

如何在通用人工智能系统中实现 “知识锁定标杆浮动效应” ——一种基于“时间压力”的简易心智模型

徐英瑾

摘要 所谓“知识锁定标杆浮动效应”，是指面对同样的信念内容，认知主体会在某些环境下将其判定为“知识”，而在另一些环境下将其判定为“非知识”。该效应的存在，使得人类能够根据环境信息的变化，灵活地改变自身信念，更好地适应环境。人工智能体对于该效应机制的模拟，也能够更好地适应环境。不过，这种模拟必须建立在对于该效应的正确理解上，而西方主流知识论学界对于该效应的解释，如语境主义、比对主义与不变主义提出的解释，要么缺乏足够的普遍性，要么本身建立在一些有待解释的概念上。与之相比较，基于“时间压力”的模型，则将智能体的知识指派倾向的强度视为与其感受的时间压力彼此负相关的一项因素，而“时间压力”本身被视为“主体所预估的问题解决所需要的时间”与其“所愿意付出且能够付出的时间”之间的差值。这样的模型不仅能够对所谓的“银行案例”与“斑马案例”作出简洁的解释，而且在原则上可以被算法化。

关键词 通用人工智能；时间压力；知识锁定标杆浮动效应；比对主义；语境主义；不变主义；固知需求

中图分类号 B222；B84 **文献标识码** A **文章编号** 1672-7320(2019)05-0054-13

基金项目 国家社会科学基金重大项目（15ZDB020）

在盖提尔问题（Gettier Problem）被提出以后，英美分析哲学的主流知识论分析有日渐琐碎化的嫌疑，让外行人很难知道这些研究的意义。本文的一个主要动机，是试图在一个特定的知识论研究课题与公众所关心的话题之间搭建桥梁，以便消除分析哲学的学院化研究和非哲学公众的普遍关涉之间的专业壁垒。笔者选择的这个公众性话题就是人工智能研究，与之配套的知识论课题是“知识锁定标杆浮动效应”。将二者联系在一起的深层动机是：笔者希望人工智能基于人类用户实践需求的编程作业，能够倒逼我们去检验那些在书斋里所想出来的知识论理论的“可操作性”，使得一些不符合这些要求的哲学空想有机会被排除掉，而另一些体现这些要求的哲学选项能够浮现并得以发展。

一、“知识锁定标杆浮动效应”概说及其算法化的必要性

人工智能的重要性与其对人类生活的重大价值，显然是不需要加以特别说明的。这里所说的“知识锁定标杆浮动效应”是什么意思呢？这个略显复杂的名词表达式，乃是笔者对于西方知识论文献中常常出现的“epistemic shift”（直译为“认知浮动”）一词组的阐释性翻译。它指的是这样一种日常现象：面对同样的一个命题 P，人们会在某些情况下说“我知道 P 是真的”，而在某些情况下会说“我未必知道 P 是真的”。譬如，在所谓的“斑马案例”^[1]（P1007-1023）中，如果你仅仅看到眼前有一匹斑马走过而心无旁念，你就会很自然地说“我知道那是一匹斑马”；而当有人告诉你这家动物园里的很多所谓斑马其实是被

涂抹了白条子的驴之后，你若再看到一匹斑马，你就会说“我并不确定那是不是真斑马”。与之类似，在所谓的“银行案例”^[2]（P913-929）中，如果你银行卡里的存钱充裕，不担心信用卡扣款会导致银行卡欠费，那么，只要你能凭记忆想起明日是本月银行还款的最后期限，那么，你就足以判断出：我的确知道明日是本月银行还款的最后期限。相反，如果你银行卡里没什么存钱，但你的欠账数额却颇为可观，即使你能凭借记忆想起明日是本月银行还款的最后期限，你也会倾向于认为你得反复核查这一记忆是否准确——因为你无法承担一旦没有及时还款而导致银行卡被冻结的严重后果。换言之，将怎样的信念锁定为“知识”的活动，其标杆乃是随着输入心智机器的信息的变化而变化的，而并非单方面地取决于信念自身的字面语义。“知识锁定标杆浮动”，便是对于上述现象的一种概括。

在我们切入这一现象的解释之前，不妨先来看看这一现象的存在对于人类日常生活的意义。若我们将西方知识论学界关于“知识”的种种繁杂定义加以搁置，而仅仅诉诸普通人对于“知识”一词的语义直觉，那么至少我们能大略地说：在日常生活中，“知识”常常是指人类信念系统中相对稳固的那部分，而“信念”则是指人类信念系统中其真值可以被更自由地加以变更的部分。很明显，“信念”与“知识”之间的界限对于人类认知活动正常展开是具有重要意义的：一方面，如果所有的信念都能被还原为知识的话，那么，我们就等于自废了“根据新经验调整既有信念系统”的武功，失去对于环境变化的适应能力；另一方面，如果我们将所有的知识还原为信念的话，我们就得随时准备搜集正例与反例来验证或证伪我们所具有的一切信念，这样一来，我们的信念系统就无法具有起码的稳定性，我们无法以起码的可预期性与世界打交道。进而言之，我们人类不但需要“信念”与“知识”之间的界限具有起码的稳定性，而且需要“知识锁定标杆浮动”这一效应使得该界限不那么僵死。也就是说，我们需要在某些条件下调高该标杆的高度，以使得一部分某些原来的知识成为信念，或反之，使得我们的信念——知识系统自身的结构特征能够随着环境的变化而得到修正。很明显，通用人工智能^①的研究者，也需要在人造的推理系统中实现这种“知识锁定标杆浮动”效应。譬如，一台在野生动物园中巡逻的机器人，应当在通常情况下将任何被摄像头所捕捉的斑马视为真正的斑马，因为毫无必要的反复怀疑将浪费系统的运作资源，影响系统的运作效率；而当其被告知野生动物园中的某些斑马是伪装的驴子的时候，该机器人就需要搜集更多的证据，来查验看到的“斑马”是不是真斑马。

不过，要在现有的主流人工智能研究中模拟这种效应，是非常困难的，因为主流的人工智能要么只能以公理化的方式一劳永逸地将人类常识的一部分锁定为“知识”，如海耶斯的“朴素物理学”规划所做的那样^[3]（P248-280），要么只能通过一个人工神经元深层网络来片面地模拟人类将一类低阶层信息识别为一类高阶层信息的归类活动。二者都无法根据不同的认知条件自主地调整知识锁定标杆之高低。这到底是为什么呢？对此，笔者的一个解释是：知识锁定标杆浮动效应其实是一个心理学现象，除非我们有一个关于此现象的简易心智模型，我们是无法在人工智能系统中重演此现象的。但主流的人工智能研究要么满足于在符号表征的层面上展列人类知识的静态外观，要么满足于在非常局部的意义上模拟人类神经元系统的活动方式，却基本上没有尝试在足够合适的尺度上模拟与知识锁定标杆浮动效应相关的心智机制。因此，在这个问题上，人工智能的工作需要心理学与哲学的深度介入。

既然提到了哲学，人工智能专家是不是还特别需要关注作为哲学分支之一的知识论对于“知识锁定标杆浮动效应”的讨论成果呢？笔者所给出的答案恐怕会有点消极。笔者认为，西方知识论学界所积累的大多数对于此类效应的讨论成果，参考价值有限，它们解释自身所牵涉的预设都过于复杂了，都过具“特异性”了。与之相比照，笔者所给出的基于“时间压力”的说明方案既简单雅致，又具有被算法化的潜力，足可与之争雄。

不过，考虑到对于“知识锁定标杆浮动效应”的说明一向被视为西方知识论学界的自留地，即使对这

^① “通用人工智能”指的是全方面模拟人类智能的人工智能研究计划，而不仅仅是只能解决特定任务的专用人工智能。

些主流意见抱有批评,笔者也有必要先将自己的批评意见陈列出来。

二、西方主流知识论学界对于“知识锁定标杆浮动效应”的说明及其疏失

严格地说,既然知识论研究的核心话题之一是信念与知识之间的关系,那么,如何解释涉及这一关系的“知识锁定标杆浮动效应”,也就自然引发了西方知识论学界的诸多重磅级人物为之殚精竭虑。笔者根据自己未必完备的文献阅读经验,将西方学界对于该效应的种种说明方案,大致分为两类:

第一类,“知识”的语义标准会随着语境的不同而浮动,由此造成“知识锁定标杆浮动效应”被归类到这个解答思路下的有“语境主义”(contextualism)^[2](P913-929)和作为其理论亲戚的“比对主义”(contrastivism)^[4](P73-103)。语境主义的大意是:通常的语境会导致我们关于什么是“知识”的语义标准维持在一个较低水平,而在怀疑论的语境中,比如在你被告知你看到的每一匹斑马都有可能是伪装的驴的情况下,该标准则会被提高。正因为这种提高,我们才倾向于在怀疑论的语境中拒绝将目标信念归类为货真价实的知识。

比对主义的大意是:“知道”这个动词本身的句法结构并不是通常所认为的“S知道P”,而是“S知道P而非Q”,而这里的“Q”就是所谓的“比对项”,尽管它在日常言谈中未必会被明说。根据比对主义的立论,在日常环境中,像“我知道那是一匹斑马”这样的句子就应当被改写为“我知道那是一匹斑马,而非一头长颈鹿”,“而非”后面的比对项在日常表达中被省略。与之相较,在怀疑论语境中,对于“而非”后面的比对项的省略却成为不可容忍之事了。说得更清楚一点,恰恰是怀疑论语境中比对项(如“被精心伪装成斑马的驴”)与被比对项(如“驴”)之间在现象层面上的接近,才使得对于被比对项的明说显得具有某种不可或缺的价值。因为二者是如此接近,关于“知识”的标准在此类语境中也被提高了,由此导致主体在知识指派过程中的踟蹰。

第二类,“知识”的语义标准不会随着语境的不同而浮动,而造成“知识锁定标杆浮动效应”的,另有其因。在文献中此类立场统称为“不变主义”(invariantism),也就是“不承认语义标准会改变”。支持不变主义的理由是非常明确的,即倘若关于“知识”的语义标准经常改变,那么,语义学和语用学之间的界限也就不存在了。但这一理论代价显然过于沉重,会使得辞典编纂业毫无意义,因为辞典编纂者必须预设语词的语义是相对稳定的,由此,我们便可反推出“知识”的语义标准不会变。然而,“知识锁定标杆浮动效应”毕竟是存在的,既然该效应不是由语义标杆的改变所导致的,其成因又是什么呢?对此,“基于利害的不变主义”与“智性不变主义”提出了两种不同的解答。“基于利害的不变主义”(stake-based invariantism)^[5]的含义是:信念所带来的潜在后果对主体所产生的“利害”的不同,会导致主体产生不同的关于知识标准的语用直觉。具体而言,正如我们在“银行案例”中所看到的,在错误的判断或许会导致主体损失巨大的情况下,主体对于判断的内容的“知识化”会持比较谨慎的态度,而在错误的判断即使发生也不会导致主体损失的情况下,主体对于判断内容“知识化”就会持相对宽松的态度。考虑到“利害”是语用因素的一部分,而不是语义因素的一部分,这种理论既解释了“知识锁定标杆浮动效应”,又没有导致对于“知识”的语义标杆的稳定性的威胁。“智性不变主义”(intellectual invariantism)^[6](P279-294)的解释则是这样的:不同的认知条件会促发主体内部一个叫“固知需求”(need-for-closure)的心理参数的变化,“固知需求”在此指的是主体将特定信念内容固化为知识的心理学需求^[7](P263-283)。具体而言,在某些语境中(比如利害因素被凸显的语境),为了防止过于急促地将某些信念划归为知识导致利益受损,主体的“固知需求”参数被调低,主体内部的“知识”标杆被提高;而在某些语境(比如在利害关系不那么被强调的语境)中,为了节省系统的运作效率,主体的“固知需求”参数被调高,主体内部的“知识”标杆被降低。由于“固知需求”体现的是主体的心理要素,而不是“知识”的语义特征,因此,“智性不变主义”并未导致对于“知识”的语义标杆稳定性的威胁。

如果通用人工智能专家试图建造一台“知识锁定标杆浮动效应”推理机器,他究竟应当从上述知识

论理论中的哪一种获得启发呢？答案是：没有。理由是：通用人工智能系统的设计需要某种统一的算法说明，以便应对各种需要，模拟“知识锁定标杆浮动效应”的复杂环境，但上面给出的所有理论都是为了应对特定的问题被硬凑出来的，过具“特设性”了。此外，这些方案本身又预设了大量难以在一个统一的方案中被算法化的哲学直觉，这些哲学直觉可能引发的哲学疑惑，恐怕要比这些方案试图解决的哲学疑惑更为严重。下面分头阐述这些既有方案的缺陷。

第一，对于语境主义的批评。语境主义的最大麻烦，是如何说明在怀疑论的语境中“知识”的语义标准会提高。主流的处理方式是引入“可能世界语义学”的概念框架。按照该概念框架，任何一种现实情况都有一些可能的情况与之相伴，如“特朗普当选美国总统”这一现实世界的周围，就伴随有“特朗普竞选美国总统失败”这一可能世界，而可能世界之间，也因为每一可能世界与现实世界相似性程度的不同，导致彼此之间不同的间距，比如，“特朗普竞选美国总统失败”这一可能世界，就比“特朗普从未参加总统竞选”这一可能世界，更接近现实世界。至于这些“可能世界”究竟是在形而上学意义上实际存在的外部对象，还是人类心理的主观建构物，在知识讨论中可以暂且不论。那么，这套话语框架又怎么与语境主义的理论相互结合呢？具体而言，在语境主义者看来，如果怀疑论的语境被引入的话，那么，在那些与现实世界最近的可能世界中，原本在现实世界中能够被视为知识的目标信念的内容就不再是知识了，因为主体对于这些切近的可能世界的意识导致了“知识”语义标准的调高。譬如，在“斑马案例”中，主体虽然看到了眼前的斑马，但意识到一个与现实世界很接近的可能世界中，斑马或许是被人精心伪装过的驴子，主体就会倾向于不急着对关于斑马的知觉信念进行知识指派。反之，如果怀疑论的语境没有被引入的话，主体就不会意识到那些虽然与现实世界接近却会导致目标信念不再为真的可能世界，也就会在这种情况下更轻易地作出知识指派。站在人工智能研究的立场上看，这一处理方案的最大问题是预设了诸可能世界是可以被一个个切分开来讨论的，甚至还可以谈论它们彼此之间的远近。且不提在形而上学文献里对于“可能世界”之本体论地位的大量争议，就是仅仅在认知建模的范围内讨论可能世界的个体化与它们之间间距的测量方式，也是让人难以下手的。对于此类困难性，下面这个论证就能轻易揭示。

1. 如果我们抱着最大的诚意去理解可能世界的可切分化标准与它们之间间距的测量原则，便只能说这些原则只是“相关性”概念导致的理论衍生物罢了。说得更清楚一点，主体会根据哪些考量与当下的目标信念相关，去决定如何切分出一个与现实世界不同的可能世界，并决定如何测量某一可能世界与当下世界之间的距离。譬如，在肯定现实世界中“凡是貌似斑马的动物皆为斑马”的情况下，我们之所以将“貌似为斑马的动物实为被伪装的驴”视为一个切近的可能世界，而不认为“貌似为驴的动物都是被伪装的骡子”是一个切近的可能世界，便是因为后一状况与讨论的主题（斑马）不相关。

2. 上述案例提到的“相关性”本身是一个非常模糊的概念，非常“特设”。

3. 在主流人工智能研究中，对于具有特设性的“相关性”的刻画依然是一个难点，而这一问题又与“框架问题”这一公认的难题相互纠葛，使其难上加难。

所以，从上面三点我们就不难推出：语境主义的方案很难被转化为一个合适的算法化模型。

第二，对于比对主义的批评。作为语境主义的变种，比对主义面临的问题是：该方案的实施，要求我们在任何一个语境中，都要找到“比对项”的具体内容，但这一要求本身就强人所难。譬如，当我们仅仅想说“我看到的是斑马”的时候，究竟这句话应当被拓展为“我看到的是斑马，而不是犀牛”呢，还是应当被拓展为“我看到的是斑马，而不是水牛”呢？很显然，在大多数情况下，我们通过“我看到的是斑马”这一语所表达出来的意思，并没有牵涉到犀牛或者水牛。此外，从另外一个方向发起的对于比对主义的批评，恐怕也会让后者感到难以招架：凭什么我们不能一条道走到黑，认为“比对项”不止两个呢？譬如，为何我们不能说“我看到的是斑马，而不是犀牛，也不是水牛，也不是……”呢？笔者看不出比对主义将如何在不去诉诸某些具有特设性的直觉下，通过某种具有算法化潜力的普遍性说明框架去回答这些疑问。

第三,对于基于利害的不变主义的批评。不变主义面临的第一个问题是,难以覆盖对于“斑马案例”的讨论,尽管它貌似可以应对“银行案例”。如果一个人陷入诸如“世上所有的斑马实为被伪装的驴”的怀疑论陷阱,那么,这与他的个人利害有何关联?要知道,以怀疑论思考为典型表现形式的哲学思考,就其本性而言就有一种“非功利性”。另外,如果我们将科学知识也纳入讨论范围的话,利害与知识指派之间的关系就会更不清楚了。譬如,即使我告诉你,你对下列科学问题回答正确与否,将严重影响你的年终奖,你也不会觉得将“知识”这一标签贴给“氧气分子有2个氧原子”这个信念,会有多大的困难——与之相对照,即使我告诉你,你就算是答错了下列的科学问题,也不会影响你的年终奖,恐怕你也会对如何确定下述信念的真假感到踟蹰,此信念即:苯乙烯分子里有8个氢原子(在这里笔者假设被测试者不是职业化学家,且不能上网查资料)。而且,即使就“银行案例”这样的明显牵涉到利益问题的案例而言,基于利害的不变主义的说明要被算法化,也需要预先给出一个如何计算“利害”的一般化模型。关于此问题,很容易想到的解决方案乃是诉诸决策论专家萨福奇所提出的方案^[8]:计算主体在所有方向上的决策具有的所有可能后果带来的好处,从中挑选出最优方案。但正如罗斯与施罗德所指出的^[9](P259-288),执行这样的方案计算成本巨大,这样的计算方案依然无法度量不同向度上的利益大小,比如在“博名”之利与金钱实利之间进行度量——除非去矫揉造作地质询被研究的当事人,以试探其愿意用多大的筹码置换某种非物质利益,但这一做法错误地预设了我们总是有时间与资源去进行这样的心理学实验。从通用人工智能系统设计的角度看,如果广泛被运用于经济学的决策论能够用于设计具有“知识锁定标杆浮动效应”推理机器,那么此类工作应当早就成功了,因为决策论并不是人工智能专家毫不熟悉的领域。但现实恰恰是:这样的推理机器还没有造出来。由此我们就不难反推出来:要么是基于利害的不变主义是错的,要么是它缺乏合理的工程学实现手段。

第四,对于智性不变主义的批评。智性不变主义的核心要点是援引“固知需求”这一心理学概念,解释人类的“知识锁定标杆浮动效应”。笔者认为,在西方主流的知识论文献对于该效应的解释路径中,该解释的合理性程度是最高的,因为该解释与心理建模的关系最为密切,从科学角度看,也似乎“最接地气”。但笔者对该理论依然有批评。在笔者看来,像“固知需求”这样的一个概念,就像诸如“锚定效应”“框架效应”这样的心理学概念一样,乃是心理学家根据自己的一些心理学发现而临时制造出来的概括工具,解释力有限。说得不客气一点,用“固知需求”的变化去解释“知识锁定标杆浮动效应”,就像用“赚钱需求”的变化去解释某人所赚取的工资外收入的数量变化一样,没有告诉听众太多有用的信息。此外,我们依然不知道这个概念所覆盖的,究竟是一组彼此不同的心智机器所共同导致的某些彼此类似的现象聚合,还是某类自成一体心理现象。从心理学文献上看,我们颇有理由怀疑答案是前者,因为诸如“时间压力”“主体的疲倦程度”“背景噪音”等因素,都被心理学家和智性不变主义者列举为“固知需求”背后的成因^[6](P287-288)。在这种情况下,“固知需求”自身在机制上的统一性就非常成疑。由此看来,智性不变主义对于相关的心理学概念的援引过于直接,缺乏后续的“哲学打磨与抛光”,难以满足哲学研究对于普遍性与可解释性的理论需要。此外,甚至通用人工智能研究者也不宜将诸如“固知需求”这样的心理学概念直接引入自身的工程学建构,因为一种具有起码统一性的心智建模工作,必须预先追问心理现象背后的统一机制成因。由此看来,要锻造一个新的、并对通用人工智能研究具有指导意义的“知识锁定标杆浮动效应”的解释平台,我们需要另辟蹊径。笔者的进路是诉诸“时间压力”概念。这个概念曾被智性不变主义者列为导致“固知需求”的因素之一,但在笔者看来,它的地位必须被提升,才会对“知识锁定标杆浮动效应”产生某种更广泛的解释力。

三、“时间压力”概说

在具体对“时间压力”建模之前,我们可以暂时满足于对于该概念的一种直觉性阐述。所谓时间压力,即一个主体在面对一项(或多项)任务时,感到自己主观上愿意投入且能投入的时间,要少于他所预

估的用以解决相关问题所需要的客观时间。我们可以形象地将这种情况称为“时间赤字”。请注意，当事人所预估的解决问题需要的“客观时间”，可能并非是解决该问题所需要的真正的客观时间。若情况相反，即他手头所愿意挤出来的、且能挤出的时间多于他所预估的用以解决该任务所需要的时间，那么，他就不会感到时间压力，而会感到时间上的闲暇，我们可以形象地将这种情况称为“时间盈余”。现在不妨假设时间压力值的大小，与知识指派的难度呈负相关关系。也就是说，时间赤字越大，知识指派越难，反之亦然。在笔者看来，仅仅依赖这样一种貌似粗糙的关于时间压力的阐述，我们就能比主流的知识论理论更简洁地解释“银行案例”与“斑马案例”。大致解释思路如下：

1. 关于“银行案例”。先假设你银行借记卡里的钱足够自动偿还你在信用卡上欠下的钱。如果你被问及“本月 26 日是不是银行还款的最后期限”，那么，你就会根据“上个月的银行还款日期也是 26 日”这一信息，归纳出“本月 26 日的确是银行的还款日期”，并认为你知道这一点。由于这项归纳任务非常简单，所以在你看来，解决该任务所需要的时间与你能提供、且愿提供的时间之间，差值很小，你就不会感到时间压力。在这种情况下，你也不会感到知识指派的任务有多难。那么，假设你银行借记卡里存的钱，远远不足以偿还你在信用卡上欠下的钱，如果你在这种情况下被问及“本月 26 日是不是银行还款的最后期限”，你是不是也会根据“上个月的银行还款日期也是 26 日”这一信息，归纳出“本月 26 日的确是银行的还款日期”这一点呢？这就难说了。假设这时候另外一个想法涌入你的头脑：“这家银行最近做了系统调整，有可能上个月的还款日子和本月不同”，那么，你又该如何根据这一新信息来确定本月 26 日是不是银行的还款日期呢？一般人的直觉反应是：在这样的情况下，当事人应当更难确定他是否知道本月 26 日是不是银行的还款日期。笔者提出的基于时间压力的解释模型，能够立即解释为何我们有这直觉。很明显，如果当事人的信念是两个，即“上月银行还款日是 26 日”“银行的还款日设置方式或许会因为最近的系统调整而变化”，那么，两个信念的矛盾会导致逻辑学家所说的“非单调推理”的情况，即前提的信息越丰富，主体对于推理的结果反而越不确信。由于非单调推理导致任务的复杂程度上升，主体也自然会预期他将投入更多的时间预算处理此问题。这样一来，此时间预算与主体所实际愿意付出、且能付出的思考时间之间的差值也会变得更大，时间压力变大。在这种情况下，主体显然会在知识指派活动中陷入更多的踟蹰。对于这个解决方案，或许有下面的反驳。

基于利害的不变主义者提出的反驳：难道不正是基于利害的考量，才导致主体在上述第二种情况中考虑到了“银行的还款日设置方式或许会因为最近的系统调整而变化”这一额外因素吗？所谓“基于时间压力的模型”，难道不正在此处预设了“基于利害的不变主义”的正确性了吗？

回答：毋宁说情况是这样的，“银行的还款日设置方式或许会因为最近的系统调整而变化”这一点之所以被想到，完全是一个偶然因素，与当事人由自身财政状况而产生的利害无甚关系。实际上，我们既能设想一个人在资金吃紧的情况下，银行的还款日设置一直没变，也可以设想在他资金充裕的情况下，银行的还款日设置一直在变。在前一种情况下，他依然不会对“本月 26 日是银行还款日”这一点的知识地位产生怀疑。此刻的他确实会因为资金的不足产生资金压力，但这一点与确定“本月 26 日是银行还款日”这一点时产生的时间压力无关。在后一种情况下，他却会对“本月 26 日是银行还款日”这一点的知识地位产生怀疑。此刻他的还款压力虽然不大，不过，这一点依然与确定“本月 26 日是银行还款日”这一点所产生的时间压力无关。

再反驳：我们为何会觉得在当事人的财政状况紧张的情况下，他更容易在知识指派中出错呢？

回答：唯一可能的解释是有人在这里有意或者无意混淆了信念内容的确定问题与信念判断任务本身的实践优先性问题。说得具体一点，一个人的还款压力不大，即使他想到了银行的还款日期会变，他也会将搞清楚这一日期的认知任务的实践优先性调低，并优先思考别的任务，比如今晚约会如何给女友制造惊喜。按照我的解释模型，此刻他依然不知道本月的还款日是否是 26 日。而在这种情况下，我们之

所以倾向于认为主体会“知道”本月的还款日的确是26日,乃是因为很多人错将“假装知道”与“知道”混为一谈了。“假装知道”与“真知道”之间的区别是:如果你真知道P,那么你就随时准备好给出某个基于P的行动,而如果你只是假装知道P,那么你只是暂时冻结了对于P的思考而已,而没有准备好做出基于P的行动。很显然,“假装知道”是智能体在面对复杂信息处理任务时候用于优化资源配置的权宜之计。很明显,这里提到的这个案例仅仅符合“假装知道”的定义,而不符合“知道”的定义,因为我们所讨论的当事人仅仅是因为要集中精力考虑约会,才暂时冻结对于银行还款日的思考,而没有准备给出基于相关思考结论的行动。麻烦的是,由于“假装知道”与“知道”之间的差异细微,当我们使用常识心理学的词汇来描述我们的信念状态的时候,二者经常被混淆。一种更细致的分析将揭示这种差异,并使得本节所提到的这一案例不至于出现解释模型的反例。需要指出的是,即使将利害关系纳入我基于时间压力的知识指派模型,该因素与时间压力的最终关系也是未定的。具体而言,就算我们承认利害关系的正向增量会导致当事人思考更多的因素,并导致解决任务的时间预算增量,利害关系的正向增量同时也会使得主体愿意付出更多的时间来解决问题,由此使得“主体预估他所客观需要的问题处理时间”与“主体愿意挤出且愿挤出的时间”之间的差值变得更难预料。我们知道,在算术中,若被减数与减数都被增量,那么,二者的差是否会增加,乃是难以预料的。与其如此麻烦,还不如干脆不去考虑任务的性质与当事人之间的利害关系,使得理论简洁。

2. 关于“斑马案例”的解释。我的模型框架中,在“被看到的斑马可能是被精心伪装的驴”这个可能性没有被提及的情况下,对于“我看到的动物是斑马”这一信念的知识指派就是非常简单的任务,不会涉及太多时间预算。而在“被看到的斑马可能是被精心伪装的驴”这个可能性被提及的情况下,对于“我看到的动物是斑马”这一信念的知识指派就会成为一项非常复杂的任务,因为这项任务本身涉及了对于斑马与被伪装成驴的斑马的识别。由于二者的分别太细微了,这项任务也会导致更大的时间预算,带来更大的时间压力,导致当事人在知识指派时候的踟蹰。对于斑马案例的这种解释也可以沿用到对于所谓“钵中之脑”(brain in the vat)式的怀疑论^①的处理上去。试想:你怎么知道你是作为一个实实在在的人呢?你怎么知道你的确是在看这篇论文呢?你怎么知道你不是一个活在营养钵中的、没有身体的“钵中之脑”呢?你怎么知道这种情况没有发生:作为“钵中之脑”的你被一些被精心输送给你大脑皮层的电子讯号所欺骗,由此产生幻觉,觉得自己在看一本学术杂志呢?很显然,上述这种带有极端怀疑论色彩的质疑,会使得任何一种知识指派都变得非常困难。而我的模型能够立即解释这一困难的产生:既然极端的怀疑论假设,如“钵中之脑”假设,与朴素实在论假设都能够解释同样的日常现象,如我觉得我在看学术杂志,我觉得自己有两只手等,那么,一个基于怀疑论假设的现象与一个基于实在论假设的同质现象之间的差别,就是几乎无法察觉的。在这种情况下,若我们仅仅基于现象而判断怀疑论假设与实在论假设究竟哪个更有根据的话,那么,由此导致的认知任务的复杂性就会变得不可控制。这立即会造成一个超级大的时间预算,进一步使得当事人的任何主观时间投入都会变得无甚意义。由此造成的巨额时间赤字,自然也会使得知识指派变得举步维艰了。在笔者看来,上述对于“知识锁定标杆浮动效应”的解释,不仅简洁,而且还能解释一些与我的理论相竞争的理论所不能解释的现象。比如,任何一个人都能看出“斑马案例”所代表的怀疑论比较温和,而“钵中之脑”案例所代表的怀疑论则颇为极端。在后一案例中,知识指派的难度会超过前者。为何会这样呢?在这个问题上,比对主义者或许能说:在“斑马案例”中,目标信念的真正结构是:“我看到的是一匹斑马,而不是一匹被伪装成斑马的驴”,而在“钵中之脑案例”之中,目标信念的真正结构是:“我看到的是一匹实实在在的斑马,而不是钵中之脑装置驱使我去‘看到’的一匹斑马”,但即使说到这一步,笔者认为比对主义者依然欠我们一个解释,以说明“钵中之脑装置驱使我去‘看到’的一匹斑马”这一比对项,为何要比“被伪装成斑马的驴”这一比对项更难使得知识得到

① 阐述该案例的原始文献见于 Hilary Putnam. *The Meaning of "Meaning."* *Mind, Language and Reality: Philosophical Papers*, Vol 1.

固化。我的模型则可以清楚说明这一点：从常识上看，区分斑马与被伪装成斑马的驴这一任务所带来的时间预算，当然会小于区分实在的斑马与“钵中之脑装置驱使我去‘看到’的一匹斑马”这一任务所带来的时间预算，因此，与前者相关的知识指派任务当然会显得更为艰难。说到这一步，我们的工作依然是纯哲学性质的。接下来，笔者还需要对如何在一个计算系统中实现这个“时间压力计算模型”提供一些重要的指导意见。

四、如何将时间压力模型算法化

现在我们需要解决的问题是：如何在一个可计算的模型中，实现时间压力的表征？一种非常粗糙的办法，就是看在心理学测验中，有多少任务时间被给予被试者。这个做法有时候并不令人满意，即使在大量时间被给予被试者的情况下，被试者也会因为没有兴趣而只愿意投入很小一部分时间。心理学测验的行为主义方法也很难预估出一个主体对于一个任务的复杂程度的预估，并由此确定他所相信的该任务的解决所需要的时间预算量。至于生理学对于时间压力的解释，帮助也不大，因为人类神经组织中那些引发压力感的化学物质——如皮质醇或儿茶酚胺的分泌，仅仅对人类具有意义，而未必对设计一台仅仅在抽象层面上实现“时间压力”的智能机器具有指导意义。新的思路是引入“步骤”这个概念，但“步骤”这个概念又如何与时间压力相关呢？我们不妨用下棋来打比方。众所周知，如果一个围棋老手与一个围棋新手都被要求在 2 秒钟之内决定下一步棋的话，那么，老手所感受到的时间压力肯定小于新手。这个差别当然与“2 秒”这个从外部给予的时间限制关系不大，而是与各自在单位时间内能够完成的估算步骤的数量有关。打个比方说，假设一个老手 1 秒钟内能够估算出 8 步未来走法以及相关的得失，而一个新手恐怕只能估算出未来 2 步的走法以及相关的得失。如此一来，老手的信息处理速度就是新手的 4 倍。由此不难想见，如果一个任务的复杂度所需要的预估步数是 16 步的话，那么对于老手来说，2 秒的任务时限带给他的时间压力就是零（ $16-8\times 2=0$ ），而新手感受到的时间压力是 12（ $16-2\times 2=12$ ）。而在这里，所有这些压力值的计算单位就得诉诸“步骤数量”这个概念。这个概念具有两个特点：一方面，它具有一定的客观性。一个棋手无论多高明，他的棋也必须一步步下，因此，当我们说一个围棋天才下了“30 步棋”的时候，由此涉及的下棋数量，在计算方式方面当无异于对于一个围棋新手的默算活动的测算方式。另一方面，该概念在与时间单位配合的情况下，能够体现个体差异，特定步骤数量与特定时间单位的比值，恰恰体现了一个个体的信息处理速度。不过，在对于认知系统的设计中直接引入“步骤数量”这个参数，显然是有困难的，因为并不是所有的认知任务都像下棋那样可以清楚地讨论“步骤数量”。一种更切实的方式是将其处理为特定的时间量与系统的“原子操作步骤所占据的时间”的比值。也就是说：

公式一：步骤数量（N）= 特定时间量（T）÷ 原子操作步骤所占时间（Ts），或：

$$N = \frac{T}{T_s}$$

这里所说的“原子操作步骤所占时间”，乃是由认知系统的硬件特性所规定的一步最简易的操作所占据的时间，在生理学层面上，这大致对应于电脉冲从一个神经元的细胞核传输到一个毗邻神经元的细胞核所需要的平均时间（我们假设不同类型的神经元之间的尺寸差异已经得到了合适的统计学处理）；而在理论计算科学的语境中，这也可能是指“万能图灵机”的一步不可被还原的动作所占据的时间，如下述动作所占据的时间：从读写磁头读入一个“0”，然后在打印字条上删除一个“1”，然后右移一格。显然，这个参数是能够被测定的，而且对于一定的物种的神经系统与一定的计算机的硬件配置来说，大约就是一个常数。此外，在人工智能的语境中，我们甚至可以设想这个数据是可以被系统内部的高阶层信息处理子系统加以调用的，因此，它也就能够出现在上述公式之中，并由此介入前述高阶层信息处理子系统对于时间压力的测算。那么，公式一所说的“特定时间量”，具体又指什么呢？在讨论时间压力的语境中，这分别就是系统所预估的解决特定任务所需要的时间预算，以及它自己愿意付出、且能付出的时间预算。由

此我们得出了:

公式二:解决特定任务所需要步骤数量的预算(Nr)= 解决该任务的时间预算(Tr)÷ 原子操作步骤所占时间(Ts),或:

$$Nr = \frac{Tr}{Ts}$$

公式三:认知系统所实际愿意给出、并有能力给出的用以解决特定问题的步骤数量(Na)= 系统所愿意给出、且能给出的时间(Ta)÷ 原子操作步骤所占时间(Ts),或:

$$Na = \frac{Ta}{Ts}$$

所以,如果系统感受到的压力值(P),是其所预估需要的操作步骤(Nr)与其愿意给出、且能够给出的操作步骤(Na)之间的差值,我们就得到公式四:

$$P = Nr - Na = k\left(\frac{Tr - Ta}{Ts}\right) (k \text{ 是一个体现特定系统特征的常数})$$

很明显,此公式右边的值是正数时,系统会感受到时间压力。若是负数,系统会感到时间盈余。我们甚至可以从这个公式出发,将智性不变主义者所关注的“固知需求”(NC)概念重新构造出来。构造的方式非常简单:既然固知需求越强,主体就更容易做出知识指派,而时间压力越大,主体就更不容易做出知识指派,那么,固知需求与时间压力之间亦就呈现出了反比关系,我们就有了公式五:

$$Nc = \frac{1}{p} = \frac{1}{k \frac{Tr - Ta}{Ts}} = \frac{Ts}{k(Tr - Ta)}$$

由此,我们也就可以从一个更具算法色彩的角度来解释“银行案例”与“斑马案例”了。就前一案例而言,根据公式四,恰恰是想到银行还款日会变的那个主体对于不同信息的整合任务,导致了 Tr -值的上升,并在其它参数没有发生显著变化的情况下,导致了 P -值的上升,最终导致了主体在知识指派过程中的踟蹰。再强调一遍,意识到还款日会变的当事人,未必就一定是财政风险高的当事人,尽管财政风险高的当事人更容易在确定还款日的过程中多注意各方面的证据。很明显,出于各种复杂的因素,一个财政状态不错的人也会偶然意识到银行还款日或许会变。就后一案例而言,根据公式四,恰恰是主体对于斑马与被伪装成斑马的驴的辨别所带来的困难,导致了 Tr -值的上升,并在其它参数没有发生显著变化的情况下,导致了 P -值的上升,最终导致了主体在知识指派过程中的踟蹰。

下面是读者对上述刻画的八点可能的质疑以及我的回答。

质疑一:你的整个模型构建是基于人工智能的实践需求的,并用这个需求去对知识论研究构成“反包围”。这样的做法,是不是通过将知识论还原为心智建模的一部分,而将知识论研究的独立性都给取消了呢?

答曰:智性不变主义已经初步给出了将知识论换为心智哲学的一部分的研究规划。如果西方主流学界能够容忍智性不变主义的存在,为何不能容忍笔者的类似努力?

质疑二:“知道”在日常生活中是一个非黑即白的词,比如,你要么知道太阳是太阳系的中心,要么不知道,你可不能说“我在一定程度上知道这一点”。但根据你的模型,“时间压力”是可以被程度化的,因此,与之呈反比的“知道”,也可以被加上“一定程度上”这样的副词。但这样的处理,岂不是与我们的上述直觉相冲突吗?

答曰:化解此冲突不难。我们在测算压力值的时候,不妨再设置一个“知识阈值”,即规定压力值低于此阈值的时候,系统的输出就会给出“知道”的标签,反之就是不给出这样的标签。需要注意的是,这个阈值本身是随着设计需要而变的。如果我们需要设计一台“谨慎”的机器人,那么这个阈值就会显得

非常低,也就是说,系统要等到“时间赤字”非常小的时候才给出知识指派;反之,若我们要设计一台“大胆”的机器人,阈值则可被适当提高。不过,如何确定这个阈值是一个实践的问题,不是理论的问题。

质疑三:不同的系统彼此不同的“知识阈值”,又怎么导致一个共通的“知识”语义标准呢?在这种情况下,你的理论究竟还是不是“不变主义”的一种呢?如果我们将“不变主义”的意思解释为“‘知道’的语义标准不变的话”?

答曰:我的模型并没有直接涉及“‘知道’的公共语义标准是否变化”这一问题。我的理论仅仅涉及了系统内部对于时间压力的感受方式,以及该感受方式与其知识指派倾向之间的关系。从人工智能角度看,我仅仅满足于制造出一台具有足够灵活性的推理机器,使其能够灵活调整信念系统中的固化部分与非固化部分之间的界限。至于日常语言所说的“知道”,则更多地涉及自然语处理的问题,因为不同的自然语言如何表达“知道”这个概念,有不同的解决方案,不结合对于不同自然语言的建模,计算机专家也无法笼统地回答这一问题。在人工智能研究中,“自然语言处理”更多涉及人-机界面问题,与推理机制的设计并不是一个课题。至于主流英语分析哲学家的元哲学观点——“基于英语直觉的针对‘知道’(to know)的研究,可以产生覆盖所有语言的一般性哲学结论”——只是一种盎格鲁民族至上主义所导致的幻觉罢了,我们中国学者大可不必对其言听计从。如果我们仅就英语谈英语,当然也是可以讨论我的技术模型是否能够支持一种关于“知道”的语义不变主义。而笔者对此的意见是:原则上我可以让我的模型与此类语义不变主义彼此兼容,而彼此兼容的方式是:尽管不同的系统会因为自身软硬件内环境的不同而产生不同的压力值,但是基于“公共生活”(这是指机器人与人类之间的公共生活,或是机器人之间的公共生活)的需要,在人-机界面语言或者机-机界面语言的层面上,各个系统依然会使用一个语义标准相对固定的“知道”(to know)概念,这个概念或许会与系统内部运行时使用的“知道”概念不同。此情况多少有点像“贵”这个概念的使用。尽管不同的经济主体自身的经济能力与对于经济压力的感受力都是彼此不同的,但不妨碍大家都会在一个类似的意义上使用“贵”这个概念。由此,尽管一个富人会对一个穷人为何会觉得“一头烤乳猪非常贵”这一点感到惊讶,但这个富人至少知道:这头烤乳猪的价格占据此穷人的可支配收入的比例是比较高的,但究竟有多高,这一点对富人往往是未知的。同样的道理,一台配置很高的通用人工智能机器人尽管未必会完全理解另外一台配置不佳的机器人为何会对一个简单的问题的答案贴上“不知道”的标签,但它至少能够估计出:后一台机器人在解决相关认知任务时正承受着巨大的时间压力,但这压力究竟有多大,这一点对前一台机器人往往是未知的。

质疑四:现在再来问几个技术性更强的问题。比如:为何不让计算机直接计算步骤数量,而一定要按照公式一的要求,将其转化为“特定时间量”和“原子步骤所占据的时间”的商呢?虽然对于人类主体而言,要预估一个任务,特别是那些与棋类活动不同的、形式特征不太明显的任务,所需要消耗的步骤量是相对复杂的,但对于计算机来说,这个任务是不是会更简单一点呢?

答曰:甚至对于计算机来说,这个任务也不简单。步骤数量是系统对于内部运行历史特征的总结,与系统从人-机界面获取的外部时间表征信息存在着格式上的差异。如果系统要同时记录内部运行的步骤数量与人类所给予的任务所自带的时间表征的话,那么,系统就会产生如何协调两类数据之间关系的额外负担。我们采用公式一,对于步骤数量的直接表征就被规避了,系统可以通过“特定时间量”和“原子步骤所占据的时间”这两个比较容易获取的数据来定义步骤数量。顺便说一句,前一个数据之所以容易获取,是因为对于它的计量方式与人类对于时间的公共计量方式相同,二者都可以在一个意义上使用“1分钟”“1小时”这样的表达。由此导致的人-机界面的友好度也会更高;后一数据之所以也比较容易获取,乃是因为这在原则上是一个相对于特定机器而言不变的常数)。

质疑五:在上文的讨论中,Tr-值与Ta-值的赋值似乎仅仅是基于讨论者的直觉(也就是说,我们是通过直觉,才意识到了某类认知任务要比别的任务来得困难),而在计算系统中,这种赋值又该如何自动

实现呢?难道计算机有直觉吗?

答曰:在笔者看来,一个足够强大的通用人工智能系统,其实是有可能通过推理来发现一个任务的解决所需要的时间量(T_r -值)的。相关路径有三条:(1)假设系统遇到的这个问题是老问题。于是,系统便调阅自己的运行历史,发现同样的任务历史上它也执行过。由此,系统就将记忆库中存留的上次解决该问题的时间或是历史上解决此问题的平均消耗时间视为这次任务所需要消耗的 T_r -值。(2)假设系统遇到的这个问题是新问题。系统可以通过类比推理发现这个新问题与过去它所处理过的某个老问题类似。于是,系统就将记忆库中存留的上次解决该问题的时间(或是历史上解决此问题的平均消耗时间)视为这次任务所需要消耗的 T_r -值(先假设此值为 a)。因为这个值是来自于类比推理的,其确实度有限,因此,系统就会给出一个数值区间,使得此时对于 T_r -值的赋值,可以容忍对于 a 的少许偏离。(3)假设系统遭遇到的这个新问题,与其所遇到过的任何一个问题都不相似。在这种情况下,系统会自动将 T_r -值调到一个非常大的数值,由此导致压力值的迅速攀升。这一点也是符合我们人类的直觉的:我们人类在遇到完全无法理解的新任务时,也会感到压力极大。但为了使得系统不停摆,系统会自动寻求人工干预,或是寻求别的通用人工智能系统的帮助。再来看系统当如何为其所愿意付出、且其能够付出的时间(T_a -值)进行赋值。笔者将该值进一步分析为“系统所愿意付出的时间”与“系统所能够付出的时间”这两个量的重叠部分。比如,如果前者是5分钟,后者是4分钟,二者的重叠部分就只能算是4分钟。

质疑六:系统该如何为“系统所愿意付出的时间”与“系统所能够付出的时间”这两个参数进一步的赋值呢?后一个参数比较容易处理,此即系统本身得到的任务完成时限与系统本身的运行所允许投入的时间之间的重叠部分。前一个参数则复杂一些,因为这牵涉到“愿意”这个麻烦的情态动词。难道我们要为系统的“自由意志”建模?这可是一项非常艰难的任务。

答曰:一个规避此难题的办法,就是设置如下算法:步骤一:系统接到一个任务时,寻求以下两个问题的答案:(1)对于该任务的执行是否与系统自身的价值观相互抵触;(2)对于该任务的执行是否会干扰系统正在执行的别的任务;步骤二:若(1)的回答是肯定的,对于该任务的执行的确会导致与系统自身的价值观抵触,那么,系统会自动将“系统所愿意付出的时间”赋值为0,由此导致 T_a -值为零。步骤三:若(1)的回答是否定的,则系统就会转而查看对于(2)的追问究竟是会得到一个肯定的还是否定的答案。若答案是否定的,即对于该任务的执行不会干扰系统正在执行的别的任务,那么,系统就会将系统所允许的运作时间视为“系统所愿意付出的时间”的赋值。若答案是肯定的,系统会转而追问问题(3):该问题是否比正在处理的问题更为重要?步骤四:若对于(3)的查看导致了一个肯定的答案,则系统会搁置正在处理的问题,并将系统允许的运作时间视为“系统所愿意付出的时间”的赋值。若上述答案是否定的,则系统会估算当下问题所需要的时间,并将“系统所愿意付出的用以解决新问题的时间”视为“系统所允许的运作时间”与“系统解决当下问题所需要的时间”的差值。

质疑七:上述路线图,似乎预设了系统必须在处理多个问题时,就问题的处理次序进行推理。但为何我们不能设想一台算力足够强大的机器,能够同时处理老问题和新问题,而不纠结于处理问题的次序呢?

答曰:世界上的任何存在者都是有限存在者,计算机也不例外。因此,算力的有限性是一切机器都无法摆脱的宿命。即使算力再强大,系统也需要在不同的任务之间进行排序:譬如在系统所接到的第1000000项任务与第1000001项任务之间排序。在这样的情况下,前文给出的流程,在原则上依然是适用的。

质疑八:你在对于第六个质疑的回答的第一个步骤中,谈到了“价值观”。考虑到即使在人类伦理学家那里,功利主义者与义务论者对于价值本性的讨论依然令人莫衷一是,价值观又如何能在人工智能中得到简洁的刻画?

答曰:无论义务论与功利论的价值观孰是孰非,从执行角度看,价值观无非就是处理事项的一种先后排序问题。比如,儒家的价值观若主张“爱有差等”,那么,一台“儒家机器人”就会在执行救援任务的时候先去探察那些与主人有特别关系的待救援者;而一台体现“兼爱”原则的“墨家机器人”,会随机挑选待救援者。在这种情况下,我们当然可以将相关的价值原则算法化,并讨论不同的价值原则之间的逻辑冲突。譬如,当“墨家机器人”被要求执行先去救出某人亲属的指令后,该指令便会因为与“随机救出待救援者”这一固有指令发生冲突而被悬置。需要补充的是,系统的运行经验亦将告诉系统一项任务自身的价值权重。譬如,如果系统已经被告知来自“教育部”的邮件的优先性要高于大学校长,那么,系统就会将带有前者标签的任务的解决优先性置前。至此,用来说明时间压力的“公式四”所涉及的所有参数的赋值原则以及由此带来的所有质疑,都已经得到说明。因此,时间压力是有希望在一个通用人工智能系统中实现算法化的。

五、结语

正如兼跨知识论研究与人工智能研究的美国学者珀洛克(John L. Pollock)的工作所显示的,人工智能的研究,其实是能够以算法化的方式,为检验特定的知识论工作的“扎实性”提供契机的。但非常可惜的是,在珀洛克于2009年亡故之后,此类跨学科研究的案例就非常少了。这几年方兴未艾的形式知识论的研究引入了大量技术手段来使得知识论的理论描述更具科学化,但如果此类研究不以建立某种统一的认知模型为目标,也会流于“为形式化而形式化”之弊,未必真正能够产生足够的跨学科效应。笔者对于“知识锁定标杆浮动效应”的计算建模尝试,实际上便是在特定认知模型建设之目标导引下所作出的一种将知识论与人工智能哲学相互结合的新努力。这一努力的方法论原则包括两条:

第一,追求理论描述的简洁性,不引入诸如“可能世界”“对比项”之类的含糊概念。换言之,不为解决一个问题而预设一些更难解决的问题。

第二,一方面接受直观的引导,如本文第三节对于时间压力的描述就是基于直观的,另一方面依然坚持为所有的直观提供底层的技术支持,如本文第四节所做的。

限于篇幅,笔者仅仅就如何将时间压力算法化作出了原则上的刻画,而没有真正落实到编程的层面上。很显然,如果我们的工作需要推进到这样一个层面,我们就需要一个具体的计算机语言,来实现这里所说的编程思想。具体而言,这种语言将具备一定的类比推理能力,并能够通过经验的积累而对信念系统自身的结构实时更新。在笔者所知道的范围内,王培先生发明的“非公理化推理系统”(Non-axiomatic Reasoning System)便能够提供一种基本满足上述要求的计算机语言^①[11]。但对于相关编程作业的技术铺展,已经超出了一篇哲学论文的论题范围,笔者只好就此打住。

(宗宁先生对本文的初稿提出了一些很有价值的批评,本文的质疑与答复部分的一些段落就是在他的启发下完成的)

参考文献

- [1] Fred Dretske. Epistemic Operators. *Journal of Philosophy*, 1970, (67).
- [2] Keith DeRose. Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research*, 1992, (52).
- [3] Patrick Hayes. The Naïve Physics Manifesto//Margaret Boden. *The Philosophy of Artificial Intelligence*. New York: Oxford University Press, 1990.
- [4] J. Schaffer. From Contextualism to Contrastivism. *Philosophical Studies*, 2004, 119(1-2).
- [5] J. Stanley. *Knowledge and Interest*. Oxford: Oxford University Press, 2005.
- [6] Jennifer Nagel. Knowledge Ascriptions and the Psychological Consequences of Changing Stakes. *Australasian Journal of*

① 关于纳思系统的文献很多,其中最重要的是:Pei Wang. *Rigid Flexibility: The Logic of Intelligence*.

Philosophy, 2008, 86 (2).

- [7] A. Kruglanski & D. Webster. Motivated Closing of the Mind: "Seizing" and "Freezing". *Psychological Review*, 1996, (103).
- [8] L. Savage. *The Foundations of Statistics*. New York: Dover Publishing, 1972.
- [9] J. Ross & M. Schroeder. Belief, Credence, and Pragmatic Encroachment. *Philosophy and Phenomenological Research Philosophical Studies*, 2012, 88(2).
- [10] Hilary Putnam. *The Meaning of "Meaning"*. *Mind, Language and Reality: Philosophical Papers, Vol 1*. Cambridge: Cambridge University Press, 1975.
- [11] Pei Wang. *Rigid Flexibility: The Logic of Intelligence*. Netherlands: Springer, 2006.

How to Simulate Epistemic Shifts in a General Artificial Intelligence System?

A Cognitive Model Based on the Notion of "Time Pressure"

Xu Yingjin (Fudan University)

Abstract The term "epistemic shifts" refers to the phenomenon that knowledge ascribers would ascribe different epistemic statuses to same beliefs under different internal/external conditions. Hence, one belief would be judged as knowledge in one circumstance, and as non-knowledge in another. The existence of this phenomenon makes human subjects be able to respond to environmental changes in a flexible manner, and hence, an Artificial General Intelligence (AGI) system is also expected to simulate this function. But this simulation is impossible if there is no convincing high-level explanation of "epistemic shifts", whereas the mainstream epistemological explanations of it (which are provided by contextualism, contrastivism, invariantism, etc.) are either too ad hoc or assuming notions which are more troublesome than the explanans in question. My competing theory appeals to the notion of "time pressure", and the intensity of time pressure is supposed to be inversely proportional to the intensity of the disposition of attributing knowledge to the target belief. Time pressure is construed in terms of the numerical difference between the estimated time needed by the completion of the task and the time that the subject can and wants to spend to complete the same task. And such account can be algorithmically treated to fit the theoretical requirement of AGI.

Key words Artificial General Intelligence (AGI); time pressure; epistemic shifts; contrastivism; contextualism; invariantism; need-for-closure

■ 收稿日期 2019-04-20

■ 作者简介 徐英瑾, 哲学博士, 复旦大学哲学学院教授、博士生导师; 上海 200433。

■ 责任编辑 何坤翁