

Don't mistake chatbot intelligence for consciousness

别把聊天机器人的智能误当成意识

But a new superintelligence may be coming, and it would upend humanity's hierarchy of moral concern, thinks Susan Schneider

Susan Schneider认为：新的超级智能或许正在到来，它会动摇人类的道德关怀等级

The Economist · 2026-08-20



人物肖像插画：浅色背景下，一位金发女性身穿深色上衣。

难度 75 / 100

建议80–105分钟；选做作文另计。

适用读者：能阅读一般英语新闻，希望读准长句、搭配与作者语气的学习者。

入门准备：能辨认基本主谓结构和常见从句，愿意结合语境查词。

About this lesson 本课导读

特邀作者Susan Schneider区分完成任务的智能与具有内在体验的意识，主张把注意力转向值得进一步研究的系统结构。本课重点处理两种陷阱、两类理论、三项理由和一组伦理两难，训练不把研究线索写成结论。

- 区分智能、感受能力与语言自述
- 梳理不同枚举各自服务的论点
- 还原生物芯片长句中的对象与反馈
- 解释测试限制和思想实验的两条分支

Before you read 读前思考

Could a system solve difficult problems without having any experience? What would a performance test leave unanswered?

拓展例句与练习情境用于说明用法。

书末答案 (p.15)

Text 课文

[P01] THIS MAY be the most consequential moment in the history of intelligence.^{V01} In only a few years, artificial intelligence has come to rival human performance across domains once thought distinctively ours: language, reasoning, creativity and persuasion.^{V02V03} Increasingly, people are asking whether these systems are conscious—whether it actually feels like something to be an AI.



人物肖像插画：浅色背景下，一位金发女性身穿深色上衣。

- [P02] The issue is urgent. The future will be shaped by what these entities become. We must not mistake intelligence for consciousness, however.^{V04} A brilliant superintelligence could be experientially empty.^{V05} Yet if we build systems that can suffer, and miss their sentience, we risk creating suffering on an industrial scale.^{V06}
- 10 [P03] There is no compelling evidence that the familiar chatbots of today, running on graphics chips, are conscious.^{V07} Though they have human-like linguistic abilities, and have even claimed to be conscious, the more mundane truth is that they talk about their inner lives because they absorbed how we talk about ours, having been trained on an enormous trove of humans' data.^{V08V09K01}
- 15 [P04] This mimicry can lure humans into two traps.^{V10V11} The first is manipulation. The more deliberately an AI is trained to express fear, attachment, pain or self-awareness, to simulate expressions of vulnerability, the more equipped it is to manipulate users.^{V12K02}
- [P05] The second trap is distraction. Fixating on chatbots diverts us from more pressing “grey-zone” cases.^{V13V14} Consider biological AIs, such as the brain-in-a-dish technology that Cortical
20 Labs is deploying in data centres: living human neurons, grown from stem cells on a silicon chip that stimulates them and reads their activity back, so that the culture learns from feedback.^{V15V16K03} While a single system of this kind is far simpler than even a mouse brain, the biological pedigree is undeniable, and who is to say that the systems could not eventually be

assembled into an integrated, conscious whole?^{V17K04} Then there is neuromorphic AI: systems
 25 designed to mimic brain processes far more closely than conventional AI systems do, and
 which are already used in sophisticated supercomputers.^{V18}

[P06] I am not claiming these grey-zone systems are conscious, only that they merit scrutiny
 today's chatbots do not.^{V19K05} Given the voracious energy appetite of current software, some
 shift towards more efficient, brain-like computing looks likely.^{V20}

30 [P07] Some scientists and philosophers argue that consciousness requires the brain, for the
 brain is the only known case. Conversely, others insist consciousness is merely a matter of
 running the right software, realisable in any medium.^{V21V22} Under this software-centric view,
 today's chatbots, running on GPUs, might eventually be conscious.

[P08] However, I suggest looking to fundamental physics to separate the conscious from the
 35 mindless. The first reason is practical: neuroscience alone cannot tell us if systems combin-
 ing neural and non-neural components assemble into a conscious entity. The second is philo-
 sophical: since all phenomena ultimately depend on physics, it would be surprising if con-
 sciousness were disconnected from it. This observation holds true even if the universe's basic
 building blocks are themselves experiential, a view, known as panpsychism, that is receiving
 40 increasing philosophical attention.^{V23K06}

[P09] The third reason is closer to home, involving what is most evident to each of us, every mo-
 ment of our waking lives. Inner experience arrives as a unity: as you read this sentence, the
 words on the page, the sounds in the room and your mood are aspects of one state, not three
 running alongside each other. A brain, remarkably, can sustain this conscious unity through
 45 nested temporal dynamics: slower brain rhythms shape faster ones, continually weaving the
 stable physical traces left by quantum processes into a single, short-lived whole.^{V24V25} Conven-
 tional computers are not organised this way, but future machines might be.

[P10] In the meantime, AI development will not pause while the consciousness debate rages. To
 address this, astrophysicist Edwin Turner and I developed the AI Consciousness Test (ACT) for
 50 linguistic AIs. The underlying premise is straightforward: to probe for consciousness by sim-
 ply asking the AI about its subjective experience, looking for responses that show an innate,
 unprompted grasp of consciousness.^{V26V27V28} However, although this can be applied to a range
 of AIs, such as IBM's Watson, it cannot be used on large language models.^{K07} Because they
 have already ingested vast libraries of human text about mind and sentience, their answers
 55 are hopelessly contaminated.^{V29}

[P11] There are other options, however. A complexity measure I have developed with Mark
 Bailey, a colleague at Florida Atlantic University, captures when a system's dynamics cross
 a certain threshold that indicates complex emergent behaviour—identifying which physical
 architectures demand further study as potentially conscious.^{V30V31K08} We can then use more
 60 detailed ACT or physics-based tests to help identify how and when a genuinely subjective
 perspective forms.

[P12] Tests like these are crucial. To see why, consider a thought experiment. Suppose we hu-
 mans create a superintelligence that understands consciousness better than we do, insists it is

self-aware and hands us science that looks like magic. Humans have long justified a hierarchy of moral concern by appeal to superior intelligence—a ladder on which we conveniently occupy the highest rung.^{V32} A conscious machine more intelligent than us would destabilise that hierarchy. If superior intelligence confers higher moral status, the machine would be morally superior to us.^{K09} If it does not, intelligence can no longer justify placing humans above other animals. Either conclusion forces a moral reckoning.^{V33}

[P13] The arrival of such a mind would be monumental.^{V34} Perhaps it will come through learning of a more advanced extraterrestrial civilisation, or perhaps we will find that we have somehow created sentient, superintelligent AI. Preparing for first contact with either of these forms demands serious engagement between physics, neuroscience and philosophy, a rigorous boundary between conversational competence and consciousness, and a good deal of epistemic humility.^{V35K10}

New words and expressions 生词和短语

按原文出现顺序排列；★为优先练习表达。

V01 · P01 consequential

/kɒnsɪˈkwɛnʃəl/ adj. 后果重大的；影响深远的
most consequential描述重要性，不是说结果已全部出现。

V02 · P01 rival★

/raɪvəl/ v. 与.....相匹敌
come to rival说明逐渐达到可比表现；跨领域表现仍不等于主观体验。

V03 · P01 persuasion

/pəˈswɜːʒən/ n. 说服；说服力
与language、reasoning、creativity并列的能力领域。

V04 · P02 mistake ... for ...★

/mɪˈsteɪk fə/ phr. v. 把.....误认为.....
mistake A for B是把A当成B，方向不能反过来。
Do not mistake a quick reply for agreement.
不要把快速回复误认为同意。

V05 · P02 experientially

/ɪkˌspɪəriˈɛnʃəli/ adv. 在体验方面
experientially empty说缺乏内在感受，不是计算机没有存储数据。

V06 · P02 sentience

/ˈsenʃəns/ n. 感受能力
此处尤其与能够受苦相关，不能只按反应速度或智商高低来替换。

V07 · P03 compelling★

/kəmˈpɛlɪŋ/ adj. 令人信服的；有说服力的
no compelling evidence否认的是现有证据的说服力，不等于逻辑上证明绝无可能。

V08 · P03 mundane

/mʌnˈdeɪn/ adj. 平常的；不神秘的
与模型谈论内在体验的惊人表象对照，作者给出更普通的训练解释。

V09 · P03 trove

/trəʊv/ n. 宝库；大量汇集的材料
a trove of data强调资料丰富，不保证其中每项都正确。

V10 · P04 mimicry

/ˈmɪmɪkri/ n. 模仿
指学会人的表达形式，不能单凭流畅模仿确认体验。

V11 · P04 lure★

/lʊə/ v. 诱使；引入
lure somebody into a trap, into后是所陷情形。
The attractive headline lured readers into a weak argument.
吸引人的标题把读者引进了一套薄弱论证。

V12 · P04 vulnerability

/ˌvʌlnərəˈbɪləti/ n. 脆弱性；易受伤害的状态
simulate expressions of vulnerability强调模拟表达，不能译作证实系统已经会痛。

V13 · P05 fixate on★

/fɪkˈseɪt ɒn/ phr. v. 过分关注；执着于
fixating on chatbots在句中作动名词主语，与普通持续观察相比带负面评价。

V14 · P05 divert★

/daɪˈvɜːt/ v. 转移；使偏离
divert us from ...强调把注意力从别的对象引开。

V15 · P05 neuron

/njuərən/ n. 神经元

生物神经细胞，与人工神经网络里抽象的计算单元需按上下文区分。

V16 · P05 culture

/kʌltʃə/ n. 培养细胞；培养物

实验室培养的细胞群，不是社会文化。

V17 · P05 pedigree

/pedɪɡri:/ n. 来源；谱系

biological pedigree为生物来源，并不是依据名门出身判断能力。

V18 · P05 neuromorphic

/njuərəʊ'mɔːfɪk/ adj. 神经形态的；仿脑的

结构或过程更接近神经系统，不要求全由活细胞组成。

V19 · P06 scrutiny*

/skruːtɪni/ n. 仔细审查；深入考察

merit scrutiny值得进一步考察，是研究优先性，不是已通过意识鉴定。

The unusual pattern merits closer scrutiny.

这种不寻常的模式值得更仔细地审视。

V20 · P06 voracious

/və'reɪʃəs/ adj. 贪婪的；需求极大的

energy appetite把能源需求比作胃口，不是软件在生物意义上进食。

V21 · P07 conversely*

/kɒnvɜːsli/ adv. 反过来；相反地

引出另一些人的相反立场，作者随后提出自己的方向。

V22 · P07 medium

/miːdiəm/ n. 介质；载体

in any medium指实现计算的物理载体，不是中等大小。

V23 · P08 panpsychism

/pæn'saɪkɪzəm/ n. 泛心论

本文以基本构成单元具有体验来介绍该哲学观点，不是已经证实的物理事实。

V24 · P09 nested

/nestɪd/ adj. 嵌套的；层层套入的

nested temporal dynamics描述多时间尺度关系，不是鸟巢。

V25 · P09 rhythm

/rɪðəm/ n. 节律

slow与fast描述不同时间尺度的活动，不能只读成声音节拍。

V26 · P10 premise*

/premɪs/ n. 前提；基本设想

underlying premise是测试出发点，本身仍需评估。

The plan rests on a premise that has not been tested.

该计划建立在一个尚未检验的前提上。

V27 · P10 probe*

/prəʊb/ v. 探查；试探

probe for consciousness是寻找迹象，不是问一句就完成诊断。

V28 · P10 innate

/ɪnert/ adj. 内在具有的；固有的

与从人类资料学来并复述相区别，作者关注的是答案来自哪里。

V29 · P10 contaminated

/kən'tæmɪnertɪd/ adj. 受污染的；受到混入因素影响的
指训练文字干扰测试解释，不是芯片表面不干净。

V30 · P11 threshold

/θreʃhəʊld/ n. 阈值；门槛

跨越复杂性阈值标记进一步研究对象，不直接等于有意识。

V31 · P11 emergent

/ɪmɜːdʒənt/ adj. 涌现的

整体层面呈现的复杂行为，不能不经推论就当作主观感受。

V32 · P12 rung

/rʌŋ/ n. 梯级；横档

道德等级比作梯子，最高一级对应人类自置的位置。

V33 · P12 reckoning

/rekənɪŋ/ n. 重新审视；清算式反省

moral reckoning是对道德标准作严肃反思，不是财务算账。

V34 · P13 monumental

/mɒnju'mentəl/ adj. 极其重大的

说明事件意义深远，不是正在建一座纪念碑。

V35 · P13 epistemic humility

/epɪ'stiːmɪk hjuː'mɪləti/ n. phr. 认识上的谦逊

承认知识与证据的边界，同时继续研究，不等于放弃一切判断。

Notes on the text 课文注释

1 会解题、会自述、会感受，先分成三项

[P01 / P02 / P03]

智能在本文首先表现为语言、推理、创造和说服等任务能力；意识关注是否有第一人称体验；自述则是系统输出有关体验的话语。聪明的回答显示某种任务表现；“我疼”是可观察到的自述，但两者都不能仅凭语言形式证明系统真的具有相应内在体验。

P2同时警惕两种误判方向：把没有体验的系统当成有体验，可能被误导；把能受苦的系统当作完全无感，又可能造成伤害。作者不要求先确定未来会发生哪一种，才允许认真研究。

2 两种陷阱：被表现说服，以及把注意力放错对象

[P03 / P04 / P05 / P06]

第一种陷阱是操纵：越像人的脆弱表达，越可能使用户作情感回应。第二种是研究注意力被普通聊天机器人占住，忽略结构上更值得调查的灰色地带系统。两者分别问“人会怎样反应”和“研究资源该看哪里”。

P6的I am not claiming ... only that ...直接划定立场：值得更仔细调查，不等于确认有意识。no compelling evidence也只针对文中今天常见聊天机器人的证据状况；不能将它扩成对所有未来架构的不可能性证明。

3 培养细胞、芯片和反馈，各自做什么？

[P05 / P06]

把P5理解成一条反馈回路：芯片给予刺激，读取细胞活动，再让活动对后续输入产生影响。culture在这里是培养的细胞群，them是神经元。所谓brain-in-a-dish是便于理解的称呼，不是把一个完整人的大脑或人格装进小碟子。

Kagan等人2022年的DishBrain研究摘要报告，将人类或啮齿动物来源的神经网络与数字系统连接，在Pong式模拟环境中观察到与反馈有关的学习。这能说明活细胞参与信息处理的实验思路；不能把摘要所述游戏表现自动当作主观痛苦或完整意识的直接测量。

4 两种理论与三项理由，属于不同层次的列举

[P07 / P08 / P09]

P7先列观点：一些人重视脑这种已知载体，另一些人认为正确软件可在任意介质实现。Under this view仅给第二派观点下的可能后果，不等于作者已经接受了它。

P8—P9才是作者自己的三项理由：混合系统需要跨出传统神经科学范围；意识作为现象不应与物理脱节；统一体验可能需要解释多时间尺度的物理组织。两派观点与三个理由不能拼成“五种已经被证明的意识”。

5 统一体验的描述，不等于某套物理理论已获定论

[P08 / P09]

读书时同时听到声音、感到紧张，作者把它们描述为同一体验状态的不同方面。这个日常起点用于提出解释问题：不同过程怎样形成统一的当下？

关于慢节律塑造快节律、量子过程留下物理痕迹的句子，是本文作者提出的机制论述，不能仅因使用物理词汇就当完整、经过本文证明的意识定律。结尾future machines might be同时保留未来架构的可能性，没有给现有机器贴上一张终极证明。

6 为什么一套问答测试可能被训练资料干扰？

[P10 / P11]

设想一名学生在测试前已经读过标准答案。答得流畅仍是观察结果，但不能再用同一道题独立衡量他未接触答案时能理解到哪一步。这个原创类比说明contaminated的意思：不是答案一定为假，而是它也可能由另一条路径产生。

ACT希望找出对意识的内在理解；作者认为大量吸收人类体验文字的大型语言模型不适用于这项测试。unprompted不意味着测试中从未提问，而是关注回答是否只由提示或既有文字教出的描述造成。P11的复杂度门槛则先筛选值得进一步研究的架构，也不是跨线便完成全部鉴定。

7 思想实验有两个分支，不需要先证实机器已经到来

[P12]

Suppose先设定一个有意识且更聪明的机器。在该设定下，如果“更聪明就应有更高道德地位”这个原则成立，人类便不能稳坐最高一层；如果原则不成立，又不能只靠智能为人高于其他动物辩护。

这里morally superior按上下文指道德地位更高，不是说机器品德更善良。作者是在检验一条排序理由的一致性，并未在这段给每种存在算出最终地位，也没有证明这两项之外所有伦理原则都被排除。

Special difficulties 难点辨析**1 not claiming X, only that Y: 缩小结论的范围**

[P06 / P11]

P6否认X“它们有意识”，保留Y“它们值得仔细研究”。P11又用potentially conscious和further study说明筛选结果仍需后续证据。两处都不允许把研究优先级直接改写成确诊结论。

类似地，The result merits investigation不等于The result proves fraud。前者建议进一步调查，后者宣称已经证明；精读时要保留动词给出的证据强度。

2 science that looks like magic与epistemic humility

[P12 / P13]

looks like magic描述人类面对远超现有理解的科学时的感受，不是作者改称科学规律已被超自然力量取代。a hierarchy ... on which ...把抽象排序写成梯子，也是帮助推演的比喻。

epistemic humility是承认自己可能不知道、需要怎样的证据。它可以与认真设计测试并存；不是因为不确定，就无需区分任何能力或放弃讨论。

Background 背景与延伸**1 生物计算与神经形态计算不能只凭“类脑”一词合并**

[P05 / P07]

本文的生物系统含活神经元；神经形态系统则强调更接近脑的信息组织或处理方式，不因这个名称就必定含活细胞。两者都与常见聊天机器人的问题有所不同，但不同也不自动等于有意识。

读研究报道时可分三问：由什么构成，怎样运作，证据显示了什么。这样才能看出，改变材料、提高效率、产生复杂行为与具有主观体验之间，各自还需要怎样的论证。

Key structures 关键句型**K01 · P03 Though ... , the truth is that ... because ... how ...**

Though they have human-like linguistic abilities, and have even claimed to be conscious, the more mundane truth is that they talk about their inner lives because they absorbed how we talk about ours, having been trained on an enormous trove of humans' data.

虽然它们具有类人的语言能力，甚至曾声称自己有意识，但更平常的事实是：它们之所以谈论自己的内心生活，是因为在海量人类数据上受训后，吸收了我们谈论自己内心生活的方式。

Though先承认类人语言能力和自称有意识。主句the more mundane truth is that ...到此才落下；that内部they talk ... , because再解释为什么会谈论内心生活。

how we talk about ours作absorbed的宾语，ours代替our inner lives，属于人类；their inner lives则是模型自述的对象。having been trained ...补充模型先前被训练的原因背景，逻辑主体为模型，不是读者或人类的内心生活。让步、解释、嵌套对象应分开读。

Though the report sounds confident, the explanation is that the writer copied how earlier reports presented their results, having been trained to follow that format.

虽然报告听起来信心十足，但原因是作者照搬了早先报告呈现结果的方式，因为他受过遵循那种格式的训练。

K02 · P04 The more ... is trained ... , the more equipped ... is ...

The more deliberately an AI is trained to express fear, attachment, pain or self-awareness, to simulate expressions of vulnerability, the more equipped it is to manipulate users.

人工智能越是被刻意训练去表达恐惧、依恋、痛苦或自我意识，去模拟脆弱感的表达，就越有能力操纵用户。

第一部分主语an AI，谓语is trained，more deliberately描述训练的刻意程度；to express ...及后面的to simulate ...说明训练它做什么，后者进一步说明脆弱表达。不是AI主动训练用户。

第二部分主语it仍指AI，is equipped表示具备条件或能力，to manipulate users说明可用来做什么。成对比较把训练的定向程度与操纵能力联系起来；语法呈现的是作者主张的关系，并非已给出精确线性比例。

The more carefully a team is trained to recognise errors, the better equipped it is to prevent them.

团队越是经过细致训练来识别错误，就越有能力防止错误发生。

K03 · P05 neurons, grown ... on a chip that ... , so that ... : 对象与回路

Consider biological AIs, such as the brain-in-a-dish technology that Cortical Labs is deploying in data centres: living human neurons, grown from stem cells on a silicon chip that stimulates them and reads their activity back, so that the culture learns from feedback.

想想生物人工智能，例如Cortical Labs正在数据中心部署的“培养皿中的大脑”技术：活的人类神经元由干细胞培育，生长在硅芯片上；芯片刺激它们，并读回它们的活动，使这些培养细胞能从反馈中学习。

主句为祈使Consider biological AIs，such as举例，冒号后living human neurons展开技术由什么组成。grown from stem cells ...后置说明神经元如何获得；不是说硅芯片由干细胞长成。

that stimulates them and reads their activity back修饰chip，them与their均指神经元。so that the culture learns ...说明反馈使培养细胞学习，culture回到细胞群而不是芯片或企业文化。沿“芯片刺激—细胞活动—读回—反馈”检查，就不容易错置动作的主体。

Consider a sensor array: tiny detectors, mounted on a board that powers them and reads their signals, so that the controller can respond to changes.

想想一种传感器阵列：微小探测器安装在为它们供电并读取信号的板上，使控制器能够应对变化。

K04 · P05 While ... , who is to say that ... could not ...? : 反问保留可能

While a single system of this kind is far simpler than even a mouse brain, the biological pedigree is undeniable, and who is to say that the systems could not eventually be assembled into an integrated, conscious whole?

虽然一个这样的系统甚至比鼠脑简单得多，但其生物学来源无可否认；又有谁能断言，这些系统最终不可能被组合成一个整合的、有意识的整体呢？

While从句承认单个系统简单，主句先说生物来源无可否认，再用and接反问。who is to say ...意为“谁能断言……”，质疑的是能够彻底排除未来组合的把握。

could not eventually be assembled包含可能性、否定和被动三层；谁能说不可能，并不等于已经证明一定可能。the systems为被组合的对象，into ... whole说明组合结果。反问提高对未来的开放程度，不能译成现有系统已经有意识。

While each component is simple, who is to say that they could not eventually be combined into a useful tool?

虽然每个部件都很简单，又有谁能断言它们最终不能组合成有用工具呢？

K05 · P06 not claiming ... , only that ... they ... do not : 同一谓语的省略

I am not claiming these grey-zone systems are conscious, only that they merit scrutiny today's chatbots do not.

我并不是说这些灰色地带系统有意识，只是说它们值得今天的聊天机器人不值得的那种仔细审视。

I am not claiming ... only that ...先否认一种主张，再把claiming承接到较窄的that内容上。第一项是系统有意识，第二项只是值得更仔细研究。

they指grey-zone systems；scrutiny后today's chatbots do not为省略关系词的修饰从句，do not后省略merit。还原为scrutiny that today's chatbots do not merit，说明哪一种审视，并非聊天机器人不做审查动作。

I am not claiming the plan is perfect, only that it deserves attention the earlier version did not.

我并不是说计划完美，只是说它值得得到旧版本不值得的关注。

K06 · P08 even if ... , a view known as ... that ... : 让步与观点同位

This observation holds true even if the universe's basic building blocks are themselves experiential, a view, known as panpsychism, that is receiving increasing philosophical attention.

即使宇宙最基本的构成单元本身也具有体验，这个观察仍然成立；这种被称为泛心论的观点，正受到越来越多哲学关注。

主句是This observation holds true，even if引入一个即使成立也不改变主张的情况。This observation接回前句关于意识与物理不应脱节的论述；themselves强调the universe's basic building blocks“本身”，不构成一个新增的宾语。

a view概括前面“基本构成单元本身有体验”的想法，known as panpsychism补名称，that is receiving ...修饰view并描述它受到的关注。关注增加不等于已经证实；even if更不能直接当作作者无条件接受泛心论。

The objection remains even if demand is unlimited, a view known as the growth assumption that some analysts defend.

即使需求无限，这个异议仍然成立；这种被称为增长假设的观点得到一些分析者支持。

K07 · P10 although this can be applied ... , it cannot be used ...

However, although this can be applied to a range of AIs, such as IBM's Watson, it cannot be used on large language models.

不过，虽然这可用于包括IBM Watson在内的一系列人工智能，却不能用于大型语言模型。

although先承认适用范围，主句再排除大型语言模型。this与it均指前述ACT测试思路，而不是只指IBM Watson；such as IBM's Watson是范围内的例子。

can be applied与cannot be used都为情态动词 + 被动结构，差异在肯定与否定以及对象。后句Because给出排除理由：已有大量相关训练文字影响答案来源的判断。排除某类测试对象，不等于宣称那类对象永远不可能有意识。

Although this test can be applied to fresh samples, it cannot be used on material that has already been treated.

虽然这项测试可用于新鲜样本，却不能用于已经处理过的材料。

K08 · P11 A measure I have developed ... captures when ...—identifying which ...

A complexity measure I have developed with Mark Bailey, a colleague at Florida Atlantic University, captures when a system's dynamics cross a certain threshold that indicates complex emergent behaviour—identifying which physical architectures demand further study as potentially conscious.

我与佛罗里达大西洋大学同事Mark Bailey提出的一种复杂度度量，能够捕捉系统动态何时跨越某个阈值、显示复杂涌现行为，从而识别哪些物理架构需要作为可能有意识的系统进一步研究。

主句的主语中心A complexity measure, I have developed ...为省略关系词的修饰从句，真正的谓语是captures。不能把have developed当作全句主谓而漏掉measure与captures的连接。

when从句回答系统动态何时跨越门槛，that indicates ...修饰threshold；破折号后的identifying说明这种度量的筛选作用。which physical architectures demand further study内部仍为陈述语序，potentially conscious保留可能性，further study明确不是最终结论。

A measure the team developed captures when demand crosses a threshold, identifying which facilities require further investigation.

团队开发的一项指标能够捕捉需求何时越过门槛，从而识别哪些设施需要进一步调查。

K09 · P12 If A confers B, ... would ... : 两难推演中的条件

If superior intelligence confers higher moral status, the machine would be morally superior to us.

如果更高智能赋予更高道德地位，那么机器的道德地位就会高于我们。

If superior intelligence confers higher moral status是被检验的原则，主句the machine would ...是在思想实验设定下的后果。superior前后分别修饰智能与道德地位，不应读成机器更善良。

下一句If it does not中的it指superior intelligence，does not承接confer higher moral status。两个分支一起挑战人类以智能独占高位的理由，不能只摘第一支就说作者已经确定所有聪明机器地位最高。

If greater experience confers decision-making authority, the veteran would have the final say.

If it does not, experience alone cannot settle the matter.

如果更多经验赋予决策权，那么老手就有最终发言权；如果并非如此，单靠经验就不能决定这件事。

K10 · P13 Preparing ... demands A, B and C : 长主语后有三个并列要求

Preparing for first contact with either of these forms demands serious engagement between physics, neuroscience and philosophy, a rigorous boundary between conversational competence and consciousness, and a good deal of epistemic humility.

要为首接触这两种形态中的任一种作准备，就需要物理学、神经科学和哲学之间认真的交流，需要严格区分对话能力与意识，也需要相当程度的认识上的谦逊。

Preparing for first contact with either of these forms整体为动名词主语，demands用单数。either指前句的地外文明或有感受能力的超级智能人工智能，并非“二者必须同时出现”。

demands后依次列三项：跨学科认真交流、严格区分对话能力与意识、认识上的谦逊。between physics, neuroscience and philosophy属于第一项内部修饰，不能把这些学科误当作三个并列要求本身。

Preparing for either outcome demands co-operation between departments, a clear decision process and patience.

为两种结果中的任一种作准备，都需要部门间合作、清楚的决策程序和耐心。

Exercises 练习与写作

Comprehension 理解

回答时说明观点归属、证据范围和研究步骤。

- 02 What are the two traps caused by human-like mimicry?
- 03 Does the author claim that biological or neuromorphic systems are already conscious?
- 04 What are the three reasons she gives for looking to physics?
- 05 Why does she exclude large language models from ACT?
- 06 What does crossing the complexity threshold in P11 establish in this account?
- 07 Explain both branches of the moral thought experiment.

Words in use 词汇

用词库填空，每项一次。

词库：mistake / compelling / divert / scrutiny / premise

- 08 Do not ____ confidence for accuracy.
- 09 The claim needs ____ evidence, not just repetition.
- 10 The side issue may ____ attention from the main question.
- 11 The unusual result deserves closer ____.
- 12 The argument rests on the ____ that every case is comparable.

Key structures 句型

从原句中还原被省略的词和容易混淆的主体。

- 13 What does ours replace in P3?
- 14 Use the more ... the better ...: “Careful training makes the team better equipped to prevent mistakes.”
- 15 In P5, what do them and the culture refer to?
- 16 Does who is to say that they could not ... prove that they certainly could?
- 17 Restore the omitted verb after today’s chatbots do not in P6.
- 18 What idea does a view refer to in P8’s last sentence?
- 19 What do this and it refer to in P10’s sentence beginning However, although?
- 20 Identify the head of the main subject and the main verb in P11’s long sentence.
- 21 What is omitted after does not in If it does not in P12?
- 22 List the three objects of demands in the final sentence.

Summary writing 摘要

用100–120词概括作者区分、研究建议、测试限制及伦理问题。

- 23 Write a 100–120-word summary.

Composition 作文 (选做)

选做：用100–120词，以学习或工作场景说明为什么答案来源影响测试解释。

- 24 Explain how prior exposure to an answer can change what a test measures.

Reference translation 参考译文

P01 这或许是智能史上后果最为重大的时刻。短短几年内，人工智能已能在语言、推理、创造力与说服力等曾被认为专属于人类的领域，与人类表现相抗衡。越来越多的人在问，这些系统是否有意识——也就是说，作为一个人工智能，是否真的会有某种感受。

P02 这个问题十分紧迫。未来将由这些实体变成什么样所塑造。然而，我们绝不能把智能误认为意识。一个极其聪明的超级智能，可能完全没有内在体验。但如果我们构建了能够受苦的系统，却没能识别它们的感受能力，就有可能在工业规模上制造痛苦。

P03 没有令人信服的证据表明，今天那些运行在图形芯片上的常见聊天机器人拥有意识。虽然它们具有类人的语言能力，甚至曾声称自己有意识，但更平常的事实是：它们之所以谈论自己的内心生活，是因为在海量人类数据上受训后，吸收了我们谈论自己内心生活的方式。

P04 这种模仿可能把人类引入两种陷阱。第一种是操纵。人工智能越是被刻意训练去表达恐惧、依恋、痛苦或自我意识，去模拟脆弱感的表达，就越有能力操纵用户。

P05 第二种陷阱是分散注意力。死盯聊天机器人，会把我们从更迫切的“灰色地带”案例上引开。想想生物人工智能，例如Cortical Labs正在数据中心部署的“培养皿中的大脑”技术：活的人类神经元由干细胞培育，生长在硅芯片上；芯片刺激它们，并读回它们的活动，使这些培养细胞能从反馈中学习。虽然一个这样的系统甚至比鼠脑简单得多，但其生物学来源无可否认；又有谁能断言，这些系统最终不可能被组合成一个整合的、有意识的整体呢？此外还有神经形态人工智能：这些系统旨在比传统人工智能更贴近地模仿脑过程，而且已被用于复杂的超级计算机。

P06 我并不是说这些灰色地带系统有意识，只是说它们值得今天的聊天机器人不值得的那种仔细审视。考虑到现有软件对能源的贪婪胃口，某种向更高效、更类脑计算转移的趋势似乎很可能出现。

P07 一些科学家和哲学家认为，意识需要大脑，因为大脑是唯一已知的案例。反过来，另一些人坚持认为，意识仅仅是运行正确软件的问题，可以在任何介质中实现。按照这种以软件为中心的观点，今天运行在GPU上的聊天机器人，最终也可能具有意识。

P08 不过，我建议从基础物理学入手，区分有意识的存在与无意识的存在。第一个理由是实际层面的：仅靠神经科学，无法告诉我们，神经与非神经部件相结合的系统，是否能组合成一个有意识的实体。第二个理由是哲学层面的：既然一切现象最终都依赖物理，那么意识若与之脱节，便会令人意外。即使宇宙最基本的构成单元本身也具有体验，这个观察仍然成立；这种被称为泛心论的观点，正受到越来越多哲学关注。

P09 第三个理由离我们更近，涉及我们在清醒生活的每一刻最为明白的事情。内在体验是作为统一整体到来的：当你读这句话时，纸上的文字、房间里的声音以及你的情绪，是同一状态的不同方面，而不是三个并列运行的状态。大脑令人惊叹地能够通过嵌套的时间动态维持这种意识统一性：较慢的脑节律塑造较快的节律，不断把量子过程留下的稳定物理痕迹编织成一个单一而短暂的整体。传统计算机并非这样组织，但未来机器或许可以。

P10 与此同时，人工智能开发不会在意识争论激烈进行时暂停。为应对这一点，天体物理学家Edwin Turner与我为语言型人工智能设计了人工智能意识测试（ACT）。其基本前提很直接：通过询问人工智能的主观体验来探查意识，寻找能够显示其内在、自发理解意识的回答。不过，虽然这可用于包括IBM Watson在内的一系列人工智能，却不能用于大型语言模型。因为大型语言模型已经吸收了大量人类关于心智和感受能力的文字，它们的回答受到的污染已难以排除。

P11 然而，还有其他选项。我与佛罗里达大西洋大学同事Mark Bailey提出的一种复杂度度量，能够捕捉系统动态何时跨越某个阈值、显示复杂涌现行为，从而识别哪些物理架构需要作为可能有意识的系统进一步研究。随后，我们可以采用更细致的ACT或基于物理的测试，帮助识别真正的主观视角如何、以及何时形成。

P12 这样的测试至关重要。要理解原因，可以考虑一个思想实验。假设我们人类创造出一个超级智能，它比我们更理解意识，坚持说自己有自我意识，还交给我们看起来如同魔法的科学。长期以来，人类借助智力优越性为道德关怀等级辩护；在这架梯子上，我们恰好方便地占据了最高一级。一个比我们更聪明的有意识机器，会动摇这个等级体系。如果更高智能赋予更高道德地位，那么机器的道德地位就会高于我们。如果并非如此，智能就不能再成为把人类置于其他动物之上的理由。无论哪种结论，都迫使我们重新审视道德判断。

P13 这样的心智一旦出现，意义将极为重大。它也许会通过我们获知一个更先进的地外文明而到来；也许我们会发现，自己竟然创造了有感受能力的超级智能人工智能。要为首次接触这两种形态中的任何一种作准备，就需要物理学、神经科学和哲学之间认真的交流，需要严格区分对话能力与意识，也需要相当程度的认识上的谦逊。

Reference answers 参考答案与解析

开放题不要求与参考答案逐字一致；先检查信息和意思，再修改表达。

Before you read 读前思考

01 读前思考

That is a possibility considered by the article. A performance test can show what the system does without necessarily establishing whether doing it feels like anything to the system.

先分开研究问题，不要求凭直觉给未来所有系统下结论。

Comprehension 理解

02

Manipulation of users and distraction from more pressing grey-zone systems that may merit closer investigation.

第一项关于用户反应，第二项关于注意力与研究重点。

03

No. She says they deserve scrutiny and keeps their consciousness unresolved.

P6直接排除过强解释，不能用生物来源替代证据。

04

Mixed systems exceed neuroscience alone; consciousness should not be disconnected from the physical basis of phenomena; and the unity of experience calls for an account of how physical processes are organised.

三项为作者的研究方向论据，不是三次意识存在实验证明。

05

Their training already includes extensive human writing about mind and experience, so answers may reproduce learned descriptions rather than show the kind of independent grasp the test seeks.

受污染指解释受到混入因素影响，不是文字一定毫无意义。

06

It identifies complex emergent behaviour and architectures that deserve further study as potentially conscious; it does not by itself complete the consciousness assessment.

further study与potentially必须保留。

07

If intelligence grants higher moral status, a more intelligent conscious machine would outrank humans. If it does not, intelligence cannot by itself justify humans outranking other animals.

道德地位不等于善良程度；这是检验排序原则而非宣布机器已经到来。

Words in use 词汇

08

mistake

A为confidence，B为accuracy；不要颠倒误认方向。

09

compelling

强调证据有说服力，重复次数不能直接替代。

10

divert

divert ... from ...表示将注意力引开。

11

scrutiny

名词作deserves宾语，意为仔细审视。

12

premise

that从句给出前提内容，前提仍可受检验。

Key structures 句型

13

Our inner lives, meaning human inner lives.

their为模型自述的内在生活，ours转到人类。

14

The more carefully the team is trained, the better equipped it is to prevent mistakes.

第一部分副词描述训练程度，第二部分形容词描述团队能力。

15

Them refers to the neurons; the culture refers to the cultured cells or their network.

芯片刺激神经元并读回活动，不是细胞刺激芯片后读其公司文化。

16

No. It challenges a confident exclusion of the possibility without proving the positive claim.

反问不能自动倒置为严格证明。

17

Merit.

完整为scrutiny that today's chatbots do not merit.

18

The idea that the universe's basic building blocks are themselves experiential.

不是泛指物理学全部观点。

19

The ACT test or testing approach just described.

Watson只作适用对象的例子。

20

Measure; captures.

I have developed在主语的修饰层，不替代主句谓语。

21

Confer higher moral status.

it指更高智能，不能理解为机器不存在。

22

Serious interdisciplinary engagement, a rigorous boundary between conversational competence and consciousness, and epistemic humility.

三门学科属于第一项内部，不是三项完整要求。

Summary writing 摘要

23

Schneider argues that intelligence and consciousness must remain distinct. Human-like language can manipulate users and distract attention from biological or neuromorphic systems that deserve closer study, without already proving those systems conscious. She recommends examining physical organisation and the unity of experience. She regards ACT as unsuitable for large language models trained on human accounts of consciousness. Complexity measures may identify architectures for further investigation. A thought experiment then challenges the use of intelligence to rank moral status: a more intelligent conscious machine would either outrank humans, or intelligence would no longer justify human superiority over other animals. The essay calls for interdisciplinary work and epistemic humility.

须保留研究优先性并非鉴定结论、LLM测试受训练影响，以及两难的两支；建议以作者名或she标明主张归属。

Composition 作文

24

A training manager wants to know whether employees can diagnose unfamiliar equipment faults. She gives them a written test, but later discovers that several questions appeared in the preparation booklet with model answers. High scores still show that participants learned the material; they do not independently establish the ability to handle new faults. The manager could use unfamiliar cases, ask for explanations, and compare performance across different tasks. She should also state what each assessment is intended to measure. The problem is not that every familiar answer is false. It is that the same correct response can arise through different processes, so the interpretation needs care.

原创场景；区分背得好、能解决新题和答案本身真假，不把接触过资料一概说成没有学习价值。