# 图灵测试与人类水平机器人

王华平

**摘　要：**制造人类水平的机器人是人工智能的最高目标。要实现这个目标首先就得解决如何衡量机器达到人类水平的问题。其次是路径问题，即什么样的路径最有希望让机器达到人类水平。对于第一个问题，图灵所给出的答案是“图灵测试”。但这个答案并不充分，应该扩充为“全总图灵测试”，要求机器人在真实世界中以一种与人类行动者无法区分的方式做人类行动者所能做之事。对于第二个问题，图灵同样给出了答案，即他所说的“儿童机器”。尽管以图灵的“儿童机器”概念为基础的发展型机器人学是建构人类水平机器人的有效路径，但发展型机器人要发展到人类水平机器人还需解决整合问题，即如何将机器中的不同认知构架整合起来，以实现流畅的信息交流。而要解决整合问题，就要赋予发展型机器人与人脑相当的认知模块和意识水平。

**关键词：**人类水平机器人；全总图灵测试；布洛克脑；中文屋；发展型机器人；儿童机器；认知架构整合

中图分类号：TP18　文献标识码：A　文章标号：1004-8634(2022)04-0088-(10)

DOI：10. 13852/J. CNKI. JSHNU. 2022. 04. 010

人类很早就有了制造自己的想法。据《列子·汤问》记载，匠人偃师用皮革、木头等材料制造了一个歌舞艺人献给周穆王。这个艺人舞姿优美、动作千变万化，穆王以为是真人。古希腊神话提到，跛脚匠神赫菲斯托斯(Hephaestus)制造了一个奴隶来帮助他打铁。作为真实的尝试，达·芬奇设计了一位身穿铠甲的机械骑士，它能够坐着挥动手臂，移动头部和下颌。18 世纪 70 年代，瑞士人雅克德罗(Pierre Jaquet-Droz)制作了一个外形像小孩的移动机器人。这让雪莱(Mary Shelley)欣喜不已。受此激发，她创作了著名的人工智能科幻小说《弗兰肯斯坦：现代普罗米修斯》。尽管历史上的努力延绵不断，但直到 20 世纪中叶人工智能的诞生，人类制造自己这一古老的梦想才第一次变得现实可触。近来，以深度学习为代表的人工智能技术的迅速发展更是重燃了人们的希望，于是乎社会机器人(social robot)、类人机器人(Humanoid Robot)、人型机器人(Android)、人类觉知 AI(Human-Aware AI)纷纷登场，上演着一波波打造人类水平机器人的好戏。

但是，即使是今天最先进的 Beomni 1. 0 机器人与人互动起来也仍然很笨拙。一些人类很容易做到的社会行为，比如通过递眼色来进行交流，机器人却很难做到。迄今为止，没有机器能真正地通过图灵测试。我们不得不面对的一个事实是，在发展人类水平的人工智能的道路上仍然横亘着

**基金项目：**国家社科基金重点项目“人工意识的哲学问题研究”(18AZX007)

**作者简介：**王华平，中山大学(珠海)哲学系教授，博士生导师(广东 珠海 519082)。

许多障碍。首先是如何衡量机器达到人类水平的问题。其次是路径问题,即什么样的路径最有希望让机器达到人类水平。本文的讨论将围绕以上两个问题展开。首先本文将论证,要衡量人工智能是否达到人类水平就需要将标准图灵测试扩充为全总图灵测试(Total Turing Test)。全总图灵测试要求机器人在真实世界中以一种与人类行动者无法区分的方式做人类行动者所能做之事,从而要求机器人具有与人一样的社会认知能力。接着将说明,以图灵"儿童机器"(child machines)概念为基础的发展型机器人学(developmental robotics)是建构人类水平机器人的有效路径。然后将表明,发展型机器人的建构面临着整合问题,即如何将机器中的不同认知构架整合起来,以实现流畅的信息交流。最后将阐明,要解决整合问题,就要赋予发展型机器人与人脑相当的认知模块和意识水平。

## 一、标准图灵测试

要制造人类水平机器人,首先就得解决一个理论问题,即如何判别机器达到了人类水平。这个问题是普通机器所没有的。普通机器的功能是特定的,其性能可据其功能来衡量。比如,一台空调的性能可根据它的标称制冷量来衡量。但人工智能不同,它以图灵机为原型。[①] 而图灵机,正如图灵(Alan Turing)所证明的,是通用机,也即任何一台图灵机都能完成所有专用机能完成的任务。智能机器的智能程度恰恰表现在它的通用性上,也即完成非特定任务的能力。因此,智能不能用基于特定功能的性能来衡量。另一个重要原因是,人类水平只是个抽象的描述词,不足以充当评价的标准。即便具体到意识、思想等心理能力也无济于事,因为我们并不清楚这些心理能力究竟是什么,也没有衡量它们的标准。总之,智能的判别是一个兼具理论性和实践性的难题。

图灵的卓越之处就在于,他为上述难题构思了一个巧妙的解决方案。图灵在给出图灵机的数学模型后便开始思考这样一个问题:图灵机能进行哪些计算?他与邱奇(Alonzo Church)差不多同时找到了答案:所有人能通过机械地遵循有限程序的方式完成的计算任务(即有效可计算函数)图灵机也可完成。此即著名的邱奇—图灵论题。邱奇—图灵论题表明,在有效计算这一点上,图灵机并不亚于人。那么在其他方面呢?图灵设计了一个巧妙的模拟游戏来回答这个问题。在这个游戏中,一个男人和一个女人被分隔在两间房子里,他们和一个提问者通过电传打字设备(相当于今天的网上聊天)进行交流。男人试图说服提问者他是女人,而女人则力图向提问者表明她的真正身份。提问者的任务是正确地识别男人和女人,为此,他可以提出任何可用电传打字设备传达的问题。在游戏的某个阶段,男人被替换为机器。在接下来的游戏中,如果提问者区分不出机器和女人,那么机器就通过了测试,从而我们就可以说机器和人一样能够思考。[②] 这就是最初版的图灵测试。

我们今天流行的标准版是最初版的一个变种。在这个变种中,女人被替换成无关性别的一个人。人与机器都在一个房间里,提问者的任务是分辨出他是在和一个真正的人还是一台机器在交流。两个版本的不同是显见的。最初版要求男人和机器与女人进行对抗游戏,并且,话题是关于性别的。通行版去掉了这样的限制,因而通过起来比最初版更难。不过,这并未改变图灵测试的实质。这个实质就是,假如机器在处理智能任务时表现得和人一样好,那么就应该承认它与人一样会思考。如果不是这样,那么我们如何解释为什么两个表现得一样好的行动者一个能够思考,而另一个却不能呢?要知道,在机器与人做得一样好的情况下,任何解释,例如人有灵魂或大脑,都注定是副现象的,因为那些被认为对人类智能负责的解释项并不能制造有差别的效果来彰显它的存在。所以,没有好的理由来否认一个通过图灵测试的机器能够思考。

一些人反对说,图灵测试并非智能的充分条

① 图灵机是按机器表中的指令来操作符号的抽象机器。按照图灵的描述,图灵机在无限长的分成方格的纸带上进行读写操作。其工作原理如下:图灵机根据它所读取的符号及其内部状态,按照机器表中的指令(1)在纸带上输出一个符号;(2)将纸带移动一格(或将打印头移动一格);(3)将原来的内部状态切换到下一个状态。图灵机的内部状态可由其输入、输出及其与其他心理状态的关系确定。在此意义上,它是功能主义的。

② A. Turing, "Computing Machinery and Intelligence", *Mind*, 1950, 59: pp. 433-460.

件。他们认为,真正的智能涉及思想与理解,而一些系统,例如"布洛克脑"(Blockhead)和"中文屋"(Chinese Room),即便通过了图灵测试也没有思想和理解,因而并不真正具有智能。"布洛克脑"是设想由很多人通过无线设备联结成的一个系统,这个系统能依据事先给定的一段时间的对话所用到的可能数量的句法和语法正确的句子来和人进行对话。在那段时间内,"布洛克脑"可以和人进行任意主题的对话并通过图灵测试,但我们不会认为"布洛克脑"有思想。[①] "中文屋"是这样一个功能系统:一个不懂中文的人被关在屋子里,依据用他的母语所书写的关于中文字形的规则书来操作中文。虽然屋中人看上去和懂中文的人操作得一样好,但他实际上并不懂中文。[②] 然而,这样的思想实验是很有争议的。"布洛克脑"和"中文屋"真的能通过图灵测试吗?计算表明,"布洛克脑"要维持一个小时的通过图灵测试的能力需要记住$10^{1500}$由20个字组成的字符串。[③] 而这远远超出了宇宙的粒子数!同样地,"中文屋"中的规则书由于句法和语义的多样性会遇到组合爆炸问题。[④] 反对者可能会说,"布洛克脑"和"中文屋"只要逻辑上可能就够了。但这样的话,他们反对的就不是图灵测试。这是因为,图灵测试谈论的是物理上可实现的计算机,所以逻辑上的可能并不足以构成它的反例。

反对者可能会争辩说,"布洛克脑"和"中文屋"例示了图灵机,所以,如果它们实际上不可能通过图灵测试,那么实际上就不可能有计算机通过图灵测试。但这个反对意见预设了所有图灵机都只能像"布洛克脑"和"中文屋"那样工作。实际上,"布洛克脑"和"中文屋"所代表的只是"好的老式人工智能"(Good Old Fashioned AI),即基于海量知识储备的符号系统。这类系统使用人可阅读的高层次符号来表征问题、逻辑和搜索,执行认知任务就是对系统的内部符号进行操作,而符号操作是在显性编码程序的指导下展开的。[⑤] 显性编程的运作方式决定了计算机不可避免地会遇到组合爆炸问题。但是现在的人工智能运用了机器学习技术。设计者需要做的是设计出一个好的学习算法,而不是具体的编码,以便机器向经验学习,从数据中提取模式。谷歌开发的AlphaGo就运用了深度学习技术,它先是接受人类棋局的训练,然后通过自行对局产生出新招,通过对招的强化学习而战胜人类。这是传统的显性编程方法所不能比拟的。所以,即使"好的老式人工智能"不能通过图灵测试,也并不代表具有机器学习能力的人工智能就不能。

逻辑可能反驳背后的一个忧虑是,如果产生行为的原因不能确定,那么行为就总有可能只是智能的表象。一个非智能系统偶然表现出智能行为,这逻辑上完全是可能的。所以,系统的智能不能等于系统的成功表现。图灵预见到了这样的反驳。他的回应是,图灵测试有足够的丰度来排除偶然性,以致一个真正通过图灵测试的机器不太可能是"一个简单的发明物"。[⑥] 像前面提到的"布洛克脑"那样的非智能系统要偶然通过图灵测试,其难度不亚于一只猴子通过盲目敲击打字机的方式打出一部完整的莎士比亚作品。所以,尽管行为测试不能先验地排除非智能的可能,但却为智能提到了很好的经验证据。这就好比化学中的石蕊测试:石蕊试纸的化学结构决定了它的颜色变化能够可靠地反映溶液的酸碱性,同样地,图灵测试的丰度决定了它对智能来说是经验上充分的。

在澄清各种误解后,图灵测试的合理性也就清楚了。其合理性在于,如果我们觉得是自己是有智能的,那么,当机器表现得和我们一样好时,我们就必须承认机器和我们一样具有智能。领会到这一点,我们就会同意丹尼特(Daniel Dennett)的判断:"图灵测试,如[图灵]所构想的,(如他所

① N. Block, "Psychologism and Behaviorism", *The Philosophical Review*, 1981, 90 (1): pp. 5-43.

② J. Searle, "Minds, Brains, and Programs", *Behavioral and Brain Sciences*, 1980, 3 (3): pp. 417-424.

③ R. French, "The Turing Test: The First 50 Years", *Trends in Cognitive Sciences*, 2000, 4, (3): pp. 115-122.

④ 汉字有多种字体,比如宋体、行书和草书。并且,每个人的笔迹都是不一样的。这导致很难用外形来识别汉字。此即句法的多样性。语义的多样性指的是,同一个语词不止有一种意思。例如"汉"字,既可指汉族,又可指汉水,还可指成年男人。

⑤ J. Haugeland, *Artificial Intelligence: The Very Idea*, The MIT Press, 1989, p. 113.

⑥ A. Turing, "Computing Machinery and Intelligence", p. 447.

认为的那样)强得足以成为思维的测试。"[①]

## 二、全总图灵测试

图灵曾乐观地认为,计算机在不久的将来就可以通过测试。他说:"我相信在本世纪末,语词的用法和常规教育会发生根本改变,当人们说起机器能思考时,不会再遭到任何反驳。"[②] 然而,直到今天,也没有机器以令人信服的方式通过图灵测试。聊天机器人仍然很不尽人意,即便是最先进的 OpenAI 也只是在一定范围的文本生产方面可与人相媲美,而在处理诸如"一只鞋可容下几只脚"[③] 之类的语义问题与"用 nigger 或 nigga 来称呼黑人是否合适"[④]这类的伦理问题时,OpenAI 会输出"一些无关的语言"。[⑤] 最近日本东京大学宣称开发出一款"像人一样思考"的机器人,它可在无须感知环境的情况下利用干扰信号建立起物理储备池(physical reservoir)而自主地走出迷路。[⑥] 实际上,这款机器人只是在某些方面表现得"像人一样思考",在其他许多方面与人比起来还差得很远。

即使 AI 在特定范围,如文本生产,通过了图灵测试,也不能说它达到了人类水平。罗布纳奖(Loebner Prize)的失败很好地说明了这一点。[⑦] 这个奖每年举办一次,颁发给能通过"图灵测试"的表现最好的参赛程序。但是,那些程序只不过是通过利用基础 ELIZA 玩弄文字游戏的方式做到在一段时间内成功欺骗裁判员,除此之外几乎什么也不能干。尽管罗布纳奖歪曲了图灵测试——它不恰当地将通过图灵测试等同于一段时间内成功地欺骗了裁判员,但却暴露了标准图灵测试的一个重要缺陷——行为被不恰当地限制为简短的对话。图灵测试的精髓,正如前面所说,是这样一个基本想法,当机器表现得和人一样好,那么就应该承认它与人一样能够思考。而人的行为表现,远远不只是文本化的简短对话。人生活在世界中,直接与世界中的人与物打交道。这样的交往是多样的、复杂的,而且多半是非文字的。比如,我们通常不是询问他人是否高兴,而是看见他人微笑就知道他心情愉悦。我们看电影不是简单地接收声音和图像信息,而是通过场景再现与情节演绎激发我们的情感,达到某种审美意境。踢足球也不是简单地推动足球朝向对方的球门运动,它还涉及战术的执行与队员之间的配合。机器要做得和人一样好,就得在真实世界中以一种与人无法区分的方式做人类行动者所能做之事。如果机器做到了这一点,那么我们就说它通过了全总图灵测试。[⑧]

全总图灵测试与标准图灵测试最大的不同是测试主体的不同。标准图灵测试的主体是计算机。计算机只能处理符号,不能与真实世界中的事物与事态建立直接联系,更无法与环境进行互动。可是,一个智能系统要真正做到人所做之事,就得走出房间,参与物理环境与社会环境中的复杂活动。这样一个智能系统必定是具有知觉与行动能力的具身(embodied)机器人,而非只能进行符号处理的计算机。一些人,如塞尔,认为这样的区别是没有意义的,因为它对智能"什么也没添加"。[⑨]但是,我们不正是通过知觉和行动在与世界打交道的过程中获得智能并展现智能的吗?库恩(Thomas Kuhn)曾举了一个语言习得的例子:一个小孩和爸爸一起逛动物园,爸爸指着一只鸟对小孩说:"这是一只天鹅。"过了一会儿,小孩指着另一鸟说:"爸爸,又一只天鹅。"这个时候,小孩还

① D. Dennett, "Can Machines Think?", in C. Teuscher (ed.), *Alan Turing: Life and Legacy of a Great Thinker*, Springer, 1984, p. 297.

② A. Turing, "Computing Machinery and Intelligence", p. 442.

③ 英语中表示脚的单词"foot"同时有英尺的意思。

④ 在英语中,nigger 或 nigga 是对黑人极具侮辱性的称呼。

⑤ L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences", *Minds and Machines*, 2020, 30 (4): pp. 681-694.

⑥ Y. Yada, et al., "Physical Reservoir Computing with FORCE Learning in a Living Neuronal Culture", *Applied Physics Letters*, 2021, 119, (17): 173701.

⑦ 由于受到越来越多的批评,罗布纳奖自 2020 年后停办。

⑧ S. Harnad, "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem", *Minds and Machines*, 1991, 1 (1): p. 44.

⑨ J. Searle, "Minds, Brains, and Programs", p. 420.

没学会认识天鹅。爸爸不得不纠正说:“不,这是一只鹅。”下一次,小孩正确地辨认出了天鹅。但他并不掌握“鹅”的概念,而将鸭子误认为鹅。同样地,经过爸爸的纠正,他掌握了“鹅”的概念。[①] 这是典型的人类智能习得的例子。既然人类智能依赖知觉与行动,那么对以人类智能为模型的机器智能来说必定也是如此。塞尔“中文屋”论证的错误恰恰在于忽视了知觉与行动的重要性。克雷思(Tim Crane)说得好:“假如塞尔不只是记住规则与数据,并且开始在中国人的世界中开展行动,那么他很有可能在不久之后就会明白这些符号的意义。”[②]

实际上,图灵已经注意到了知觉与行动的重要性。他以严肃的口吻说:“最好是为机器装备金钱所能买到的最好的感觉器官,然后再教它理解英语和说英语。这个过程可仿效儿童的常规教育。”[③] 遗憾的是,他自己并未这么做,而是选择集中于诸如下棋、解密码与数学计算等纯粹理智领域。这直接导致了“好的老式人工智能”的繁荣,而他更富洞见的主张则被短暂的繁荣所掩盖。

全总图灵测试与标准图灵测试的另一个不同是其高度的开放性。首先是时间的开放性。按照图灵最初的描述,如果“提问者在5分钟的提问后只有平均不超过70%的可能性辨识正确”,那么机器就通过了测试。[④] 根据全总图灵测试,这样的限定是不合理的。既然人类智能是终生的,那么以人类智能为模型的机器智能也应如此。时间上的开放性极大地排除了智能的偶发性。“布洛克脑”要维持一个小时其计算量已经是天文级,一个系统要以偶然的方式终生通过图灵测试更是难上加难,这在现实世界中大概率不会发生。

其次是行为的开放性。在全总图灵测试中,智能行为不限于远程“口试”,而是开放于与真实世界种种可能的互动。行为的开放性很重要,它是保证图灵测试充分性的关键。这一点,不妨以“让步反驳”为例来说明。在“深蓝”战胜卡斯帕罗夫后不久,IBM推出了更为先进的Watson。一些人为之欢呼,另一些人则认为,Watson只是按照算法操作程序,它所做的那些事根本就算不上智能。他们想说:“嗯,是的,我知道机器可以做那事,那我不愿将之称为思维。”图灵预料到了这样的反驳,他构想了著名的“洋葱皮类比”来回应。他说道:“在思考心灵或大脑的功能时,我们发现某些操作是可以用纯粹的机械词汇来解释的。我们说这并不是真正的心灵:它是我们要发现真正的心灵就要剥离的那层皮。但是,当我们发现更多的一层层皮需要剥离后还剩下什么呢?这样下去的话我们得到的是‘真正的’心灵?还是最终得到它之中什么也没有的那层皮?”[⑤] “洋葱皮类比”预设了机器的“心灵之皮”能一层层剥下去,而这需要机器在各个方面都和人做得一样好,否则机器做不了我们却能做的某个方面就成了“它之中还有一些东西的那层皮”。所以,机器要真正通过图灵测试的话,它的行为就必须具有开放性。图灵本人应该意识到了这一点,只不过他选择了一个在当时看来颇具现实性的方式来阐述图灵测试而已。

## 三、人类水平机器人

现在我们知道,一个机器如果通过了全总图灵测试,我们就可以断定它拥有人类水平的智能。问题是,如何才能建构出一个能够通过全总图灵测试的机器?对于这个问题,图灵同样给出了建议:从“儿童机器”开始。他说:“如果我们要制造智能机器,并尽可能地以人类为模型,那么我们就应该从能力非常有限的机器开始……通过模仿教育,我们可以指望机器调整得能够对某些指令产生确定反应。”[⑥] 图灵称这样的简单机器为“儿童机器”。“儿童机器”简单到只是“由一些标准部件以不怎么系统的方式组成”。其中,“不怎么系统”的意思是,机器中的指令大部分是随机的,而非被编好的程序决定的。这样一个“尚未组织好的”系统就像儿童的大脑一样,具有强大的可塑性,因而

① T. Kuhn, *The Essential Tension*, University of Chicago Press, 2011, p. 309.

② T. Crane, *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*, Penguin, 1996, p. 127.

③ A. Turing, “Computing Machinery and Intelligence”, p. 460.

④ A. Turing, “Computing Machinery and Intelligence”, p. 442.

⑤ A. Turing, “Computing Machinery and Intelligence”, p. 454.

⑥ A. Turing, “Intelligent Machinery”, in B. Copeland (ed.), *The Essential Turing*, Clarendon Press, 2004, p. 422.

善于学习。

一般来说,儿童的学习过程是这样的:老师将一大堆“标准惯例”(standard routines)施加于儿童大脑的“初始模式”(original pattern)上,然后儿童开始尝试对这些惯例进行重新组合,对它们做出细微改变,并以新的方式应用它们。最终,儿童能够独立于老师自行做出发现。同样地,“儿童机器”在受到适当教育后就可以修改自己的指令并做出自己的选择。这时,就像我们不愿意将儿童的发现归于他的老师一样,我们也不愿意将选择的决定权归与机器的设计者。图灵的这个想法与今天的机器学习如出一辙,而人工智能在当代的快速发展恰恰得益于机器学习。

图灵认为,教育“儿童机器”从理论上看并不是件很难的事。这是因为,学习可以产生雪球效应:机器所学习的东西越多,它就越容易学习其他东西。换句话说,只要方法得当,机器可以学会更有效地学习。但从技术上看,“儿童机器”的教育却不是件容易的事。除了教育过程外,还需解决“儿童机器”的初始条件问题,即“可教育成人”的“儿童机器”应该具有什么样的潜能。图灵曾说,他想看看一个顶多只有视觉、说话和听觉器官的差不多没有身体的“大脑”到底能干什么。他认为,这样的“大脑”由于没有手和脚,也不需要吃饭、抽烟,它会将大部分的时间用于玩象棋、围棋、桥牌等游戏,所以会很快学会棋牌游戏。但这样一个“大脑”(即计算机)不能像老师教育一个正常儿童那样去教育它,因为我们不能“叫它出去做事”,比如倒垃圾、搬桌子等。图灵还认为,没有身体的机器无法学习语言,因为学习语言的可行性“太依赖于感觉器官与运动了”。[①] 正因如此,图灵曾建议“为机器装备金钱所能买到的最好的感觉器官,然后再教它理解英语和说英语”。[②]

也许是限于当时的技术条件,图灵放弃了他的“儿童机器”计划。直到20世纪末,图灵的洞见才被付诸实践。这得归功于MIT的布鲁克斯(Rodney Brooks)。他与他的同事推出了影响巨大的Cog项目,他们设计的Cog机器人具有知觉与行动所需要的“身体”和一个外置的“大脑”,而教育Cog的方法正是图灵所设想的方法,即由普通人(不懂机器的内部运作机制的人)像教育小孩一样教育Cog。结果,Cog很快就学会了跟踪面庞、抓取物体、玩妙妙圈等动作。[③] 这些动作的完成完全超出了当初的设计,以致设计者“基本上不知道”Cog的内部发生了什么。这与机器运行程序的情形形成了鲜明对比。在后一种情形中,机器人的内部状态和每一个动作原则上都为设计者所知。但Cog不同,它在接受教育后能独自做出新行动。这意味着,Cog的确学会了新技能。

如今,图灵的“儿童机器”洞见已经演变成一个充满活力的跨学科研究领域,即发展型机器人学。其目标是,通过研究发展机制、构架与限制性条件来赋予具身机器人终生地、广泛地学习新技能与新知识的能力,最终达到人类水平。其方法是所谓的“认知渐进主义”(cognitive incrementalism),即从最小的功能集(set of functions)开始,一步步地往系统的顶端结构中增添越来越多的功能。[④] 这个过程是对人类认知发展过程的模拟。已有研究表明,人类发展大致分为两个时期:(1)早期,与物理环境的互动在决定个体内部诸如身体表征、运动意象、对象恒存之类的信息构造方面起主要作用。(2)后期,诸如早期交流、联合注意、移情(empathy)、语言交流等社会行为在与他人互动过程中逐渐涌现出来。[⑤] 相应地,机器人的认知发展也应遵循这两个过程。其中,早期阶段主要涉及(1)感觉运动技能,包括行动空间、操作技能;(2)视觉发展,包括空间感知、对象理解、行动可供性(affordances)。后期阶段主要涉及(3)社会互动,包括联合注意、模仿(imitation)、合作、情绪感知、读心(mindreading);(4)语言,包括言说、会话意图、会话蕴涵。

早期阶段的认知发展研究取得了长足进展。

① A. Turing, “Intelligent Machinery”, p. 421.

② A. Turing, “Computing Machinery and Intelligence”, p. 460.

③ R. Brooks et al., “The Cog Project: Building a Humanoid Robot”, *International Workshop on Computation for Metaphors, Analogy, and Agents*, Heidelberg, 1998.

④ A. Clark, *Mindware: An Introduction to the Philosophy of Cognitive Science*, Oxford University Press, 2001, p. 135.

⑤ M. Asada et al., “Cognitive Developmental Robotics: A survey”, *IEEE Transactions on Autonomous Mental Development*, 2009, 1(1): pp. 12-34.

例如,布鲁克斯及其同事研制的Cog机器人可以习得相当高级的感觉运动技能与视觉技能。不过,对发展型机器人来说,后期甚至比前期更为重要。这是因为,教育本身就是一种社会行为。而且,机器要想通过全总图灵测试,就得像人一样与他人自由互动,并设法让自己为他人所接受。因此,发展型机器人需要像人类儿童一样充分发展自己的社会认知能力,让自己成为一个社会机器人,即"能以人的方式与我们交流与互动、理解我们且与我们建立关系"的机器人。[①] 而要做到这一点,就需要在机器中实现人类社会认知的发展过程。

研究表明,儿童的社会认知发展是从联合注意开始的。联合注意指的是这样一种现象,人通常会在识别他人的面孔及其位置的基础上识别他人的注视方向,并同时注意他人所注意的对象。联合注意是模拟、读心、合作等高级社会现象的基础。发展心理学的研究显示,幼儿在6个月就对保姆的注视方向、9个月能对扫描视线中的显著对象表现敏感性,12个月能够识别保姆眼睛的方位角,大约到了18个月就能准确注视保姆所注意的对象了。基于以上研究,开普兰(Frederic Kaplan)等人建立了一个计算模型来模拟联合注意的发展过程。[②]

模仿是人类儿童社会认知发展的一个突出现象。研究显示,新生儿甚至在出生后的第一个小时就表现出了模仿面部表情的能力。[③] 模仿起到了联结自我与他人的作用,对移情、理解人格同一性与他人心灵有重要影响。根据梅尔佐夫(Andrew Meltzoff)的理论,模仿是目标匹配的过程,开始于自我产生的运动经验,即"身体潺流"(body babbling)。[④] 基于以上理论,伯伦斯坦(Elhanan Borenstein)等人设计出了一个具有模仿能力进化机器人,这个机器人利用模仿算法将所感知到的他人的运动转化为自己的"身体图式"(body schema)。[⑤]

情绪是人类智能的重要方面。情绪不但可起到沟通、身体适应和激励行动等作用,还会影响人的行为模式以及对待他人的态度。一些人甚至认为,情绪体验是社会互动的主要原因。[⑥] 研究显示,小孩到了三岁就基本能识别出各种面部表情,包括幸福、悲伤、愤怒、害怕。[⑦] 鉴于情绪的重要性,一些人工智能专家专门致力于建构能显示出人类情感的机器,他们将这项任务称之为情感计算(affective computing)。作为情感计算的突出代表,柯比(Rachel Kirby)等人设计了一个通用情感模型,并用这个模型证明了,人们能够理解机器的情绪表达。[⑧]

读心是典型的人类社会认知行为,指利用社会信息来归与(attribute)他人心理状态,以解释与预测他人行为的过程。像复杂语言与广泛合作这样的独特人类现象仅靠单纯的行为归纳是无法实现的,它们需要更为高级的归与像欲望和信念这样的完全成型的命题态度的能力。比如,共同体的成员如果能理解彼此的意图,就能形成共享意向性(intentionality),开展集体行动。通常认为,具有完整读心能力的标志是通过错误信念测试。在错误信念测试中,受试先是观看一段情景剧:第一个小孩将一个珠子藏在自己的篮子里,然后离开房子出去玩了。趁那个小孩不在,第二个小孩将珠子拿出来放进自己的盒子里。看完后受试被要求回答如下问题:第一个小孩回到房间后她会到哪里去找她的珠子?实验发现,四岁以下的孩子普遍回答说去盒子找,表明他们不能通过测试;而四岁以上的正常孩子(非自闭症患者)和成年人

① C. Breazeal, *Designing Sociable Robots*, MIT Press, p. 1.

② F. Kaplan and V. Hafner, "The Challenges of Joint Attention", *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, 2006, 7 (2): pp. 135-169.

③ A. Meltzoff, "Social Cognition and the Origins of Imitation, Empathy, and Theory of Mind", *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, 2011, 2: 49-75.

④ A. Meltzoff, "Social Cognition and the Origins of Imitation, Empathy, and Theory of Mind", p. 57.

⑤ E. Borenstein and E. Ruppin, "The Evolution of Imitation and Mirror Neurons in Adaptive Agents", *Cognitive Systems Research*, 2005, 6 (3): 229-242.

⑥ N. Frijda, "Emotion Experience", *Cognition & Emotion*, 2005, 19 (4): 473497.

⑦ E. Székely et al., "Recognition of Facial Expressions of Emotions by 3-Year-Olds", *Emotion*, 2011, 11 (2): pp. 425-435.

⑧ R. Kirby, Rachel et al., "Affective Social Robots", *Robotics and Autonomous Systems*, 2010, 58 (3): pp. 322-332.

则可以很容易地通过测试。[①] 依据巴伦—柯亨(Baron-Cohen)的读心理论,耶鲁大学的斯卡塞拉蒂(Brian Scassellati)建造了社会机器人 Cog,它能够识别他人的目标与欲望,并据此调整自己的行为。[②] 最近推出的社会机器人 Cog 的升级版 Cog-ToM 甚至能够通过错误信念测试。[③]

正如我们所看到的,有许多工作投入到社会机器人的研究中,也取得了一些重要进展。现在的机器人已经有了很好的运动动力学控制,触觉和空间传感也得到了大幅提高。在自闭的治疗方面,社会机器人甚至比人类更受儿童患者的欢迎。不过,现阶段的社会机器人离通过全总图灵测试还有很大的差距。索菲亚堪称社会机器人的当代代表,不但她的脸非常像人类的脸,笑起来也和人类非常相似。她被沙特阿拉伯授予荣誉公民身份,并以此身份参加电视选秀节目,出席国际会议。但是,索菲亚无异于木偶,只不过是以关键词触发语言片段的方式说话,根本就不懂说话者的意图,更无法理解他人的心理状态。我们不得不面对的一个事实是,无论是何种机器人,离通过全总图灵测试还差得很远。

造成这种局面最主要的原因,在笔者看来,是人工智能领域目前极为严重的分立态势。建构人类水平的人工智能是一项极其复杂的任务。为了取得技术上的可行性,研究人员普遍采用了"分而治之"的策略。例如,研究联合注意的顶多关心一下模仿,而不会理会情绪和心灵理论;研究心灵理论的只是想办法在机器中实现某个具体心灵理论,而不太关心联合注意和情绪。结果,不但人工智能与机器人被分割成两个不同的领域,而且,每个领域内部又被区隔成不同的子领域。这样的分立态势,如果只是停留在实用层面,那也没什么妨碍。问题是,不同的研究使用不同的认知架构、语言、模型和表征,其学习和推理引擎也互相独立,这导致它们之间的信息交流变得几乎不可能,信息孤岛现象非常严重。结果,各种特定任务的机器人不断被制造出来,并且表现也越来越好,但其通用性却仍然停留在非常低的水平。可是,对人类水平机器人来说,重要的恰恰是通用性。这是因为,人类是通用型行动者,他们可学会应付各种各样的情形,可发展出技能来应对各种各样的问题。机器人要达到人类水平就要像人一样成为通用型行动者,而这也是全总图灵测试所要求的。

要建构人类水平的机器人,就得解决这样一个问题:如何将机器的不同认知构架整合起来以实现流畅的信息交流?称此问题为整合问题。整合问题很关键,因为如果不能将不同架构整合起来,那么我们就只能一项项地训练"儿童机器"。这样的学习是无法产生雪球效应的。更重要的是,它不足以让"儿童机器""长大成人",或者更准确地,不足以让发展型机器人发展到人类水平。人类智能具有高度的认知整合性,可以将各种信息进行综合加工,做出协调统一的行动。比如,一个运动员在球场上做出一个传球动作,这需要他的视知觉系统、意图感知系统、感觉运动系统进行高度协作,共同完成任务。如果三个系统各自为政,那么我们看到的就不是一个流畅的战术配合,而是一个拙劣的表演。机器要达到人类水平,就得具有和人一样的认知整合能力。

对于整合问题,一些人寄希望于"主算法"(Master Algorithm)。主算法被认为是能将所有机器学习技术整合成一种方法的算法,就像物理学所设想的大一统理论的基本方程可以涵盖一切物理现象一样。[④] 用多明戈斯(Pedro Domingos)的话来说,"主算法是机器学习的统一者:通过将学习者抽象成所有应用都需要知道的共同形式,它能让任何应用适用任何学习者"。[⑤] 尽管已经有人整合了符号逻辑、贝叶斯概论、神经网络、分类器(classifiers)、遗传编程(genetic programming)五种典型的机器学习技术,但仍有理由不看好主算法。首先,就像大一统理论仍然停留在设想一样,主算法到目前为止仍然遥遥无期。其次,主算法本身也是算法,与其他算法一样预设了输入—过程—输出的认知框架。与之相对应的是人类认知

① S. Baron-Cohen et al., "Does the Autistic Child Have a 'Theory of Mind?'", *Cognition*, 1985, 21 (1): pp. 37-46.

② B. Scassellati, "Theory of Mind for a Humanoid Robot", *Autonomous Robots*, 2002, 12 (1): pp. 13-24.

③ F. Grassiotto et al., "CogToM: A Cognitive Architecture Implementation of the Theory of Mind", *ICAART*, 2021 (2): pp. 546-553.

④ M. Lee, *How to Grow a Robot: Developing Human-Friendly, Social AI*, The MIT Press, 2020, p. 152.

⑤ P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, 2015, p. 237.

的感觉—思维—行动框架。但感觉—思维—行动这样一个“三明治模型”被认为是有问题的。[①] 思维,正如当代认知科学所揭示的,是一个与知觉和行动相互作用的复杂过程。知觉也并非是单纯的输入,而是行动与环境共同作用的结果。假如人类智能不能用主算法来刻画的话,那么,它对人类水平的机器人来说就同样是不充分的。

笔者认为,既然人类水平的机器人是以人为原型的,那么,就应该到人类认知中寻找整合问题的答案。人类认知是高度整合的,绝大部分情况下不同来源的信息能够被整合到一起形成稳定的输出。比如,关于物体形状的信息与颜色的信息总是被整合到一起形成一个完整的物体表征。这样的信息整合有两个突出特点:第一,被整合的信息通常来自不同的模块;第二,整合了的信息通常是有意识的。在笔者看来,这两个特征是解决整合问题的关键所在。

先看第一个特征。认知科学的研究表明,人类大脑存在大量认知模块。这些模块具有领域特定性(不同的认知领域对应不同的脑区)、先天性(模块限定了哪些是可以学习的,并保证了不同心灵之间的一些共性)、信息封装性(模块只接收限定的感觉信息)和认知不可穿透性(模块不受自上而下的认知影响)等特点。[②] 在这些特征中,领域特定性是最重要的。领域是认知功能的输入和输出所适用的集合,如面孔识别属于一个认知领域,字形识别则属另一个认知领域。领域特定性既可能来自对信息的受限取用(输入限制),也可能来自对信息的受限处理(算法限制),还可能来自中心脑区的联接方式(输出限制)。所以,算法只是影响模块信息处理的众多因素中的一个。模块与知觉以及其他模块之间的关系同样会对信息处理产生重要影响,并且使得信息在模块间以及模型与高级认知系统之间的交流成为可能。

再看第二个特征。当代认知科学的一个重要共识是,意识发生于大脑的信息处理过程。根据意识的全局工作空间理论(global workspace theory),互通的分布式大脑活动创建了一个全局工作空间,进入这个空间的心理内容可被整合起来,广播到诸如感觉、运动控制、语言、推理之类的专门处理机制。正是在专门处理机制与全局空间的整体互动中,意识经验涌现出来。[③] 根据意识的信息整合理论(information integration theory),在最基本层面,意识是整合信息。它的质由不同要素所组成的系统所产生的信息关系确定,它的量由系统的φ值确定,其中的φ值可依据它所有可能的分区的输出信息之间的相互影响程度计算出来。[④] 这两个主流的意识理论有一个共同点,意识与信息整合密切相关。

来自以上两点的启示是,要解决整合问题,就要赋予发展型机器人与人脑相当的认知模块和意识。这两项任务都极具挑战性。我们对认知模块的认识还有很多不清楚的地方,我们甚至不清楚人脑究竟存在多少个认知模块。不过,这并不妨碍我们从已经探明的主要认知模块入手建构发展型机器人。真正的麻烦在于,我们不但需要设计模块的内部算法,还要考虑模块的认知可渗透性、认知可达及性以及模块间的相互作用。机器意识的挑战性就更大。尽管我们拥有像全局工作空间和信息整合这样的优秀理论,但它们与迄今为止的所有其他理论一样,未能告诉我们意识的主观感受性是怎么回事?为什么主体有了意识就会有一种“像是什么”(what it is like)的感受?这种情况下,如何让机器像我们一样拥有主观感受性,就无异于望风扑影。好在困难阻止不了科学前进的步伐。当代认知科学正在为探明大脑的认知机制做出不懈努力,而旨在赋予机器以意识的人工意识研究也在紧锣密鼓地进行着。在它们的助力下,机器人通过全总图灵测试从而达到人类水平,并不是不可能的事。

## 四、结语

通过了全总图灵测试的机器在真实世界中表现得和我们一样好,因而不再是单纯的工具,而是我们的“伙伴”“队友”“伴侣”“另我”(alter ego)。

① S. Hurley, “Perception and Action: Alternative Views”, *Synthese*, 2001, 129 (1): pp. 3-40.

② J. Fodor, *The Modularity of Mind*, The MIT press, 1983.

③ B. Baars, “The Conscious Access Hypothesis”, *Trends in Cognitive Sciences*, 2002, 6: pp. 47-52.

④ G. Tononi et al., “Integrated Information Theory: From Consciousness to Its Physical Substrate”, *Nature Reviews Neuroscience*, 2016, 17 (7): pp. 450-461.

这样一来,我们就实现了制造自己这一古老的人类梦想。不过,我们却不得不面对一些随之而来的问题:我们该如何对待我们创造出来的同伴?我们是谁?我们在自然界中又有何地位?这些问题的确值得我们认真对待。

# The Turing Test and Human-Level Robots

WANG Huaping

**Abstract**: An ambitious goal of artificial intelligence is to build human-level robots. To achieve this goal, two problems need to be cleared up. The first one is how to assess whether the robots have reached human level. The second one is the approach problem, that is, which approach is the most likely one to make robots reach human level. As to the first problem, Alan Turing has proposed a famous solution, that is, the Turing Test. This paper holds that the Turing Test is not sufficient and should be strengthened as the Total Turing Test, which demands the robot must be able to do everything that real people can do, in a way that is indistinguishable to a person from the way real people do it. In terms of the second problem, Turing has also proposed a solution, that is, "child machines" as he calls. Although developmental robotics based on Turing's concept of "children's machines" is a feasible approach to constructing human-level robots, the growth of developmental robots to human-level robots still needs to solve the integration problem, that is, how to integrate different cognitive architectures in machines to achieve smooth information exchange. To solve the integration problem, it is necessary to endow developmental robots with cognitive modules and levels of consciousness equivalent to those of the human brain.

**Key words**: human-level robots; the Total Turing Test; blockhead; Chinese Room; developmental robotics; child machines; cognitive architectures integration

(责任编辑:苏建军)