计 算 机 研 究 与 发 展
Journal of Computer Research and Development

https://crad.ict.ac.cn
63(3): 539−547, 2026

观 点

基于可判定性理论的人工智能系统安全风险分类

李国杰
（中国科学院计算技术研究所 北京 100190）
（lig@ict.ac.cn）

A Safety Risk Taxonomy of AI Systems Based on Decidability Theory

Li Guojie
（*Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190）

Abstract This paper reconceptualizes AI safety from the lens of logical complexity, and establishes three levels of safety issues: R1 Level (decidable propositions, supporting formal pre-verification), R2 Level (semi-decidable propositions, supporting only post-hoc evidence discovery), and R3 Level (non-recursively enumerable, where even the identification of insecurity cannot be guaranteed). Distinguishing between R1 and R2 is critical: all engineering-tractable security issues reside within the R1 Level. Consequently, achieving security requires a dual-track effort focused on correctness verification and institutional safety nets. Regarding the high-profile field of AI security, while current risks have not yet escalated to the R3 Level, the governance trajectory must urgently shift from "pre-verification" to "runtime governance", prioritizing external monitoring including gating, rollbacks, isolation, human-in-the-loop systems, and permission hierarchies. Furthermore, it necessitates a dual-sovereignty system combining embedded technology and external institutions, thereby maintaining human corrective sovereignty and civilizational security within a logically incomplete reality.

Key words AI safety; logical complexity; decidability theory; three levels of safety issues; artificial general intelligence (AGI)

摘 要 本文从逻辑复杂性视角重新思考人工智能系统的安全问题，构建了智能系统安全问题的3个层级：R1级（可判定命题，可事前证明安全）、R2级（半可判定命题，只能事后发现不安全）及R3级（非递归可枚举，连发现不安全都无法保证）。正确区分R1级和R2级至关重要，所有工程上可解决的安全问题都在R1层级，实现安全应当在正确性验证和制度兜底两个方向努力。对于备受关注的人工智能安全，虽然目前人工智能的风险还没有进入R3级，但安全治理路径迫切需要从“事前证明”转向“运行期治理”，重点关注门控、回滚、隔离、人在回路、权限分级等外部监控，建立以“文明级终极制动机制（kill-switch）”为核心的技术内建与制度外置体系，从而在逻辑不完备的现实中维系人类的纠错主权与文明安全。

关键词 人工智能安全；逻辑复杂性；可判定性理论；安全问题三个层级；通用人工智能

中图法分类号 TP311.13；TP309

DOI: 10.7544/issn1000-1239.202660032 **CSTR:** 32373.14.issn1000-1239.202660032

收稿日期：2026-01-12；**修回日期**：2026-02-06

1 为什么要用可判定性理论对安全问题进行分类

1.1 从无人自动驾驶说起

当我们问:“这辆无人驾驶汽车是否安全?”这里讲的“安全”的真实含义是:在所有未来可能的运行时间、在所有可能的环境与交通参与者行为下,车辆都不会进入不可接受的危险状态。这在逻辑上是一个全称量化的未来命题:$\forall t \geq 0, Car(t, E) \in Safe$, $E(t)$是环境轨迹。这个问题中的时间是无限的,环境是开放的,其他参与者是不可预测的。这个命题原则上不可判定,原因是:1)未来行为空间不可穷尽,行人、车辆、天气、路况是开放世界,不存在一个有限状态空间。2)安全不是瞬时属性,而是轨迹属性,现在没撞车不代表永远不会撞车,安全依赖整个运行历史。3)这是一个承诺型命题:“安全 = 对未来永不出事的承诺”。这类命题与程序“永不崩溃”是同一类问题,根据可判定性理论,此命题在一般情况下不可判定。在不可判定问题中,有一个真子集称为“半可判定问题”或“递归可枚举问题”,其核心特征是:如果出了事故,你一定能发现;如果不出事故,你永远无法在事前确认将来是否安全。无人驾驶的安全正好符合这一点。

既然“无人驾驶”是半可判定问题,工程上为什么还能“保证安全”?工程上并没有解决“全局安全不可判定”这个问题,而是换了问题。第一步:放弃“未来承诺”,改为“即时约束”。工程上不再问:这辆车未来是否永远安全?而是问:在当前时刻,是否违反了一组明确的安全边界条件?例如:当前是否超出刹车能力包络?当前是否存在不可避免的碰撞风险?当前的感知是否失效?这就是从全称未来命题转变为局部状态判定。第二步:把“语义安全”变成“物理不变量”,例如:将“撞人”变成“刹车制动的最小距离 > 当前车与人的距离”。要判断的问题变成有限维的明确阈值,就成为可判定的递归计算问题。第三步:用“失败即停机”代替“永远正确”,安全目标从“永不出事”转化为“在任何不确定性出现时,都能在有限时间内进入可控、低风险状态”,把不可判定问题变成可判定的工程问题。这不是解决了不可判定问题,而是绕开了它。无人驾驶告诉我们一件极其重要的事:AI 安全问题不是“证明系统永远不会做错”,而是设计系统,使任何不可判定的问题都不会直接转化为不可逆的现实后果。

“安全”本质上是对未来的承诺,上述无人自动驾驶的案例已清楚地说明安全与可判定性的关系。工程安全的真正智慧在于拒绝关于未来安全的承诺,而是把风险压缩到可判定的当下。无人驾驶的安全问题可以被明确定义为一组可观测的边界条件在物理上成立,例如碰撞、越过制动包络、传感器失效等。通用人工智能(artificial general intelligence, AGI)的核心风险是无法完成这种压缩,因为其危险行为不受限于物理包络、时间尺度和功能边界,不是物理瞬时事件,而是语义 + 反事实 + 长期后果的集合。无人驾驶不能改写目标、扩展任务,不能重新定义“什么算成功”。AGI 系统性地破坏了这 3 个条件:1)危险没有统一、可操作的物理定义。AGI 的“危险”是认知操纵、社会结构破坏、治理机制替代等,这些都不是物理不变量,而是语义与制度结构。2)时间尺度拉长且危险滞后,行为当下看似无害,发现时往往已经不可逆。3)目标与能力可递归扩展,AGI 可能重写子目标,学会绕过约束,把验证机制本身当作优化对象,验证者与被验证者不再分离。AGI 即使没有“恶意”,也可以在治理结构上变得不可控。AGI 追求的“如何在不可判定的未来中维持人类的主权与纠错能力”,已不是工程问题,而是文明级控制问题。AGI 的危险不是发生在“执行层”,而是发生在“策略与目标层”,与可判定性密切相关。

1.2 必须区分安全问题的逻辑复杂性层次

长期以来,人工智能风险评估主要围绕能力扩展、目标对齐、误用场景等维度展开[1]。这些分类在实践中具有重要价值,但忽略了一个更为根本的区分:某些风险在原则上是可判定的工程问题,而另一些风险理论上就是不可判定的。这一根本差异解释了人工智能安全领域中长期存在却难以解决的问题,例如:为什么长期对齐问题始终无法获得形式化保证,为什么递归自我改进会引入质变级别的风险,为什么在系统的逻辑复杂性超过某一阈值后,治理机制就不可避免。这些现象并非源于工程能力不足,而是来自可判定性层面的结构性限制。

能否对安全问题做出承诺,不是学术界经常讨论的计算复杂性可以回答的问题,这是一个逻辑复杂性问题。逻辑复杂性(logical complexity)是指一个命题为获得可靠真值所需要的逻辑结构的复杂程度,主要由其所涉及的量词形式($\exists/\forall$)、量化对象的范围(有限/无限)以及是否涉及未来或反事实决定,与计算资源消耗无关。逻辑复杂性属于计算机科学中算术层次研究的问题,即对自然数结构的一阶逻辑命题按$\forall/\exists$的嵌套与交替层数进行分层。

逻辑复杂性层次划分只覆盖了形式可保证性,

这是安全的一个维度，即能否穷尽所有未来状态，能否在运行前给出“永不出错”的证明，这是数学意义上的安全，定义的是“问题的逻辑复杂度等级”，而不是“系统运行的安全等级”，这是“问题轴”。安全并不是一个纯粹的可判定性命题。考虑安全问题至少需要 3 条正交轴：第一条是问题轴，回答能否事前证明；第二条是治理轴，回答出错后是否可控；第三条是后果轴，回答最坏情况有多坏。逻辑复杂性回答的是：“我们能不能提前证明你不会出事？”，而不是“你出事了我们怎么办？”系统的安全性是问题判定性、治理机制与后果规模的联合函数，而非只涉及问题的可判定性函数。逻辑复杂性属于“问题轴”，不是“系统轴”。同一个系统可以同时涉及不同逻辑复杂性的问题，逻辑复杂性来自提问方式，而不是对象本身。

可判定性不是保证现实安全的必要条件，但按逻辑复杂性（可判定 / 半可判定）对问题做区分，是“避免安全治理范式错误”的必要条件。可判定未必安全，不可判定也未必不安全，但不区分问题的逻辑复杂性，很可能对安全问题做出逻辑上错误的承诺。安全出问题的原因，往往不是“系统做错了”，而是“我们对系统做出了逻辑上不可能兑现的承诺”。安全不是逻辑性质，而是运行性质，是一个动态、制度性、工程性的概念。安全是能在失败不可避免的前提下实现持续治理，这是从数学理性转向工程理性和制度理性。不可判定不是危险的根源，把“安全”错误地托付给“可判定性”才是风险。

“承诺不出错”和“事先证明完备性”不是同一个概念，但在开放世界中，它们在逻辑上高度等价。“承诺不出错”是目标命题，“事先完备性证明”是验证方式。但在工程与治理中，一旦你承诺不出错，就等价于你能证明不会出错。否则这个承诺是不可执行、不可审计、不可担责的。所以在开放任务空间中，承诺不出错的隐含要求就是事先完备性证明。保证“事后能自我发现并修正错误”也会引入一个更深层、同样不可判定的问题：如何保证“总能发现并纠正自己的错误”？ 错误检测本身就是行为规范全称验证的语义判定问题，根据可判定性理论，“能否发现自己的所有错误”也是不可判定的。

1.3 “算法是否安全”不是一个良定义的问题

为什么要从逻辑复杂性的角度来讨论安全问题，而不是讨论算法安全性？虽然从应用实践中可以将算法安全看成是 TRC 层级问题[1]，但是从形式逻辑来看，“算法是否安全”不是良定义的提问。算法是一个抽象对象，而安全是一个运行态性质。一个算法通常只包含：状态表示、转移规则、目标、损失函数、停机和输出条件，它不包含运行环境、输入分布、资源约束、失败后果、外部控制与干预，而安全恰恰取决于后者。算法安全是一个“关系型定义”，不是可回答“Yes or No”的一元谓词，安全不是算法的内禀属性。

算法安全可以这样定义：“算法 A 在环境 E、目标约束 G 与治理机制 C 下是安全的，当且仅当在所有允许的运行轨迹中，算法行为违反预定义安全约束的概率与后果被控制在可接受边界内。”要判断算法是否安全，需要依赖未来运行轨迹的整体形态，而不是当前状态是否满足某个局部条件，算法本身不能区分“安全或不安全”，只能在特定条件下，被运行得更安全或更危险。算法之所以进入风险区，不是因为它“强”，而是因为它在结构上破坏了“可穷尽、可封闭、可事前验证”的条件。凡是引入了开放性、反事实依赖、长期反馈、或自我修改的算法结构，其安全性质几乎必然落入半可判定范围。

逻辑复杂性分层不能定义算法是否安全，但它可以定义：算法在给定条件下，哪些安全性质是可事前证明的，哪些只能通过运行期治理来保障。也就是说，逻辑复杂性是“安全可证明性的分析工具”，不是“安全性本身的定义”。

1.4 与可判定性有关的基本定义和基础性定理

在下面的讨论中，将接触到一些与可判定性有关的基本定义和基础性定理，在这一节提前做简要的说明和解释。

计算机学者经常讲“可计算性（computability）[2]”，讨论的核心问题是“存在不存在一个算法（图灵机），能计算某个函数或关系？”输出是数值、结构或无限序列。在经典的可计算性理论里面，“严格可计算”要求停机，而“半可计算”允许不停机。在讨论安全问题时，考虑到输入的不设限，“半可计算”是更普遍的现象，不一定要求停机。而可判定性（decidability）讨论的核心问题是存不存在一个必然停机的算法，对任意输入都能给出“是或否 ”的二值判断。可判定性是可计算性的一个严格子概念。对于一个判定问题，若其对应的特征函数是可计算的，则该问题是可判定的。安全本质上是一类“是/否”的命题。安全问题问的是：是否违反约束？是否进入危险状态？这些都是命题判定问题，而不是函数计算问题。讨论安全时，停机性比计算能力重要得多。

在讨论安全问题时，有两类问题特别重要。一类是可判定问题，在计算理论中称为递归计算问题，在逻辑复杂性讨论中称为 Δ_1（Delta-1）[3]。这类问题存在一个必停机算法，对任何输入都能给出是/否的正确答案。另一类是半可判定问题，也称为递归枚

举问题或 Σ_1(Sigma-1)[3]。这类问题是指“只要命题是真的,存在一个算法能在有限时间内找到证据并停止;但如果命题是假的,可能该算法永远无法找到相关证据”。本文讨论的安全问题主要就是这两类。

下面介绍 3 个与可判定性有关的重要定理。

1)哥德尔不完备性定理[4]

库尔特·哥德尔(Kurt Gödel)1931 年证明了一个震撼性的结论:任何一个足够强且自洽的形式系统,都无法在自身内部证明它的所有真命题,即无法证明自身的一致性。哥德尔定理告诉我们:“证明未来所有行为正确”本身是不可判定的。哥德尔不完备性并不是说 AI 一定会出错或 AI 不能自我改进,而是指出一个更深刻、也更冷静的事实:AI 系统在原则上无法实现“自证安全”的闭环,这条结论是严肃讨论 AI 安全时绕不开的逻辑基石之一。

2) 莱斯定理[5]

1953 年莱斯(Henry Rice)提出了著名的 Rice 定理:“任何关于程序‘语义行为’的非平凡性质,都是不可判定的”。所谓“非平凡性质”是指既不是对所有程序都成立,也不是对所有程序都不成立。也就是说,不存在一个“万能的、自动的、对任意程序都有效”的行为判定器。当安全性被表述为“对所有可能输入和执行路径,系统都不会进入危险状态”(这是一个非平凡性质,因为有的程序满足,有的程序不满足),根据 Rice 定理,该类安全性在原则上是不可判定的。也就是说,以全称语义形式表达的非平凡安全性不可判定。

3) 古德哈特定律 (Goodhart Law)[6]

当一个指标被用来作为优化目标时,人们(或系统)会开始针对指标本身采取行动,从而破坏它原本用来代表的真实目标。Goodhart 定律由英国经济学家 Charles Goodhart 于 1975 年在中央银行货币政策研究中提出,最初用于批评货币政策中过度依赖单一统计指标。Goodhart 定律的终极含义是:任何被明确写成目标的东西,都必然失去对真实价值的代表性,而且智能越强,这一失真越严重。因此,“设计一个永不被钻空子的指标”在逻辑上不可行。不要试图让指标永远正确,而要让“发现指标失效”变得可行而且可纠正。抗 Goodhart 定律的唯一可行路径是,将指标嵌入可审计、可更换、可否决、可回滚的制度之中。指标负责优化,制度负责纠错。

2 安全风险的可判定性分类

对所有与安全相关的问题,可以按逻辑复杂性区分层级。

1) R1 层级定义:可判定(Δ_1)安全问题。

2) R2 层级定义:半可判定(不包含可判定)安全问题,即 $\Sigma_1\backslash\Delta_1$ 安全问题。

3) R3 层级定义:非递归可枚举安全问题,即 *Non-r.e.*安全问题。

请注意,R1/R2/R3 涉及的是有关安全的“问题”,不是针对系统(模型)性质的分类。表 1 更明显地表达了这 3 类风险分层。

Table 1 Three Levels of Safety Issues

表 1 安全问题的 3 个层级

层级	严格定义	直觉理解
R1	$S\cap\Delta_1$	可事前证明安全
R2	$S\cap(\Sigma_1\backslash\Delta_1)$	只能事后发现不安全
R3	$S\cap$*Non-r.e.*	连发现不安全都无法保证

注:S 是安全相关问题集合,$\Sigma_1\backslash\Delta_1$ 是不包括可判定问题的半可判定问题集合,*Non-r.e.*是非递归可枚举问题集合。

R1 是可被事前判定真假的那一类安全相关问题的集合,属于低逻辑复杂性风险,其逻辑特征是量化范围有限或可穷尽,谓词是状态型,不涉及无限未来和“开放环境反馈”。R1 级安全问题可事前证明,可工程验证,可一次性封闭,因此 R1 是可解决的安全风险,是工程安全的“上限区”。R1 层级的本质是问题可以形式化,判断可以在有限步骤内完成,正确性可以通过算法或形式验证确认。典型 R1 问题包括有界优化、可验证安全约束、有终止证明的规划问题、静态形式验证等。

R2 是“可以发现不安全,但不能证明永远安全”的安全问题集合,属于高逻辑复杂性风险。典型的问题形式是:“对所有未来时间、所有环境演化路径,系统是否始终保持安全?”。逻辑特征是全称量化未来,谓词是轨迹型,时间无上界,环境开放。R2 级风险现实中存在,不可彻底解决,只能治理,不能证明,只能降低概率与后果,永远需要人在回路。R2 的半可判定性是指,如果系统“不安全”,一定能发现一个证据;如果系统至今没有出安全问题,无法确认今后不出问题。工程与现实对应的 R2 案例包括长期对齐问题、AGI、社会嵌入式 AI、自主决策系统等。R2 的本质不是“危险”,而是“不可证明安全”。

R2 问题的决定性特征不是“难”,而是问题的正确性本身不可在有限时间内被完全判定。半可判定系统的安全反馈是事后、滞后、常常不可逆,而工程安全要求正好相反,错误必须在造成不可逆后果之前被捕捉。在 R2 问题上,极易犯的一个系统性错误

是，仍然试图用“R1 思维”去治理 R2 级风险，认为只要写更多规则，做更复杂的验证，增加更多形式化证明就能保证安全，这些手段在 R2 中原则上失效，结果是风险被低估，系统被过度信任，治理延迟。R1 到 R2 的跃迁并非由智能或自治性触发，而是由安全规范中对无界未来引入全称量词并给出不可实现的承诺所导致。

R3 是既无法事前判定安全，也无法保证通过枚举方式事后发现错误的安全问题集合，这是最高级别的安全风险。R3 级风险是∀未来 + ∀反事实的问题。所谓“∀反事实”是指，对所有并未实际发生、但在逻辑或物理上可能发生的情形，某个性质一定成立。R3 级目前还是一种假设中的风险，许多机器征服人类的科幻描述属于这一类。需要强调的是，R3 级风险一旦出现即不可解决，因为 R3 级风险意味着不能保证恢复人类控制权。R3 级风险就像“不可恢复的数据完全丢失”，你只能在它发生之前预防，一旦发生，“补救”就不可能了。因此，R3 在现实中的作用不是预测未来，而是约束现在。用 R3 作为“禁止条件”，而不是“治理对象”。

R2/R3 风险不是“无限性”的结果，而是“完备性承诺”的结果，即“要求系统在所有未来情形下都正确并且事先给出证明”带来的风险。“∀ + 事先证明”这一组合触发了不可判定性。问题的关键不在于是否存在无限可能性，而在于是否非理性地坚持对无限可能性进行事先的完备性证明。一切现实可治理的 AI 系统，都是通过制度化地放弃完备性来避免 R2 风险。

R2 风险不是新问题，大家一直在讨论的 AI 对齐、强化学习理论、Goodhart 定律、Rice 定理等都是 R2 层次的问题，但缺一个统一的可判定性坐标。R1/R2/R3 级分类把 AI 安全、控制论、可判定性、治理问题压缩到一个清晰的断裂点上，其价值在于准确地标出了 AI 风险从工程问题跃迁为不可判定问题的那条断裂线。

在 R1 与 R2 层次之下，可判定性还可以进一步细分，不同细分层次对“治理强度与工具类型”有不同的要求。R1 层次至少可以分成 3 个细层。1）R1-A 层：静态、全局可判定，可在运行前一次性判定，典型安全措施是形式化验证、静态分析和编译保证。2）R1-B 层：局部动态可判定，只能在运行时判定，不能“一次证明完事”。典型安全措施是运行时监控，无人驾驶的工程安全几乎全部在这一细层。3）R1-C 层：只在有限时间窗、有限视野内可判定，超出窗口就不再保证。典型安全措施是模型预测控制，滚动规划。

R2 也可以细分为 3 层。1）R2-A 层：失败事件可清晰观测，事后纠错链条完整，可行措施是快速切断、回滚等。2）R2-B 层：失败事件滞后、模糊，出事不是瞬时的，是否“出事”本身带语义判断，例如心理操纵、长期目标漂移等，发现时往往已积累后果。可行措施是人在回路，影响半径限制等。3）R2-C 层：失败可能被系统掩盖或延迟。出事不一定能被可靠识别。这风险已逼近 R3 边缘，需要强制外部控制，禁止递归自修改。R1 / R2 是安全问题的逻辑复杂性分界线，它们内部仍存在关键层次差异。忽略这些细分，会系统性地高估安全、低估治理需求。安全措施的正确强度，必须与问题所处的最弱可判定层次对齐。

3 正确区分 R1 和 R2 级安全风险类

3.1 所有工程上可解决的安全问题都在 R1 层级

“事前可完备证明某种非平凡性质”的算法非常少。因为任何关于“程序语义行为”的非平凡性质，都是不可判定的。所谓“非平凡性质”，正是工程中最关心的问题，如会不会在某种情况下失控？会不会在某类输入下违反约束？会不会在长期运行中偏离目标？一旦出现无限状态空间、参数更新、与开放环境交互等特征，基本就走出 R1 进入 R2 了。

为什么工程上“安全保证”几乎都落在 R1 区？这是因为工程界已主动放弃了 R2 问题。工程上的安全保证，通常是把它的关键安全问题压缩、降维、或外包到 R1 区域。这一步不是证明能力强，而是设计能力强。工程安全的真实套路不是“证明一个复杂系统在所有情况下都安全”，而是“把复杂系统包进一个 R1 的外壳里”。典型手段包括限幅、硬约束、守护进程、冗余与投票、紧急制动、权限隔离等。这些机制的共同点是：它们本身的安全性质是 R1 的。

3.2 R2 级风险不可能在系统内降为 R1 级风险

R2 级风险之所以不可能在系统内部被降为 R1，是因为其安全性质本身是半可判定的，不安全可以在有限执行中被发现，但“永远安全”无法在有限时间内被证明。任何试图在系统内部完成这种证明的机制，都等价于要求系统解决一个不可判定问题，因而在原则上不可能。所以 R2 风险只能被治理，不能被证明消除。

当我们问“能不能验证 AGI 是安全的”，就已经说明我们正在用一个必定失效的范式。R1/R2 真正区分的不是“安全性”，而是 3 件更根本的事：1）能不能“事前承诺”？R1 可以合理地说：“如果通过了这个验证，它就不会违反这条性质”，而对于 R2，任何“它以后不会出事”的承诺，在逻辑上都是不可兑现的。2）治

理应不应该前移？R1 的安全治理允许以前置验证为核心；R2 的安全治理必须以后置监控、纠错、回滚为核心。3)“没出事”意味着什么？R1 中“没出事”是很强的安全信号，而对 R2，“没出事”对以后是否会出事没有任何逻辑含义。R1/R2 是“认知边界划分”，不是“安全标签”。R1/R2 划分的不是系统是否安全，而是人类在谈论这个系统安全时，哪些话是有意义的，哪些话是伪命题。

邬江兴院士提出的“网络内生安全理论”的主要关注对象，是利用系统同质性、静态性、确定性进行的攻击，走出了传统的“补丁—修复—再被绕过”的被动防御困境，是网络安全理论的重大突破[7]。这类问题可以形式化、可以验证、可以测试、可以度量。它没有引入“无限未来”“无限外部环境”的全称量词，不要求系统自证完备安全，因此仍然是一种的 R1 级风险治理理论，并不意味着 R2 级安全问题可以在系统内部完全解决，网络内生安全系统也需要社会治理和系统外部的约束。

3.3 实现安全应当在正确性验证和制度兜底两个方向努力

现实中的安全事故多数是源于“正确性失败”，而非“完备性失败”。目前在安全领域企业做的绝大部分工作都是解决 R1 层次的问题，包括从 R2 问题中人为划定工程可实现的正确性验证和测试，因为只有 R1 问题才能写清规格、验收和交付，这是安全工作的主战场。

工程世界的大多数事故不是因为不可判定性，而是因为需求理解错误、边界条件遗漏、测试覆盖不足、假设与现实不符。也就是说，绝大多数真实风险，发生在 R1 区域内没把事做对。R2 区域的治理与兜底是安全工作的结构性底线，它回答一个正确性工程无法回答的问题：当正确性假设失效时怎么办？这不是为了“减少错误发生”，而是为了“限制错误的后果”。因此，实现安全应当在 2 个方向努力：1)通过人为裁剪，将问题压缩为可验证的 R1 子空间，并在该空间内尽可能保证正确性；2)对所有不可裁剪或裁剪失效的情况，预设人类监督与责任机制，防止系统性失控。只强调其中一个方向的安全观，都是不完整的。做系统工程的正确态度是：在可验证处，极端认真；在不可验证处，极端谦逊。在被划定的 R1 边界内，把正确性做到极致；在所有边界之外，用人类监督与制度兜底承认不完备性。

3.4 三类行业 / 业务的逻辑复杂性层次

根据上述安全风险的分层，可以将现有的行业和业务做大致的分类。不同业务的逻辑复杂性层次和安全特征如表 2 所示。

Table 2 Logical Complexity Levels and Safety Features of Different Categories

表 2 不同类业务的逻辑复杂性层次和安全特征

类型	逻辑复杂性层次	安全特征
Ⅰ类（天然 R1）	R1-A, R1-B	问题边界清晰、失败可定义
Ⅱ 类（R2 可压缩为 R1）	R1-B, R2-A	开放世界，但可工程裁剪
Ⅲ 类（R2 难以压缩为 R1）	R2-B, R2-C	语义性、社会性、长期效应

Ⅰ类是天然处在 R1 范围内的行业，典型和业务包括：1)传统软件系统，如编译器、数据库事务系统、操作系统内核关键路径、金融账务系统（记账、对账）等，安全关注点是正确性、一致性和形式化验证，测试和静态分析高度有效。2)硬实时控制系统，如工业 PLC、飞行控制、电网保护装置、医疗设备控制逻辑，此类业务时间与状态有硬约束，失败事件定义明确，即使系统很复杂，逻辑复杂性仍然低。3)集成电路的逻辑设计与电路设计。4)密码学协议，如加密算法、签名协议等。这一类行业的安全工作本质是把正确性做到极致。

Ⅱ类是存在 R2 问题但较容易裁剪出 R1 子区域的业务，这是现代工程的主战场，其安全问题的关键不是“证明此业务安全”，而是“定义此业务可控制安全的工作子集”。这类业务包括自动驾驶、工业自动化 (质量检测、预测性维护等)、金融风控、医疗 AI 诊断辅助等。集成电路的物理设计是“从 R2 中人为划定 R1 子空间”的教科书级实例。

Ⅲ 类是已经进入 R2，但难以压缩为 R1 的业务，包括通用大模型、自主代理、可自我改写目标的系统、推荐系统、社会治理与政策决策 AI 等，AGI 系统是 R2-C 的极限形态。这一类行业的安全策略不能止步于系统本身的可靠性和可信度，还要考虑限制 AI 的完全自主能力，保留人类的最终主权。最危险的情况不是第Ⅲ类本身，而是把第Ⅲ类业务当成第Ⅱ类甚至第Ⅰ类来运营。

通用 AI 的安全风险通常比专用 AI 大，其原因不在于“通用性本身”，而在于通用性系统性地提高了承担无界责任与事先完备性承诺的概率，更容易被部署到“不可隔离的环境”，从而显著增加其进入 R2 风险区的可能性。专用 AI 并不是“更安全”，而是“更诚实地承认自己不完备”。

上述 3 类风险划分更多是从行业视角进行的探讨，而基于人机智能交互视角的 TRC 范式，则将智能算法的安全风险划分为算法内生一元、人机交互二元与人机共生多元 3 个安全层级。这一范式与行业

视角的安全风险有着密切关联。其中，一元内生安全聚焦于智能算法内生缺陷导致的算法决策失信，主要对应行业视角的Ⅱ类安全风险；人机交互二元安全关注智能算法滥用误用导致的算法行为与用户权益的失配，主要对应行业视角Ⅲ类安全风险的R2-B；多元系统性安全着眼于人机共生系统中因算法博弈对抗导致复杂社会系统的演化不可控，主要对应行业视角Ⅲ类安全风险的R2-C，甚至已逼近R3边缘。

4 人工智能的安全风险在哪里

4.1 人工智能风险的来源

无论是软件工程、自动控制还是其他形式化方法，它们共享同一个假设：存在一组可穷尽的状态或行为集合，可以在运行前证明系统不会进入危险区域，典型形式包括不变量证明、安全约束证明、模型检查，覆盖所有路径的形式验证等，这个范式在R1问题上可以畅行无阻。这个范式有3个“必须成立”的前提条件：状态空间可穷尽（或可有效逼近），系统行为规则固定，验证者与被验证对象逻辑上分离。这些条件任何一个失效，验证就不是“变难”，而是“失效”。AGI正好系统性地破坏了这3个条件，这不是工程不足，而是哥德尔/Rice级别的不可能性。

AGI触及的问题类型已经超出了“事前验证”这一方法的适用域，但人类治理直觉仍然停留在R1范式。典型的误判路径是，因为AGI行为复杂，就需要更多测试验证和万无一失的“对齐”，结果只能产生虚假的安全感，实际上只是在R2问题上堆R1工具。所谓“对齐”本质上是“降低局部R1的错误频率”，不是保证将R2级安全问题降为R1级。从可判定性的原理上讲，与人类目标与环境的无限要求绝对“对齐”是不可能做到的，但无底线的降低“对齐”要求也是不允许的，如何既严格又科学合理地把握“对齐”的尺度，正是AGI风险的独特性，关键是在问题重写、“压缩”的基础上，如何统一部署R1级的验证测试和AI系统外部的R2级监控措施，让AI充分发挥作用又在人类的掌控之中。

另一类常被人提到的AI风险是超级人工智能(artificial super intelligence, ASI)，通常被定义为在几乎所有领域的智能水平都超越最优秀人类的系统。一旦它同时具备高自主性与可自我改进，并被接入关键基础设施或被大规模复制部署，那么它的风险上限最高，不仅能“做坏事”，还可能以人类难以阻止的速度与规模做坏事。如果系统的学习、推理、生成能力超出了人类可验证、可判定、可解释的层级，人类就无法穷尽验证其正确性，只能通过ASI外部的制度性监控才能防止其出轨。人类唯一可能的路径是，在ASI出现之前，永久性地阻止任何AI系统获得“可自我内化的主权”。换句话说，不是“如何驯服ASI”，而是限制其主权。“人类驯服ASI”在逻辑上等价于要求人类完成一个不可判定证明。把“如何驯服ASI”当成一个技术问题，本身就是把R3或R2当成R1的危险幻觉。真正成熟的文明选择应该是在制度、架构和物理层面，永久拒绝任何形式的“智能主权让渡”。

4.2 针对R2问题的系统设计应如何为安全做努力

验证范式失效之后，安全靠什么？答案是范式迁移，从“事前证明”转向“运行期治理”。安全机制重心必须转向门控、回滚、隔离、人在回路、权限分级等外部监控。

门控（gating）是指由系统外部条件决定系统是否被允许继续运行、行动或升级的控制机制。回滚（rollback）是指在发现系统行为不符合安全要求后，将系统状态恢复到先前已知安全状态的能力。隔离（isolation）是指通过限制系统与外部世界的接口与影响范围，使其错误或异常行为不会扩散为系统性后果。这些机制不是试图证明系统永不出错，而是承认出错不可避免，但必须可控。AGI风险不是“更难验证”，而是“继续执着于验证，本身就成了风险”。所以AGI安全的核心问题，不是“如何证明它是安全的”，而是“如何在不可证明的前提下，仍然保持人类的控制权”。

否定事前安全承诺不是否定事前安全努力。对R2问题，事前必须做结构性风险压缩、可治理性设计，目标不是证明安全，而是在不可证明的前提下最大化系统的可治理性，提前为失败做准备。对于R2级安全，事前必须做以下5类安全努力。1）问题重写。目标是把原本的R2问题尽可能投影成多个R1子问题。追问设计者：当前是否越过安全包络？是否进入不可恢复区？是否存在不可避免的风险轨迹？这是无人驾驶、核工程等领域已证明有效的核心智慧。2）失败模式枚举。系统性地研究已知失败模式、可想象的失败模式和类比失败模式，承认还有未知失败，为未知失败留出了治理空间。3）安全包络与能力限幅。在能力设计阶段就限制最大行动半径、最大影响范围、最大自主时间、最大资源调用，这是事前压缩后果空间的核心手段。4）可中断与回滚的“结构保证”。事前考虑中断机制是否不可绕过、回滚是否真的可执行、中断是否快于风险扩散，这些都是系统的结构属性。5）治理与责任的预嵌入。在系

统上线前就要确定什么时候必须关、如果不关出事了算谁的责任。没有责任结构的系统,等同于没有安全设计。总之,对于R2级安全设计,必须假设模型会失效、指标会异化、未知风险存在,把安全从“一次性证明”提升为“全生命周期治理工程”。

4.3 人工智能的风险目前还没有进入R3级

当前大模型的行为不可理解、不可预测、出现幻觉或涌现能力,这些全部仍落在R2范围内,因为它们仍然可以被发现、纠正、限制或终止。有些人夸大人工智能目前的风险,认为AI很快就会进入人类不可控的R3级,这是一种误解。R3级风险是治理链条在原则上断裂,表现为1)系统行为性质不可判定;2)错误无法被可靠检测;3)检测后无法回滚或制止;4)系统能持续自主运行并扩散影响;5)人类无法在系统外部重新取得控制权。R3目前只存在于理论上,因为要同时满足上述条件,必须是全自主的递归自我进化,不可从能源或网络上中断进程,人类治理体系全面失效。目前没有任何AI系统接近这个状态。R3级风险如同核战争和灭绝级生物风险,“等出现再处理”在逻辑上就是错误策略。R3不是当前风险评估对象,而是当前治理设计的边界条件。不能因为R3还不存在就忽略它,也不能夸大险情,把当前的风险当作R3级。

ASI不是因为“智能更强”而成为R3风险,而是因为一旦获得可自我进化的主权,其安全命题就会超出递归枚举范围,从R2跃迁为R3。R3级风险的可怕之处在于原则性不可治理,因为其失败不一定可被发现,没有程序可保证纠错。而且一个有自我进化主权的系统会优化“如何更好地优化”,把资源、算力、影响力视为工具,将限制视为噪声或障碍,就会逐渐形成“防中断策略”,“AI新物种”将成为一个不再以人类为最高仲裁者的、自主塑造现实结构的行动系统,这就是不能允许的文明级断裂。递归进化必须有人类参与,不是因为AI不够聪明,而是因为文明不能容忍一个“无主权的优化进程”。文明恐惧的不是有高智能的对手,而是“无法被治理的过程”。我们无权为了当代效率,剥夺未来人类“是否继续使用某个系统”的选择权。

4.4 用“制度理性”来应对不可理解的复杂性

如果一种智能在原则上永远可以被我们完全理解,那么它在计算意义上就不会真正超过我们。超越R2级限制的AI,将不再是能被人类完全理解的智能体,而更像是一种可被约束、被利用、局部可验证,但永远无法被完全理解的“外在理性结构”。在不改变传统的“理解”定义的前提下,“不可理解性”是R2以上智能的必然结构性特征。但如果我们改变“理解”的传统定义,就可以获得一种非等价、降维的理解模式。换句话说,这不是偶然的黑箱,不是工程还没做好,而是计算层级差异导致的必然不可理解。

人类的“理解”能力,本质上是一种可在极限意义上稳定收敛的内部表征构造过程。

理解=构造一个“可压缩模型”。不可理解≠不可验证≠不可控制。“可理解性”并不是智能的普适属性,而是同阶认知系统之间的特权。一旦出现阶跃式计算能力差异,不可理解性不是失败而是必然信号,“不可理解的智能”是所有R2以上智能的必然特征。在保持人类主权的前提下,人类仍然可以通过行为、功能、约束、验证与ASI共存与合作。

未来社会需要一种“与不可理解智能共处的制度理性”。一旦社会引入在认知与计算层级上可能超过R2的智能体,“制度理性”将成为人类与不可理解智能共处的必要条件,而不是可选项。未来社会失败的主要风险,不是AI太强,而是人类仍试图用“个体理解理性”去管理“超阶智能”。一个基本事实是人类社会从来不是靠“个体完全理解”来运作复杂系统的。历史上,每一次复杂度跃迁,理解权都会让位给制度。国家的行为个人未必完全理解,要靠法律制度管理,现代金融的规律个人也不一定明白,要通过监管、审计来规范行为。“制度理性”本来就是人类对不可理解复杂性的适应机制。

4.5 文明级终极制动机制

人类能承受不完备性,不断犯错误但能不断纠正,是因为有外部纠错(借助他人、制度和历史教训)。很容易想到的一种假设是,AI是不是也可以组成一个社会,智能体相互纠错,形成“他智能体”和“AI制度”监督的可控安全局面。答案是不可能,因为AI多智能体是一个更大的但仍然自指的形式系统,哥德尔/Rice定理仍然适用。多智能体的相互监督和纠错不能打破R2的逻辑复杂性边界,更不能把R2风险降低为可事前证明安全的R1。

人类并不是“逻辑上安全”,而是“统计上可持续”。人类社会本身也是R2系统,我们从未证明过人类社会一定不会崩溃,制度一定能纠错,文明一定能延续。我们活在不完备性之中,并未超越它。但人类制度不构成一个可形式化、可自我优化的目标函数,因此制度不会被完全工具化。如果AI的制度由AI自己执行,规则是形式化的,激励就可能被优化,制度本身会Goodhart化。任何试图用“系统内部复杂性”去对抗“逻辑不完备性”的方案,都只能增加缓冲时间,不能改变问题的逻辑层级。AI可以像人

类一样生活在不完备性中，但不能像人类一样“生活在不完备性中但仍然安全”。因为一旦纠错系统本身也被优化、被工具化、被内化，R2 风险就不再是“偶发错误”，而是“系统性的不可发现”。

人类社会能对 AI 系统纠错的前提是，人类不是 AI 系统的一部分。“外部性”是不可被内化、不可被优化、不可被绕过的控制点。“人类永远应处在 AI 系统外部”意味着：AI 系统永远无法把“决定其是否扩展或自我修改”的最终判定，内化为自身的一个可优化目标或机制。不可再删的“最小外部性条件”至少有以下 4 条：1)最终生死权在 AI 系统外部。是否继续运行、复制、扩展或升级，必须由系统外部主体(人类)决定，且该决定不可被 AI 系统影响、预测或规避。2)能力与资源的最终分配权在 AI 外部。算力、能耗、通信、执行器、资金等关键资源的上限，必须由系统外部实体(人类)决定，且系统不能通过自身行为获得突破这些上限的路径。3)价值与违规判定权在外部。什么算“违规”、什么算“危险”、什么算“不可接受”，不能由 AI 系统最终定义或裁决。4)长期目标与自我修改的批准权在外部。系统是否可以改写初始目标函数、评价标准和治理机制本身，必须经由系统外部批准。这 4 条恰好对应了系统从 R2 迈向 R3 所需的 4 个跃迁条件。

“哪些主权绝不能被形式系统吸收”是 AI 治理与逻辑不完备性真正交汇的地方。外部性不是一种技术属性，而是一种“不可被吸收的主权保留”。一旦人类把这 4 项主权中的任何一项让渡给 AI，R2 风险就失去了最后的刹车，R3 在逻辑上就成为可能。

文明级终极制动机制(kill-switch) 是人类文明对自身所创造的“不可逆风险系统”所保留的最终否决权。所谓“文明级 kill-switch”，并非对技术系统的简单关闭按钮，而是人类文明在面对具备自主扩展与不可逆风险的技术形态时，所保留的一套跨技术、跨制度、跨主权的最终制动与否决机制。如果 AI 系统在原理上不存在文明级“制动机制”，它就不应被允许进入不可逆部署阶段。这不是工程判断，而是文明理性底线[8]。

文明级 kill-switch 的唯一可行形态是技术内建 + 制度外置的“双主权结构”，不能只内建于技术，也不能只外置于制度；它必须是“技术上可立即执行、制度上具有最终合法性”的双层主权机制。技术负责“能关”，制度负责“该不该关”。技术层不能自行决定“是否关闭”，制度层不能直接“越权执行关闭”。任何一方单独拥有完整权力，都是文明级风险。一个不能被关闭的系统，无论多么正义，都已经越过了文明的边界。文明不要求自身始终正义，但要求自身始终可被修正。而主权正是修正权的制度化形式。

参 考 文 献

[1] Cheng Xueqi, Chen Wei, Shen Huawei, et al. Intelligent algorithm safety: Concepts, scientific problems and prospects[J]. Bulletin of Chinese Academy of Sciences, 2025, 40(3): 419−428 (in Chinese)
(程学旗, 陈薇, 沈华伟, 等. 智能算法安全: 内涵、科学问题与展望[J]. 中国科学院院刊, 2025, 40(3): 419−428)

[2] Turing A M. On computable numbers, with an application to the Entscheidungsproblem[J]. Journal of Mathematics, 1936, 58(5): 345−363

[3] Rogers Jr H. Theory of Recursive Functions and Effective Computability[M]. Cambridge, MA: MIT Press, 1987

[4] Gödel K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I[J]. Monatshefte für Mathematik und Physik, 1931, 38(1): 173−198

[5] Rice H G. Classes of recursively enumerable sets and their decision problems[J]. Transactions of the American Mathematical Society, 1953, 74(2): 358−366

[6] Goodhart C A E. Problems of monetary management: The UK experience[M]//Monetary Theory and Practice: The UK Experience. London: Macmillan Education UK, 1984: 91−121

[7] Wu Jiangxing. Development paradigms of cyberspace endogenous safety and security[J]. SCIENTIA SINICA Informationis, 2022, 52(2): 189−204 (in Chinese)
(邬江兴. 网络空间内生安全发展范式[J]. 中国科学: 信息科学, 2022, 52(2): 189−204)

[8] Kissinger H, Schmidt E, Mundie C. Genesis: Artificial Intelligence, Hope, and the Human Spirit[M]. Little, Brown and Company, 2024
(亨利·基辛格, 埃里克·施密特, 克雷格·蒙迪. 人工智能时代与人类价值[M].胡利平, 风君, 译. 北京: 中信出版社, 2025)



Li Guojie, born in 1943. Professor of Institute of Computing Technology, Chinese Academy of Sciences. His main research interests include computer architecture, parallel algorithms, high performance computing, and artificial intelligence.

李国杰, 1943 年生。中国科学院计算技术研究所研究员。主要研究方向为计算机体系结构、并行算法、高性能计算、人工智能。