武漢大學學報(哲学社会科学版)
WUHAN UNIVERSITY JOURNAL (Philosophy & Social Science)
第71卷第6期 2018年11月
Vol.71 No.6 Nov. 2018 079~092
DOI:10.14086/j.cnki.wujss.2018.06.008

如何设计具有自主意图的人工智能体
——一项基于安斯康"意图"哲学的跨学科探索

徐英瑾

摘 要 具有自主意图、只依赖小数据运作的通用人工智能系统的出现,并不会像有些人所预估的那样导致"机器奴役人类"的局面出现,因为此类技术对于小数据的容忍可以大大增加此类技术的潜在用户的数量,并使得体现不同用户价值观的通用人工智能系统能够大量出现。这样一来,具有不同意图的通用人工智能系统彼此之间的对冲效应,最终会使得任何一种具有特定意图的通用人工智能系统都无法占据主宰地位。相反,由于作为专用人工智能技术代表的深度学习技术的运用在原则上就需要大量数据的喂入,其对于民众隐私权的侵犯就成为一种难以被全面遏制的常态,因此,此类技术的发展在原则上就会加强一部分技术权贵对于大多数民众的统治地位。不过,要在通用人工智能系统里实现对于意图的工程学建模,就需要我们在哲学层面上首先厘清关于意图的种种哲学迷思。在这个问题上,美国女哲学家安斯康的意图理论是一个比较好的讨论起点。具体而言,安斯康关于"意图是在欲望驱使下做某事的理由"的观点,是可以在通用人工智能的语境中被实现的,但是她关于信念与意图之二元对立的观点,却在不少地方有失偏颇。而"非公理化推理系统"(纳思系统),则将为吸纳安斯康意图论的合理部分提供相应的工程学手段。

关键词 通用人工智能;非公理化推理系统;纳思系统;深度学习;公众隐私;安斯康
中图分类号 B222;B84 **文献标识码** A **文章编号** 1672-7320(2018)06-0079-14
基金项目 国家社会科学基金重大项目(15ZDB020);国家社会科学基金一般项目(13BZX023)

一、从"人工智能是否会奴役人类"谈起

随着近几年以来人工智能技术在工程学层面上的不断进步,关于"人工智能是否会在未来统治人类"的担忧,也日渐被人提起。但在笔者看来,这个问题本身已经包含了诸多语言混乱。如果不预先对这些混乱加以厘清,我们将很难对这一问题作出严肃的应答。具体而言,该问题所涉及的第一重语言歧义即:这里所说的"人工智能"究竟是指专用人工智能(即只能用于特定工作目的的人工智能系统),还是通用人工智能(即能够像人类那样灵活从事各种工作的人工智能系统)?有人或许会说,抓住这一点歧义不放乃是小题大做,因为所谓通用人工智能技术,无非就是既有的专用人工智能技术的集成。但持此论者却没有意识到如下三个问题:

(甲)就既有专业人工智能技术中发展最快的深度学习系统而言,此类系统的运作其实是需要大量的数据输入为其前提的。因此,深度学习系统并不具备根据少量数据进行有效推理的能力——换言之,

它们缺乏“举一反三”的智能,尽管这种智能乃是任何一种理想的通用人工智能系统所不可或缺的。不得不提到的是,在“迁移学习”这一名目下,目前不少深度学习研究者都在研究如何将在一个深度学习网络中已经获得的网络权重分布“迁移”到一个新的网络中去。这姑且可以被视为某种最初步的“举一反三”。然而,这种意义上的迁移学习必须预设深度学习网络所从事的新任务与旧任务之间有足够的相似性,而无法模拟人类在非常不同的领域(如“孙子兵法”与商业活动)之间建立起类比推理关系的能力。

(乙)现有的深度学习架构都是以特定任务为导向的,而这些任务导向所导致的系统功能区分,既不与人类大脑的自然分区相符合(譬如,我们人类的大脑显然没有一个分区是专门用于下围棋的,而专门用于下围棋的“阿尔法狗”系统的内部结构则是为下围棋量身定做的),也缺乏彼此转换与沟通的一般机制。因此,深度学习系统自身架构若非经历革命性的改造,其自身是缺乏进阶为通用人工智能系统的潜力的。

(丙)目前真正从事通用人工智能研究的学术队伍,在全世界不过几百人,这与专业人工智能研究的庞大队伍相比,可谓九牛一毛[①]。

有鉴于特定技术流派的发展速度往往与从事该技术流派研究的人数成正比关系,所以,除非有证据证明通用人工智能的研究队伍会立即得到迅速扩充,否则我们就很难相信:通用人工智能研究在不久的将来就会取代专用人工智能研究,迅速成为人工智能研究的主流。而这一点又从另一个侧面印证了专用人工智能与通用人工智能之间的差异性。

除了上述差异性之外,“人工智能是否会在未来统治人类”?这个问题包含的另一重歧义便是:此问中作为宾语而出现的“人类”,究竟是指“智人”这个生物学概念所指涉的所有个体,还是某一类特定人群,如城市中产阶级或是贫民阶层?有人或许认为这样的提问依然是在小题大做,因为从字面上看来,该问题的提出者显然关心的是人类总体,其判断根据则如下:在该问题中作为主语出现的“人工智能”,显然与作为宾语的“人类”构成了排他性关系,因此,此主语本身应当不包含人类的任何一个成员,而此宾语也由此可以“独占”所有人类个体。不过,在笔者看来,上述问题是有漏洞的,因为“人工智能是否会在未来统治人类”一语的核心动词“统治”在正常情况下显然是需要一个人格化主体作为其主词的,而“人工智能”是否是一个人格化主体,则又取决于这个词组指涉的是专用人工智能,还是通用人工智能。假设它指涉的是专用人工智能(并因此不是一个人格化主体),那么“统治”这词显然就无法在字面上被解读,而只能被视为一个隐喻性表达。在这样的情况下,我们恐怕就不能认为“人工智能”本身与“人类”彼此构成了某种排他关系了,正如在“资本主义正在奴役人类”这句同样具有隐喻色彩的判断中,作为主语的“资本主义”与作为宾语的“人类”亦没有构成排他关系一样[②]。换言之,在这种情况下,我们就只能将“人工智能”视为一个与人类个体成员有相互交叉的复合概念——比如“掌握人工智能技术的一部分技术权贵与这些技术本身的结合体”——并由此将原来的问题改变为这样一个样子:“掌握人工智能技术的一部分人,会在未来奴役另外一部分人类吗?”

经过对于上述两重语言歧义的澄清,我们原来的问题——“人工智能是否会奴役人类”——就会立即有四个变种,其中每一个变种,都由“专用人工智能—通用人工智能”与“所有人类—部分人类”这两个对子各自的构成因素两两组合构成:

变种甲:人类技术权贵与专用人工智能技术(特别是深度学习技术)的结合,是否会导致另一部分人类受到奴役?

---

① 从 2007 年开始,世界通用人工智能协会都会在世界各地进行专业学术会议,讨论通用人工智能的各种研究方案,并定期出版《通用人工智能会议记录》(以 Artificial General Intelligence 为名,在 Springer 出版社定期出版)。相关的旗舰性期刊乃是《通用人工智能杂志》(Journal of Artificial General Intelligence)。但是,根据该期刊的执行主编王培先生的介绍,真正全力进行通用人工智能研究的人士,在全世界也就几百人而已。

② 很显然,“资本主义”作为一种生产关系,是无法脱离具体的人而存在的,否则我们就只能视其为一种神秘的柏拉图式对象了。

变种乙：人类技术权贵与通用人工智能技术的结合，是否会导致另一部分人类受到奴役？

变种丙：专用人工智能技术，是否可能奴役人类全体？

变种丁：通用人工智能技术，是否可能奴役人类全体？

在这四重可能性之中，首先需要被剔除的乃是对于“变种丙”的肯定回答，因为正如我们刚才所提到的，“奴役”这个主词所需要的乃是一个具有真正人格性的主体：这样的主体能够理解“奴役”的含义，并能够理解进行这种“奴役”的目的。而“变种丙”显然难以满足这样的形式要求，因为所谓的专用人工智能，在实质上与我们所使用的便携式计算器一样，都不会产生自己的欲望与意图，遑论“奴役人类”这样的高度抽象的意图。而在上述四个变种之中，最难以被剔除的乃是对于“变种甲”的肯定回答，因为“变种甲”对于作为人类个体的技术权贵的涉及，显然使得“奴役他人”这一意图的承载者得到了落实。同时，目下深度学习技术所预设的“顶级数据采集者”的功能定位，实际上也是为前述技术权贵量身定做的。此外，一部分人利用技术优势对另一部分人进行统治，也是人类历史上常见的现象，因此，如若未来真有人使用人工智能技术对另外一部分人类进行深入的奴役的话，也不会让我们感到过于吃惊。

不过，从伦理角度看，我们依然希望技术的发展最终能够像马克思所预言的那样，带来全人类的解放，而不是加深人类的异化。因此，从这个角度看，“变种甲”所指涉的人类发展方向虽然会有很大的概率成为现实，却非吾人之所欲。在这种情况下，我们不妨再来看看，“变种乙”与“变种丁”是否带给我们更多的希望。

从表面上看，“变种乙”似乎比“变种甲”更不可欲，因为“变种乙”对于更强大的人工智能机制的诉求，及其对于这种机制与人类特定成员的结合的希冀，似乎会造就更为严重的技术异化。但更为仔细的考量，将使得我们发现“变种乙”所蕴含的某种对技术权贵不利（并因此对普罗大众有利）的因素。这就是通用人工智能技术自身。如果这种技术能够发展到让机器产生自身的意图的水准的话，那么我们就难以防止如下两个层面的事件发生了：机器对权贵要其执行的命令产生了怀疑（这种怀疑可能是基于对于相关命令的可实践性的顾虑，甚至可能是基于对于相关命令自身合法性的怀疑），或者说，机器甚至对于自己是否要继续效忠权贵产生了怀疑。换言之，实现“变种乙”所指涉的社会发展方向，势必会对技术权贵本身构成反噬效应。

有的读者或许会对这种“反噬效应”真会发生有所怀疑。他们或许会说：足够狡猾的技术权贵可以让通用人工智能产品的技术水准达到“既能展现灵活性，又不至于破坏忠诚性”的地步，而由此压缩这种“反噬效应”产生的逻辑空间。笔者并不否认这种“小聪明”或许会有一定的施展空间。然而，从根本上看，智能的核心要素就是对于环境的高度适应性，而人类所处的自然与人文环境又是高度复杂的。从这两点中我们就不难推出，除非技术权贵能够严格控制通用人工智能系统的所有信息输入，否则我们便很难设想具有不同利益背景的不同技术权贵竟然会为自己的通用人工智能系统灌输同样的“价值观”——而具有不同“价值观”的通用人工智能系统之间的斗争（其实质是不同利益集团之间的斗争），显然也就会为弱势群体利用这种矛盾寻找更大的利益诉求空间提供可能。同时，在信息多元化的社会背景下，前面提到的“严格控制通用人工智能系统的信息输入”这一要求自身也是难以被满足的。此外，从技术角度看，对于输入信息的全面控制，自然会对通用人工智能系统自主获取信息的行为构成限制，而这种限制又反过来会使得机器在行为上缺乏足够的灵活性，并由此使得其变得不再那么智能。因此，要兼得鱼与熊掌，恐怕不是那么容易的。

对于“变种乙”的分析，自然将我们导向“变种丁”。笔者看来，“变种丁”指涉的那种可能性实现的机会，甚至还要远远小于“变种乙”，因为用户环境使用的多样性，会立即造就不同的通用人工智能系统之间的差异性，并由此使得“机器联合起来对抗整个人类”的场面变得更为遥不可及。

由本节的分析，我们不难看出，要阻止一部分技术权贵凭借人工智能技术奴役人类（即“变种甲”所

指涉的那种可能性),最好的办法就是使得大量的机器产生彼此不同的自主意图,由此对冲掉大量机器被少数人的意图控制的恐怖场面。但这就牵涉到了一个更根本的问题:如何使得我们未来的人工智能系统具有自主意图呢?

二、意图与信念之间的关系

很显然,要回答"如何使得我们未来的人工智能系统具有自主意图"这个问题,我们就难以回避"关于意图的一般理论为何"这样一个问题。而考虑到我们的最终目的是将这个理论施用到目前还没有真正实现的通用人工智能系统上去,该理论就必定会具有一定的抽象性,以便使得它能够在面对不同的技术实现手段时都能够具有一定的覆盖力。这也就是我们在此一定要诉诸相对抽象的哲学讨论的原因。

不过,有鉴于关于意图的哲学讨论在战后英语哲学圈中是非常兴盛的,而本文的工作语言又是汉语,因此,在正式展开本节的讨论之前,笔者还是有必要对"意图"在英文中与汉语中的区别与联系进行阐述。首先,在汉语与英文中,"意图"都可以作为名词出现。比如,我们既可以在汉语中说"玛丽抱着去喝水的意图而去拿起杯子",也可以在英文中说"Mary picks up the cup with the intention of drinking water"。第二,作为名词的"意图"在英文中可以通过加上特定词缀成为副词"具有意图地"(intentionally),而在汉语中,"具有意图地"则是一个非常别扭的表达。如果我们将这个副词短语缩略为"有意地"的话,虽然语气上显得更顺了,但意思却改变了(汉语中"有意地做某事"包含了意图本身具有负面价值意蕴的语义,但英文中的"intentionally"的价值色彩则较为中立)。第三,"意图"可以在英语里轻松转化为动词"intend",但在现代汉语中,"意图"必须与"做"联合成短语"意图做",才能够承担动词的功能属性。第四,无论在汉语或英语中,"意图"虽然都与"欲望"有着深度的勾联,但都比单纯的欲望具有更明确的所指对象。欲望可以是某种前命题层面上的情绪(比如某种模糊的野心),但意图则必须被具体化为相关的命题内容才能变得有意义。比如,当孙权问张昭"曹贼进犯江东的意图是什么"的时候,张昭可不能笼统地回答说:"曹贼志向不小"(因为这是人尽皆知的废话),而要将意图的内容加以清楚地陈说。

既然无论在汉语语境还是在西语语境中,"意图"都可以被视为某种对象被明确化了的欲望,那么,对于意图的讨论自然就会勾联到一个更为宽广的哲学史争议的背景,此即"理性一元论"与"欲望—理性二元论"之间的争议。具体而言,黑格尔便是典型的理性一元论者。他将"欲望"与"生命"视为被"概念"统摄的下层环节,并由此完成理性世界的大一统;而与之相对比,叔本华则在康德的启发下,将"生存意志"视为康德式"自在之物"的替代品,并以他独自的方式维护了康德在"现象界"与"自在之物"之间的二元对立(比如主张在现象界可以被感受到的"人生意义",在自在之物的层面上乃是彻底的虚无;在现象界能够感受到的时空关系,乃是人类认知架构自我反射所导致的假象,而与自在之物无关,等等)。尽管全面地涉及这些哲学史争议的细节并非本文主旨,但上述提示足以向通用人工智能的研究者提出一个问题:未来的通用人工智能系统应当遵循的是黑格尔式的"一元论"的思路,即做出一个能够将信念系统与意图—欲望系统相互融合的某个统一的大系统,还是应当遵循叔本华式的"二元论"的思路,即先预设信念系统与意图—欲望系统之间的彼此独立,然后再做出一个将二者合成的混合式系统呢?

笔者本人的立场是处在黑格尔与叔本华之间的。笔者同情叔本华的地方在于,笔者也认为终极生存欲望的产生具有个体认知架构无法解释的神秘性,因此只能将其作为给定事实而加以接受。而在研制通用人工智能系统的语境中,这些神秘的给定事实就包括:为何系统具有在物理世界中自我保护的倾向,而不是趋向于自我毁灭以及为何一台机器是以电力为驱动方式,而不是以蒸汽为驱动方式的,等等(很显然,不同的物理驱动方式就决定了机器将以怎样的方式来实现自我保护)。那么,为何说这些给定事实具有神秘性呢?这主要取决于我们评判时所采取的立场。如若我们采取的是系统设计者的立场,那么上述这些被给定事实的产生机制自然是毫无神秘性可言的。但若从机器自身的立场上看,情况就非常不

同了。说得更具体一点,这些使得机器得以运作的基本前提,很可能并不会在机器自身的表征推理系统中出现,而很可能是作为一种隐蔽的思维逻辑而出现在机器的硬件配置之中的。举个例子说,一台由蒸汽推动的机器,未必会在操作界面上写明“本机器由蒸汽提供动力”这句话。这就使得整台机器运作的某种深层动力因与目的因,成了某种类似于“自在之物”般的存在者,并由此落在了系统的自主推理系统的视野之外。而意识到了这一点的人工智能设计者,也必须要试图通过机器硬件配置的方式来完成对于这些深层动力因与目的因的物理实现,而不能试图首先在代码编纂的层面上解决这些问题。

不过,黑格尔的理性一元论依然有其可取之处。欲望必须与理性相互结合才能构成行动,而二者的结合显然需要一个结合点。这样的结合点便是作为欲望之具体化或命题化形式的意图。从这个角度看,虽然欲望本身的确是难以在系统的表征语言中以明晰的方式得到展现的,但是作为其替代者的意图却必须被明晰化,否则行动自身的统一性就无法达成。意识到这一点的通用人工智能研究者,也必须设法在编程语言的层面上构造一种能够使得信念系统与意图系统彼此无缝对接的推理平台。

笔者的上述立场最后就导致了一个非常明显的行动哲学层面上的推论: 信念与意图之间的界限是相对的, 是一个统一表征系统内部的分界, 而不是现象界与自在之物之间的那种隔绝式分界。与之相比较, 在战后英美行动哲学圈中因研究“意图”而名声大噪的女哲学家安斯康(Gertrude Elizabeth Margaret Anscombe, 1919-2001),则主张扩大信念与意图之间的裂痕。有鉴于安斯康的意图理论与笔者立论之间的竞争关系以及她的理论对于另外一些重要哲学家(如约翰·塞尔)的巨大影响,下面笔者将对她的理论进行一番批判性评估。

安斯康的“意图”理论有如下几个要点(笔者将在每一要点后附加自己的批评性文字)[①]:

1. *意图乃是欲望驱动下做某事的理由。* 安斯康当然意识到意图与欲望之间的紧密联系与微妙差异。二者之间的联系,乃在于意图是欲望的具体化,而二者之前的差异则在于: 意图必须具体到“理由”的层面,而欲望则否。举个例子来说,如果茶圣陆羽感到口很渴,并有解渴的欲望,而他相信喝茶能够解渴,那么他就会有理由去喝茶,或者说,“去喝茶”这一意图就成了陆羽去解渴的理由。很显然,在这种情况下,“喝茶”这个行动就作为意图的内容或对象而出现了。而如果外部环境没有任何因素阻止这样的行动得到展现的话,那么,陆羽就会去执行这个行动。与之相比较,倘若没有任何意图起到“将欲望本身加以具体化”的作用的话,那么,欲望就不会得到任何管道以便通向行动的。因此,我们也可以将意图视为欲望与行动之间的转换环节。

现在我们就立即转入对于这一要点的简短评价。需要指出的是, 虽然笔者对安斯康的意图理论的不少方面有所批评,但是对于“意图乃是欲望驱动下做某事的理由”这一点,笔者大致上也是赞同的。但需要注意的是,正因为意图在本质上是一种理由,而被意图持有人所意识到的理由显然就是一种信念状态,因此,安斯康的这一理论在客观上马上就会导致信念与意图之间界限的模糊化。比如,一个人所持有的信念自身的非理性状态,会立即导致在内容上与之相关的意图的可执行性(简言之,由一个愚蠢的信念所导致的意图肯定是不可被执行的,尽管一个不那么愚蠢的信念未必就一定会导致一个可以被执行的意图[②])。这种模糊化状态当然不是说意图与信念之间是没有区别的,因为毕竟不是所有的信念都像意图那样既勾联着欲望,又牵连着行动(比如,“刘备意图娶孙尚香”是一回事,而“刘备相信他已经娶了孙尚香”又是一回事)。但尽管如此,二者之间的界限依然不能被绝对化,因为一个不基于任何信念的意图其实是不可能发生的。比如,“刘备意图娶孙尚香”这一点的确是基于“刘备相信他能够通过与孙尚

① 安斯康研究意图问题的代表作就是《意图》一书, 其版本信息是: Gertrude Elizabeth Margaret Anscombe. Intention. Oxford: Basil Blackwell, 1957; 2nd edition, 1963。考虑到安斯康的话语方式对于不熟悉分析哲学的读者来说会显得比较晦涩,在下面的转述中笔者将根据自己的理解,运用汉语文化中的案例对其原先的案例进行大量的替换。

② 举例来说,倘若陆羽愚蠢地相信喝海水能够解渴,那么他当然就有某种理由去喝海水,并在这种情况下产生喝海水的意图。但正是因为这个信念本身是不合理的,陆羽通过喝海水而解渴的意图肯定会落空。

香结婚巩固孙刘联盟”这一点的。

现在的问题就冒出来了:虽然意图与非意图信念之间的区分是不容抹杀的,但我们又应当在多大程度上勘定二者之间分界带的宽度呢? 正如前文所指出的,笔者的意见是尽量缩小这一宽度,而安斯康的意见是尽量拓宽之。而她进一步拓宽该分界带的理由,则又牵涉到她关于意图本质的另外几个观点 (编号续前):

2. 意图并非预测。 预测性信念,如刘备持有的“与孙尚香结婚有利于巩固孙刘联盟”这一信念,显然是与意图最接近的一类信念,因为二者的时间指向都是面向未来的。因此,对于预测与意图之间界限的勘定,显然有助于拓宽信念与意图之间的分界带。而安斯康用于区分意图与信念的基本理由如下:对于一个预测的支持,需要的是证据——比如,预测者若要评估孙刘联姻对于巩固孙刘政治联盟的作用到底有多大,他就需要观察历史上的政治婚姻的后效,并评估孙权这一特定结盟对象的政治信用,而所有这些评估都是基于证据的。但需要注意的是,上述评估活动与评估者本人的兴趣并无直接关系。也就是说,一个人即使对孙刘联盟没有直接兴趣,他也能够评估二者之间通过姻亲来结盟的可行性。与之相比较,如果刘备本人对孙刘联盟本身没有兴趣的话,那么,即使他相信自己能够通过与孙尚香结婚巩固孙刘联盟,他也无法产生“娶孙尚香”这一意图。换言之,作为“欲望驱动下做某事的理由”,意图自身与特定欲望的直接勾联,使得它能够有别于单纯的预测。

现在我们就立即转入对于这一要点的简短评价。很显然,正如安斯康所指出的,单纯的预期并不能构成意图,因为意图本身的确要有深层的欲望作为其基底。但由此认为意图与预期不搭界,则显得有点矫枉过正了,因为意图本身毕竟是基于预期的。举个例子,如果陆羽有通过喝茶来解渴的意图,那么他就肯定有一个关于“喝茶能够解渴”这一点的预期,否则我们就难以解释一个连自己都不相信喝茶可以解渴的人,竟然能自愿地产生“通过喝茶来解渴”这样的意图。这也就是说,为意图奠基的那些信念自身的证据支持力,也会在相当程度上成为相关意图的证据支持力,并因此使得我们完全有资格去讨论“一个意图的合理性是否有证据支持”这样的议题。

不过,不得不承认的是,尽管“基于一定的预期性信念”这一点的确是意图的构成要素,但构成意图的另外一个要素——与特定欲望的勾联——显然是与证据支持这一议题无关的。举个例子来说,陆羽感到口渴了就是口渴了,他没有必要为这种欲望本身的产生寻找除了相关现象感受之外的任何额外证据(尽管对于陆羽之外的另外一个观察者来说,他的确是需要额外的证据来支持“陆羽感到口渴”这一信念的)。然而,对于这一点的肯定并不会导致我们将意图本身与信念分割,因为作为欲望的明晰化的意图,其本身并不是欲望。

3. 意图具有“事从于心”的符合方向,而信念具有“心从于事”的符合方向。 这是安斯康心目中信念与意图之间的另一重区分。在她看来,要让一个信念成真,信念内容就要符合外部事实,而如果信念内容错的话,责任不在于事实,而在于信念。譬如说,如果孙权错误地相信曹操参与赤壁之战的兵力有 83 万而不是实际上的 20 万人,那么需要修正的是孙权的信念,而不是事实。因此,信念本身就有一种“心从于事”的符合方向。与之相对比,孙权如果具有一个“消灭曹操的 20 万大军”的意图,而实际上该意图并没有得到满足,那么需要改变的乃是外部事实(即孙权需要“火烧赤壁”这样实打实的操作来对曹操的军队进行真正意义上的物理消灭),而不是孙权的信念本身。因此,意图就具有“事从于心”的符合方向。

应当看到的是,关于“符合方向”的讨论乃是安斯康用以区分信念与意图的一个关键性论点,并对以后塞尔的意向性理论产生了巨大影响。但笔者却对该分论点非常怀疑。笔者的批评在于:

第一,从认知系统的表征活动来看,并不存在着真正意义上的“外部事实”。我们前面已经看到,预期乃是一个意图的构成要素,而从这个角度看,如果一个意图没有得到满足的话,真正发生的事情乃是该

意图中的预期性信念与主体最新获取的外部环境报告之间的矛盾（如曹操关于“孙权会投降”的预期与“孙权已经回绝了劝降信”这一报告之间的矛盾）。很显然，从人工智能系统设计的层面上看，除了对于预期自身的时间因子刻画所带来的技术性问题之外，这一矛盾与非意图性信念之间的矛盾并没有本质的不同，因此，我们完全没有必要为“意图的满足或不满足”开创出一套与“真”或“假”不同的评价性谓词（尽管在日常语言的层面上，“真—假”区分的确与“满足—不满足”区分有所分别）。

第二，我们很难说在意图没有得到满足的时候，需要为之负责的乃是外部世界，而意图本身不需要修正。举个例子，我们完全有理由说丰臣秀吉所产生的“通过朝鲜征服明国”的意图自身是荒谬的，但丰臣秀吉的侵略军遭遇中朝联军激烈抵抗的时候，对于丰臣秀吉本人来说，合理的方式是撤销这一意图本身，而不是投入更多兵力来继续贯彻原来的意图。或说得更抽象一点，当一个意图是奠基在一定的预期之上，而该预期本身又是缺乏证据的（如“日本的兵力足以征服明国”这样的愚蠢的预期），那么意图的持有者本人就得为持有意图这件事情负责。在这种情况下再去苛求外部世界，乃是不合理的。

然而，笔者上述的分析并不否认：如果一个使得意图本身得到奠基的期望是的确得到证据支持的，而该意图又没有得到满足，那么主体就有理由继续去改造世界。即使在这种情况下，基于上面提到的第一点意见，笔者依然不能认为改造世界这一活动会牵涉到一种与信念逻辑完全不同的新推理逻辑。

不过，关于安斯康的意图理论，如下面的论点，笔者也有赞成之处。

4. 理由不是原因。 前面已经说过，意图乃是作为“欲望驱动下做某事的理由”存在于安斯康的理论中的。因此，对于意图的说明，就难以回避对于“理由”的说明。“原因”是一个形而上学概念，而“理由”则是在知识论、伦理学与行动哲学的背景中被使用的。譬如，当我说“太阳晒石头是石头热的原因”的时候，这句话并不意味着“太阳晒石头”是“石头热”的理由，因为太阳不是意志或伦理的主体，谈不上对于理由的持有。但我却可以说“我之所以判断这石头会热的理由，乃在于我认为太阳会将其晒热”，因为作为判断的主体，我本人是有完全的资格去拥有一个理由的。由此看来，一般意义上的理由也好，作为意图的理由也罢，它们都必须处在认知主体的表征系统中，并与一定的描述面相发生关联。譬如，当孙尚香起床在院落里舞剑的时候，她的意图若仅仅处在“我要通过舞剑来保持武艺”这一描述面相之下的话，那么即使她不小心吵醒了正在酣睡的刘备，“通过舞剑的声音唤醒自己的丈夫”这一点显然无法构成她舞剑的理由或是意图——尽管这一点的确构成了促使刘备醒来的原因。

从哲学史角度看，安斯康对于原因与理由的区分方案遭到了哲学家戴维森的批评。后者从一种在内涵与外延上都得到拓展的“原因”概念出发，将“理由”也视为一种广义上的“原因”，这一想法本身又是导缘于亚里士多德的“四因说”的[1]（P685-700）。但从通用人工智能系统的设计者来说，安斯康的方案似乎更为可取，因为对于机器的运作程序的设计的确需要暂时“悬搁”使得系统得以运作的外部物理原因，并仅仅从系统内部的操作逻辑入手来进行工程学的构建。很显然，一个被设计出来的系统，当其在特定知识背景下产生一个做某事的理由的时候，支撑该理由的核心要素亦将不直接与系统运作的外部原因相关联——除非这些原因可以被转化为系统内部的信念。

5. 意图的意义内容会渗入相关的实现手段。 按照一般人的理解，由欲望驱动的意图在转向行动的过程中，还需要经历另外一个环节，即实现意图的手段。说得更具体一点，同样的意图可以经过很多不同的手段来实现，而某人之所以选择了这个而不是那个手段来满足其意图，主要也是因为被偏好的手段的实现成本比较低。

但安斯康的意见却与之不同。学界将安斯康的相关意见以及在该问题上与之意见相同的戴维森的意见，统一称为所谓的“安斯康—戴维森论题”：

> 安斯康—戴维森论题：若某人通过做乙事来做甲事的话，那么，他做甲事的行为，就是其做乙事的行为。[2]

很多人或许会认为该论题是反直观的,因为如果“拿蓝色茶杯装的水来解渴”与“拿白色茶杯装的水来解渴”是实现“喝水”这一意图的两个手段的话,那么该意图就不可能与其中的任何一个手段相互同一——否则,按照“同一关系”所自带的传递性(即:A 若与 B 同一,B 若再与 C 同一,则 A 与 C 同一),这两个手段也会彼此同一。但既然蓝色与白色不是一种颜色,“拿蓝色茶杯装的水来解渴”与“拿白色茶杯装的水来解渴”又怎么可能彼此同一呢? 因此可反推出:整个安斯康—戴维森论题就是错的。

安斯康本人则通过对于一个与之平行的案例的分析来为自己的观点做辩护。作为一个天主教徒,她为“夫妻同房时若需避孕,当通过自然避孕法,而非人工避孕法”这一天主教教义进行了哲学捍卫。乍一看这一教义是非常荒谬的,因为既然自然避孕与人工避孕的终极目的乃是一样的,而且,既然安斯康并不否认避孕这一终极意图是可以被接受的,那么,采取何种方式方法避孕,就纯粹是一个取决于当事人方便的琐碎问题。但安斯康的反驳是:与自然避孕和人工避孕这两个手段相互伴随的执行意图(有别于刚才所说的“避孕”这一终极意图)是彼此不同的,或说得更具体一点,与人工避孕相伴随的执行意图并不包含着婚姻的责任,而与自然避孕相伴随的执行意图却伴随着对于婚姻的承诺。因此,非常严格地说,与这两个手段相互对应的,其实是两个不同的意图[3](P41-51)。

有的读者或许会认为安斯康在这里是在狡辩,因为站在当下中国文化的立场上看,我们当然可以设想人工避孕措施的实施未必会导致对于婚姻责任的放弃。但如果我们将避孕的案例置换为喝水的案例的话,或许就能规避文化差异所导致的上述困惑。安斯康想表达的,毋宁说是这个意思:用什么手段(包括什么颜色的杯子)来喝茶的问题,其实并不仅仅涉及手段自身,而且也涉及了与之相伴随的意图。譬如,用某种特定颜色的茶杯来喝茶能够带来的特定审美体验,也是相关意图的特定组成部分。而既然这种特定意图在意义上已经包含了对于手段的指涉,那么在“实现意图”与“实施手段”之间划出一条清楚的界限来,也就变得没什么必要了。

笔者之所以认为安斯康的这一见解对通用人工智能研究有借鉴意义,乃是基于如下考量:主流人工智能研究已经过于习惯于将手段视为外在于任务目标的工具了,而没有意识到意图对于手段的渗透作用。譬如,在司马贺(Herbert Alexander Simon)等人设计的“通用问题求解器”所包含的“目标—手段”进路中,方法库中的一个手段之所以被选中,仅仅是因为对于它的虚拟执行所带来的与目标状态的接近程度,恰好能够越过某条被预先设置的“及格线”,而不是因为伴随该手段自身的意图得到了某种精细的表征。这也就是说,按照这样的思路设置出来的人工智能系统,是很难像人类那样区分两个貌似相似的意图之间的微妙差异的,因此也就无法执行一些需要此类“意向微调”的精密智力活动。

思维锐利的读者或许会反驳说:要实现这种需要此类“意向微调”的精密智力活动,我们就不得不将意图的表征内容变得非常冗长,并由此削弱整个系统的运作效率。但需要指出的是,设计系统的时候,我们未必真的需要将对于相关手段各方面特征的表述全部放在桌面上。实际上,特定意图与特点手段之间的关联,完全可以以非命题的方式表达为展现意图的特定语义节点簇与展现手段的特定语义节点簇之间的共激发关系,由此一来,伴随手段出现的特定意图就可以转化为一个边界模糊的动态结构局域网。在下节的分析中我们就会看到,实现这种技术企图的计算平台,其实还是能够在通用人工智能研究的现有武器库里被找到的。

综合本节笔者的结论是:安斯康对于信念与意图之间的区分是不可取的,因为意图本身就是基于信念的。她对于信念与理由之间的区分则是可取的,而她对于特定意图与特定手段之间的同一性的断定亦有一定的启发意义。下面我们就将基于这些观察,来看看通用人工智能研究应当如何将上述哲学见解在工程学层面上予以转化。

三、通用人工智能语境中的意图刻画

首先需要指出的是，目前的主流人工智能研究（其实质乃是专用人工智能研究）是难以对意图进行哪怕最初步的工程学建模的。比如，主流的符号人工智能、神经网络—深度学习技术，其实都难以在工程学的层面上落实我们在哲学层面上所给出的关于意图的最基本勾画。下面就是对于这一论点的简要展开。

首先，传统的符号 AI 系统是难以落实我们对于意图的一般哲学刻画的。最典型的符号 AI 系统乃是所谓的“专家系统”，其要点是在系统中预存大量已经经由计算机语言整编的人类专家领域知识，并通过某些“生产规则”来衍生出切合特定的问题求解语境的知识推论。很显然，在专家系统中得到预存的知识，显然已经预设了设计意图——如一个关于医疗诊断的专家系统显然已经预设了“治疗病患”这样的意图。但需要注意的是，与人类医生相比，此类系统无法自主地产生“治疗病患”的主观意图，因为它不可以选择不治疗病患（而一个医生却完全可以选择离开医院而去报考公务员）。从这个意义上说，安斯康对于“意图”的定义——“意图乃由欲望驱动去做某事的理由”——并不能通过专家系统而得到落实。

有的读者或许会发问：对于一个用于医疗目的的专家系统来说，为何我们要让其产生“不再做医疗诊断”这一意图呢？难道让其稳定地服务于人类，不正是我们原初的设计目的吗？对于这个问题的应答其实非常简单：能够自主产生意图乃是任何智能系统都应当具有的某种最一般的认知能力，因此，如果一个认知系统在一开始就被剥夺产生“不去做医疗诊断”这样的意图的潜在能力的话，那么它也就不会产生在特定的情况下自主地产生“去实施这样的（而不是那样的）治疗方案”的意图。在这样的情况下，这样的系统至多只能根据过往的医疗数据去预测：在特定的医疗方案被实施后，病患被治好的概率有多大——但正如我们在分析安斯康的意图理论时就已经看到的那样，预测本身并不是意图，因为意图乃是预测性信念与特定欲望的复合体，而通常意义上的专家系统则是没有任何特定的、属于其自身的欲望的。

有的读者或许会说：即使在哲学层面上我们不能将特定的欲望赋予专家系统，但是只要人类设计者预先将“治病救人”的隐含目的寓于整个系统之中，难道系统不是照样可以根据其预测来为病人做诊断吗？在这样的情况下，我们在人工系统中表征特定意图的实践目的又是什么呢？

为了回答这一疑惑，请读者思考下面这个案例。某个用于医疗目的的专家系统根据既往数据，预测出：某病患的肠癌病灶部分如果做大面积切除的话，患者术后生活质量会大大降低。不过，此手术本身的失败率不是很高，只有 20%。而若只做小面积切除的话，患者的术后生活质量则不会受到太大影响——但麻烦的是，这样的话，手术成功的失败率会提高到 40%。那么，系统究竟应当推荐大面积切除的手术方案，还是小面积切除的手术方案呢？

很显然，这是一个关于“要手术成功率还是术后生存质量”的艰难选择。不难想见，即使广义上的“治病救人”的目的已经通过设计者预装到了系统的知识库中，这样的抽象的目的指向依然不足以向系统告知：在“优先考虑手术成功率”与“优先考虑术后生存质量”之间，哪项选择与“治病救人”这项目的更为相关。很显然，对于这一问题的回答，将牵涉到在特定语境中对于病患个体生命价值观的考量，而这些考量的具体结果，又是很难通过某些预先给定的程序设计而被予以一劳永逸地规定的（因为我们完全可以设想一部分病患更在乎术后的生命尊严，而不是手术成功率——而事先被编制的程序显然难以预料到具体病患的具体情况）。而要解决此问题的唯一出路，就是使得系统自己能够通过自己的欲望以及所获取的信息（包括对于当下病患的观察）产生自己的意图——比如自己产生出某种更偏向于提高患者术后生存质量的意图，等等）。然而，这些要求显然已经超出了目下的专家系统所能做的极限。

与符号人工智能相比，基于联接主义或深度学习技术的人工智能系统，离“自主产生意图”这一目标更远。此类系统的基本工作原理，就是通过大量的数据训练，经由一个内部参数可以被调整的大型人工

神经元网络系统的运作，而形成特定输入与特定输出之间的稳定映射关系。而由于此类数据训练工作其实是由人类程序员提供理想的输入—输出关系模板的，所以，人类程序员自身的偏见就很容易被移植到系统上，由此使得系统自身也成为人类偏见的放大器。比如，人类程序员完全可能在设计一个人脸识别系统时预先规定哪些人脸特征具有犯罪倾向，并由此构成对社会中某些特定族裔的不公平的高压态势——而系统本身则根本无法察觉到此类意图的存在，只能按照类似的模板去运作。更麻烦的是，与运用于装备检测或医疗目的的专家系统不同，深度学习技术与广告营销、用户推广等更具资本气息（却也因此更缺乏伦理气息）的运用方式具有更紧密的贴合度，这就使得此类系统甚至可能连“治病救人”这样的最抽象层面上的目的指向都不具备，遑论在这种大的目的指向下产生更为精细的意向生成能力，以便在“重视生命质量”与“延长生命时间”之间进行自主抉择。

所谓的基于“能动主义（enactivism）”思想的人工智能系统，也并不在“自主产生意图”方面有任何推进。能动主义的核心哲学理念是：无论对于人工智能体还是人类而言，认知的实质乃是行动中的有机体与特定环境要素之间互动关系的产物。而对具体的人工智能研究来说，这样的哲学口号一般落实为对于设计机器人的外围传感设备与行动设备的工作的高度重视以及对于中央信息处理系统的设计问题的相对轻视。但这种做法的本身显然会立即使得任何意图自身所依赖的信念系统本身失去了着落（因为信念系统本身就是中央信息处理系统的一部分）；同时，该技术路径对于外部环境因素施加于机器人传感器的因果效力的高度依赖，则又会使得“理由”与“原因”之间的界限变得非常模糊（然而，正如我们所看到的，按照安斯康的看法，作为“理由”的意图在实质上并不能被还原为任何一种“原因”）。此外，也正因为中央语义系统的缺失，任何基于能动主义的人工智能系统在原则上都不可能将具有微妙内容的意向投入到特定的行动中去，因此，这样的系统在原则上就不可能实现前文所说的“安斯康—戴维森论题”。

要在人工智能系统中真正实现对于意图的工程学建模，我们显然需要另辟蹊径。很显然，这样的技术路径需要从根本上处理信念系统、欲望系统之间的相互关系，并在这种基础上实现对于意图的刻画。而有鉴于信念系统与欲望系统之间的互动关系会在不同的问题求解语境中产生不同的意图，这样的技术路径就不可能仅仅局限于特定的问题求解语境，而一定得具有鲜明的“通用人工智能”意蕴。而在这方面，国际通用人工智能活动的代表之一、华裔计算机科学家王培先生发明的“非公理推演系统”——简称为“纳思系统”——便是一个具备被升级为具有自主意图的人工智能体之潜能的计算平台[4]。下面笔者就将相关的技术路线图，以一种相对浅显的方式予以勾勒。

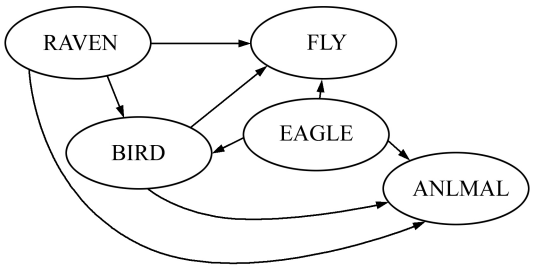

图 1　纳思语义网

由于意图的产生乃是信念系统与欲望系统相互作用的产物，我们就不得不首先介绍一下在纳思系统中信念系统的表征方式。在纳思系统中，一个最简单的判断或信念是由两个概念节点构成的，比如，“乌鸦”（RAVEN）和“鸟”（BIRD）。在纳思系统的最基本层面 Narese-0 上，这两个概念节点由继承关系（inheritance relation）加以联接，该关系本身则被记作“→”。这里的“继承关系”可以通过以下两个属性而得到完整的定义：自返性（reflexivity）和传递性（transitivity）。举例来说，命题“RAVEN →

RAVEN”是永真的(这就体现了继承关系的自返性);若“RAVEN → BIRD”和“BIRD → ANIMAL”是真的,则“RAVEN → ANIMAL”也是真的(这就体现了继承关系的传递性)。这里需要注意的是,在继承关系中作为谓项出现的词项,就是作为主项出现的词项的“内涵集”中的成员(因此,在上述判断中,“鸟”就是“乌鸦”的内涵的一部分),而在同样的关系中作为主项出现的词项,就是作为谓项出现的词项的“外延集”中的成员(因此,在上述判断中,“乌鸦”就是“鸟”的外延的一部分)。换言之,与传统词项逻辑不同,在纳思的推理逻辑中,“内涵”并不代表某种与外延具有不同本体论地位的神秘的柏拉图对象,而仅仅是因为自己在推理网络中地位的不同而与“外延”有所分别。

大量的此类纳思式主—谓判断,则由于彼此分享了一些相同的词项而构成了纳思语义网,如图1。

需要注意的是,这样的一个纳思语义网自身的内容与结构都不是一成不变的,而能够随着系统的操作经验的概念而得到自主更新(这一点将首先在纳思系统的 Narese-1 层面上实现,Narese-1 本身则代表了一种比 Narese-0 更复杂的计算机语言构建)。而其中最重要的一项更新措施,就是根据一个纳思式判断获取的证据量的变化来改变自身的真值,由此改变网络中特定推理路径的权重。说得更具体一点,纳思判断的真值,是由两个参数加以规定的:“频率”(frequency)值和“信度”(confidence)值。现在我们将前一个值简称为 f 值,后一个值简称为 c 值。前者的计算公式如下式所示:

$$f = w^{+}/w \tag{1}$$

说明:在此,w 就意味着证据的总量,而 w^{+} 则意味为正面证据。比如说,若系统观察到 100 只乌鸦,其中 90 只为黑,10 只为白,则命题 “$RAVEN \rightarrow BLACK$” 的 f 值 $= 90/(90+10) = 0.9$

后者的计算公式则如下所示:

$$c = w/(w+k) \tag{2}$$

(例子:在常数 $k=1$ 的情况下,假设系统已经观察到了 100 只乌鸦,则 $c = 100/101$ 0.99)

也就是说,根据系统所获得的外部证据的不同,系统会自行调整推理网络中相关路径的权重值,由此形成不同的推理习惯。而通过对于系统所获得证据数量与种类的相对控制,人类程序员也可以实现按照特定目的“教育”纳思系统的目的。但需要注意的是,与对于人类婴儿的教育一样,在通用人工智能的语境中,对于纳思系统的“教育”并不意味着对于系统的输入的全面人为控制。系统自身自主探索外部环境的相对自由,将始终得到保留。

现在我们再来讨论一下系统的“欲望”系统。与人类的基本生物学欲望(如饥渴)类似,一个人工智能系统也会因为电量不足等原因产生充电的“欲望”,或者因为任务负载过多而产生“休息”的“欲望”。不过,对于系统内部运作状态的此类表征并不会自动产生相关的意图,除非系统已经通过如下步骤完成了“意图”塑造过程:

第 1 步:经过一段时间的学习,系统已经获得一个小型知识库,以便获知使得系统自身能正常运作的一系列条件(如关于电量水平与运作流畅度之间关系的知识)。

第 2 步:系统将由此获得的一般知识施用于对于当下的内部状态的评估,以便得知其当下的状态是否正常。

第 3 步:假设目前系统发现自己目下的状态并不正常。

第 4 步:根据系统自身的推理能力,系统发现:如果某个条件 P 被满足,则当下的状态就能够变得正常。

第 5 步:系统发现没有证据表明 P 已经被满足了。

第 6 步:系统现在将“满足 P”视为备选考虑的意图内容。

第 7 步:系统计算具有怎样的质量与数量规定性的证据,才能够使得 P 成真。这样的证据集被简称为 W。

第 8 步:系统对其以往的操作经历进行回溯,并对当下的任务解决资源 R(如时间、剩余电量)进行评估,以便获知:R 本身是否能够支持系统引发特定的行动 A,以使得 W 能够成真。

第 9 步:如果上一步的评估结果是正面的,则系统会产生这样的意图:通过特定行动 A,使得 W 成真,由此最终使得 P 成真。

第 10 步:如果第 9 步的评估结果是负面的,于是系统本身会检查:是不是有什么别的目标,能够比原本的目标 P 要求更少的操作以便使得系统的状态恢复正常(此即“目标调整”)。如果有,则得到新目标 P',并将第 6 到第 9 步再执行一遍。否则再执行本步骤的头一句话,除非系统发现:此前进行的目标调整的行动,已经穷尽了系统知识库中所有的推理路径,或系统已经没有足够的资源进行这种目标调整活动。而一旦系统有了这种发现,它将转向执行下一步。

第 11 步:系统寻求人类或者其它通用人工智能体的外部干预。

关于这 11 个步骤,笔者还有如下说明:

第一:在纳思系统中,信念系统与意图系统之间的界限是不清晰的。系统对于某意图是否可以得到满足的评估,在相当程度上取决于系统对于使得该意图所对应的状态成真的证据集 W 自身的“可供应性”的评估,因此,关于意图的推理逻辑只是一种更为复杂的关于信念评估的推理逻辑而已(二者之间的微妙差别在于:在对意图的评估中,这些证据本身是作为一种虚拟的存在而被表征的,而证据本身的可供应性又是建立在对于使得这些证据得以被供应的操作的可执行性之上的;与之相比照,在对于信念的评估中,证据本身则是已经被直接给予了)。因此,纳思系统在设计原理上并不那么亲和于安斯康将信念与意图截然二分的意图分析路数。

第二,在纳思系统中,任何系统内部或外部的物理原因都不能够直接构成系统在内部表征中进行推理的理由。举个例子来说,系统内部的电量不足问题,必须被转化为一个能够在纳思语义网中能够被表征出来的词项或判断,才可能进入纳思的推理过程。在这个问题上,纳思系统的设计原理是接近安斯康的意图理论的相关描述的。

第三,正因为在纳思系统中,意图的产生是依赖于信念系统的运作的,而信念系统自身所依赖的推理网络又是系统自身运作经验的结晶,那么,不同的纳思系统自然就会因为自身不同的推理习惯构成不同的意图产生倾向。而不同的人类程序员也将通过对于系统的输入的有限调控,来使得系统自身的具有个性的推理习惯得以产生。

第四,在上面给出的 11 步流程中,为了简化表达,笔者只是预设了纳思系统只关心自身运作的安全(如自身电量的充足性),而不关心人类用户或者其它通用人工智能系统的安全。但是我们完全可以设想纳思系统已经经过程序员的“调教”而形成了这样的推理倾向:一旦系统发现了某个或某类的特定人类用户的安全性受到威胁,系统就会努力寻找方法来消除这些威胁。虽然从伦理学角度看,对于自身安全的优先性考虑会导致利己主义而对于他者安全性的优先性考虑会导致利他主义,但是从纳思系统的设计原则上看,利他主义的推理所牵涉到的意图产生模式,并不会导致与执行利己主义思路的纳思系统根本不同的技术实现路径。

第五,在最一般的意义上,我们当然希望一个通用人工智能系统既能保证为之服务的人类用户的利益,也能够兼顾其自身的利益,正如著名的“阿西莫夫三定律”所表示的那样。然而,在纳思系统的设计过程中,我们并不鼓励通过命题逻辑的方式预先将系统的优先考虑对象锁死,因为这会降低系统在处理特定道德二难处境时的灵活性。一种更值得推荐的方式,便是在一些展示此类二难处境的教学性案例中

让系统学习人类用户的类似处理方式，以便系统能够在面对新的二难推理处境时，根据自带的类比推理能力来自主解决问题。

第六，纳思系统能够更好地实现前文所说的安斯康—戴维森论题，其相关理由如下：我们已经看到，在纳思系统中，对于一个意图目标的可满足性的评估将被转换为对于相关虚拟证据的可兑现性的评估。现在假设有两个证据集（即 W 与 W'），且对于它们的兑现都能够使得相关目标得到满足。但严格地说，既然它们是两个不同的证据集，二者在纳思语义网中所牵涉的内部推理关系就不可能完全重合，因此，通过 W 来满足目标，就会与通过 W' 来满足目标产生不同的推理后效，由此达到不同的目标。换言之，手段与目标之间的截然二分在纳思系统中之所以是不存在的，乃是因为纳思语义网中特定手段与特定目标之间的特定推理路径，显然已经破坏了这种二元性。

四、消除“人工智能体产生恶意”威胁的途径

对于笔者在上文中提出的“通过纳思系统设计具有自主意图的通用人工智能体”的意见，有的读者或许会质疑说：既然纳思系统将通过自己的操作经验获取自身的意图产生习惯，我们又怎么能防止具有不良企图的纳思系统出现呢？

笔者对于这个问题的答案是直截了当的：没有办法防止这一点，因为纳思系统在原则上就允许不同的程序员根据自己的价值观训练自己的系统。因此，正如一个更尊重患者术后生命质量（而不是手术成功率）的医生可以调教纳思系统也按照他的价值观进行推理一样，一个具有邪恶动机的人当然也可以调教纳思系统去行恶。但需要注意的是，只要纳思系统的用户足够多，这种局部的恶很可能会被更广泛语境中的对冲力量所中和，因为大量的纳思系统的使用者会各自将不同的价值观输入到各自所掌握的系统中去，由此使得带有邪恶价值观的纳思系统的作恶行为得到遏制。而这种中和效应之所以可能发生，乃是因为纳思系统的运作本身是不需要用户进行海量的数据搜集的，而这一点在相当程度上就会大大降低用户的使用门槛，并使得利用纳思系统进行价值观博弈的主体在数量上大大增加。换言之，纳思系统使用的低门槛性，自然会使得少数技术权贵很难通过对于此项技术的垄断来进行市场垄断，并由此使得全社会的文化多样性能够得到保存。

有的读者或许还会问：我们是否可以通过立法的方式来阻止通用人工智能技术被别有用心的人所利用呢？笔者的应答是：相关的立法难度很大，因为从法律上看，你很难预先确定哪些人会犯罪；而且，你也很难通过立法去打击那些运用通过人工智能技术去做违背公德（却恰好没有违背法律）的事情。而对于此类法规的无限诉求最后很可能会导致对于整个通用人工智能技术的法律禁止令——但正如笔者在本文第一节中所指出的，这种对于通用人工智能技术的偏见，最终反而会方便打着“专用人工智能技术”名头的技术权贵通过大数据收割机来将普通民众的隐私收割干净，并由此制造出一个更大范围内的伦理上的恶。换言之，除了“在局部容忍恶”与“容忍更大范围内的恶”之间，我们其实并没有第三条出路可走，除非我们要学习美国的阿米什族人，彻底向现代数码技术告别，但这条出路本身也是不具有现实性的，因为几乎没有任何力量可以劝说世界上的大多数人口放弃计算机技术带来的便利。

最后需要指出的是，正如前文中关于如何在纳思系统中产生自主意图的 11 步法所展示的那样，在纳思系统中，意图的产生将取决于系统对于相关实现手段的可行性的评估。因此，一个足够理性（但的确十分邪恶）的纳思系统即使产生了要通过核弹来杀死十万人的邪恶念头，这个念头也不足以形成一个能够兑现为行动的意图，因为它会根据推理发现它根本就无法实现对于核弹（甚至核材料）的拥有（在这里我们假设世界上所有的核弹或相关敏感物质都在各自政府的严格监控下）。换言之，邪恶的念头本身无法杀人，除非它与特定的物质条件相互结合——而幸运的是，对于关涉到公众安全的敏感物质的管控，其实各国政府都有相对成熟的规章制度可以遵循。因此，某些科幻小说所描绘的通用人工智能系统通过超级武器奴役人类的场面，其实是不太可能出现的。由此不难推出，公众与其为未来的人工

智能系统是否会具有自主意图而忧心忡忡并因为这种过度的担心而对人工智能专家的学术自由指手画脚,还不如敦促各自的政府更加严密地看管与公众安全密切相关的敏感物质,使得恶念(无论是产自于自然人的,还是产自于人工智能体的)始终没有机会在物理世界中得到实现。

参考文献

[1] Donald Davidson. Actions, Reasons, and Causes. Journal of Philosophy,1963, 60 .

[2] Donald Davidson. Essays on Actions and Events. Oxford: Oxford University Press,1980.

[3] Gertrude Elizabeth Margaret Anscombe. Contraception and Chastity. The Human World, 1972,(9).

[4] Pei Wang. Rigid Flexibility: The Logic of Intelligence. Netherlands: Springer, 2006.

How to Design an Artificial General Intelligence System Bearing Intentions

An Interdisciplinary Inquiry Based on Anscombe's Philosophy of Intention

Xu Yingjin (Fudan University)

Abstract Whether Artificial Intelligence (AI) will enslave human beings is very relevant to the meaning of "AI" itself. If "AI" means "Artificial General Intelligence" (AGI) and the AGI system in question can work well by tolerating a small size of inputs, then the potential number of the users of this technology will be significantly increased. And if these AGI systems can have their habits of producing intentions in accordance with different users' values, then varieties of AGI systems imbued with different values will make it tough for any single type of machine to dominate the society. Hence, machines cannot enslave human-beings if machines are AGI systems in this sense. In contrast, if "AI" only means deep learning systems which cannot function well without exploiting large quantity of data, then the issue on how to protect the human privacy will always be salient, and in this sense, AGI systems with the capacities of having their own intentions is ethically superior to their deep learning counterparts. As to how to produce intentions in an AGI system, G. E. M. Anscombe's philosophy of intention may offer many inspirations, although some of her claims on the nature of intentions may be controversial. Pei Wang's Non-axiomatic Reasoning System (NARS) will offer a technical realization of the plausible part of Anscombe's theory of intention.

Key words artificial general intelligence (AGI); non-axiomatic reasoning system; NARS; deep learning; public privacy; Anscombe

---

■ **收稿日期** 2018-09-15

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

■ **责任编辑** 何坤翁