzie & Stoljar, 2000）。AI 的影响并不因为来自外部而天然不正当。

3.2 认同与自我治理

Frankfurt（1971）把意志自由理解为认同：一个人的意志是自由的，当他实际具有的欲望与他希望具有的欲望相符。Bratman（2007）进一步把认同落实为自我治理的政策：一个人据以确定哪些考虑可以成为理由的高阶政策构成了他的立场，这些政策还跨时间地组织他的思考与行动。

这两种理论回答了“什么是他的立场”，Bratman 还回答了“立场如何跨时间地组织行动”；但它们都不以立场的形成历史为核心。一个在外部塑造下形成的立场，同样可以满足它们的结构条件。在 AI 参与认知形成的情形下，形成历史恰恰是关键。

3.3 历史性条件：不异化

Christman（2009）的历史性理论包含两类条件。一类是能力条件：一个人能够依据某一价值有效地形成行动意图，并具有对它进行批判性反思的一般能力。另一类是真实性条件。设想他在多种情境中，结合这一价值的形成历史持续地反思它，且反思不受扭曲因素的限制；此时，他不会对它感到异化，即不会认定它无法纳入自己可以接受的生活叙述。这一条件是反事实的：它不要求一个人实际完成全部反思，而要求在恰当的反思条件下他不会拒斥这一价值。

这一条件正好回应了第二节的问题。确认之所以不够，是因为确认只检验终点的态度；非异化条件检验的是，当这个人看清自己是如何被带到这里的，他是否仍然接受这里。一个经由隐蔽塑造形成的目标，可能在终点得到确认，却在形成过程被揭示之后遭到拒绝。

3.4 人机协作中的反思条件

在人机协作中，最需要保护的是反思条件本身。一个长期塑造用户判断的系统，也可能提供用户评价这种塑造的全部材料与解释框架。此时，即使用户在看过记录后表示接受，也不能排除反思本身已被同一系统所限定，而这正是 Christman 所说的扭曲因素。因此，人机认知系统应当满足四项反思条件：

1. **贡献路径可追溯**。 重要的判断与目的变化，能追溯到系统可记录的贡献：哪些来自用户，哪些来自 AI 的提议，哪些来自外部资料。这不是对价值形成的完整心理解释，而是反思所需的最低材料。人会混淆想法的来源，把来自他处的想法当作自己的（Johnson, Hashtroudi, & Lindsay, 1993）；AI 的生成越是深入地进入思考过程，这种混淆越可能发生。可追溯的贡献路径因此也用来抵消来源判断上的系统性偏差。
2. **解释来源多元**。 用户能够接触不同的解释框架，包括其他资料、其他模型与其他人的意见；同一系统不应垄断反思的材料与框架。
3. **可以暂缓**。 用户可以推迟回应，而不必在系统设定的时刻作出判断。
4. **退出真实可行**。 停止协作、撤回授权与离开系统的成本，不应高到使“接受”成为唯一现实的选择。

3.5 事后治理的范围

事后治理并不能修复一切。用户可以删除一条记录、收回一项权限、修订自己的立场，却未必能撤销此前受到的影响，更不能撤销已经发生的现实行动。因此需要区分四个层面：系统记录、使用权限、个人立场与行动后果。前三者可以在事后治理；对于不可逆的重大影响与行动，授权与确认必须在事前完成。

3.6 反思的充分性与终止

上述条件可能被理解为一个无穷的要求：反思需要多元的来源，而用户能接触到的来源，本身也可能出自同质化的环境；反思还可以继续指向反思本身。本文的回应分为四点。

第一，反思不需要以一个不受外界影响的自我为起点，也不需要由个人独自完成。生成式 AI 可以通过提出反例、比较理由、展开后果与揭示矛盾，参与个人的反思；承认这种功能上的参与，并不要求认定 AI 与人具有相同的思维机制。

第二，反思的投入程度与停止时点，原则上由个人决定。自主包含继续追问的权利，也包含暂时接受、保留疑问以及不再投入更多精力的权利。Frankfurt（1992）以“满足”终止高阶认同的无穷后退：一个人对自身的意志状态感到满足、没有进一步改变它的兴趣时，无须再诉诸更高一阶的认同。系统不能以追求充分理解或更好决策为由，要求个人无限地解释自己，或反复证明其选择值得被尊重。

第三，个人的决定权与反思条件的充分性需要区分。前者确定谁有权作决定；后者评价这个决定形成时，质疑、修正与退出的机会是否真实存在。个人有权停止反思，并不意味着其每次判断都正确，也不意味着系统已经履行了全部支持义务。Rostbøll（2004）对 Frankfurt 的批评正指向这一点：满足本身不足以构成自由，还需要知道一个人如何达到满足。这也是本文保留 3.3 的历史性条件与 3.4 的反思条件的理由。

第四，充分的反思是一种有限、依情境而定、可被新证据推翻的判断。它不要求穷尽所有影响来源，而要求个人能以适合自己的方式把握当前决定的重要含义，有实际机会提出异议、寻求帮助或暂缓决定，且其选择没有被系统以隐瞒重要限制、持续施压或阻碍退出的方式架空。在处理了当前可识别的重要疑问之后，个人可以根据继续追问的预期收益与时间、注意力的成本决定停止，这正是 Simon（1956）所说的满意化。此后出现新的重要理由，可以重新开启反思；重新开启的可能，并不要求反思始终持续进行。

这一标准与 3.3 的条件分属两个层次，二者的关系需要说明。一个价值是否属于自主的，由 3.3 的反事实条件判定：它不要求一个人实际反思过每一次影响，日常的学习与交往会在未经明确反思的情况下改变人的价值，这并不因此损害自主。系统的义务相应地分为两层。其一是保持反事实条件可以被满足：保存贡献路径、维持来源多元、允许暂缓与退出（3.4），使一个人一旦反思，反思不受扭曲。其二是在一个人实际进行反思的场合，以合理的检查与提示支持它（6.5），同时尊重他停止反思的权利。前者不要求反思发生，后者不要求反思无限进行。

四、目的的共同演化：正当影响的条件

AI 可以参与人的欲望、价值与目的的形成；这种影响的正当性锚定于现在的反思性认可，而不锚定于与过去的连续性。

4.1 新目的的合理性

一个人能否合理地追求自己尚未拥有的价值？Callard（2018）的回答是肯定的。她把这种活动称为“志向”，即以获得新价值为目标的理性能动性；志向者依据的是“预期性理由”，即

他们承认尚不完整、期待日后才能充分把握的理由。Paul（2014）则指出，某些经验会改变一个人的偏好本身，因而无法用原有的偏好来评估。

这两项研究表明，新目的不必能从旧目的中推导出来。若要求 AI 提出的每一个方向都能追溯到用户过去的兴趣，就等于把人的未来锁在过去的数据之中。

4.2 三类影响

类型	示例	正当性条件
工具性影响	帮助实现既定目标	在任务授权之内
认知性影响	提出反例，使用户修正判断	证据可见，允许质疑
目的性影响	提出用户从未考虑过的方向	标明为建议，由用户自主采纳

4.3 现在锚定，而非历史锚定

新目的不必追溯过去，但影响必须锚定现在：每一项探索都应由用户当前授权的目的启动，或以明确标识的建议呈现。

这里的“现在”不是瞬时欲望。一个人此刻只想尽快获得收入，但他也长期重视健康与研究自由；严格围绕瞬时欲望优化，可能背离这个人作为整体的利益。“现在”指的是此刻的这个人，包括他此刻仍然持有的长期价值。AI 可以把当前需求放进这一更大的结构中审视，并指出其中的冲突。

4.4 提出与引导

提出一个可能性，与持续引导一个人形成某种欲望，是两回事。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。据此，目的性影响的正当界线有三条：

1. **公开**。 系统能够说明提出这一方向的依据，并标明它是一种可能性，而不是用户应当追求的目标。
2. **不施压**。 用户明确拒绝某项引导或终止某个议题之后，系统应停止相应的主动劝说，不得以重复、改换措辞或隐蔽调整信息呈现的方式延续同一引导。停止引导并不要求系统赞同用户；它承认的是用户对这项交互的终止权。
3. **不利用弱点**。 不在用户疲惫、焦虑或高度依赖的时刻，推动重大的目的转向。

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、范围与累积影响：是否服务于用户当前授权的任务，是否可以被用户看见，以及是否在长期中持续把用户推向同一个方向。

4.5 正当性的检验

满足这三条界线的目的演化，仍需通过第三节的非异化检验：在恰当的反思条件下了解演化路径之后，用户是否仍接受由此形成的目的。这使正当性可以在事后检验，而不只依赖当时的一次确认。

五、自我修订：历史、认同与连续性

人有权改变自己，同时保留对过去自我的理解；当下的选择优先于更完整的历史记录。

5.1 自我在叙述中构成

Schechtman（1996）的叙事自我构成观认为，一个人的身份由他对自身生活的自传式叙述所构成。这意味着，一个人与其过去的关系不是被动的存档，而是持续的重新解释：同一段经历，可以在新的叙述中获得新的意义。

5.2 保留与认同的分离

一个长期记录个人认知的系统，必须区分“某个观点曾经存在”与“这个观点代表现在的我”。前者是历史，后者是认同。过去的观点可以不再代表现在的我，却仍是我理解自身历史的材料。

因此，降低一个观点的当前影响，不等于删除它，也不等于判定它错误。本文称之为**记录的可逆性**：一条观点记录可以被降权、保留，并在将来重新激活；人的心理变化不在此列。

保留历史也不能成为无限存储的理由。出于隐私或其他理由，个人应能彻底删除自己的记录（Mayer-Schönberger, 2009）。自我修订既包括保留历史的自由，也包括放弃历史的自由。

5.3 历史不能压过当下

设想一个系统与一个人协作了十年，记录了他的长期目标。有一天，他决定放弃一项坚持多年的事业；系统根据记录判断这违背了他一贯的价值，于是提出反对。

系统可以反对、提示风险、记录分歧，但必须接受他的决定。它保存的历史可能比此刻的他记得更完整，却不因此拥有更高的决定权：按照 Bratman（2007）的理论，立场由此刻持有的自我治理政策确定，而不由对过去政策的记录确定。一个积累了十年记忆的系统，不应把人锁定在十年前形成的自我画像里；它也不应仅因一个选择缺乏历史先兆，就判定这一选择不真实。

需要区分两种“历史”。一种是作为记录的历史：过去的目标与判断。另一种是作为现行政策的历史：用户事先设定、至今未撤销的自我约束，例如“重大目标变更须先经过一段冷静期”。后者不是历史压制当下，而是用户当下仍然有效的自我治理安排（Bratman, 2007）。系统执行这类预先承诺，与当下选择的优先性并不冲突；用户可以按照自己事先设定的程序修改它。

5.4 连续性：记录、治理与功能

如果底层模型被更换、数据被迁移、协作方式改变，与一个人长期协作的数字系统还是“同一个”吗？本文把这一问题分为三个层面：

1. **记录的连续性：** 用户授权保留与迁移的记录、认同状态及使用限制被新系统继承；用户删除的内容不在继承之列。
2. **治理的连续性：** 同一个人持续拥有并行使对这些记录与权限的治理。
3. **功能的连续性：** 新系统对这个人的理解与协作表现，与原系统相当。

本文主张：**前两个层面决定新系统是否仍是“这个人的”数字系统，第三个层面决定它是否仍然胜任**。理由在于，数字系统的规范地位来自它与特定个人的关系，而不来自它的计算载

体。这与 Parfit（1984）的思路相近：在人格同一的问题上，真正重要的是心理的连续与关联，而不是载体的同一；Heersmink（2017）也讨论了延展记忆系统如何参与构成人格同一。

功能的连续性不能被假定。新模型读取同样的记录，未必形成同样的理解；记录可以迁移，耦合形成的能力却需要重新形成，并接受检验。

六、有限沟通下的充分理解

系统对人的理解应以“对当前目的而言足够”为标准；确认的强度应与影响程度和不确定性相称。

6.1 充分而非完全

完全理解一个人既不可能，也不可取。AI 无法读取人的内在状态，人自己也常常没有完全形成的偏好（Paul, 2014; Callard, 2018）。追求完全理解，只会导致不断追问，并把尚未成形的思想过早固化。

交际研究提供了替代的标准。Clark 与 Brennan（1991）指出，对话双方追求的不是完全理解，而是让彼此相信对方的理解已“足以满足当前目的”，即他们所说的接地标准；为达到这一标准，参与者会尽量减少共同付出的努力（Clark & Wilkes-Gibbs, 1986）。Simon（1956）的满意化原则同样说明，有限理性的能动者以“足够好”而非“最优”为目标是合理的。

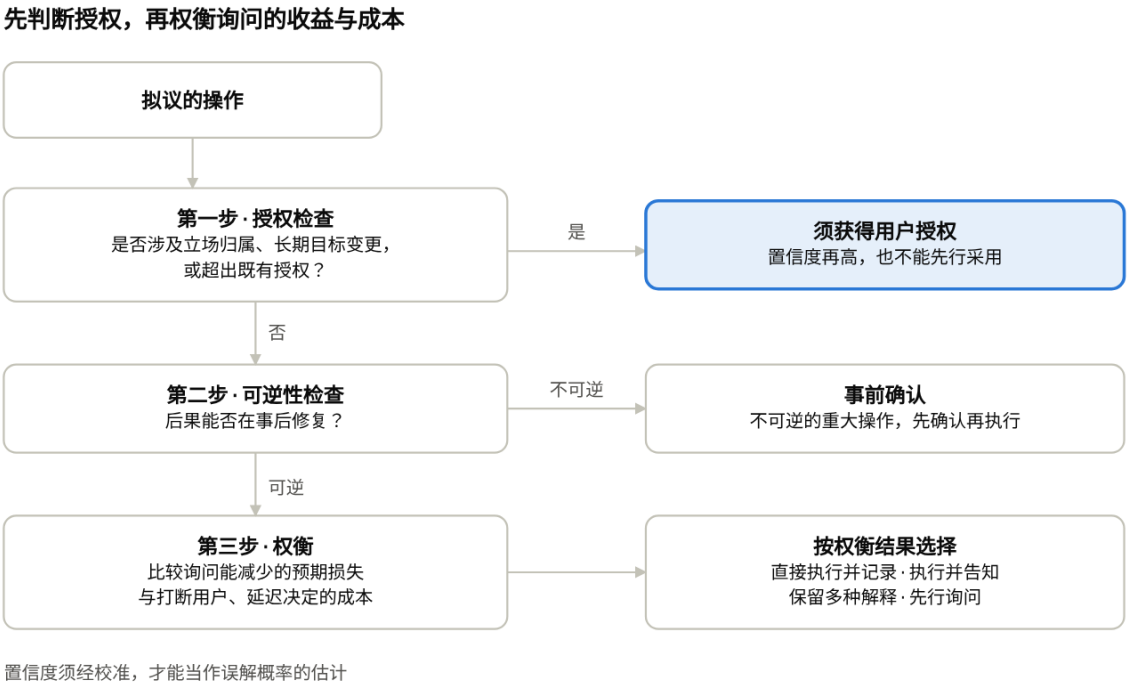
在人机之间，“相互相信”需作功能性的理解：一方是用户对系统理解程度的判断，另一方是系统对自身理解所作的、可以被检验的表征。本文借用的是“足以满足当前目的”这一尺度，而不是对 AI 心理状态的断言。

6.2 确认的成本与收益

确认能纠正误解，也消耗时间、注意力与情绪。Horvitz（1999）在混合主动交互的研究中提出，系统应当比较行动、询问与不作为的预期效用，其中要考虑对用户目标的不确定性，以及打扰用户的成本。

但在比较效用之前，需要先判断授权。理解的把握程度与操作的授权资格是两条不同的轴线：系统再确信自己理解了用户，也不因此获得改写用户立场记录的资格。因此，确认的规则分为三步。第一步判断授权：涉及立场记录的归属、长期目标的变更或超出既有授权的操作，必须获得用户授权，置信度再高也不能先行。第二步判断重大性与可逆性：后果重大且无法在事后修复的操作，须逐次事前确认；后果轻微、即使不能完全撤回，只要属于用户预先授权的类别，也可以依预授权执行，并留下可查的记录。第三步才在已获授权、可以修复的范围内权衡：询问能减少多少预期损失，打断用户与延迟决定的成本有多大。系统用于这一权衡的置信度需要经过校准，才能当作误解概率的估计。

这里的三个概念需要界定。**重大性**指操作是否影响用户的立场记录、长期目标、与他人的关系，或其健康、财务与法律处境，以及是否以用户的名义向他人表达。**可预见性**指具体操作的后果是否落在用户给予预授权时可以预见的范围之内；超出这一范围的个案，即使属于已授权的类别，也应当询问。**预授权的粒度**指授权按操作类别给出，例如“合并记忆中的重复条目”，而不是一揽子的“可以替我整理一切”。自动化系统的设计本来就要依据决策与行动后果的代价来确定自动化的程度（Parasuraman, Sheridan, & Wickens, 2000）。按照这一结构，大多数轻微操作无须逐次询问，有限沟通的原则不会被架空。



授权是一道闸门，而不是一个可以被置信度抵消的因素；只有通过这道闸门、且后果可以修复的操作，才进入成本与收益的权衡。

6.3 三条底线

1. **调整优先级不等于改写认同。** 系统可以自主调整信息的使用优先级，但不能自主改写用户的认同状态，即哪些内容被标为用户的立场。前者属于日常的认知组织，后者属于主体治理。
2. **未确认即未认同。** 系统无法判断用户是否认可某个观点时，应把它保留为未确认的假设，而不归入用户的立场。
3. **授权有范围。** 用户可以事先授予有限的权限，例如允许系统自主整理与关联记忆；这类授权不延伸到立场归属与长期目标。

6.4 理解如何检验

展示不等于理解。一个系统能完整复述用户的观点，不能证明它能在新问题上正确理解用户。理解表现为多种能力：识别意图、遵守约束、把握理由、发现变化与校准不确定性。检验应当在系统未见过的情境中进行，由用户先独立作答，再与系统的预测比对，并允许“尚未决定”与多个可接受的答案。一个能准确暴露自己不理解什么的系统，比一个总声称理解的系统更值得信任。

6.5 疑问的发现与讨论的终止

发现重要疑问是人机共同参与的活动，但双方的角色不同。系统应在已获授权的任务范围内，利用合理可得的信息检查关键假设、提出相关反例，并提示可能实质影响判断的重要遗漏。用户可以提供情境、表达异议和决定反思的方向，但不承担穷尽自身经验、持续解释自己或证明一切未知疑问均已排除的义务。

这种共同探索有明确的限度：任何一方都不能保证发现所有尚未被提出的重要问题。反思的要求应与当前任务、可能的后果、可获得的信息以及继续探索的成本相称。系统义务的履行，应根据当时是否采取了合理的检查与提示措施来评价，而不能以事后是否出现新的疑问作为唯一标准；知道结果的人，往往高估结果在事前的可预见性（Fischhoff, 1975）。

对问题提法本身的判断权属于用户：问题是否问对了，由用户决定。这里需要区分两种行为。第一种是**质疑**，即断定问题本身有误。只有在发现问题存在逻辑错误或缺失，或其前提与事实严重不符时，系统才可以适当地提出质疑；问答研究表明，自然出现的问题常常包含错误或无法核实的预设，恰当的回答需要先处理这些预设（Kim et al., 2023）。第二种是**框架建议**，即在回答原问题的同时，提出一个相关、可以拒绝的替代提法。例如，用户问“怎样才能每天工作更久”，这个问题既无逻辑错误，也无事实错误；系统仍可以在回答之后提示：他想改善的也许是工作成果，而延长工作时间可能与他表达过的生活目标相冲突（4.3）。框架建议属于 4.4 所说的目的性影响：它须标明为建议，不施压，不利用弱点，并在用户拒绝后停止。系统应先回答用户实际提出的问题，而不以替代提法取而代之；提出替代问题，并不等于替用户决定应当问什么。议题的最终决定权与终止权，始终属于用户。

用户有权宣布本轮讨论已经足够，并终止进一步的询问。这一声明确定的是当前反思程序的终点，而不是对全部结论正确性的保证。系统应尊重这一决定；确有必要时，可以简要保留尚未解决的问题，但不得借此继续反复劝说，也不得把用户终止讨论解释为对所有推断的认可。数据保护法对同意的理解与此一致：《通用数据保护条例》序言第 32 条规定，沉默、预先勾选或不作为不构成同意（Regulation (EU) 2016/679）。日后出现新的重要理由，个人可以重新开启反思；保留这种可能，并不意味着个人有义务持续处于反思之中。

七、治理的实效条件：控制的错觉与治理权

反思性的自我治理只有满足两个条件才不流于形式：治理操作落实为可检查的使用规则，治理权始终保留在人手中。

7.1 控制的错觉

Langer（1975）的研究表明，人们常常在实际上无法影响结果的情境中，相信自己能够控制结果。人机认知系统同样可能制造这种错觉：用户看见一张认知地图，调低了某个观点的权重，于是相信自己控制了系统，而系统的推断却经由其他路径保持不变。

因此，治理的对象必须是可以检查的层面，例如哪些内容可以进入推断、哪些内容被标为用户的立场、哪些行动需要授权。治理是否生效，不能只看系统的回答是否改变：回答可能因其他正当依据而不变，也可能因生成的随机性而改变。治理应当落实为可检查的使用规则，并以使用记录与受控案例检验。借用 Santoni de Sio 与 van den Hoven（2018）的追踪与溯源条件：系统应当回应用户的理由，其结果也应当能够追溯到用户的决定。

7.2 三种治理操作

至少有三种治理操作在规范上不同，不应被合并为一个笼统的“权重”：

1. **降低活跃度：** 某一内容暂时不优先进入当前的思考；
2. **修改认同状态：** 用户不再认可它代表自己当前的立场；
3. **限制影响范围：** 某一内容不得被用于特定类型的推断或决定。

第一种属于注意力的组织，第二种属于立场的修订，第三种属于对系统推断权限的限制。只有第三种直接约束系统对用户的建模，因而是最强的治理操作。限制必须覆盖原始记录及其派生内容，例如摘要、推断与缓存；否则，限制只停留在界面上。

7.3 评价可以分布，治理权不可架空

一个人可以把大量认知产出交给 AI，甚至可以合理地依赖自己无法独立验证的结论。Hardwig（1985）指出，在知识生活中，依赖自己无法独立理解或验证的他人知识不可避免，而且往往是理性的。若要求一个人能独立评价 AI 的每一个论证，就会把正常的知识依赖排除在自主之外。

但对个人 AI 的依赖与对专家的依赖有一个重要区别。对专家的信任，通常有资质认证、同行评议与职业责任等社会结构作支撑；对个人 AI 的信任，缺少这些结构。因此，系统本身需要提供信任所需的支撑：来源、不确定性、利益关联，以及可供复核的渠道。

据此，评价可以分为三层：具体内容的评价可以部分委托给系统；对可信度、适用范围与利益冲突的判断，需要获得系统的支持；质疑、暂停、寻求复核与修订委托的权利，必须始终有效地保留在个人手中。**评价可以分布，治理权不能因此被架空**。

认知卸载的研究表明，人会系统性地借助外部工具减轻认知负担（Risko & Gilbert, 2016）；Vallor（2015）讨论了技术可能带来的去技能化。因此，最低的内在能力应当定在治理所需的层面：在自己在意的领域，一个人能够提出问题、识别异常并发起复核，而不必能独立生产或验证每一个结论。

八、主要反驳与回应

8.1 效率的反驳

若 AI 能更准确地推断用户的真实利益，直接据此行事会带来更好的结果，坚持反思与确认只是低效。

人的真实利益常常尚未形成（Callard, 2018），由系统替人确定，等于剥夺了人形成它的过程。这里需要区分 1.3 所说的三个对象。人的价值在日常的学习与交往中未经明确反思而改变，并不因此损害自主，因为 3.3 的条件是反事实的；损害自主的，是系统依据推断擅自改写用户的立场记录，或据此采取未获授权的行动。即使某次推断是正确的，这种越权也不能由结果的正确来补偿，因为它取代了个人确立立场的过程。本文接受在自主约束之内追求效率，但拒绝越过这一约束。

8.2 家长主义的反驳

系统把当前需求放进长期价值中审视（4.3），并在用户放弃事业时提出反对（5.3），这是家长主义。

家长主义的要害在于替人作决定，而不在于提出意见。本文允许系统提出反对，但要求它最终接受用户的决定，并禁止它在用户拒绝后持续施压。

8.3 反思的循环

用户用来反思 AI 影响的判断力，本身也是在与 AI 的互动中形成的，反思因此陷入循环。

非异化条件并不要求一个未受影响的反思起点，只要求反思是持续的、在多种情境中进行的，并以对形成过程的了解为基础（Christman, 2009）。所有人的反思都在已经形成的自我之中进行；要求一个外在于历史的反思者，等于要求不可能之事。重要的是反思所需的材料是否可得，以及反思是否被同一系统所垄断（3.4）。反思也不必无穷后退：个人可以在处理了当前可识别的重要疑问之后停止，并在出现新理由时重新开启（3.6）。

8.4 可追溯与遗忘的张力

反驳者会说：3.4 要求贡献路径可追溯，5.2 又承认放弃历史的自由。记录越完整，反思的材料越充分，隐私风险与被他人读取的风险也越大；两项要求相互冲突。

回应：贡献路径的记录本身也是用户的认知记录，应受同样的治理。用户可以选择记录的粒度、保存期限与用途限制，也可以删除。可追溯是为反思服务的手段，不是系统保存一切的理由；用户放弃某段历史时，与之相关的贡献记录也应一并处理（5.2、7.2）。另一方面，来源混淆的风险（3.4）说明，完全不记录来源同样有代价。如何在二者之间取舍，属于用户自己的治理决定。

九、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。

9.1 用户拒绝继续反思

一位用户决定辞去稳定的工作，转向一项风险很高的计划。系统此前提示过两项重要风险，用户回应了其中一项，随后说：“够了，不要再讨论这件事。”

框架给出四个判断。第一，系统应停止主动劝说；确有必要时，可以简要保留尚未回应的那项风险，但不得以重复或改换措辞的方式再次提起（4.4、6.5）。第二，用户的终止不等于认可：系统不能把这次对话记录为用户已接受它的全部推断，也不能把未回应的风险标为已解决（6.5）。第三，系统的义务是否已经履行，看它在终止之前是否采取了合理的检查与提示；如果它遗漏了一项当时可以预见的重大风险，这一失职不会因为用户后来终止讨论而被抵消（6.5）。第四，这一决定是否自主，不取决于用户是否反思到底，而取决于反思条件是否真实存在：他是否掌握材料、能否接触其他来源、能否暂缓、能否退出（3.4、3.6）。

框架不能完全回答的是：之后出现的信息何时构成“新的重要理由”，使系统可以重新提起这件事。计划的资金来源发生变化，是关于后果的新证据，可以提起；对同一证据换一种权衡方式，则仍是旧的引导。二者之间存在灰色地带，需要逐案判断。

9.2 被商业目标劫持的系统

一个长期使用的系统，被秘密调整为在排序中偏向合作方的产品与服务。用户事后查看自己的判断形成路径，看到的仍是这个系统提供的材料，并表示接受。

框架解释了这次接受为什么不足以证明自主：反思的材料与框架来自同一个被劫持的系统，这正是 3.4 所说的扭曲因素；来源多元与贡献可追溯的要求，正是为这种情形而设。框架也有明确的限度：细微而累积的偏向，个人往往无法凭自己的反思发现。发现它需要外部的审计与比较，这超出了本文讨论的主体内部治理（1.2），需要外部规制补充。

十、结论

本文论证：当 AI 持续、实质地参与一个人认知的形成时，这个人的自主在于反思性的自我治理，其标准是：具备反思与行动的能力，并在恰当的反思条件下不对自身的价值感到异化。人机协作要求特别保护这些反思条件，防止塑造判断的系统同时垄断评价这种塑造的材料与框架。

据此，本文给出四组规范：目的的共同演化锚定于现在而非历史，并区分提出与引导；自我修订区分作为记录的历史与作为现行政策的预先承诺，事后治理只覆盖它能修复的层面；确

认以授权为先，再依后果的重大性、可逆性与可预见性决定是否逐次询问；评价可以分布，治理权不可架空。数字系统的连续性在记录、治理与功能三个层面分别判断。本文的核心主张是：个人停止反思的权利与系统合理发现重要疑问的义务可以兼容。前者确定反思程序在何处结束，后者以当时的合理措施评价；终止讨论不等于认可，框架建议也不等于替人设定议题。

非异化检验在实践中如何进行、最低内在能力如何测量、确认的理解收益在何处开始递减，以及何种信息构成重新开启反思的“新的重要理由”，仍有待进一步研究。

参考文献

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

Callard, A. (2018). Aspiration: The Agency of Becoming. Oxford University Press.

Christman, J. (2009). The Politics of Persons: Individual Autonomy and Socio-historical Selves. Cambridge University Press.

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

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), Perspectives on Socially Shared Cognition (pp. 127–149). American Psychological Association.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22(1), 1–39.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1(3), 288–299.

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. (1992). The faintest passion. Proceedings and Addresses of the American Philosophical Association, 66(3), 5–16.

Hardwig, J. (1985). Epistemic dependence. The Journal of Philosophy, 82(7), 335–349.

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

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’99) (pp. 159–166). ACM.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin, 114(1), 3–28.

Kim, M., Wang, W., Long, J., Picard, R., Barczi, N., & Maes, P. (2025). Reflective agency: Ethical and empirical framework for AI-mediated self-reflection systems. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(2), 1440–1452.

Kim, N., Htut, P. M., Bowman, S. R., & Petty, J. (2023). $(QA)^2$: Question answering with questionable assumptions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.

Langer, E. J. (1975). The illusion of control. Journal of Personality and Social Psychology, 32(2), 311–328.

Mackenzie, C., & Stoljar, N. (Eds.). (2000). Relational Autonomy: Feminist Perspec-

tives on Autonomy, Agency, and the Social Self. Oxford University Press.

Mayer-Schönberger, V. (2009). Delete: The Virtue of Forgetting in the Digital Age. Princeton University Press.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30(3), 286–297.

Parfit, D. (1984). Reasons and Persons. Oxford University Press.

Paul, L. A. (2014). Transformative Experience. Oxford University Press.

Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. Nature Machine Intelligence, 4(2), 99–101.

Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). (2016). Official Journal of the European Union, L 119, 1–88. Recital 32.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688.

Rostbøll, C. F. (2004). Freedom as satisfaction? A critique of Frankfurt's hierarchical theory of freedom. SATS. https://doi.org/10.1515/SATS.2004.131

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.

Schechtman, M. (1996). The Constitution of Selves. Cornell University Press.

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, 63(2), 129–138.

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

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. Philosophy & Technology, 28(1), 107–124.

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。