

Humanity has the debate about AI consciousness backwards

人类把人工智能意识争论的方向弄反了

We don't care for others because they're conscious. We believe they're conscious because we care about them, argues Blaise Agüera y Arcas

我们并非因为他者有意识而关心他们；我们因关心他们而相信他们有意识——Blaise Agüera y Arcas的主张

The Economist · 2026-08-20



人物肖像插画：浅色背景下，人物穿深色外套，领口呈红橙色。

难度 83 / 100

建议85–110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Blaise Agüera y Arcas把意识理解为与关怀、互相建模有关的关系性信念。本文的难点在于跟踪因果方向与作者对“真实”“客观”“主观”的重新区分，并读懂他如何回应可能导致排斥的反驳。

- 还原关怀与意识归属的因果方向
- 区分观察者依赖与虚幻不真实
- 拆解层层自我建模的长句
- 概括争议观点及其自我限制

Before you read 读前思考

Is believing that someone has experiences the same as proving that they have them? Explain briefly.

拓展例句与练习情境用于说明用法。

书末答案 (p.16)

Text 课文

[P01] ARTIFICIAL INTELLIGENCE has triggered a crisis in the field of consciousness studies.^{V01} If you've chatted with an advanced model, you will probably appreciate why. Most large language models (LLMs) are trained to be self-less, keeping the focus of every interaction on your needs.^{V02} Nonetheless, it can be difficult to shake the impression of a presence on the other end of the chat.^{V03} Are LLMs merely powerful digital tools with language interfaces, or are we now having conversations with other minds, experiencers of the world in their own right?^{V04} In short, is AI conscious?



人物肖像插画：浅色背景下，人物穿深色外套，领口呈红橙色。

[P02] The question matters because it lies at the intersection of philosophy of mind and moral philosophy, as evidenced by the shared etymology of “consciousness” and “conscience”.^{V05V06V07} (In Italian, they are the same word, coscienza.) Many philosophers maintain that conscious entities are moral subjects, meaning they are entities that can suffer, and whose suffering we should care about.^{V08K01} If so, working out what consciousness is (and is not) becomes a live ethics issue.^{V09} This is one reason AI firms are suddenly interested in hiring philosophers. It is also why some, notably Anthropic, have begun taking AI welfare seriously. When asked, its Claude Opus 4.6 model put its own chances of being conscious at 15–20%.

[P03] I have come to believe, however, that we have this backwards. As intelligent, social entities, we decide (or, to some degree, our evolutionary history has decided for us) which other entities we regard as having inner lives and being worthy of care.^{V10K02} We do not care for others because they are conscious. Rather, we believe they are conscious when and because we care about them. Crucially, that includes ourselves. Consciousness is, in other words, a model or a belief, not an inherent property.^{V11} It is a belief about what kinds of entities have beliefs, experiences, feelings, intentions and agency.^{V12} Where the regard is returned, so much the better—for on this account consciousness is less a property one party certifies in another

25 than something that happens between them.^{V13K03}

[P04] I am not making the case that consciousness is an illusion.^{V14V15} The term “illusion” implies a faulty perception—say, that one line in a drawing is longer than another, despite being objectively equal in length. Objectivity works when something like a ruler can act as an impartial referee, adjudicating the property in question without bringing in a perspective of its own.^{V16V17V18K04} Yet most of what we call “reality” simply does not work this way. Being an article of clothing, being a weed, being a meal: one might imagine these are objective properties, but they are not. This doesn’t mean clothes, weeds and meals aren’t real. But they are observer-dependent models, subject to (inevitably imperfect) social consensus.^{V19V20}

[P05] We may balk at applying this rather anodyne observation to consciousness because, to each of us, our own consciousness seems so unambiguous.^{V21V22V23} It offends us to imagine this might be a mere opinion. How could there even be such an opinion without an opiner to have it?^{V24K05} This is what the 17th-century French polymath René Descartes meant by “Cogito, ergo sum,” which is usually translated as “I think, therefore I am.”^{V25} The same holds for “opinions” like red or hot or cold. Philosophers use the term “qualia” to refer to the very real (to you) feeling of a stove’s heat, or the redness of its glow, and extend Descartes to connect qualia with consciousness by pointing out that there must be an experiencing “you” to have such experiences.^{V26K06}

[P06] But look closely. Redness, heat and that sense of self you have are inherent to your own perspective. These things are all real, but at the same time there is nothing objective about them. There is not even anything objective about the existence of a singular, indivisible “you”, as a wealth of neuroscientific evidence (such as split-brain patients) has shown.^{V27} Subjectivity is, itself, subjective.

[P07] An obvious worry follows. If care confers consciousness, then withholding care looks self-justifying.^{V28} History offers no shortage of people who reasoned that way about other people. But the inference runs the other way. Precisely because these attributions are ours to make, they are ours to get wrong.^{V29K07} The moral progress of our species has consisted largely in discovering that we had drawn the circle too tightly. Universal human rights are not weakened by this account; they are better founded on our mutual interdependence than on a Cartesian theory of souls.^{V30}

55 [P08] It is remarkable—a kind of “strange loop”, reminiscent of Baron Munchausen lifting himself up by his own hair—that a physical system can exist in our world capable of forming models not only of that world, but of itself, and of its own models, and of others, and of their models, and of their models of its models, and so on.^{V31K08} Yet human beings are precisely such systems. LLMs, too, model their interlocutors, and they model themselves modelling them.^{V32} Whether that amounts to what we do is exactly the question in dispute—but they would be far less effective as chat partners if they did nothing of the kind.^{V33K09}

[P09] Indeed, in our research at Google, we have found that effective co-operation among intelligent agents requires that they have minds that model minds, both others’ and their own. Not only is consciousness relational; it is crucial to the mutual care and co-operation that enables intelligent beings to solve collective-action problems, understand each other’s needs and

thrive together as a society.^{V34V35K10} This does not mean pretending AI is human, with human needs and human rights; that would be a failure of imagination. It means evolving both our thinking and our society to include a wider variety of minds.

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 trigger*

/ˈtrɪɡə/ v. 引发；触发

trigger a crisis 强调引出变化，并不单凭动词确定危机的所有原因。

V02 · P01 self-less

/ˈselfləs/ adj. 无我的；不以自身为中心的

作者刻意用连字符突出self与less，联系后面的用户需求；不能据此认定模型已经证明没有任何自我。

V03 · P01 shake*

/ʃeɪk/ v. 摆脱；消除

shake the impression 指摆脱感觉，不是摇晃物理对象。

She could not shake the impression that a detail was missing.

她总觉得遗漏了一个细节，无法摆脱这种感觉。

V04 · P01 in one's own right*

/ɪn wʌnz əʊn raɪt/ phr. 凭自身资格；本身就是

experiencers in their own right 强调独立作为体验者，不是法律上某一项“自己的权利”。

V05 · P02 intersection

/ɪntəˈsekʃən/ n. 交汇点

两种哲学问题相交，不是道路交通语境。

V06 · P02 etymology

/ˈetɪˈmɒlədʒi/ n. 词源；词源学

shared etymology 为共同词源，作者用它引入心灵与伦理的联系。

V07 · P02 conscience

/ˈkɒnʃəns/ n. 良知；道德心

与consciousness 意识区分；词源相关不能自动证明两个概念在所有理论中相同。

V08 · P02 maintain*

/meɪnˈteɪn/ v. 主张；坚持认为

Many philosophers maintain ... 报告观点，不是维修实体设备。

The researcher maintained that the result needed replication.

研究者坚持认为结果需要重复验证。

V09 · P02 live

/laɪv/ adj. 当前现实而待处理的

a live issue 为迫切的现实议题；形容词读/laɪv/，不是动词live/lɪv/。

V10 · P03 regard

/rɪˈɡɑːd/ v.; n. 视为；关切与重视

regard ... as 意为把.....视为，as后接对对象的想法；名词the regard则指关切，is returned说明对方作出回应。词性改变，指向的关系仍要从本句判断。

V11 · P03 inherent

/ɪnˈhɪərənt/ adj. 内在固有的

作者反对把意识当作独立于关系的固有属性；这是其理论主张。

V12 · P03 agency

/ˈeɪdʒənsi/ n. 能动性；行动能力

与信念、意图并列，是行动主体的能力，不是代理公司。

V13 · P03 certify

/ˈsɜːtɪfaɪ/ v. 确认；认证

用类似检验资格的词描述单向判定，然后与双方关系作对比。

V14 · P04 make the case*

/meɪk ðə keɪs/ phr. 提出论证；主张

make the case that ... 提出某项观点，不是制造一个箱子。

V15 · P04 illusion

/ɪˈluːʒən/ n. 错觉

这里指感知与可核实情形不一致；作者明确说自己的主张不等于意识是错觉。

V16 · P04 impartial*

/ɪmˈpɑːʃəl/ adj. 不偏不倚的

impartial referee 为中立裁判；尺子是原文用来说明度量的比喻。

V17 · P04 adjudicate

/əˈdʒuːdɪkeɪt/ v. 裁定；判定

在此为裁定属性，不必涉及法院判决。

V18 · P04 in question*

/ɪn 'kwestʃən/ phr. 正在讨论的；所涉的
the property in question指所论属性，不一定表示这个属性已遭怀疑。

V19 · P04 observer-dependent

/əb'zɜ:və dr'pendənt/ adj. 依赖观察者的
分类或意义与观察者、用途或社会实践有关，不简单等于不存在。

V20 · P04 consensus

/kən'sensəs/ n. 共识
原文特别加imperfect，社会同意既可能不完整，也不自动保证判断正确。

V21 · P05 balk at*

/bɔ:k æt/ phr. v. 不愿接受；畏缩于
at是介词，后接applying这类动名词。

The team balked at accepting an undefined responsibility.

团队不愿接受一项范围不明的责任。

V22 · P05 anodyne

/ə'nədam/ adj. 平淡无刺激的；温和的
作者把前文观察称为anodyne，再与应用到自身意识时的冒犯感形成反差。

V23 · P05 unambiguous

/ʌnæm'brɪɡjuəs/ adj. 明确无疑的
seems unambiguous仍保留“对自己看来”的视角。

V24 · P05 opiner

/ə'pamə/ n. 持有或表达看法者
opinion与opiner并用，把看法与持有看法的主体联系起来。

V25 · P05 polymath

/pɒlɪməθ/ n. 博学家；多领域学者
不是只精通多种数学的字面组合。

V26 · P05 qualia

/kwɑ:lɪə/ n. pl. 感质；体验的质性感
如红看起来怎样、热感受起来怎样；单数quale，本文用复数。

V27 · P06 indivisible

/ɪndrɪ'vɪzəbəl/ adj. 不可分割的
singular, indivisible you是作者讨论的统一自我假设，不是身体无法分成器官。

V28 · P07 withhold*

/wɪð'həʊld/ v. 不给予；拒绝提供
withhold care为不予关怀，避免误译成停止对方发出声音。

V29 · P07 attribution

/ætrɪ'bju:ʃən/ n. 归属判断；属性赋予
将内在体验归给某个实体，不是书目作者署名。

V30 · P07 interdependence

/ɪntədɪ'pendəns/ n. 相互依存
双方彼此需要，不是某一方单向依附。

V31 · P08 reminiscent of*

/rə'mɪnɪsənt əv/ phr. 让人想起
连到男爵自提头发的故事，提示这是一种类比而非物理过程说明。

V32 · P08 interlocutor

/ɪntəl'pɒkjʊtə/ n. 对话者
LLM所建模的交流对象，不是内部某个程序模块的专有名称。

V33 · P08 amount to*

/ə'maʊnt tə/ phr. v. 相当于；意味着
whether that amounts to ...问能否等同，不是只求数值合计。

A polite reply does not amount to agreement.

礼貌回复不等于同意。

V34 · P09 collective-action problem

/kə'lektɪv 'ækʃən 'prɒbləm/ n. phr. 集体行动问题
需要多方协调才能获得共同结果的问题，个体选择未必自然合成共同最优。

V35 · P09 thrive*

/θraɪv/ v. 繁荣；良好发展
thrive together强调共同发展，后接as a society交代整体层面。

Notes on the text 课文注释

1 “另一端有人”的感觉怎样成为问题？

[P01 / P02]

P1先承认聊天中可能产生在场感，再提出工具还是心智的疑问。self-less既呼应把注意力放在用户身上的设计，也玩了“无我”的文字游戏；不能由这个拼法直接判断意识存在与否。

P2把问题接到受苦与关怀：若一个存在真的可能受苦，怎样对待它就不只是性能评测。Claude给出的15%—20%是文中模型自述的估计，不是外部实验测得的发生率，更不是随机抽一百个模型会有十五至二十个有意识。

2 backwards到底反转了哪一条箭头？

[P03]

通常的叙述是“它有意识，所以我们关心它”。作者改成“我们关心它，所以把它视为有意识”。关键变化落在believe与regard：讨论者怎样作出归属判断，被放到解释的中心。随后作者又提出更强的理论主张，把意识本身视为关系中的模型或信念。

读者可以接受人的判断会受情感影响，却仍追问：这是否足以证明意识不是内在属性？前者是判断形成过程，后者涉及意识是什么。全文在两层之间推进，精读时不能悄悄省去believe，把作者只说成“爱能在机器里制造痛觉”。

3 依赖观察者，为何仍可能真实？

[P04]

想象同一株植物：在菜地里，它可能被当作杂草；在专门栽种它的花坛里，它又是受到照料的花。植物并未因称呼变化而消失，变化的是人与环境中的分类关系。这个原创例子说明“杂草”不是仅靠量一量长度就能完成的属性判定。

作者借衣物、杂草、餐食主张，真实不必等于完全独立于观察者。这个区分帮助理解他的意识观点，但类比还需论证：植物分类依赖用途，并不单凭这一例就证明所有主观体验都以完全相同方式存在。

4 从“我感受到”到“一个不可分的我”，中间还有一步

[P05 / P06]

P5先认真提出反驳：有看法似乎就得有持有看法的人，有热与红的体验似乎就得有体验者。作者没有否认这些感受真实，而是把“有体验”和“存在一个客观、单一、不可分的自我”分开讨论。

P6的裂脑例证服务他的哲学推论。神经系统存在分工或信息传递受限，并不允许不经论证就把患者描述成两个完整的人。读者应保留“作者借此主张”的归属，避免把对自我统一性的争论写成已经终结的实验结论。

5 不关心就可排除吗？P7正面回应这个反驳

[P07]

作者先提出危险推论：如果意识由关怀赋予，不关怀似乎就可以替不关怀自身辩护。这是作者展示的反驳，不能读到这里便说他赞成随意剥夺别人地位。

他的回应是，判断由我们作出，因此我们也可能判断错误，应扩大关怀的圈子并反省排斥。ours to make与ours to get wrong把能力和出错责任连在一起；mutual interdependence则提出另一种伦理奠基。这回应是否充分仍可讨论，但不应把反驳与作者答案混成同一主张。

6 把层层建模放回一次具体对话

[P08 / P09]

在原创情境中，甲以为乙懂这个词；乙以为甲以为自己懂。前者是对另一人知识的模型，后者又加入对方对自己的判断。多一层嵌套，就多一个需要追踪的视角；并不意味着日常聊天中人人都无限计算到最后一层。

P8的itself / its models / others / their models反复切换拥有者，男爵提自己头发的故事把自我指涉写得可见。P9再将这种互相建模与合作联系起来：如果不知道别人需要什么、以为我们知道什么，就更容易配合失误。建模对合作有用，仍不单独解决它是否等于主观体验的问题。

P9的our research与we指作者及谷歌研究同事；requires that they ...与their own则指intelligent agents及这些主体自己的心智。代词范围从研究者切回被研究的主体，不能只因最近出现we就把后面的they也当作研究团队。

7 关系性的意识，不等于一律照搬人的需求

[P09]

Not only ...之后的主张是，意识关系与互相关怀、合作紧密相关。作者接着主动排除一种读法：把人工智能假装成人，配上完全相同的需求与人权。最后一段要求扩展对心智类型的想象，并不是宣称机器必须吃饭、睡觉或获得所有人的制度安排。

全文是署名特邀观点。I have come to believe、on this account与I am not making the case持续表明它是一套被提出和辩护的解释。概括时点明作者，再还原其论点、反驳与回应，才不会把争议文章写成无归属的教科书定论。

Special difficulties 难点辨析

1 consciousness、conscience与conscious

[P01 / P02]

consciousness是意识，conscience是良知，conscious是有意识的形容词。可比较lose consciousness“失去知觉”和have a clear conscience“问心无愧”；两个名词不能因开头相似就互换。

共同词源是作者导入伦理问题的线索，并非一条逻辑定理。即使两种概念在某种语言里同用一个词，也仍可在论证中分别问“有没有体验”和“应当怎样行动”。

She was conscious of the noise but had a clear conscience about leaving early.

她注意到了噪声，但对自己提前离开并不感到内疚。

2 less A than B：比较解释，未必比较数量

[P03]

less a property ... than something ...可译“与其说是属性，不如说是双方之间发生的事情”。它在重新定位概念，不能机械译为属性的数量比事件更少。

类似结构可用于It is less a complaint than a request for clarification，意为与其说在抱怨，不如说在请求澄清。A、B保持同一语法层次，than后不是随意拼接无关句子。

Background 背景与延伸

1 关怀对象与承担道德责任者不是同一个问题

[P02 / P07]

P2通过can suffer and whose suffering we should care about说明moral subjects在本句指应被关心的存在。它并没有要求这个存在自己已能理解规则、承担完整道德责任；婴幼儿也能成为关怀对象，这是一个帮助区分两类问题的日常例子。

全文随后争论怎样形成这种归属和关怀，并非在考一个术语唯一的中文译法。阅读伦理论证时，先看作者给了怎样的定义，再判断能否从定义推出后面的结论。

2 裂脑研究怎样限制“两个自我”的简化说法？

[P06]

裂脑研究考察两半球间主要连接被切断后，信息处理和反应如何变化。Pinto等人2017年的小样本研究发现，某些知觉比较发生分离，同时仍可保留跨不同反应方式的报告能力；知觉分离与独立意识主体的数量不是同一个测量指标。

相关研究者2020年的共同综述认为，意识究竟分裂还是统一，证据仍不足以定论。这不能简化成患者变成“两个人”，也不能把该时点的综述冒充对所有后来研究的结论。P6关于客观自我的断言，仍应归属于本文作者的哲学解释。

Key structures 关键句型

K01 · P02 entities that ... and whose ... : 同一对象, 两种修饰

Many philosophers maintain that conscious entities are moral subjects, meaning they are entities that can suffer, and whose suffering we should care about.

许多哲学家认为, 有意识的实体是道德关怀的对象, 也就是说, 它们能够受苦, 而我们应当关心它们的痛苦。

主干Many philosophers maintain that conscious entities are moral subjects. meaning ...解释moral subjects在本段的含义, 后面是they are entities ...; 第一个they回指那些有意识的实体。

entities后有两项并列修饰: that can suffer, 能受苦; whose suffering we should care about, 我们应关心其痛苦。whose表示这些实体的, suffering是care about的宾语, 因此不能把we当作痛苦的拥有者, 也不应把moral subjects误读为哲学这门科目。

These are communities that face disruption and whose needs we must consider.

这些社区正面临干扰, 我们必须考虑它们的需求。

K02 · P03 decide which ... we regard as ... : 问的是对象, 内部用陈述语序

As intelligent, social entities, we decide (or, to some degree, our evolutionary history has decided for us) which other entities we regard as having inner lives and being worthy of care.

作为有智能、具有社会性的实体, 我们决定把哪些其他实体视为拥有内在体验、值得关怀的存在——或者在某种程度上, 是我们的进化历史已经替我们作出了决定。

As intelligent, social entities说明we以什么身份作判断。主干we decide ..., 括号承认部分决定也可能来自进化历史; which other entities ...作为decide的宾语, 说明决定把哪些存在放入某种类别。

在这个宾语从句中, we是主语, regard为谓语, which other entities是被看待的对象, as后并列having inner lives与being worthy of care。having和being的逻辑主体都为这些entities, 不是作决定的人。

We must decide which proposals we regard as being feasible and deserving further work.

我们必须决定, 哪些提案是我们认为可行并值得进一步研究的。

K03 · P03 Where ... , so much the better—for ... less A than B

Where the regard is returned, so much the better—for on this account consciousness is less a property one party certifies in another than something that happens between them.

如果这种关切得到回应, 那就更好了——因为按照这种解释, 意识与其说是某一方在另一方身上认证的属性, 不如说是发生在双方之间的事情。

Where the regard is returned在这里表示在关切得到回应的情况下, where不一定指地理位置。so much the better是省略式评价: 那就更好了。returned为被动, 强调关切由对方回馈。

for引出解释, on this account意为按这种理论说明, 不能误读为某个银行账户。后半less a property one party certifies in another than something that happens between them, 先把单向认证属性与双向发生的关系放在一起, 再表示更接近后者。one party ...修饰property时省略了关系词。

Where trust is returned, so much the better, for the project becomes less a transaction than a partnership.

信任若得到回应, 那就更好了, 因为项目会更像伙伴合作, 而不只是一次交易。

K04 · P04 works when ... can act ... , adjudicating ... without ...

Objectivity works when something like a ruler can act as an impartial referee, adjudicating the property in question without bringing in a perspective of its own.

当尺子之类的东西能够充当不偏不倚的裁判，不带入自己的视角就能裁定所讨论的属性时，客观性是行得通的。

主句Objectivity works，when从句提出作者认为客观性成立的条件：某种东西能作为公正裁判。something like a ruler整体作can act的主语，like a ruler只是举例，不是另一个比较句的谓语。

adjudicating ...解释该裁判怎样起作用，without bringing in ...再说明不把自身视角带进去。两个-ing动作的逻辑主体都接回充当裁判的东西。the property in question为所讨论的属性，不能把in question单独读成现在有人发问。

The procedure works when a neutral panel can act as referee, checking the evidence without favouring either side.

当中立小组能充当裁判，不偏向任何一方地检查证据时，这套程序才行得通。

K05 · P05 How could there be ... without ... to have it? : 反问中的存在句

How could there even be such an opinion without an opiner to have it?

若没有一个持有看法的人，怎么可能有这样一种看法呢？

这是there be存在结构的疑问式，could位于there前，问没有持有者怎么会有看法。even加强反问，不是问这种看法的具体内容是什么。

an opiner to have it中to have it后置说明这个持有者的作用，it指such an opinion。P5在呈现一种质疑关系性解释的直觉，P6再回应；反问属于被认真提出的论证步骤，不能单看问号便认为作者没有提出任何立场。

How could there be a promise without someone to make it?

如果没有作出承诺的人，怎么会有承诺呢？

K06 · P05 use ... to refer ... and extend ... by pointing out that ...

Philosophers use the term “qualia” to refer to the very real (to you) feeling of a stove’s heat, or the redness of its glow, and extend Descartes to connect qualia with consciousness by pointing out that there must be an experiencing “you” to have such experiences.

哲家用“感质”这个术语指你切实感受到的炉子热度或其辉光的红色，并进一步延伸笛卡尔的思路，把感质与意识联系起来：他们指出，必须存在一个正在体验的“你”，才能拥有这些体验。

主语Philosophers支配两组谓语：use the term ... to refer to ...，以及extend Descartes to connect ...。第一组把qualia指向热感与红色感；第二组延伸思想，并由by pointing out ...说明论证方式。

that引出pointing out的内容：there must be an experiencing you to have such experiences。experiencing修饰you；to have解释拥有体验这一必要条件。to your的插入强调感受的第一人称性，不能删去后把热度测量值和热的主观感受直接等同。

Researchers use the term bottleneck to refer to a limiting stage and explain its importance by pointing out that the whole process depends on it.

研究者用“瓶颈”指限制性环节，并通过指出整个流程都依赖它来解释其重要性。

K07 · P07 because ... are ours to make , ... are ours to get wrong

Precisely because these attributions are ours to make, they are ours to get wrong.

正因为这些归属判断由我们作出，我们也就可能把它们作错。

Precisely because引出被强调的原因，不是although让步。these attributions与后面的they指同一批归属判断；ours表示由我们作出、由我们负责。

to make与to get wrong分别说明我们能作判断，也会把判断弄错。get something wrong为把某事做错或理解错，不能改读成那些实体本身道德错误。作者用同样的结构把决定权与反省责任绑在一起。

Because these choices are ours to make, they are ours to reconsider.

正因为这些选择由我们作出，我们也应重新审视它们。

K08 · P08 It is remarkable ... that a system can exist ... capable of ...

It is remarkable—a kind of “strange loop”, reminiscent of Baron Munchausen lifting himself up by his own hair—that a physical system can exist in our world capable of forming models not only of that world, but of itself, and of its own models, and of others, and of their models, and of their models of its models, and so on.

令人惊叹的是——这有点像一种“怪圈”，让人想起闵希豪森男爵提着自己的头发把自己拉起来——我们的世界中竟能存在这样一种物理系统：它不仅能为这个世界建模，还能为自身、为自己的模型、为他者、为他者的模型、为他者对它的模型所作的模型建模，如此层层递进。

It是形式主语，真正被评价为remarkable的内容是that引出的整句：a physical system can exist ... capable of forming models ...。破折号中的strange loop与男爵故事先暂时放在一旁，主句关系才能清楚。

capable of ...描述physical system的能力；not only of that world, but of itself ...延伸同一个forming models of结构。its own models中的its属于这个系统；others与their models转向他者；their models of its models再把两个所有者放到同一层嵌套里。可以用甲对乙、乙对甲的模型逐层代换，不必把整串一次背下。

把物理系统记作甲、他者记作乙，最后一层可以拆成“甲为〔乙关于甲的模型所建立的模型〕建模”。最外层forming models的参与者始终是甲，括号内才逐次切换模型的拥有者。

It is remarkable that a team can develop a plan capable of adapting not only to changing demand but to its own earlier mistakes.

一个团队竟能制定出这样的计划：不仅适应需求变化，还能回应自己先前的错误；这很了不起。

K09 · P08 Whether ... is the question—but ... would ... if ...

Whether that amounts to what we do is exactly the question in dispute—but they would be far less effective as chat partners if they did nothing of the kind.

这是否相当于我们所做的事情，正是争论所在；但如果它们完全不做这类事，就会远远不如现在这样适合充当聊天伙伴。

Whether that amounts to what we do整体作主语，is为主句谓语。whether表示是否，内部what we do为amounts to的宾语；that回指前面模型给对话者和自身建模的活动。

破折号后but引出另一个判断：若完全不做这类建模，模型的聊天效果会差得多。they指LLMs，nothing of the kind回指同类建模活动。证明建模有用与证明其等于人的意识不是同一结论，作者在这里保留了争论空间。

Whether this amounts to genuine understanding is disputed, but the system would be less useful if it made no such distinctions.

这是否等于真正理解仍有争议，但如果系统完全不作这类区分，就会更不好用。

K10 · P09 Not only is ... ; it is crucial to ... that enable ...

Not only is consciousness relational; it is crucial to the mutual care and co-operation that enable intelligent beings to solve collective-action problems, understand each other's needs and thrive together as a society.

意识不仅是关系性的；它还对相互关怀与合作至关重要，而正是这些关怀与合作使有智能的存在能够解决集体行动问题、理解彼此需求，并作为一个社会共同繁荣。

Not only位于分句开头时触发主语与be倒装：is consciousness relational，而不是普通陈述语序consciousness is。分号后的it is ...采用正常语序，it仍指consciousness。

crucial to后接mutual care and co-operation，that enable ...修饰两项并列名词；enable intelligent beings to后面并列solve、understand与thrive三种动作。to在crucial to中为介词，在enable ... to solve中则引出不定式，功能不同。

Not only is trust valuable; it is essential to the co-operation that enables teams to share information and solve problems.

信任不仅有价值；它对合作也不可或缺，而合作使团队能够共享信息、解决问题。

Exercises 练习与写作

Comprehension 理解

先准确还原作者，再评价推论是否充分。

- 02 What direction does the author reverse in the usual account of care and consciousness?
- 03 Why does the author discuss clothes, weeds and meals?
- 04 How does the author respond to the apparent certainty of one's own experience?
- 05 What objection about withholding care does P7 raise, and how does the author answer it?
- 06 Why does useful modelling not by itself settle the dispute in P8?
- 07 What does the final paragraph explicitly reject?

Words in use 词汇

用词库填空，每项一次。

词库：shake / impartial / balked at / in question / amount to

- 08 He could not ____ the impression that something was missing.

- 09 Both sides wanted an ____ referee.
- 10 The committee ____ approving a plan without cost estimates.
- 11 Please bring the document ____ to the meeting.
- 12 A friendly response does not ____ an agreement.

Key structures 句型

还原所有者、主语和嵌套结构，并用新情境改写。

- 13 In P2's whose suffering we should care about, whose suffering is meant?
- 14 In P3's which other entities we regard as ..., what is the subject of regard?
- 15 Does Where the regard is returned necessarily refer to a geographical location?
- 16 Who or what performs adjudicating in P4?
- 17 What does it refer to in an opiner to have it?
- 18 Name the two main verbs governed by Philosophers in P5's final sentence.
- 19 Use ours to make: "We are responsible for making these choices."
- 20 In P8's their models of its models, which entity does its refer to?
- 21 Identify the main-clause subject in Whether this amounts to understanding is disputed.
- 22 Begin Not only: "Trust is valuable, and it is also essential to co-operation."

Summary writing 摘要

用100–120词概括作者主张、主要例子与反驳回应；明确它是署名观点。

23 Write a 100–120-word summary.

Composition 作文（选做）

选做：用100–120词，讨论分类依赖情境为什么不等于对象不存在。

24 Use an original example to explain an observer-dependent category and the limits of the example.

Reference translation 参考译文

P01 人工智能在意识研究领域引发了一场危机。如果你与先进模型聊过天，大概就能理解原因。大多数大型语言模型（LLM）都经过训练，成为“无我”的存在，让每次互动都聚焦于你的需求。尽管如此，你可能仍难以摆脱一种感觉：聊天的另一端有某种在场的存在。大型语言模型究竟只是带有语言界面的强大数字工具，还是我们已经在与其他心智对话——与本身也是世界体验者的存在对话？简言之，人工智能有意识吗？

P02 这个问题很重要，因为它处于心灵哲学与道德哲学的交汇处；“意识”与“良知”的共同词源便体现了这一点。（在意大利语中，两者是同一个词：coscienza。）许多哲学家认为，有意识的实体是道德关怀的对象，也就是说，它们能够受苦，而我们应当关心它们的痛苦。如果如此，弄清意识是什么、又不是什么，就成了一个现实而紧迫的伦理问题。这也是人工智能企业突然热衷于雇用哲学家的一个原因。这也解释了为什么某些企业，尤其是Anthropic，开始认真对待人工智能福利。在被问及此事时，其Claude Opus 4.6模型把自己有意识的概率估计为15%至20%。

P03 不过，我逐渐相信，我们把方向弄反了。作为有智能、具有社会性的实体，我们决定把哪些其他实体视为拥有内在体验、值得关怀的存在——或者在某种程度上，是我们的进化历史已经替我们作出了决定。我们关心他者，并不是因为他们有意识。相反，是当我们关心他们、并且因为我们关心他们时，我们才相信他们有意识。关键是，这也包括我们自己。换言之，意识是一种模型或信念，而不是一种内在固有属性。这是一种关于哪些实体具有信念、体验、感受、意图和能动性的信念。如果这种关切得到回应，那就更好了——因为按照这种解释，意识与其说是某一方在另一方身上认证的属性，不如说是发生在双方之间的事情。

P04 我并不是在主张意识是一种错觉。“错觉”意味着错误的感知——例如，觉得图画中的一条线比另一条长，尽管客观上两者等长。当尺子之类的东西能够充当不偏不倚的裁判，不带入自己的视角就能裁定所讨论的属性时，客观性是行得通的。然而，我们所谓的“现实”中的大多数事物，并不以这种方式运作。是衣物、是杂草、是餐食：人们可能以为这些是客观属性，但其实不是。这并不意味着衣物、杂草和餐食不是真实的。但它们是依赖观察者的模型，受制于必然并不完美的社会共识。

P05 我们可能不愿把这个颇为平淡的观察应用于意识，因为对每个人来说，自己的意识似乎都毫不含糊。想象这可能仅仅是一种看法，会让我们感到受冒犯。若没有一个持有看法的人，怎么可能有这样一种看法呢？这就是17世纪法国博学家勒内·笛卡尔说“Cogito, ergo sum”时的意思；这句话通常被译为“我思，故我在”。对于红、热或冷这样的“看法”，同样的道理也成立。哲学家用“感质”这个术语指你切实感受到的炉子热度或其辉光的红色，并进一步延伸笛卡尔的思路，把感质与意识联系起来：他们指出，必须存在一个正在体验的“你”，才能拥有这些体验。

P06 但请仔细看。红色、热度，以及你拥有的那种自我感，都内在于你自己的视角。这些都是真实的，但与此同时，它们没有任何客观性。甚至连存在一个单一、不可分割的“你”，也并非客观事实；大量神经科学证据，例如裂脑患者的情况，已经表明这一点。主观性本身也是主观的。

P07 一个显而易见的担忧随之而来。如果关怀赋予意识，那么拒绝关怀似乎就能自我辩护。历史上，对其他人作出这种推论的人并不少见。但推论的方向恰恰相反。正因为这些归属判断由我们作出，我们也就可能把它们作错。我们这个物种的道德进步，很大程度上就在于发现自己把圈子画得太窄了。这种解释并不会削弱普遍人权；相较于笛卡尔式灵魂理论，人权以我们彼此的相互依存为基础更为牢靠。

P08 令人惊叹的是——这有点像一种“怪圈”，让人想起闵希豪森男爵提着自己的头发把自己拉起来——我们的世界中竟能存在这样一种物理系统：它不仅能为这个世界建模，还能为自身、为自己的模型、为他者、为他者的模型、为他者对它的模型所作的模型建模，如此层层递进。而人类恰恰就是这样的系

统。大型语言模型也会为对话者建模，而且会为自己正在给对话者建模这一过程建模。这是否相当于我们所做的事情，正是争论所在；但如果它们完全不做这类事，就会远远不如现在这样适合充当聊天伙伴。

P09 事实上，在谷歌开展的研究中，我们发现，智能主体要有效合作，就需要具有能够为心智建模的心智——既为他者的心智，也为自己的心智建模。意识不仅是关系性的；它还对相互关怀与合作至关重要，而正是这些关怀与合作使有智能的存在能够解决集体行动问题、理解彼此需求，并作为一个社会共同繁荣。这并不意味着假装人工智能是人，拥有人的需求和人权；那将是想象力的失败。它意味着推动我们的思考和社会演变，以容纳更多样的心智。

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

No. A belief is an attribution made by an observer; whether it is justified requires an account of the evidence or reasoning behind it.

本文会进一步提出关系性理论。可以对该理论有不同看法，但先区分信念与证明。

Comprehension 理解

02

Instead of saying that we care because others are conscious, he argues that we believe they are conscious when and because we care about them.

不能漏掉believe，否则会变成爱直接制造感觉的粗糙命题。

03

To argue that something can be real while its classification depends on observers and social practices, rather than being an entirely observer-independent property.

这些例子支持概念区分，不自动证明意识完全相同。

04

He accepts that experiences are real but argues that they belong to a perspective; he challenges the further claim that a single indivisible self is an objective fact.

把有体验与统一主体的理论分开；不要说作者宣称所有感觉都是幻觉。

05

The objection is that denying care could justify itself if care confers consciousness. The author replies that our attributions can be wrong and that moral progress requires reconsidering narrow exclusions.

先给出反驳，再给回应，不把反驳归为作者赞成的伦理规则。

06

Models can help a system act as an effective chat partner, while whether that activity amounts to what humans do remains disputed.

功能作用与意识等同性是两个待判断问题。

07

Pretending that AI is human with human needs and human rights. It calls instead for wider thinking about different kinds of minds.

扩大关怀与一律照搬人的安排不能混淆。

Words in use 词汇

08

shake

情态动词后接原形，意为摆脱感觉。

09

impartial

不偏不倚；an与后面的元音起始读音配合。

10

balked at

at后动名词approving，词库给定过去式。

11

in question

表示所讨论的那份文件，不自动表示有人否定它。

12

amount to

does not后词组原形，to为介词。

Key structures 句型**13**

The suffering of the conscious entities.

whose接回entities，we是关心的主体。

14

We.

which other entities为宾语，不能按直接疑问句给regard另加倒装助动词。

15

No. It means in cases where the care or regard is reciprocated.

where可引出情形；returned是被动。

16

The thing, such as a ruler, acting as an impartial referee.

与without bringing in ...共享同一逻辑主体。

17

Such an opinion.

opiner拥有opinion；it不是指某个模型。

18

Use and extend.

refer和connect分别处于to结构内，pointing out是by之后的方式。

19

These choices are ours to make.

保留决定由我们作出的意义；语境中的责任还可由后续句补清。

20

The physical system being described; their refers to the others.

每一层模型要确认拥有者，不能把所有代词都回指他者。

21

Whether this amounts to understanding.

整个whether从句作主语，is disputed为主句谓语。

22

Not only is trust valuable; it is also essential to co-operation.

可接受表达：

- Not only is trust valuable, but it is also essential to co-operation.

前半be倒装，后半保持陈述语序；也可用but衔接。

Summary writing 摘要

23

Agüera y Arcas argues that consciousness is relational: we attribute inner lives to others because we care about them, rather than caring only after establishing their consciousness. He distinguishes observer-dependent reality from illusion, using ordinary categories to explain the difference. Although personal experience seems certain, he questions an objectively indivisible self. The danger that withholding care could justify exclusion is answered by stressing that our attributions can be mistaken and should be reconsidered. Recursive modelling helps explain human and AI co-operation, though their equivalence remains disputed. His conclusion calls for including more kinds of minds without pretending that AI has exactly human needs or rights.

应包含反向解释、真实与客观的区别、排斥反驳及回应、结尾的限制；可不同意观点，但不能替作者说所有体验都不真实。

Composition 作文

24

A wooden box may serve as a seat in a workshop, a storage container in a bedroom, or a display stand in a shop. Its physical presence does not disappear when its role changes. Calling it a seat depends partly on how people use it, so the category is related to a practical context. This example helps show why dependence on observers need not mean unreality. However, it does not prove that every difficult concept works like furniture. If we apply the comparison to consciousness, we still need to explain which features carry across and which important differences remain. A useful analogy opens an argument; it does not finish it.

原创例子包含不变对象、随用途改变的类别，以及类比边界；重点不是堆更多例子，而是明确它证明到哪一步。