苏菲拉底 THE SOPHIE LAB

本期追问

为什么我们会觉得意识存在一个无法解释的难题?

如果解释了我们谈论意识的行为，意识本身还剩下什么?

我们对自身意识的直觉判断，可信到什么程度?

视频精读 VIDEO STUDY

The Meta-Problem of Consciousness | Professor David Chalmers | Talks at Google

来源：YouTube 视频（https://www.youtube.com/watch?v=OsYUWtLQBS0）

节目发布：2019-04-02 · 全文 123 段 · 页边注释 81 条 · 生词词组 100 条

好视频，值得精读

目录 CONTENTS

导读　视频总结

本期讲者

大卫·查尔默斯（David Chalmers）　澳大利亚裔哲学家、认知科学家，纽约大学教授。1995年提出「意识的难问题」，著有《有意识的心灵》，并与安迪·克拉克合著著名论文《延展心灵》。

观众（Google 员工）　Talks at Google 现场的工程师与研究人员，在问答环节就机器意识、哥德尔定理、IIT 理论与延展心灵等话题提问。

第一部分　全文对照 · 附页边注释与生词标注

01　开场：为何研究意识问题

0:00

[MUSIC PLAYING]

[音乐播放]

0:06 大卫·查尔默斯

DAVID CHALMERS: Thanks for coming out. It's good to be here. As Eric said, I am a philosopher thinking about consciousness. Coming from a background in the sciences and math, it always struck me that the most interesting and hardest unsolved problem in the sciences was the problem of consciousness. And way back 25 years ago when I was in grad school, it seemed to be the best way to come at this from a big picture perspective was to go into philosophy and think about the foundational issues that arise in thinking about consciousness from any number of different angles, including the angles of neuroscience and psychology and AI.

大卫·查尔默斯：感谢大家前来。很高兴来到这里。正如埃里克所说，我是一名思考意识问题的哲学家。我出身于理科和数学背景，一直觉得科学中最有趣、最难的未解难题，就是意识问题。早在25年前我读研究生的时候，我觉得从宏观视角切入这个问题的最佳方式，是进入哲学领域，去思考从各种不同角度思考意识时所产生的那些基础性问题，包括神经科学、心理学和人工智能的角度。

大卫·查尔默斯（David Chalmers），澳大利亚裔哲学家，纽约大学教授，1994—1995年正式提出「意识的难问题」一词，此概念界定了此后三十年意识科学与心灵哲学的核心争论。

0:52

In this talk, I'm going to present a slightly different perspective on the problem after laying out some background, the perspective of what I call the meta-problem of consciousness. I always liked the idea that you approach a problem by stepping one level up, taking the metaperspective. I love this quote, "Anything you can do, I can do meta." I have no idea what the origins was. I like the fact this is attributed to Rudolf Carnap, one of my favorite philosophers. But anyone who knows Carnap's work, it's completely implausible he would ever say anything so frivolous.

在这场演讲中，我会先铺垫一些背景，然后从一个略有不同的视角来谈这个问题，也就是我称之为“意识的元问题”的视角。我一直喜欢这样一种想法：你通过往上跳一层、采取元视角来处理一个问题。我很喜欢这句话：“你能做的任何事，我都能做得更‘元’。”我完全不知道它的出处是什么。我喜欢它被归到鲁道夫·卡尔纳普名下这件事，他是我最喜欢的哲学家之一。但任何了解卡尔纳普著作的人都知道，他绝不可能说出这么轻佻的话。

鲁道夫·卡尔纳普（Rudolf Carnap，1891–1970），逻辑经验主义与维也纳学派代表人物，以极端严谨的形式化哲学著称——因此查尔默斯打趣说他绝不可能讲这种俏皮话。

frivolous /ˈfrɪvələs/ *adj.*
轻佻的，不严肃的

implausible /ɪmˈplɔzəbəl/ *adj.*
难以置信的，不像真的

1:26

It's also being attributed to my thesis advisor, Doug Hofstadter, author of "Godel, Escher, Bach" and a big fan of the metaperspective. But he assures me he never said it either. But the metaperspective on anything is stepping up a level. The meta-problem, as I think about it, is it's called the meta-problem because it's a problem about a problem. A metatheory is a theory about a theory. Meta-problem is a problem about a problem. In particular, it's the problem of explaining why we think there is a problem about consciousness.

这句话也被归到我的博士导师道格·侯世达名下，他是《哥德尔、埃舍尔、巴赫》的作者，也是元视角的忠实拥趸。但他向我保证，他也从没说过这话。不过，对任何事情采取元视角，就是往上跳一个层次。在我看来，元问题之所以叫元问题，是因为它是一个关于问题的问题。元理论是关于理论的理论。元问题就是关于问题的问题。具体来说，它是要解释：我们为什么会认为意识存在一个问题。

道格拉斯·侯世达（Douglas Hofstadter），《哥德尔、埃舍尔、巴赫》作者，该书1980年获普利策奖，主题正是自指与心智。查尔默斯在印第安纳大学随他攻读博士。

02 意识的难问题与现象意识

2:02

So there's a first-order problem, the problem of consciousness. Today, I'm going to focus on a problem about it. But I'll start by introducing the first-order problem itself. The first-order problem is what we call the hard problem of consciousness. It's the problem of explaining why and how physical processes should give rise to conscious experience. You've got all of these neurons firing in your brain, bringing about all kinds of sophisticated behavior. We can get it to be [INAUDIBLE] explaining our various responses, but there's this big question about how it feels from the first person point of view.

所以存在一个一阶问题，也就是意识问题。今天，我要聚焦的是一个关于它的问题。但我会先介绍这个一阶问题本身。这个一阶问题就是我们所说的“意识的难问题”。它要解释的是：物理过程为什么以及如何产生有意识的体验。你的大脑中有这么多神经元在放电，带来各种各样复杂精巧的行为。我们可以做到用它来解释我们的各种反应，但还有一个大问题：从第一人称角度来说，这一切感觉起来是怎样的

「难问题」的正式表述：物理过程为何以及如何产生主观体验。与之相对的「简单问题」只需解释行为与认知功能。这一区分是全场讲座的出发点。

2:42

That's the subjective experience. I like this illustration of the hard problem of consciousness. It seems to show someone's hair catching fire, but I guess it's a metaphorical illustration of the subjective perspective. So the hard problem is concerned with what philosophers call phenomenal consciousness. The word consciousness is ambiguous 1,000 ways. But phenomenal consciousness is what it's like to be a subject from the first person point of view. So a system is phenomenally conscious if there's something it's like to be it.

「现象意识」（phenomenal consciousness）是哲学术语，指从第一人称视角「成为某物是什么感觉」，区别于「取用意识」等意识一词的其他含义。

3:14

A mental state is phenomenally conscious if there's something it's like to be in it. So the thought is there are some systems-- so there's something it's like to be that system. There's something it's like to be me. I presume there's something it's like to be you. But presumably, there's nothing it's like to be this lectern. As far as we know, the lectern does not have a first person perspective. This phrase was made famous by my colleague, Tom Nagel at NYU, who back in 1974 wrote an article called "What Is It Like To Be A Bat?".

托马斯·内格尔（Thomas Nagel），NYU哲学家，1974年发表《成为一只蝙蝠是什么感觉？》，是20世纪被引用最多的哲学论文之一，确立了「主观视角不可从客观描述推出」的经典论证。

lectern /ˈlɛktə-n/ *n.*
讲台，演讲桌

3:49

And the general idea was, well, it's very hard to know what it's like to be a bat from the third person point of view, just looking at it as a human who has different kinds of experience. But presumably, very plausibly, there is something it's like to be a bat. The bat is conscious. It's having subjective experiences, just of a kind very different from ours. In human subjective experience, consciousness divides into any number of different kinds or aspects, like different tracks of the inner movie of consciousness.

蝙蝠靠回声定位感知世界，体验形态与人类根本不同：我们无法从第三人称「知道」它的感受，但它很可能「有」感受。这是主观性论证的核心直觉。

4:23

We have visual experiences like the experience of, say, these colors, blue and red and green from the first person point of view and of depth. There are sensory experiences like the experience of my voice, experiences of taste and smell. They're experiences of your body. Feeling pain or orgasms or hunger or a tickle or something, they all have some distinctive first person quality. Mental images like recalled visual images, emotional experiences like a experience of happiness or anger. And indeed, we all seem to have this stream of a current thought or at the very least, we're kind of chattering away to ourselves and reflecting and deciding.

5:07

All of these are aspects of subjective experience, things we experience from the first person point of view. And I think these subjective experiences are, at least on the face of it, data for the science of consciousness to explain. These are just facts about us that we're having these subjective experiences. If we ignore them, we're ignoring the data. So if you catalog the data that, say, the science of consciousness needs to explain, there are certainly facts about our behavior and how we respond in situations.

关键方法论主张：主观体验本身就是意识科学要解释的「数据」，忽视它们就是忽视数据。这是查尔默斯反驳「只需解释行为」的还原论的基本立足点。

on the face of it *phr.*
乍看之下，从表面上看（常暗示深入后可能不同）

03 简单问题与解释鸿沟

5:39

There are facts about how our brain is working. There are also facts about how subjective experiences, and on the face of it, they're data. And it's these data that pose what I call the hard problem of consciousness. But this gets contrasted with the easy problems, the so-called easy problems of consciousness, which are the problems of explaining behavioral and cognitive functions. Objective things you can measure from the third person point of view typically tied to behavior. Perceptual discrimination of a stimulus, I can discriminate two different things in my environment.

「简单问题」并非真的容易，而是方法论上清晰：找到执行相应功能的神经或计算机制即可，例如刺激辨别、信息整合、行为控制与口头报告。

6:18

I can say, that's red, and that's green. I can integrate the information about the color and the shape. I can use it to control my behavior. Walk towards the red one rather than the green one. I can report it, say that's red, and so on. Those are all data too for science to explain. But we've got a bead on how to explain though they don't seem to pose as big a problem. Why? We explain those easy problems by finding a mechanism, typically a neural or computational mechanism that performs the relevant function to explain how it is that I get to say there's a red thing over there or walk towards it.

got a bead on *phr.*
（get a bead on）瞄准；摸清了……的门路

7:00

Well, you find the mechanisms involving perceptual processes and action processes in my brain that leads to that behavior. Find the right mechanism that performs the function you've explained what needs to be explained with the easy problems of consciousness. But for the hard problem, for subjective experience, it's just not clear that this standard method works. It looks like explaining all that behavior still leaves open a further question. Why does all that give you subjective experience? Explain the reacting, the responding, the controlling, the reporting, and so on.

7:43

It still leaves open the question, why is all that accompanied by subjective experience. Why doesn't it go on in the dark without consciousness, so to speak? There seems to be what the philosopher Joe Levine has called a gap here, an explanatory gap, between physical processes and subjective experience. At least our standard kinds of explanation, which work really well for the easy problems of behavior and so on, don't obviously give you a connection to the subjective aspects of experience. And there's been a vast amount of discussion of these things over-– I mean, well, for centuries, really.

「解释鸿沟」（explanatory gap）由哲学家约瑟夫·莱文（Joseph Levine）于1983年提出，指物理机制描述与主观体验之间在解释上的断裂，是难问题的另一种表述。

8:20

But it's been particularly active in recent decades, philosophers, scientists, all kinds of different views. Philosophically, you can divide approaches to the hard problem into at least two classes. One is an approach on which consciousness is taken to be somehow irreducible and primitive. We can't explain it in more basic physical terms, so we take it as a kind of primitive. And that might lead to dualist theories of consciousness where consciousness is somehow separate from and interacts with the brain.

非还原进路：把意识当作不可再分析的「初始项」，可导向二元论——意识独立于大脑并与之交互，属笛卡尔传统的当代变体。

irreducible /ˌɪrɪˈduːsəbəl/ *adj.*
不可还原的，不可简化的（哲学术语）

dualist /ˈduəlɪst/ *adj./n.*
二元论的；二元论者（主张心灵与物质分立）

8:49

Recently very popular has been the class of panpsychist theories of consciousness. I know Galen Strawson was here a while back talking. He very much favors panpsychist theories where consciousness is something basic in the universe underlying matter. And indeed, there are idealist theories where consciousness underlies the whole universe. So these are all extremely speculative but interesting views that I've explored myself. There are also a reductionist theories of consciousness from functionalist approaches, where consciousness is just basically taken to be a giant algorithm or computation, biological approaches to consciousness-- my colleague Ned Block was here, I know, talking about neurobiology-based approaches, where it's not the algorithm that matters, but the biology it's implemented in-- and indeed, the kind of quantum approaches that people like Roger Penrose and Stuart Hameroff have made famous.

一段浓缩的理论地图：泛心论（意识是物质底层的基本属性，Galen Strawson主张）、唯心论、功能主义（意识即算法）、生物学进路（Ned Block）、量子进路（彭罗斯与Hameroff的理论）。

panpsychist /pænˈsaɪkɪst/ *adj./n.*
泛心论的；泛心论者（认为意识是万物基本属性）

idealist /aɪˈdiəlɪst/ *adj./n.*
唯心论的；唯心论者

speculative /ˈspɛkjələtɪv/ *adj.*
思辨性的，推测性的

reductionist /rɪˈdʌkʃənɪst/ *adj./n.*
还原论的；还原论者（主张高层现象可还原为基础过程）

04 元问题：一条新的研究进路

9:44

I think there's interesting things to say about all of these approaches. I think that right now, at least, most of the reductionist approaches leave a gap. But the non-reductionist approaches have other problems in seeing how it all works. Today, I'm going to take a different kind of approach, this approach through the meta-problem. One way to motivate this is to-- I often get asked, well, you're a philosopher. It's fine. You get to think about these things like the hard problem of consciousness.

10:15

How can I, as a scientist or an engineer or an AI researcher-- how can I do something to contribute, to help get this at this hard problem of consciousness? Is this just a problem for philosophy? For me to work on it as a AI researcher, I need something I can operationalize, something I can work with and try to program. And as it stands, it's just not clear how to do that with the hard problem. If you're a neuroscientist, there are some things you can do. You can work with humans and look at their brains and look for the neural correlates of consciousness, the bits of the brain that go along with being conscious.

「意识的神经相关物」（NCC）是1990年代由克里克与科赫推动的神经科学主流纲领：寻找与意识状态伴随出现的脑活动。但它预设了被试有意识，且回避「为什么」的问题。

operationalize /ˌɑpəˈreɪʃənəlaɪz/ *v.*
使可操作化，把抽象概念转为可测量的指标

neural correlates of consciousness *phr.*
意识的神经相关物（NCC，与意识伴随出现的脑活动）

10:55

Because at least with humans, we can take as a plausible background assumption that the system is conscious. For AI, we can't even do that. We don't know which AI systems we're working with that are conscious. We need some operational criteria. In AI, we mostly work on modeling things like behavior and objective functioning. For consciousness, those are the easy problems. So how does someone coming from this perspective make a connection to the hard problem of consciousness? Well, one approach is to work on certain problems among the easy problems of behavior that shed particular light on the hard problem.

此处点出AI研究者的困境：AI擅长建模行为与客观功能（即简单问题），却没有判定系统是否有意识的操作性标准——这正是元问题进路要提供的切入口。

11:31

And that's going to be the approach that I look at today. So the key idea here is there are certain behavioral functions that seem to have a particularly close relation to the hard problem of consciousness. In particular, we say things about consciousness. We make what philosophers call phenomenal reports, verbal reports of conscious experiences. So I'll say things like, I'm conscious, I'm feeling pain right now, and so on. Maybe the consciousness and the pain are subjective experiences. But the reports, the utterances, I am conscious, well that's a bit of behavior.

「现象报告」（phenomenal reports）指关于意识体验的言语报告，如「我现在感到疼」。关键一步：报告本身是行为，属于原则上可用机制解释的简单问题。

utterances /ˈʌtərənsəz/ *n.*
话语，言语表达（utterance）

12:14

In principle, explaining those is among the easy problems. It's objectively measurable response. We can find a mechanism in the brain that produces it. And among our phenomenal reports, there's the special class we can call the problem reports, reports expressing our sense that consciousness poses a problem. Now admittedly, not everyone makes these reports. But they seem to be fairly widespread, especially among philosophers and scientists thinking about these things. But furthermore, it's a sense that it's fairly easy to find a very wide class of people who think about consciousness.

「问题报告」(problem reports)是现象报告的子类：表达「意识构成难题」这种感受的言论。元问题要解释的正是这类报告从何而来。

12:53

People say things like, there is a problem of consciousness, a hard problem. On the face of it, explaining behavior doesn't explain consciousness. Consciousness seems non-physical. How would you ever explain the subjective experience of red and so on? It's an objective fact about us-- at least about some of us-- that we make those reports. And that's a fact about human behavior. So the meta-problem of consciousness then, at a second approximation, is roughly the problem of explaining these problem reports, explaining, you might say, the conviction that we're conscious and that consciousness is puzzling.

元问题的初步定义：解释我们为何做出问题报告、为何坚信自己有意识且意识令人困惑。注意它本身不预设意识是否真实存在。

conviction /kənˈvɪkʃən/ *n.*
坚定的信念，确信

13:34

And what's nice about this is that although the hard problem is this airy fairy problem about subjective experience that's hard to pin down, this is a puzzle ultimately about behavior. So this is an easy problem, one that ought to be open to those standard methods of explanation in the cognitive and brain sciences. So there's a research program. There's a research program here. So I like to think of the meta-problem as something we could play that role. I talked about earlier, if you're an AI researcher thinking about this, the meta-problem is an easy problem, a problem about behavior, that's closely tied to the hard problem.

核心策略转换：难问题关于飘渺难抓的主观体验，元问题却关于可测量的行为——因此是「简单问题」，可以用认知科学与脑科学的标准方法研究。这是整场讲座的枢纽论点。

airy fairy /ˌɛri ˈfɛri/ *adj.*
（英式口语）虚无缥缈的，不切实际的

pin down *phr. v.*
确切界定，把……说清楚

14:09

So it's something we might be able to make some progress on using standard methods of thinking about algorithms and computations or thinking about brain processes and behavior while still shedding some light, at least indirectly, on the hard problem. It's more tractable than the hard problem. But solving it ought to shed light on the hard problem. And today, I'm just going to kind of lay out the research program and talk about some ways in which it might potentially shed some light. This is interesting to a philosopher because it looks like an instance of what people sometimes call genealogical analysis.

「谱系学分析」（genealogical analysis）：不直接问某物是否存在，而是追问我们关于它的信念从何而来。这为下文引出「祛魅论证」与幻觉主义做铺垫。

tractable /ˈtræktəbəl/ *adj.*
易处理的，可解决的（反义：intractable）

genealogical /ˌdʒiniəˈlɑdʒɪkəl/ *adj.*
谱系学的，追溯起源的

05 谱系学分析与幻觉主义

14:41

It goes back to Friedrich Nietzsche on the genealogy of morals. Instead of thinking about what's good or bad, let's look at where our sense of good or bad came from, the genealogy of it all in evolution or in culture or in religion. And people think a genealogical approach to God, instead of thinking about does God exist or not, let's look at where our belief in God came from. Maybe there's some evolutionary reason for why people believe in God. This often leads, not always, but often leads to a kind of debunking of our beliefs about those domains.

尼采1887年《论道德的谱系》开创此方法：考察善恶观念在进化、文化、宗教中的起源。对上帝信念的进化论解释是当代最常用的类比案例。

debunking /diˈbʌŋkɪŋ/ *n./v.*
揭穿，祛魅（证明信念缺乏根据）

15:14

Explain why we believe in God in evolutionary terms, no need for the God hypothesis anymore. Explain how moral beliefs and evolutionary terms, maybe no need to take morality quite so seriously. So some people, at least, are inclined to take an approach like this with consciousness too. If you think about the meta-problem explaining our beliefs about consciousness, that might ultimately debunk our beliefs about consciousness. This leads to a philosophical view, which has recently attracted a lot of interest, a philosophical view called illusionism, which is the view that consciousness itself is an illusion.

「祛魅论证」（debunking argument）：若一种信念可以完全用与其对象无关的机制解释，该信念的合理性就被削弱。幻觉主义正是把这一逻辑应用于意识。

15:55

Or maybe that the problem of consciousness is an illusion. Explain the illusion, and we dissolve the problem. I take that in terms of the meta-problem, that view roughly comes to solve the meta-problem. It will dissolve the hard problem. Explain why it is that we say all these things about consciousness, why we say, I am conscious, why we say, consciousness is puzzling. If you can explain all that in algorithmic terms, then you'll remove the underlying problem because you'll have explained why we're puzzled in the first place.

幻觉主义（illusionism）的核心主张：解决元问题即「消解」（dissolve）难问题——解释清楚我们为何困惑，困惑背后就不再有真问题剩下。

dissolve /dɪˈzɑlv/ *v.*
消解（问题）；使消失。注意与 solve（解决）对举

16:29

Actually, walking over here today, I noticed that just a couple of blocks away, we have the Museum of Illusions, so I'm going to check that out later on. But if illusionism is right, added to all those perceptual illusions is going to be the problem of consciousness itself. It's roughly an illusion thrown up by having a weird kind of self model with a certain kind of algorithm that attributes to ourselves special properties that we don't have. So one line on the meta-problem is the illusionist line.

16:59

Solve the meta-problem, you'll get to treat consciousness as an illusion. That's actually a view that has many antecedents in the history of philosophy, one way or another. Even Immanuel Kant and his great critique of pure reason had a section where he talked about the self or the soul as a transcendental illusion. We seem to have this indivisible soul. But that's the kind of illusion thrown out by our cognitive processes. The Australian philosophers, Ullin Place and David Armstrong, had versions of this that I might touch on a bit later.

康德在《纯粹理性批判》中论证：不可分的灵魂是认知过程投射出的「先验幻相」。普莱斯（Place）与阿姆斯特朗（Armstrong）是1950—60年代澳大利亚唯物主义（心脑同一论）代表。

antecedents /ˌæntəˈsidənts/ *n.*
先例，前身，思想渊源

transcendental /ˌtrænsɛnˈdɛntəl/ *adj.*
先验的（康德哲学术语）

indivisible /ˌɪndəˈvɪzəbəl/ *adj.*
不可分割的

17:32

Daniel Dennett, a leading reductionist thinker about consciousness has been pushing for the last couple of decades the idea that consciousness involves a certain kind of user illusion. And most recently, the British philosopher, Keith Frankish, has been really pushing illusionism as a theory of consciousness. He has a book centering around a paper by Keith Frankish on illusionism as a theory of consciousness that I recommend to you. So one way to go with the meta-problem is the direction of illusionism.

丹尼尔·丹尼特（Daniel Dennett）借用计算机「用户界面幻觉」比喻意识；基思·弗兰基什（Keith Frankish）2017年主编《幻觉主义》文集，使该词成为正式理论标签。

18:05

But one nice thing about-- many people find illusionism completely unbelievable. They find, how could it be that consciousness is an illusion? Look, we just have these subjective experiences. It's a data about our nature. And I confess, I've got some sympathy with that reaction. So I'm not an illusionist myself. I'm a realist about consciousness in the philosopher's sense, where a realist about something is someone who believes that thing is real. I think consciousness is real. I think it's not an illusion.

18:33

I think that solving the meta-problem does not dissolve the hard problem. But the nice thing about the meta-problem is you can proceed on it-- to some extent, at least in initial neutrality-- on that question, is consciousness real or is it an illusion. It's a basic problem about our objective functioning in these reports. What explains those? There's a neutral research program here that both realists, illusionists, people of all kinds of different views of consciousness can explain. And then we can come back and look at the philosophical consequences.

方法论卖点：元问题研究纲领在「意识是否真实」上保持中立——实在论者与幻觉主义者可以共享同一实证纲领，做完研究再回头争论哲学后果。

19:06

So I'm not an illusionist. I think consciousness is real. I've got to say, I do feel the temptation of illusionism. I find it really intriguing and in some ways attractive view. It's just fundamentally unbelievable. Nevertheless, I think that the meta-problem should be a tractable problem. Solving it, at the very least, will shed much light on the hard problem of consciousness even if it doesn't solve it. If you can explain our conviction that we're conscious, somehow the source, the roots of our conviction that we are conscious, must have something to do with consciousness especially if consciousness is real.

值得注意的坦诚：作为难问题提出者，查尔默斯承认幻觉主义「迷人且有吸引力」，只是「根本无法让人相信」。这种对对手立场的同情理解贯穿全场。

19:42

So I think it's very much a good research program for people to explain. So then I'll move on now to just outlining the research program a little bit more and then talk a bit about potential solutions and on impact on theories of consciousness before wrapping up with a little bit more about illusionism. So this meta-problem, which I've been pushing recently, opens up a tractable empirical research program for everyone, reductionists, non-reductionists, illusionists, non-illusionists. We can try to solve it and then think about the philosophical consequences.

06 问题直觉的类型与实证研究

20:22

Now what is the meta-problem? Well, the way I'm going to put it is it's the problem of topic-neutrally explaining problem intuitions or else explaining why that can't be done. And I'll unpack all the pieces of that right now. First, starting with problem intuitions. What are problem intuitions? Well, there are the things we say. There are things we think I say. Consciousness seems irreducible. I might think consciousness is irreducible. People might be disposed, have a tendency to say or think those things.

元问题的精确表述：「以论题中立的方式解释问题直觉，或解释为何做不到」。「论题中立」指解释中不引用意识本身，例如一个纯算法层面的解释。

disposed /dɪˈspoʊzd/ *adj.*
有……倾向的（be disposed to do）

20:55

Problem intuitions all take to be roughly, that tendency. We have dispositions to say and think certain things about consciousness. What are the core problem intuitions? Well, I think they break down into a number of different kinds. There is the intuition that consciousness is non-physical. We might think of that as a metaphysical intuition about the nature of consciousness. There are intuitions about explanation. Consciousness is hard to explain, explaining behavior doesn't explain consciousness.

dispositions /ˌdɪspəˈzɪʃənz/ *n.*
倾向，性向（disposition）

metaphysical /ˌmɛtəˈfɪzɪkəl/ *adj.*
形而上学的，关于存在本性的

21:25

There are intuitions about knowledge of consciousness. Some of you may know the famous thought experiment of Mary in the black and white room who knows all about the objective nature of color vision and so on, but still doesn't know what it's like to see red. She sees red for the first time. She learns something new. That's an intuition about knowledge of consciousness. There are what philosophers call modal intuitions about what's possible or imaginable. One famous case is the case of a zombie, a creature who is physically identical to you and me but not conscious.

「玛丽房间」思想实验由弗兰克·杰克逊1982年提出：黑白房间中通晓颜色科学全部物理事实的玛丽，初见红色时仍学到新东西——用以论证物理知识不穷尽现象知识。

modal /ˈmoʊdəl/ *adj.*
模态的（哲学中关于可能性与必然性的）

21:56

Or maybe an AI system, which is functionally identical to you and me, but not conscious. That at least seems conceivable to many people. So this is the philosophical zombie. Unlike the zombies and movies, which have weird behaviors and go after brains and so on, the philosophical zombie is a creature that seems, at least behaviorally, may be physically like a normal human, but doesn't have any conscious experiences. All the physical states, none of the mental states. And it seems to many people that's at least conceivable.

哲学僵尸（philosophical zombie）：物理上与常人完全相同却毫无主观体验的假想存在，与影视僵尸无关。其「可设想性」是查尔默斯反物理主义论证的关键前提。

conceivable /kənˈsivəbəl/ *adj.*
可设想的，可想象的

22:25

We're not zombies. I don't think anyone here is a zombie-- I hope. But nonetheless, it seems that we can make sense of the idea. And one way to pose the hard problem is, why are we not zombies. So this imagined ability of zombies is one of the intuitions that gets the problem going. And then you can go on and catalog more and more intuitions about the distribution of conscious, maybe the intuition that robots won't be conscious. That's an optional one, I think. Or consciousness matters morally in certain ways, and the list goes on.

22:57

So I think there is an interdisciplinary research program here of working on those intuitions about consciousness and trying to explain them. Experimental psychology and experimental philosophy-- a newly active area-- can study people's intuitions about consciousness. We can work on models of these things, computational models or neurobiological models, of these intuitions and reports. And indeed, I think there's a lot of room for philosophical analysis. And there's just starting to be a program of people doing these things in all these fields.

实验哲学（experimental philosophy）是2000年代兴起的领域，用问卷与实验方法研究普通人的哲学直觉。此处被指定为元问题纲领的实证工具之一。

interdisciplinary /ˌɪntəˈdɪsəplənɛri/ *adj.*
跨学科的

23:28

I mean, it is an empirical question, how widely these intuitions are shared. You might be sitting there thinking, come on, I don't have of these intuitions. Maybe this is just you. My sense is-- from the psychological study to date-- it seems that some of these intuitions about consciousness are at least very widely shared, at least as dispositions or intuitions, although they are often overridden on reflection. But the current data on this is somewhat limited. Although there is a lot of empirical work on intuitions about the mind concerning things like belief, like when do kids get the idea that your beliefs about the world can be false, concerning the way your self persists through time-- could you exist after the death of your body-- where consciousness is concerned, there's work on the distribution of consciousness.

实证现状：现有研究表明这些直觉分布相当广泛，但常在反思后被推翻。儿童心智理论（如错误信念任务）与自我跨时间同一性是相邻的成熟研究领域。

overridden /ˌoʊvəˈrɪdən/ *v.*
被推翻，被覆盖（override 的过去分词）

24:12

Could a robot be conscious? Could a group be conscious? Here's a book by Paul Bloom, "Decartes' Baby" that catalogs a lot of this interesting work, making the case that many children are intuitive dualists. Thinks they're naturally inclined to think there's something non-physical about the mind. So far, most of this work has not been so much on these core problem intuitions about consciousness, but there's work developing in this direction. Sara Gottlieb and Tania Lombrozo have a very recent article called "Can Science Explain The Human Mind" on people's judgments about when various mental phenomena are hard to explain.

保罗·布鲁姆（Paul Bloom），耶鲁心理学家，《笛卡尔的婴儿》（2004）主张儿童是「天生的二元论者」，天然倾向认为心灵中有非物理的东西。

inclined /ɪnˈklaɪnd/ *adj.*
倾向于……的（be inclined to think）

24:48

And they seem to find that yes, subjective experience and things to which people have privileged first person access seem to pose the problem big time. So there's the beginning of a research program here. I think there's room for a lot more. The topic neutrality part-- when I say we're looking for a topic neutral explanation of problem intuitions, that's roughly to say an explanation that doesn't mention consciousness itself. It's put in neutral terms. It's neutral on the existence of consciousness.

Gottlieb与Lombrozo的研究发现：人们恰恰认为具有「第一人称特权通达」的主观体验最难被科学解释——为问题直觉的普遍性提供了初步实证证据。

privileged /ˈprɪvəlɪdʒd/ *adj.*
特许的，独享的；privileged access 特权通达（哲学术语）

big time *adv. (口语)*
非常，严重地（pose the problem big time 问题极其突出）

25:18

The most obvious one would be something like an algorithmic explanation. Now here is the algorithm the brain is executing that generates our conviction that we're conscious and our reports about consciousness. There may be some time between an algorithm and consciousness, but to specify the algorithm, you don't need to make claims about consciousness. So the algorithmic version of the meta-problem is roughly find the algorithm that generates our problem intuition. So that's, I think, in principle a research program that maybe an AI researchers in combination with psychologists-- the psychologist could help isolate data about the way that the human beings are doing it, how these things are generated in humans.

元问题的算法版本：找出大脑生成「我有意识」这一确信的算法。分工设想：心理学家提供人类如何产生直觉的数据，AI研究者尝试在机器中实现该算法。

07 候选解答：自我模型与内省

26:01

And the AI researcher can try and see about implementing that algorithm in machines and see what results. And I'll talk about a little bit of research in this direction in just a moment. OK now I want to say something about potential solutions to the problem. Like I said, this is a big research program. I don't claim to have the solution to the meta-problem. I've got some ideas, but I'm not going to try and lay out a major solution. So here are a few things, which I think might be part of a solution to the problem, many of which have got antecedents here and there in scientific and philosophical discussion.

26:38

Some promising ideas include retrospective models, phenomenal concepts, introspective opacity, the sense of acquaintance. Let me just say something about a few of these. One starting idea that almost anyone is going to have here is somehow models of ourselves are playing a central role here. Human beings have models of the world, naive physics, naive psychology, models of other people, and so on. We also have models of ourselves. It makes sense for us to have models of ourselves and our own mental processes.

introspective /ˌɪntrəˈspɛktɪv/
adj.
内省的，反观自身心理的

acquaintance /əˈkweɪntəns/ *n.*
亲知，直接知晓（哲学术语）；相识

27:11

This is something that the psychologist Michael Graziano has written a lot on. We have internal models of our own cognitive processes, including those tied to consciousness. And somehow something about our introspective models explains our sense, A, that we are conscious and B, that this is distinctively problematic. And I think anyone thinking about the meta-problem, this has got to be at least the first step. We have these introspective models. If you were an illusionist, they'll be false models.

迈克尔·格拉齐亚诺（Michael Graziano），普林斯顿神经科学家，其「注意图式理论」主张大脑为自身注意过程建立简化模型，正是该模型让我们自认拥有主观体验。

27:41

If you're a realist, they needn't be false models. But at the very least, these introspective models are involved, which is fine. But the devil's in the details. How do they work to generate this problem? A number of philosophers have argued we have special concepts of consciousness, introspective concepts of these special subjective states. People call these phenomenal concepts, concepts of phenomenal consciousness. And one thing that's special is these concepts are somehow independent of our physical concepts.

「现象概念策略」是物理主义阵营的重要方案：我们有两套彼此独立的概念系统——建模外部世界的物理概念与建模自身心灵的内省概念，概念上的独立被误认为存在上的独立。

the devil's in the details
idiom
魔鬼藏在细节里；难点全在细节

28:13

They explain we've got one set of physical concepts for modeling the external world. We've got one set of introspective concepts from modeling our own mind. And these concepts, just by virtue of the way they're designed, are somewhat independent of each other. And that partly explains why consciousness seems to be independent of the physical world intuitively. So maybe that independence of phenomenal concepts could go some distance to explaining our problem reports. So I think there's got to be something to this as well.

by virtue of *phr.*
凭借，由于

28:44

At the same time, I don't this goes nearly far enough because we have concepts of many aspects of the mind, not just of the subjective experiential past but things we believe and things we desire. And so when I believe that Paris is the capital of France, that's part of my internal self model. But that doesn't seem to generate the hard problem in nearly the same way in which the experience of red does. So a lot more needs to be said about what's going on in cases like having the experience of red and having the sense that generates a gap.

关键反驳：我们同样有关于信念、欲望的自我模型，但「相信巴黎是法国首都」不产生难问题，「看见红色」却产生——说明现象概念策略的解释力不够。

29:16

So it doesn't generalize to everything about the mind. Some people have thought that what we might call introspective opacity plays a role, that when we introspect what's going on in our minds, we don't have access to the underlying physical states. We don't see the neurons in our brains. We don't see that consciousness is physical. So we see it as non–physical. Most recently, the physicist Max Tegmark has argued in this direction, saying somehow consciousness is substrate–independent. We don't see the substrate.

内省不透明性：内省看不到底层神经机制，于是把意识「看成」非物理的。物理学家马克斯·泰格马克的「基质独立性」说法与此呼应。

opacity /oʊˈpæsəti/ *n.*
不透明性（引申指无法看清内部机制）

substrate /ˈsʌbstreɪt/ *n.*
基质，底层载体（如大脑或硅芯片）

29:47

So then we think maybe it can float free of the substrate. Armstrong made an analogy with the case of someone in a circus where–– the headless person illusion where someone's there with a veil across their head, and you don't see their head. So you see them as having no head. Here is a 19th century booth at a circus, so–called headless woman. There's a veil over her head. You don't see the head so somehow, it looks–– at least for a moment–– like the person doesn't have a head. So Armstrong says maybe that's how it is with consciousness.

阿姆斯特朗的「无头女人」类比：马戏团用幕布遮住头部，观众便「看到一个没有头的人」——把「没看到物理性」误当作「看到了非物理性」，是同一种推理错误。

veil /veɪl/ *n.*
面纱，遮盖物

30:23

You don't see it as physical, so you see it as non–physical. But still the question comes up, how do we make this inference. There's something that's special goes on in cases like color and taste and so on. The color experience seems to attribute primitive properties to objects like redness, greenness, and so on, when, in fact, in the external world at the very least, they have complex reducible properties. Somehow, our internal models of color treat colors like red and green as if they are primitive things.

更深一层：颜色体验把红、绿表征为世界中的「原始简单属性」，而它们实际上是复杂可还原的物理属性。我们的知觉模型系统性地把对象「原始化」。

inference /ˈɪnfərəns/ *n.*
推论，推断

30:56

It turns out to be useful to have these models of things. We treat certain things as primitive, even though they're reducible. And it sure seems that when we experience colors, we experience greenness as a primitive quality even though it may be a very, very complex reducible property. That's something about our model of colors. The philosopher Wolfgang Schwartz tried to make an analogy with sensor variables in image processing. You've got some visual senses and a camera or something you need to process the image.

沃尔夫冈·施瓦茨的类比：图像处理中把传感器读数当作原始维度最为高效，无需表征光子层面的细节——内省或许同样把意识状态当作「传感器变量」来处理。

reducible /rɪˈdusəbəl/ *adj.*
可还原的，可简化为更基本成分的

31:27

Well, you've got some sensor variables to represent the sensory inputs that the various sensors are getting. And you might treat them as a primitive dimension because that's the most useful way to treat them. You don't treat them as certain amounts of lights or photons firing. You don't need to know about that. You use these sensor variables and treat them as a primitive dimension. And all that will play into a model of these things as primitive, maybe taking that idea and extending it to introspection.

31:57

These conscious states are somehow like sensor variables in our model of the mind. And somehow, these internal models give us the sense of being acquainted with primitive concrete qualities and of our awareness of them. This is still just laying out. I don't think this is still yet actually explaining a whole lot. But it's laying out–– it's narrowing down what it is that we need to explain to solve the meta–problem. But just to put the pieces together, here's a little summary. One thing I like about this summary is you can read it in either an illusionist tone of voice, as an account of the illusion of consciousness–– so this is how false introspective models work–– or in a realist tone of voice, as an account of our true correct models of consciousness.

32:45

But we can set it out in a way which is neutral on the two and then try and figure out later whether these models are correct, as the realist says, or incorrect, as the illusionist says. We have introspective models deploying introspective concepts of our internal states that are largely independent of our physical concepts. These concepts are introspectively opaque, not revealing any of the underlying mechanisms. Our perceptual models perceptually attribute primitive perceptual qualities to the world.

deploying /dɪˈplɔɪŋ/ *v.*
运用，部署（deploy）

33:17

And our introspective models attribute primitive mental relations to those qualities. These models produce the sense of acquaintance, both with those qualities and with our awareness of those qualities. Like I said, this is not a solution to the meta-problem, but it's trying, at least, to pin down some parts of the roots of those intuitions and to narrow down what needs to be explained. To go further, you want, I think, to test these explanations, both with psychological studies to see if this is plausibly what's going on in humans-- this is the kind of thing which is the basis of our intuitions-- and computational models to see if, for example, we could program this kind of thing into an AI system and see if it can generate somehow qualitatively similar reports and intuitions.

检验路径有二：心理学研究验证这是否为人类直觉的真实来源；计算建模验证能否让AI系统产生类似的报告与直觉——呼应开头对AI研究者的邀请。

qualitatively /ˈkwɑlətetɪvli/ *adv.*
在性质上，定性地

08 AI 建模与人工意识测试

34:06

You might think that last thing is a bit far fetched right now, but I know of at least one instance of this research program, which has been put into play by Luke Muehlhauser and [INAUDIBLE] two researchers at Open Philanthropy very interested in AI and consciousness. They actually built-- they took some ideas about the meta-problem from something I'd written about it and from something that the philosopher Francois Kammerer had written about it. A couple of basic ideas about where problem intuitions might come from.

Open Philanthropy是资助有效利他主义方向的机构，Luke Muehlhauser长期研究AI与意识的道德地位问题，2017年发表过关于意识与道德受体的长篇报告。

far fetched /ˌfɑr ˈfɛtʃt/ *adj.*
牵强的，不太可信的

34:40

And they tried to build them into a computational model. They built a little software agent, which had certain axioms about colors and how they work. There's the red and there's green and certain axioms about their own subjective experiences of colors. And then they combined it with a little theorem prover. And they saw what did this little software agent come up with. And it came up with claims like, hey, well, my experiences of color are distinct from any physical state, and so on. OK they cut a few corners.

这一小实验的要点：给软件智能体一些关于颜色及自身体验的公理，加上定理证明器，它自己推出「我的颜色体验不同于任何物理状态」——问题直觉可被算法生成的初步演示。

axioms /ˈæksiəmz/ *n.*
公理（axiom）

theorem prover *phr.*
定理证明器（自动从公理推导结论的程序）

cut a few corners *idiom*
（cut corners）走捷径，省略应做的步骤

35:17

This is not a yet truly a convincing sophisticated model of everything going on in the human mind. But it shows that there's a research program here of trying to find the algorithmic basis of these states. And I think as more sophisticated models develop, we might be able to use these to kind of provide a way in for AI researchers in thinking about this topic. Of course, there is the question, you model all this stuff better and better in a machine, then is the machine actually going to be conscious or is it just going to have found self models that replicate what's going on in humans.

35:55

So some people have proposed an artificial consciousness test. Aaron Sloman, Susan Schneider, Ed Turner have suggested somehow that if a machine seems to be puzzled about consciousness in roughly the ways that we are, maybe that's actually a sign that it's conscious. So if a machine actually looks to us as if it's puzzled by consciousness, is that a sign of consciousness? These people-- this is suggested as a kind of Turing test for machine consciousness. Find machines which are conscious like we are.

「人工意识测试」由苏珊·施奈德（Susan Schneider）与天体物理学家Edwin Turner提出（称ACT）：若机器自发地像人类一样对意识感到困惑，可视为其有意识的证据——一种针对意识的图灵测试。

36:27

Of course, the opposing point of view is going to be no, the machine is not actually conscious. It's just like machine that studied up for the Turing test by reading the talk like a human book. It's like, damn, do I really need to convince those humans that I'm conscious by replicating all those ill–conceived confusions about consciousness. Well I guess I can do it if I need to. Anyway, I'm not going to settle this question here. But I do think that if we somehow find machines being puzzled, it won't surprise me that once we actually have serious AI systems, which engagement in natural language and modeling of themselves and the world, they might well find themselves saying things like, yeah, I know in principle I'm just a set of silicon circuits, but I feel like so much more.

反方立场：机器可能只是「背熟了像人一样说话的教材」，复述人类关于意识的困惑而无真实体验——该测试面临与图灵测试相同的作弊质疑。查尔默斯预测大语言模型式的系统很可能自发说出这类话。

ill–conceived /ˌɪl kənˈsivd/ *adj.*
构想拙劣的，考虑不周的

09 用元问题检验意识理论

37:15

I think that might tell us something about consciousness. Let me just say a little bit about theories of consciousness. I do think a solution to the meta–problem and a solution to the hard problem ought to be closely connected. The illusionist has solved the meta–problem. You'll dissolve the hard problem. But even if you're not an illusionist about consciousness, there ought to be some link. So here's a thesis. Whatever explains consciousness should also partly explain our judgments, now reports about consciousness.

讲座后半的关键论题：无论什么解释了意识，它也应当部分解释我们关于意识的判断与报告——否则意识的基础与我们谈论意识的行为将诡异地互不相干。这成为检验意识理论的新标准。

37:47

The rationale here is it would just be very strange if these things were independent, if the basis of consciousness played no role in our judgments about consciousness. So they can use this as a way of evaluating or testing theories of consciousness. For theory of consciousness says mechanism M is the basis of consciousness, that M should also partly explain our judgments about consciousness. Whatever the basis is ought to explain the reports. And you can use this. You can bring this to bear on various extant theories of consciousness.

rationale /ˌræʃəˈnæl/ *n.*
理据，根本原因

extant /ˈekstənt/ *adj.*
现存的，尚存的（书面语）

bring this to bear *phr.*
(bring sth to bear on)
把……运用于，施加于

38:24

Here's one famous current theory of consciousness, integrated information theory developed by Giulio Tononi and colleagues at the University of Wisconsin. Tononi says the basis of consciousness is integrated information, a certain kind of integration of information for which to and he has a measure that he calls phi. Basically, when your phi is high enough, you get consciousness. A consciousness is high phi, and there's a mathematical definition. But I won't go into it here. But it's a really interesting theory.

整合信息论（IIT）由威斯康星大学的朱利奥·托诺尼（Giulio Tononi）提出，用希腊字母phi（Φ）量化系统的信息整合度，主张phi足够高即有意识，是当前最有影响的意识理论之一。

39:00

So here's a-- basically it analyzes a network property of systems of units. And it's got a informational measure called phi that's supposed to go with consciousness. Question, if integrated information is the basis of consciousness, It ought to explain problem reports, at least in principle. Challenge, how does that work? And it's at least far from obvious to me how integrated information will explain the problem reports. It seems pretty dissociated from them. On Tononi's view, you can have simulations of systems with high phi that have zero phi.

对IIT的元问题挑战：按托诺尼的观点，高phi系统的计算机模拟可以phi为零，却做出完全相同的报告——意识的基础与问题报告在该理论中相互脱节。

dissociated /dɪˈsoʊsieɪtəd/ *adj.*
相分离的，脱节的

39:36

They'll go about making exactly the same reports but without consciousness at all. So phi is at least somewhat dissociable. You get systems with very high phi, but no tendency to report. Maybe that's less worrying. Anyway, here's a challenge for this theory, for other theories. Explain, not just how high phi gives you consciousness, but how it plays a central role in the algorithms that generate problem reports. Something similar goes for many other theories, biological theories, quantum theories, global workspace, and so on.

dissociable /dɪˈsoʊsiəbəl/ *adj.*
可分离的，可拆开的

10 幻觉主义正反方的僵局

40:08

But let me just wrap up by saying something about the issue of illusionism that I was talking about near the start. Again, you might be inclined to think that this approach through the meta-problem tends, at least very naturally, to lead to illusionism. And I think it can be-- it certainly provides, I think, some motivation for illusionism, the view that consciousness doesn't exist, we just think it does. On this view, again, a solution to the meta-problem dissolves the hard problem. So here's one way of putting the case for illusion.

幻觉主义论证的完整形式：若元问题有不提及意识的算法解释，则该解释无需意识也能成立，从而像进化论解释祛魅上帝信念一样，祛魅我们关于意识的信念。

40:41

If there is a solution to the meta-problem, then there is an explanation of our beliefs about consciousness that's independent of consciousness. There's an algorithm that explains our beliefs about consciousness. It doesn't mention consciousness. Arguably, it could be in place without consciousness. Arguably, that kind of explanation could debunk our beliefs about consciousness the same way that perhaps explaining beliefs about God in evolutionary terms might debunk belief in God. It certainly doesn't prove that God doesn't exist.

Arguably /ˈɑrgjuəbli/ *adv.*
可以说，按理说（表有论据支持的判断）

41:12

You might think that if you can explain our beliefs in terms of evolution, it somehow removes the justification or the rational basis for those beliefs. So something like that, I think, can be applied to consciousness too. And there's a lot to be said about analyzing the extent to which this might debunk the beliefs. On the other hand, the case against illusionism is very, very strong for many people. And the underlying worry is that some of the illusionism is completely unbelievable. It's just a manifest fact about ourselves that we have conscious experience, we experience red, we feel pain, and so on.

反幻觉主义的根本理由：我们有意识体验是关于自身的「明白事实」，否认它等于否认数据本身——呼应第9段「主观体验是数据」的立场。文中dielectric为dialectic（辩证）的转写错误。

manifest /ˈmænəfɛst/ *adj.*
显而易见的（a manifest fact 明白的事实）

41:48

To deny those things is to deny the data. No, the dielectric here is complicated. The illusionist will come back and say, yes, but I can explain why illusionism is unbelievable. These models we have, these self models of consciousness, are so strong that they were just wired into us by evolution. They're not models we can get rid of. So my view predicts that my view is unbelievable. And the question is, what-- the dialectical situation is complex and interesting. But maybe I could just wrap up with two expressions of absurdity on either side of this question, the illusionist and the anti-illusionist, both finding absurdity in the other person's views.

精彩的辩证转折：幻觉主义者回应说「我的理论恰恰预测了它自己不可信」——意识自我模型被进化深深固化、无法摆脱，因此理论的不可信反而成了它的证据。

wired into *phr.*
（被）内置于，天生固化于（常指进化或本能）

dialectical /ˌdaɪəˈlɛktɪkəl/ *adj.*
辩证的；论辩攻防上的（dialectical situation 论辩局势）

42:31

Here's Galen Strawson, who was here. Galen's view is very much that illusionism is totally absurd. In fact, he thinks it's the most absurd view that anyone has ever held. There occurred in the 20th century, the most remarkable episode in the whole history of ideas, the whole history of human thought, a number of thinkers denied the existence of something we know with certainty to exist, consciousness. He thinks this is just a sign of incredible philosophical pathology. Here's the rationalist philosopher, Eliezer Yudkowsky, and something he wrote a few years ago on zombies and consciousness and the epiphenomenalist view that consciousness plays no causal role, where he was engaging some stuff I wrote a couple of decades ago.

盖伦·斯特劳森（Galen Strawson），英国哲学家、泛心论者，称否认意识是「整个思想史上最不可思议的事件」——代表反幻觉主义一侧的极端表述。

pathology /pəˈθɑlədʒi/ *n.*
病理，病态（此处喻「哲学病」）

43:16

He said, "this zombie argument"-- the idea we can imagine zombies physically like us but without consciousness-- "may be a candidate for the most deranged idea in all of philosophy. The causally closed cognitive system of trauma's internal narrative is malfunctioning, in a way, not by necessity but just in our own universe miraculously happens to be correct." And here he is expressing this debunking idea that on this view, there's an algorithm that generates these intuitions about consciousness. And that's all physical.

埃利泽·尤德科夫斯基（Eliezer Yudkowsky），理性主义社区与机器智能研究所（MIRI）创始人，批评僵尸论证是「全部哲学中最疯狂的想法」：若意识不起因果作用，大脑算法竟恰好正确描述它，需要奇迹般的巧合。

deranged /dɪˈreɪndʒd/ *adj.*
疯狂的，精神错乱的

causally closed *phr.*
因果封闭的（物理主义术语：物理事件的原因都是物理的）

miraculously /məˈrækjələsli/ *adv.*
奇迹般地

43:47

And there's also this further layer of non-physical stuff. And just by massive coincidence, the physical algorithm is a correct model of the non-physical stuff. That's a form of debunking here. It would take a miracle for this view to be correct. So I think both of these views are onto-- these objections are onto something. And to make progress on this on either side, we need to find a way of getting past these absurdities. You might say, well, there's middle ground between very strong illusionism and very strong epiphenomenalism.

onto something *idiom*
（be onto something）发现了有价值的线索，说到了点子上

44:20

It tends to slide back to the same problems. Other forms of illusionism, weaker forms don't help much with the hard problem. Other forms of realism are still subject to this. It takes a miracle for this view to be correct critique. So I think to get beyond absurdity here, both sides need to do something more. The illusionist needs to do more to explain how having a mind could be like this, even though it's not at all the way that it seems. They need to find some way to recapture the data. Realists need to explain how it is that these meta-problem processes are not completely independent of consciousness.

查尔默斯给双方布置的作业：幻觉主义者须解释心灵为何与其显现完全不同，并「重新捕获数据」；实在论者须解释生成意识直觉的元问题过程如何植根于意识本身。

recapture /riˈkæptʃə/ *v.*
重新捕获，找回（此处指重新解释「数据」）

critique /krɪˈtik/ *n.*
批判，评论（比 criticism 更学术化）

45:01

Realists need to explain how meta-problem processes, the ones that generate these intuitions and reports and convictions about consciousness, are essentially grounded in consciousness even if it's possible somehow for them or conceivable for them to occur without consciousness. Anyway so that's just to lay out a research program. I think a solution to the meta-problem that meets these ambitions might just possibly solve the hard problem of consciousness or at the very least shed significant light on it.

grounded in *phr.*
植根于，以……为基础

11 问答：机器意识与哥德尔类比

45:34 观众

In the meantime, the meta-problem is a potentially tractable research project for everyone, and might I recommend to all of you. Thanks. [APPLAUSE] AUDIENCE: Yes, I just want to say I think it's very interesting, this concept of, we have these collection of mental models and that this collection of mental models is consciousness, basically. Consciousness defines the collection of these mental models that we have. And the problem with consciousness is that we don't understand the physical phenomenon that causes these mental models or that stimulates these mental models.

提问者的思路接近功能主义：若意识就是一组心智模型，机器人即使传感器和物理基础不同，也能实现同样的模型，从而产生「同样的意识效果」。

46:17

So we just have this belief that it's ephemeral or not real or something like that. And if you take that view, then what's interesting is that you could simulate these mental models like robot could simulate these mental models. And you could simulate consciousness as well. And even if the underlying physical phenomena that fuels these mental models is different-- robots have different sensors, et cetera-- you could still get the same consciousness effect in both cases. DAVID CHALMERS: Yeah, I think that's right.

ephemeral /ɪˈfɛmərəl/ *adj.*
短暂的，飘忽虚幻的

46:56 大卫·查尔默斯

Or at the very least, it looks like you ought to be able to get the same models, at least, in a robot. If the models themselves are something algorithmic and ought to be, you ought to be able to design a robot that has, at the very least, let's say, isomorphic models and some sense that is conscious. Of course, it's a further question-- at least by my [? lights--?] whether then the robot will be conscious. And that was the question I alluded to in talking about the artificial consciousness test.

isomorphic /ˌaɪsəˈmɔrfɪk/ *adj.*
同构的（数学术语，结构一一对应）

alluded to *phr. v.*
（allude to）间接提及，暗指

47:21

But you might think that would at least be very good evidence that the robot is conscious. If it's got a model of consciousness just like ours, it seems very plausible there ought to be a very strong link between having a model like that and being conscious. I think probably something like Ned Block-- who was here arguing against machine consciousness-- would say, no, no, the model is not enough. The model has to be built of the right stuff. So it's gotta be built of biology. And so on. But at by my [? light,?] I think if I had found an AI system that had a very serious version of our model of consciousness, I'd take that as a very good reason to believe it's conscious.

内德·布洛克（Ned Block），NYU哲学家，主张意识依赖生物基质而非算法结构——「模型还不够，得用对材料造」。查尔默斯本人则倾向把同构的意识模型当作机器有意识的有力证据。

47:55 观众

AUDIENCE: In the IIT theory, is there a estimate or plausible estimate for what the value of phi is for people and for other systems? DAVID CHALMERS: Basically, no. It's extremely hard to measure in systems of any size at all. Because the way it's defined, it involves taking a sum over every possible partition of a system. It turns out A, it's hard to measure in the brain because you've got to involve the causal dependencies set between different units on neurons. But even for a pure algorithmic system, you've got a neural network laid out in front of you.

phi难以计算的原因：定义要求对系统所有可能的划分求和，组合爆炸使超过约15个单元的系统在计算上不可行——「原则上可检验」与实际可检验之间的落差。

partition /pɑrˈtɪʃən/ *n.*
划分，分割（数学：集合的分划）

48:34 大卫·查尔默斯

It's computationally intractable to measure the fire of one of those once I get to bigger than 15 units or so. So Tononi would like to say this is an empirical theory and in principle empirically testable. But there's the in principle. It's extremely difficult to measure phi. Some people, Scott Aaronson the computer scientist, has tried to put forward counterexamples to the theory, which were basically very, very simple systems like matrix multipliers that multiply two large matrices. Turn out to have enormous phi, phi as big as you like if the matrices are big enough and therefore by Tononi's theory will not just be conscious, but as conscious as a human being.

计算机科学家斯科特·阿伦森（Scott Aaronson）2014年构造反例：结构简单的大矩阵乘法器phi可以任意高，按IIT应比人类还有意识——以此对该理论做归谬。

intractable /ɪnˈtræktəbəl/ *adj.* 难以处理的；computationally intractable 计算上不可行的

counterexamples /ˈkaʊntəɪg ˌzæmpəlz/ *n.* 反例（counterexample）

49:15 观众

And Aaronson put this forward as a reductio ad absurdum of the IIT theory. I think Tononi basically bit the bullet and said, oh yeah, those matrix multipliers are actually having some high degree of consciousness. So I think IIT is probably at least missing a few pieces if it's going to be developed. But it's a research program too. AUDIENCE: You mentioned belief as an example of something where there's another mental quality, but people don't seem to have the same sense that it is very hard to explain.

托诺尼「咬紧牙关」接受了归谬结论：矩阵乘法器确实有高度意识。在哲学论辩中，「接受看似荒谬的推论」（bite the bullet）是常见但代价高昂的辩护策略。

reductio ad absurdum /rɪ ˌdʌktioʊ æd æbˈsə-dəm/ *n. (拉丁)* 归谬法：由某主张推出荒谬结论以驳倒它

bit the bullet *idiom* （bite the bullet）硬着头皮接受；咬牙承认不利结论

49:47

In fact, it almost seems too easy where people-- like a belief about something sort of feels like just how things are. But you have to reflect on a belief to notice it as a belief. Do you think there is also or has there been research related to this question into why is that different? It seems like another angle of attack on this problem. It's just like, why doesn't this generate the same hard problem. DAVID CHALMERS: Yeah. In terms-- I'm not sure if there's been research from the perspective of the meta-problem or of theory of mind.

50:20 大卫·查尔默斯

Certainly, people have thought, in their own right, what is the difference in belief and experience that makes them so different. This goes way back to David Hume, a philosopher a few centuries ago who said basically, perception is vivid. Impression of impressions and ideas. And impressions like experiencing colors are vivid in force and vivacity, and ideas are merely a faint copy or something. But that's just the first order. And then there are contemporary versions of this kind of thing, far more sophisticated ways of saying a similar thing.

大卫·休谟在《人性论》(1739)中区分「印象」与「观念」：知觉印象生动有力（force and vivacity），观念只是其微弱摹本——这是「为何知觉比信念更成问题」的最早解释。

vivacity /vɪˈvæsəti/ *n.*
生动，活力（休谟术语 force and vivacity）

50:52

But yeah, you could, in principle, explore that through the meta-problem. Why does it seem to us that perception is so much more vivid? What about our models of the mind makes perception seem so much more vivid than belief and makes beliefs seem structural and empty whereas perception is so full of light? But no, I don't know of work on that from the meta-problem perspective. Like I said, there's not that much work on these introspective models directly. There is work on theory of mind about beliefs.

51:22 观众

Tends to be about models of other people. It may be there's something I could dig through in my literature on belief that says something about that. It's a good place to push. AUDIENCE: Thanks. AUDIENCE: I wanted to bring up Kurt Godel. You mentioned your advisor wrote "Godel, Escher, Bach". There's something that seems very like Godel-- Godelian or whatever about this whole discussion in that-- so Godel showed that given a set of axioms in mathematics, that it would either be consistent or complete but not both.

提问者引入哥德尔不完备定理（1931）：包含算术的一致公理系统必有不可判定命题（此处「要么一致要么完备」的表述稍欠精确）。丹尼特被类比为「宁要一致、容忍不完备」的体系。

51:55

And it seems like when Daniel Dennett-- Daniel Dennett seems to have a set of axioms where he cannot construct consciousness from them. He seems to be very much in this consistent camp, like he wants to have a consistent framework but is OK with the incompleteness. And I wonder if a similar approach could be taken with consciousness where we could, in fact, prove that consciousness is independent of Daniel Dennett's set of axioms, the same way they proved-- after Godel, they proved the Continuum Hypothesis was independent of ZF set theory, and then they added the axiom of choice, made it ZFC set theory.

连续统假设被证明独立于ZF集合论（哥德尔1940年证其一致、科恩1963年证其独立）。提问者设想类似地证明「意识独立于丹尼特的公理」，再寻找需要补充的最小公理。

Continuum Hypothesis *n. (数学)*
连续统假设（关于无穷集合基数的著名命题）

52:38

So I wonder if we could show that in Daniel Dennett's world, we are essentially zombies or we are either zombies or not. It doesn't matter. Either statement could be true. And then find what is the minimum axiom that has to be added to Dennett's axioms in order to make consciousness true. DAVID CHALMERS: Interesting. I thought for a moment this was going to go in a different direction and you were going to say Dennett is consistent but incomplete. AUDIENCE: Yes. DAVID CHALMERS: He doesn't have consciousness in his picture.

53:08 大卫·查尔默斯

I'm complete. I've got consciousness-- AUDIENCE: Yes. DAVID CHALMERS: --but inconsistent. That's why I say all these crazy theory. AUDIENCE: Right, yeah. DAVID CHALMERS: And you're faced with the choice of not having consciousness and being incomplete or having consciousness and somehow getting this hard problem and being forced into, at least, puzzles and paradoxes. But the way you put it was friendlier to me.

53:30

Yeah, certainly, Doug Hofstadter himself has written a lot on analogies between the Godelian paradoxes and the MindBody problem. And he thinks always our self models are always doomed to be incomplete in the Godelian way. He thinks that might be somehow part of the explanation of our puzzlement, at least about consciousness. Someone like Roger Penrose, of course, takes this much more seriously literally. He thinks that the computational aspects of computational systems are always going to be limited in the Godel way.

> 侯世达在《我是一个怪圈》等书中主张：自我模型注定以哥德尔式的方式不完备，这可能部分解释我们对意识的困惑——他把哥德尔当作类比资源而非严格论证。

54:03

He thinks human beings are not so limited. He thinks we've got mathematical capacities to prove theorems, to see the truth of certain mathematical claims that no formal system could ever have. So he thinks that we somehow go beyond the incomplete Godelian-- I don't know if he actually thinks we're complete, but at least we're not incomplete in the way that finite computational systems are incomplete. And furthermore, he thinks that extra thing that humans have is tied to consciousness. I never quite saw how that last step goes, even if we didn't have these special non-algorithmic capacities to see the truth of mathematical theorems, how would that be tied to consciousness.

> 罗杰·彭罗斯在《皇帝新脑》（1989）中的著名论证：人类能「看出」形式系统无法证明的数学真理，故心智超越算法，且这种能力与意识相关。查尔默斯指出「与意识相关」这最后一步从未讲清。

54:41 观众

But at the very least, there are structural analogies to be drawn between those two cases, about incompleteness of certain theories. How literally we should take the analogies, I'd have to think about it. AUDIENCE: Has there been some consideration that the problem of understanding consciousness inherently must be difficult because we address the problem using consciousness? I'm reminded of the halting problem in computer science where we say that in the general case, a program cannot be written to tell whether another program will halt because what if you ran it on itself.

停机问题（图灵1936年证明）：不存在能判定任意程序是否停机的通用程序，因为它无法一致地处理以自身为输入的情形。提问者以此类比「用意识研究意识」的自指困难。

inherently /ɪnˈhɪrəntli/ *adv.*
内在地，本质上

halting problem *n.（计算机）*
停机问题（图灵证明不可判定的经典问题）

55:15 大卫·查尔默斯

It can't be broad enough to include its own execution. So I wonder if there is a similar corollary in consciousness where we use consciousness to think about consciousness and so therefore, we may not have enough equipment there to be able to unpack it. DAVID CHALMERS: Yeah, it's tricky. People say it's like, you use a ruler. To measure a ruler-- well, I can do this ruler to measure many other things. But it can't measure itself. It's not [INAUDIBLE]. Well, on the other hand, you can measure one ruler using another ruler.

corollary /ˈkɔrəlɛri/ *n.*
推论，必然结果

55:46

Maybe you can measure one consciousness using another. The brain-- [INAUDIBLE] the brain can't study the brain. But the brain actually does a pretty good job of studying the brain. There are some self-referential paradoxes there. And I think that, again, is at the heart of Hofstadter's approach. But I think we'd have to look for very, very specific conditions under which systems can't study themselves. I did always like the idea that if the mind was simple enough that we could understand it, we would be too simple to understand the mind.

查尔默斯的回应策略：自指限制未必致命——一把尺子量不了自己，却能量另一把尺子；大脑事实上把大脑研究得不错。真正的任务是找出自我研究失效的精确条件。

self-referential /ˌsɛlf rɛfəˈrɛnʃəl/ *adj.*
自指的，指涉自身的

56:16

So maybe something like that could be true of consciousness. On the other hand, I actually think that if you start thinking that consciousness can go along with very simple systems, I think at the very least, we ought to be able to study consciousness in other systems simpler than ourselves. And boy, if I could solve the hard problem even in dogs, I'd be satisfied. Yeah? AUDIENCE: Hey, so I have a question about how the meta-problem research program might proceed, sort of related to the last question.

「如果心灵简单到我们能理解它，我们就会简单到无法理解心灵」是流传甚广的悖论式格言（常归于Emerson Pugh）。查尔默斯的出路：研究比我们更简单的系统（如狗）的意识。

56:43 观众

So certainly things we believe about our own consciousness, even if we all say them, probably some of them are false. Our brain has a tendency to hide what reality is like. If you look at visual perception, there's what's called lightness constancy. Our brain subtracts out the lighting in the environments so we actually see more reliably what the color of objects are. Like these viral examples of the black and gold dress is an example of this. And when you're presented with an explanation of it, it's like, huh?

明度恒常性（lightness constancy）：视觉系统自动扣除环境光照以稳定感知物体颜色。2015年病毒式传播的「蓝黑/白金裙子」正是该机制因光照线索歧义而在人群中产生分裂判断的案例。

lightness constancy *n. (心理学)*
明度恒常性：视觉系统扣除光照变化以稳定颜色知觉

57:14 大卫·查尔默斯

My brain does that? It's not something we have access to. Or Yani Laurel– – DAVID CHALMERS: Laurel Yani, yeah. AUDIENCE: ––illusion is another one where when you hear the explanation, the scientists that understand it, our own introspection doesn't include that. So how do you proceed with trying to get at what consciousness really is versus what our whatever simplified or distorted view might be? DAVID CHALMERS: Yeah, I think well one view here would be that we never have access to the mechanisms that generate consciousness, but we still have access to the conscious states themselves.

卡尔·拉什利（Karl Lashley），20世纪美国神经心理学家，名言「任何脑过程都不是有意识的」：我们能内省到状态本身，却永远内省不到产生状态的过程。

57:50

Actually, Karl Lashley said this decades ago. He said no process of the brain is ever conscious. The processes that get you to the states are never conscious. The states they get you to are conscious. So take your experience of the dress. For me, it was white and gold. So I knew that. Each of us was certain that I am–– I was certain that I was experiencing white and gold. Maybe you were certainly you were experiencing blue and black. AUDIENCE: I forget which it was. All I remember is I was right.

58:18

[LAUGHTER] DAVID CHALMERS: You were sure that, yeah, those idiots can't be looking at this right. I think the natural way to describe this, at least, is that each of us was certain what kind of conscious experience we were having, but what we had no idea about was the mechanisms by which we got there. So the mechanisms are completely opaque. But the states themselves were at least prima facie transparent. I think that would be the standard view. Even a realist about consciousness could go with that.

标准实在论回应：意识状态本身「初步看来是透明的」（我确知自己看到的是白金），产生它的机制则完全不透明。幻觉主义者要更进一步：连状态本身也可能是内省模型的虚构。

prima facie /ˌpraɪmə ˈfeɪʃi/ *adv./adj. (拉丁)*
初步看来（成立）的，表面上的

58:47

They'd say we know what conscious states where we know what those conscious states are. We don't know the processes by which they are generated. The illusionist, I think, wants to go much further and say, well it seems to you that you know what conscious state you're having. It seem to you that you're experiencing yellow and gold. Sorry, yellow and white, whatever it was. Gold and white. AUDIENCE: Black and gold is what I remember. DAVID CHALMERS: No, black and blue, I think, and-- AUDIENCE: Blue?

59:10

DAVID CHALMERS: --gold and white. It seems to you're experiencing gold and white. But, in fact, that too is just something thrown up by another model. The yellow gold was a perceptual model. Then there was an introspective model that said you are experiencing gold and white when maybe, in fact, you're just a zombie. Or who knows what's actually going on in your conscious state. So the illusionist view, I think, has to somehow take this further and say, not just the processes that generate the conscious states, but maybe the conscious states themselves are somehow opaque to us.

59:42 观众

AUDIENCE: It feels like some discussion of generality of a problem is missing from this discussion. The matrix multiplier example of having high phi is still-- it's not a general thing. Is there someone exploring the space, the intersection of generality and complexity that leads to consciousness as an emergent behavior? DAVID CHALMERS: When you say generality, there's the idea that a theory should be general, that it should apply to every system. You mean mechanisms? AUDIENCE: Generality of the agent.

提问者引入「通用性」维度：只会下井字棋的程序无论多复杂，似乎都谈不上意识——询问是否有人研究通用性与复杂性交叉处的意识涌现问题。

emergent /ɪˈmə-dʒənt/ *adj.*
涌现的（整体呈现部分所无的性质）

60:11

If I can write an arbitrarily complex program to play tic-tac-toe and all it will ever be able to do is play tic-tac-toe, it has no outputs to express anything else. DAVID CHALMERS: Yeah. So general in the sense of AGI, artificial general intelligence. Some aspects of consciousness seem to be domain general like for example, maybe insofar as belief and reasoning is conscious. Those are domain general. But much of perception doesn't seem especially domain general. Right? Color is very domain. Taste is very domain specific.

insofar as /ˌɪnsoʊˈfɑr æz/ *conj.*
在……的范围内，就……而言（正式）

60:41 大卫·查尔默斯

So it's still conscious. AUDIENCE: But if my agent can't express problem statements, like if I don't give it an output by which it can express problem statements, you can never come to a conclusion about its consciousness. DAVID CHALMERS: I'd like to distinguish intelligence and consciousness and even be able to-- even natural language and being able to address a problem statement and analyze a problem, that's already a very advanced form of intelligence. I think it is very plausible that a mouse has got some kind of consciousness, even though it's got no ability to address problem statements, and many of its capacities may be very specialized.

查尔默斯的重要区分：智能与意识是两回事。老鼠无法处理问题陈述、能力高度特化，但很可能有意识；高通用性是高智能的必要条件，却未必是意识的必要条件。

61:17

It's still much more general than a simple neural network that can only do one thing. A mouse can do many things. But I'm not sure that I see an essential-- I certainly see a connection between intelligence and generality. We want to say somehow a high degree of generality is required for high intelligence. I'm not sure there's the same connection for consciousness. I think consciousness can be extremely domain-specific as a taste and maybe vision are. Or it can be domain-general. So maybe those two across cut each other a bit.

12 问答：意识的价值与延展心灵

61:51 观众

AUDIENCE: So it seems to me like the meta-problem as it's formulated implies some amount of separation or epiphenomenalism between consciousness and brain states. And one thing that I think underlies a lot of people's motivation to do science is that it has causal import. Like predicting behaviors is clearly a functionally useful thing to do, and if you can predict all of behavior without having to explain consciousness, their motivation for explaining consciousness sort of evaporates and it feels like, yeah, well, what's the point of even thinking about that because it's just not going to do anything for me.

副现象论（epiphenomenalism）：意识由物理过程产生但对物理世界无因果作用。提问者指出其实践后果：若解释行为无需意识，科学家研究意识的动机就会蒸发。

epiphenomenalism /ˌɛpɪfəˈnɑmənəlˌɪzəm/ *n.*
副现象论：意识是物理过程的副产品，无因果作用

causal import *phr.*
因果效力，因果上的重要性

evaporates /ɪˈvæpəreɪts/ *v.*
蒸发；（动机、希望等）消失殆尽

62:34

What do you say to someone when they say that to you? DAVID CHALMERS: What is the thing that they said to me again? AUDIENCE: That maybe consciousness exists, maybe it doesn't. But if I can explain all of human behavior and all of the behavior of the world in general without recourse to such concepts, then I've done everything that there is that's useful, like explaining consciousness isn't a useful thing to do. And thus, I'm not interested in this, and it may-- DAVID CHALMERS: I see. AUDIENCE: --as well not be real.

without recourse to *phr.*
不诉诸，不借助

63:03 大卫·查尔默斯

DAVID CHALMERS: I think epiphenomenalism could be true. I certainly don't have any commitment to it, though. It's quite possible that consciousness has a role to play in generating behavior that we don't yet understand. And maybe thinking hard about the meta-problem can help us get clearer on those roles. I think if you've got any sympathy to panpsychism, maybe consciousness is intimately involved with how physical processes get going in the first place. And there are people who want to pursue interactionist ideas where consciousness interacts with the brain.

查尔默斯回应的第一步：副现象论未必为真——泛心论、交互论、还原论各自都给意识留下了因果角色，元问题研究或许恰恰能澄清这些角色是什么。

63:31

Or if you're a reductionist, consciousness may be just a matter of the right algorithm. In all those views, consciousness may have some role to play. But just say it turns out that you can explain all of behavior, including these problems, without bringing in consciousness. Does that mean that consciousness is not something we should care about and not something that matters? I don't think that would follow. Maybe it wouldn't matter for certain engineering purposes, say you want to build a useful system.

63:58

But at least in my view, consciousness is really the only thing that matters. It's the thing that makes life worth living. It's what gives our lives meaning and value and so on. So it might turn out that consciousness is not that useful for explaining other stuff. But if it's the source of intrinsic significance in the world, then understanding consciousness would still be absolutely essential to understanding ourselves. Furthermore, if it comes to developing other systems like AI systems or dealing with non-human animals and so on, we absolutely want to know.

全场最有分量的价值主张：即使意识对解释行为无用，「意识是唯一真正重要的东西」——它是生活值得过、生命有意义与价值的根源。理解意识属于理解我们自己，与工程效用无关。

intrinsic /ɪnˈtrɪnsɪk/ *adj.*
内在的，固有的（intrinsic significance 内在价值）

64:31

We need to know whether they're conscious because if they're conscious, they presumably have moral status. If they can suffer, then it's very bad to mistreat them. If they're not conscious, then you might-- I think it's very plausible-- treat non-conscious systems, we can treat how we like. And it doesn't really matter morally. So the question of whether an AI system is conscious or not is going to be absolutely vital for how we interact with it and how we build our society. That's not a question of engineering usefulness.

意识与道德地位直接挂钩：能感受痛苦的存在不可虐待。AI是否有意识将决定我们如何对待它、如何组织社会——这正是近年「AI福祉」讨论的核心问题，此判断在大模型时代愈显前瞻。

moral status *phr.*
道德地位（伦理学术语：应否被道德考量的资格）

mistreat /mɪsˈtrit/ *v.*
虐待，苛待

64:59 观众

It's a question of connecting with our most fundamental values. AUDIENCE: Yeah, I completely agree. I just-- I haven't found that formulation to be very convincing to others necessarily. AUDIENCE: Hi, thanks so much for coming and chatting with us today. I'm really interested in some of your earlier work, the extended mind [? distributed?] cognition. And you're at a company speaking with a bunch of people who do an incredibly cognitively demanding task. DAVID CHALMERS: Yeah. AUDIENCE: Most of the literature that I've read on this topic uses relatively simple examples saying like it's difficult to think just inside your head on these relatively simple things.

65:38

And if you take a look at the programs that we build, on a mundane day-to-day basis, they're millions of lines long. I've read people in the past say something like the Boeing 777 was the most complicated thing that human beings have ever made, and I think most of us would look at that and say, we got that beat. The things that large internet companies do, the size, the complexity of that is staggering. And yet if we close our eyes, everyone in here is going to say, I'm going to have difficulty writing a 10 line program in my head.

mundane /mʌnˈdeɪn/ *adj.*
平凡的，日常的

staggering /ˈstægərɪŋ/ *adj.*
令人震惊的，巨大得惊人的

66:07 大卫·查尔默斯

So I've just sort of, as an open, I'd be very interested in hearing your thoughts about how the activity of programming connects to the extended mind ideas. DAVID CHALMERS: Yeah, so this, I guess, is a reference to something that I got started in about 20 years ago with my colleague, Andy Clark. We wrote an article called "The Extended Mind" about how processes in the mind can extend outside the brain when we become coupled to our tools. And actually, our central example back then in the mid-90s was a notebook, someone writing stuff in a notebook.

《延展心灵》（The Extended Mind）由安迪·克拉克与查尔默斯合写，1998年发表，是近三十年被引最多的哲学论文之一：当工具与人紧密耦合时，认知过程可延伸到大脑之外；原始例子是记事本。

66:40

And even then, we knew about the internet, and we had some internet examples. I guess this company didn't exist yet in '95. But now, of course, our minds have just become more and more extended. And smartphones came along a few years later, and everyone is coupled very, very closely to their phones and their other devices that couple of them very, very closely to the internet. Now it's suddenly the case that a whole lot of my memory is now offloaded onto the servers of your company somewhere or other, whether it's in the mail systems or navigation mapping systems or other systems.

offloaded /ˈɔfloʊdəd/ *v.*
卸载，转移出去（offload 认知任务外包给工具）

67:23

Yeah, most of my navigation has been offloaded to maps. And much of my memory [INAUDIBLE] has been offloaded. Well, maybe that's in my phone. But other bits of my memory are offloaded into my file system on some cloud service. So certainly, yeah, vast amounts of my mind are now existing in the cloud. And if I were somehow to lose access to those completely, then I'd lose an awful lot of my capacities. So I think now we are now extending into the cloud thanks to you guys and others. The question's specifically about programming.

「认知卸载」的当代图景：记忆交给邮件与云盘、导航交给地图，失去访问权限即失去大部分认知能力——「我们的心灵正延展进云端」，对着Google员工说尤为应景。

68:07

Programming is a kind of active interaction with our devices. I mean, I think programming is something that takes a little bit longer. It's a longer timescale so the core cases of the extended mind involve automatic use of our devices, which are always ready to hand. We can use them to get information, to act in the moment, which is the kind of thing that the brain does. So insofar as programming is a slower process-- and I remember from my programming days, all the endless hours of debugging and so on-- then it's at least going to be a slower timescale for the extended mind.

ready to hand *phr.*
触手可及的，随取随用的

68:45

But still, Feynman talked about writing this way. Someone looked at Feynman's work and a bunch of notes he had about a physics problem he was thinking about. And someone said to him, oh it's nice you have this record of your work. And Feynman said, that's not a record of my work. That's the work. That is the thinking. I was writing it down and so on. I think, at least my recollection from my programming days, is that when you're actually writing a program, it's not like you just do a bunch of thinking and then code your thoughts.

费曼轶事的要点：手稿不是思考的「记录」，它就是思考本身。查尔默斯借此论证：编程不是先想好再誊写代码，写代码在相当程度上就是思考过程本身。

recollection /ˌrɛkəˈlɛkʃən/ *n.*
回忆，记忆

69:21 观众

The programming is to some very considerable extent your thinking. So is that the sort of thing you're-- AUDIENCE: Yes, absolutely. [INTERPOSING VOICES] If we, I think as people that program, start to reflect on what we do, and very few of us actually-- if you're the tech lead of a system, maybe you've got it in your head. But you would agree that most of the people on the team who have come more recently only have a chunk of it in their head. And yet, they're somehow still able to contribute. DAVID CHALMERS: Oh, yeah.

69:51 大卫·查尔默斯

This is now distributed cognition. The extended mind, the extended cognition starts with an individual and then extends their capacities out using their tools or their devices or even other people. So maybe my partner serves as my memory, but it's still centered on an individual. But then there's the closely related case of distributed cognition, where you have a team of people who are doing something and making joint decisions and carrying out joint actions in an absolutely seamless way. And I take it at a company like this, there are going to be any number of instances of distributed cognition.

分布式认知（distributed cognition）：认知不以单个个体为中心，而分布于团队、工具与流程之中，源自认知科学家埃德温·哈钦斯对舰船导航团队的经典研究。大型软件团队是典型案例。

seamless /ˈsimləs/ *adj.*
无缝的，流畅衔接的

70:24

I don't know whether the company as a whole has one giant Google mind or maybe there's just a near infinite number of separate Google minds for all the individual teams and divisions. But I think probably some anthropologist has already done a definitive analysis of distributed cognition in this company. But if they haven't, they certainly need to. AUDIENCE: Thank you. [APPLAUSE]

definitive /dɪˈfɪnətɪv/ *adj.*
权威性的，最终定论的

anthropologist /ˌænθrəˈpɑlədʒɪst/ *n.*
人类学家

第二部分　核心句型 · 值得模仿的表达

1. It always struck me that …

“It always struck me that the most interesting and hardest unsolved problem in the sciences was the problem of consciousness”

strike sb that… 表示「某事一直给我很深的印象 / 我一直觉得」，比 I always thought 更强调直觉性触动，适合引出个人学术动机或长期观察。

2. There's something it's like to be X

“A system is phenomenally conscious if there's something it's like to be it”

「成为X是有某种感觉的」，源自内格尔的哲学惯用语。what it's like to do sth 是表达主观感受的地道结构，日常也常用：Do you know what it's like to fail?

3. On the face of it, …

“On the face of it, explaining behavior doesn't explain consciousness.”

「乍看之下」，暗示后文可能翻转或深入。写议论文时用它先摆出表面印象，再引出更深分析，比 at first glance 更书面。

4. … still leaves open the question (of) why …

“It still leaves open the question, why is all that accompanied by subjective experience.”

「仍然留下了一个悬而未决的问题」。leave open 表示未封闭、未解决，学术写作中用于指出既有解释的缺口，是提出研究问题的标准句式。

5. shed (some) light on …

“While still shedding some light, at least indirectly, on the hard problem”

「阐明、帮助理解」。可加程度词灵活搭配：shed much/significant/some light on。论文和演讲中描述研究贡献的高频表达。

6. The devil's in the details.

“But the devil's in the details. How do they work to generate this problem?”

习语「魔鬼藏在细节里」：大方向容易认同，难点在具体落实。用于承认某思路正确后，转向追问其具体机制，衔接自然。

7. It would just be very strange if …

“It would just be very strange if these things were independent”

虚拟语气表达论证理据：「如果……那就太奇怪了」，以反面不可信来支持正面主张。学术辩论中委婉而有力的说理句式。

8. insofar as …

“Maybe insofar as belief and reasoning is conscious. Those are domain general.”

「在……的范围内 / 就……而言」，正式书面连词，用于限定断言的适用范围，避免过度概括。仿写：Insofar as the data allow, the theory holds.

9. bite the bullet

“I think Tononi basically bit the bullet and said, oh yeah, those matrix multipliers are actually having some high degree of consciousness.”

习语「咬紧牙关接受」：明知结论令人难堪仍然承认。哲学讨论中特指接受归谬推出的荒谬结论以保全理论，日常也指硬着头皮做难事。

第三部分　生词总表 · 复习用

词汇	音标	词性	释义	出现
acquaintance	/əˈkweɪntəns/	n.	亲知，直接知晓（哲学术语）；相识	26:38
airy fairy	/ˌɛri ˈfɛri/	adj.	（英式口语）虚无缥缈的，不切实际的	13:34
alluded to	—	phr. v.	（allude to）间接提及，暗指	46:56
antecedents	/ˌæntəˈsidənts/	n.	先例，前身，思想渊源	16:59
anthropologist	/ˌænθrəˈpɑlədʒɪst/	n.	人类学家	70:24
Arguably	/ˈɑrgjuəbli/	adv.	可以说，按理说（表有论据支持的判断）	40:41
axioms	/ˈæksiəmz/	n.	公理（axiom）	34:40
big time	—	adv.（口语）	非常，严重地（pose the problem big time 问题极其突出）	24:48
bit the bullet	—	idiom	（bite the bullet）硬着头皮接受；咬牙承认不利结论	49:15
bring this to bear	—	phr.	（bring sth to bear on）把……运用于，施加于	37:47
by virtue of	—	phr.	凭借，由于	28:13
causal import	—	phr.	因果效力，因果上的重要性	61:51
causally closed	—	phr.	因果封闭的（物理主义术语：物理事件的原因都是物理的）	43:16
conceivable	/kənˈsivəbəl/	adj.	可设想的，可想象的	21:56
Continuum Hypothesis	—	n.（数学）	连续统假设（关于无穷集合基数的著名命题）	51:55
conviction	/kənˈvɪkʃən/	n.	坚定的信念，确信	12:53
corollary	/ˈkɔrəlɛri/	n.	推论，必然结果	55:15
counterexamples	/ˈkaʊntəɹɪgˌzæmpəlz/	n.	反例（counterexample）	48:34
critique	/krɪˈtik/	n.	批判，评论（比 criticism 更学术化）	44:20
cut a few corners	—	idiom	（cut corners）走捷径，省略应做的步骤	34:40
debunking	/diˈbʌŋkɪŋ/	n./v.	揭穿，祛魅（证明信念缺乏根据）	14:41

definitive	/dɪˈfɪnətɪv/	adj.	权威性的，最终定论的	70:24
deploying	/dɪˈplɔɪɪŋ/	v.	运用，部署（deploy）	32:45
deranged	/dɪˈreɪndʒd/	adj.	疯狂的，精神错乱的	43:16
dialectical	/ˌdaɪəˈlɛktɪkəl/	adj.	辩证的；论辩攻防上的（dialectical situation 论辩局势）	41:48
disposed	/dɪˈspoʊzd/	adj.	有……倾向的（be disposed to do）	20:22
dispositions	/ˌdɪspəˈzɪʃənz/	n.	倾向，性向（disposition）	20:55
dissociable	/dɪˈsoʊsiəbəl/	adj.	可分离的，可拆开的	39:36
dissociated	/dɪˈsoʊsieɪtəd/	adj.	相分离的，脱节的	39:00
dissolve	/dɪˈzɑlv/	v.	消解（问题）；使消失。注意与 solve（解决）对举	15:55
dualist	/ˈduəlɪst/	adj./n.	二元论的；二元论者（主张心灵与物质分立）	8:20
emergent	/ɪˈmə-dʒənt/	adj.	涌现的（整体呈现部分所无的性质）	59:42
ephemeral	/ɪˈfɛmərəl/	adj.	短暂的，飘忽虚幻的	46:17
epiphenomenalism	/ˌɛpɪfəˈnɑmənəlˌɪzəm/	n.	副现象论：意识是物理过程的副产品，无因果作用	61:51
evaporates	/ɪˈvæpəreɪts/	v.	蒸发；（动机、希望等）消失殆尽	61:51
extant	/ˈɛkstənt/	adj.	现存的，尚存的（书面语）	37:47
far fetched	/ˌfɑr ˈfɛtʃt/	adj.	牵强的，不太可信的	34:06
frivolous	/ˈfrɪvələs/	adj.	轻佻的，不严肃的	0:52
genealogical	/ˌdʒiniəˈlɑdʒɪkəl/	adj.	谱系学的，追溯起源的	14:09
got a bead on	—	phr.	（get a bead on）瞄准；摸清了……的门路	6:18
grounded in	—	phr.	植根于，以……为基础	45:01
halting problem	—	n.（计算机）	停机问题（图灵证明不可判定的经典问题）	54:41
idealist	/aɪˈdiəlɪst/	adj./n.	唯心论的；唯心论者	8:49
ill-conceived	/ˌɪl kənˈsivd/	adj.	构想拙劣的，考虑不周的	36:27
implausible	/ɪmˈplɔzəbəl/	adj.	难以置信的，不像真的	0:52
inclined	/ɪnˈklaɪnd/	adj.	倾向于……的（be inclined to think）	24:12
indivisible	/ˌɪndəˈvɪzəbəl/	adj.	不可分割的	16:59

inference	/ˈɪnfərəns/	n.	推论，推断	30:23
inherently	/ɪnˈhɪrəntli/	adv.	内在地，本质上	54:41
insofar as	/ˌɪnsoʊˈfɑr æz/	conj.	在……的范围内，就……而言（正式）	60:11
interdisciplinary	/ˌɪntəˈdɪsəplənɛri/	adj.	跨学科的	22:57
intractable	/ɪnˈtræktəbəl/	adj.	难以处理的；computationally intractable 计算上不可行的	48:34
intrinsic	/ɪnˈtrɪnsɪk/	adj.	内在的，固有的（intrinsic significance 内在价值）	63:58
introspective	/ˌɪntrəˈspɛktɪv/	adj.	内省的，反观自身心理的	26:38
irreducible	/ˌɪrɪˈdusəbəl/	adj.	不可还原的，不可简化的（哲学术语）	8:20
isomorphic	/ˌaɪsəˈmɔrfɪk/	adj.	同构的（数学术语，结构一一对应）	46:56
lectern	/ˈlɛktə-n/	n.	讲台，演讲桌	3:14
lightness constancy	—	n.（心理学）	明度恒常性：视觉系统扣除光照变化以稳定颜色知觉	56:43
manifest	/ˈmænəfɛst/	adj.	显而易见的（a manifest fact 明白的事实）	41:12
metaphysical	/ˌmɛtəˈfɪzɪkəl/	adj.	形而上学的，关于存在本性的	20:55
miraculously	/məˈrækjələsli/	adv.	奇迹般地	43:16
mistreat	/mɪsˈtrit/	v.	虐待，苛待	64:31
modal	/ˈmoʊdəl/	adj.	模态的（哲学中关于可能性与必然性的）	21:25
moral status	—	phr.	道德地位（伦理学术语：应否被道德考量的资格）	64:31
mundane	/mʌnˈdeɪn/	adj.	平凡的，日常的	65:38
neural correlates of consciousness	—	phr.	意识的神经相关物（NCC，与意识伴随出现的脑活动）	10:15
offloaded	/ˈɔfloʊdəd/	v.	卸载，转移出去（offload 认知任务外包给工具）	66:40
on the face of it	—	phr.	乍看之下，从表面上看（常暗示深入后可能不同）	5:07
onto something	—	idiom	（be onto something）发现了有价值的线索，说到了点子上	43:47
opacity	/oʊˈpæsəti/	n.	不透明性（引申指无法看清内部机制）	29:16

operationalize	/ˌɑpəˈreɪʃənəlaɪz/	v.	使可操作化，把抽象概念转为可测量的指标	10:15
overridden	/ˌoʊvəˈrɪdən/	v.	被推翻，被覆盖（override 的过去分词）	23:28
panpsychist	/pænˈsaɪkɪst/	adj./n.	泛心论的；泛心论者（认为意识是万物基本属性）	8:49
partition	/pɑrˈtɪʃən/	n.	划分，分割（数学：集合的分划）	47:55
pathology	/pəˈθɑlədʒi/	n.	病理，病态（此处喻「哲学病」）	42:31
pin down	—	phr. v.	确切界定，把……说清楚	13:34
prima facie	/ˌpraɪmə ˈfeɪʃi/	adv./adj.（拉丁）	初步看来（成立）的，表面上的	58:18
privileged	/ˈprɪvəlɪdʒd/	adj.	特许的，独享的；privileged access 特权通达（哲学术语）	24:48
qualitatively	/ˈkwɑlətəteɪtɪvli/	adv.	在性质上，定性地	33:17
rationale	/ˌræʃəˈnæl/	n.	理据，根本原因	37:47
ready to hand	—	phr.	触手可及的，随取随用的	68:07
recapture	/riˈkæptʃɚ/	v.	重新捕获，找回（此处指重新解释「数据」）	44:20
recollection	/ˌrɛkəˈlɛkʃən/	n.	回忆，记忆	68:45
reducible	/rɪˈdusəbəl/	adj.	可还原的，可简化为更基本成分的	30:56
reductio ad absurdum	/rɪˌdʌktioʊ æd æbˈsɚdəm/	n.（拉丁）	归谬法：由某主张推出荒谬结论以驳倒它	49:15
reductionist	/rɪˈdʌkʃənɪst/	adj./n.	还原论的；还原论者（主张高层现象可还原为基础过程）	8:49
seamless	/ˈsimləs/	adj.	无缝的，流畅衔接的	69:51
self-referential	/ˌsɛlf rɛfəˈrɛnʃəl/	adj.	自指的，指涉自身的	55:46
speculative	/ˈspɛkjələtɪv/	adj.	思辨性的，推测性的	8:49
staggering	/ˈstægərɪŋ/	adj.	令人震惊的，巨大得惊人的	65:38
substrate	/ˈsʌbstreɪt/	n.	基质，底层载体（如大脑或硅芯片）	29:16
the devil's in the details	—	idiom	魔鬼藏在细节里；难点全在细节	27:41
theorem prover	—	phr.	定理证明器（自动从公理推导结论的程序）	34:40

tractable	/ˈtræktəbəl/	adj.	易处理的，可解决的（反义：intractable）	14:09
transcendental	/ˌtrænsɛnˈdɛntəl/	adj.	先验的（康德哲学术语）	16:59
utterances	/ˈʌtərənsəz/	n.	话语，言语表达（utterance）	11:31
veil	/veɪl/	n.	面纱，遮盖物	29:47
vivacity	/vɪˈvæsəti/	n.	生动，活力（休谟术语 force and vivacity）	50:20
wired into	—	phr.	（被）内置于，天生固化于（常指进化或本能）	41:48
without recourse to	—	phr.	不诉诸，不借助	62:34

第四部分　理解自测 · 8 题

1. 查尔默斯所说的「意识的难问题」与「简单问题」分别指什么?
2. 「元问题」是什么?它为何被认为比难问题更可操作?
3. 幻觉主义的主张是什么?查尔默斯本人持什么立场?
4. 「哲学僵尸」思想实验的内容和作用是什么?
5. 斯科特·阿伦森如何批评整合信息论（IIT）?托诺尼如何回应?
6. Open Philanthropy 的研究者做了什么实验?结果说明什么?
7. 面对「若意识不影响行为就不值得研究」的质疑，查尔默斯给出了哪两条理由?
8. 「延展心灵」理论的核心观点是什么?查尔默斯如何将它联系到编程?

参考答案

1. 难问题：解释物理过程为何以及如何产生主观体验（第4段）；简单问题：解释可客观测量的行为与认知功能，如辨别、整合、报告，找到执行相应功能的机制即可（第10—11段）。
2. 元问题是解释我们为什么认为意识成问题，即解释「问题报告」和我们的相关直觉（第21段）。因为报告是行为，属于简单问题范畴，可用认知科学、脑科学和算法建模的标准方法研究（第22段）。
3. 幻觉主义认为意识（或意识问题）是自我模型抛出的幻觉，解决元问题即可消解难问题（第25—27段）。查尔默斯是实在论者，认为意识真实存在，但承认幻觉主义迷人且有吸引力，只是「根本无法让人相信」（第30—32段）。
4. 哲学僵尸是物理上（或功能上）与常人完全相同却没有任何主观体验的假想存在。多数人觉得它至少可以设想，这种可设想性是引出难问题的核心直觉之一——「为什么我们不是僵尸」（第36—38段）。
5. 阿伦森构造反例：简单的大矩阵乘法器可以有任意高的phi，按IIT应与人一样有意识，以此归谬。托诺尼「咬紧牙关」接受结论，承认矩阵乘法器确实有高度意识（第82—83段）。
6. 他们给一个小型软件智能体设定关于颜色及自身体验的公理，配上定理证明器，它自行得出「我的颜色体验不同于任何物理状态」之类的主张——初步演示了「问题直觉」可以被算法生成（第58—59段）。
7. 第一，副现象论未必为真，意识可能有尚未理解的因果角色；第二，即使无工程用途，意识是生活意义与内在价值之源，且决定AI和动物是否有道德地位、能否被虐待，关乎社会如何对待它们（第109—112段）。
8. 当人与工具紧密耦合时，认知过程可延伸到大脑之外（1998年与克拉克合著论文，最初例子是记事本，如今是手机与云端）。他借费曼「手稿就是思考本身」的轶事指出：写代码在相当程度上就是思考过程本身，而团队开发则属于分布式认知（第115—121段）。