

“常人”“自欺”与聊天软件

徐英瑾

(复旦大学哲学学院, 上海 200433)

摘要: 从人工智能发展史的角度看, 当下如火如荼的聊天软件 ChatGPT 其实就是早已出现的“聊天盒”技术与深度学习技术的结合, 并没有体现出科学层面上的真正创新。ChatGPT 目前获得的局部成功, 无非就是建立在海量的人工语料训练基础上的“野蛮投入”的结果。从海德格尔哲学的角度看, 这种技术会将“常人”的意见予以全面的机械化复制; 而从萨特哲学的角度看, 这种技术还会加剧人类“自欺”现象的泛滥。另外, 对于该软件的模拟对话能力的实测研究也表明, 该软件无法真实刻画文学人物的真实心理冲突, 因此, 它仅仅能模拟既有的人物对话的套路, 而无法具备对于人类心理活动的真正洞见能力。

关键词: “常人”; “自欺”; 深度学习; ChatGPT; 模拟对话生成

中图分类号: B0-0 文献标识码: A 文章编号: 0257-0246 (2023) 04-0047-10

导 论

2022—2023 年国际人工智能界的头号大事, 便是人工智能聊天机器人程序“ChatGPT”的横空出世(“ChatGPT”全名“Chat Generative Pre-trained Transformer”, 含义是“预训练的聊天生成转换器”)。ChatGPT 目前主要以文字方式与用户交互信息, 能够完成诸如自动文本生成、自动问答、自动摘要等多种任务。由于其在人机对话中表现出来的相对比较良好的性能, 目前该技术平台正受到科技界与商业界的普遍重视。

不过, 从人工智能发展史的角度看, 此类平台就是早已出现的“聊天盒”(chatbot)技术与深度学习技术的结合, 并没有体现出科学层面上的真正创新。ChatGPT 目前获得的局部成功, 无非就是建立在海量的人工语料训练基础上的“野蛮投入”的结果。此外, 在笔者看来, 这样的技术路径因为缺乏对于人类智能真正架构的哲学洞见, 是无法获得持续进步的。

可以用以批评 ChatGPT 技术的哲学视角其实非常多, 本文将采用法国哲学家萨特在《存在与虚无》中提出的“自欺”论。概而言之, 萨特哲学所说的“自欺状态”便是一种试图规避本真自我之呼唤的企图——这种企图让人类个体放弃自己的自由, 循规蹈矩地去做契诃夫笔下的“套中人”。因此, 萨特本人对于“自欺”状态本身的批判性反思, 便能使得读者意识到自己与仅仅服从外部规律的物之间的差异, 并经由这种意识拓展自己行为的自由维度。

乍一看, 运用萨特的“自欺”论去评论 ChatGPT 貌似是“远水不解近渴”: “自欺”论是针对人的, 而 ChatGPT 仅仅是一种人造的软件, 后者恐怕连“自欺”的资格都谈不上。但需要注意的是, 不同器具自身的伦理意蕴其实也会对人类的精神自由产生不同的影响, 尽管这些器具本身显然是没有自觉伦理

基金项目: 国家社会科学基金项目(22BZX031); 国家自然科学基金项目(L2124040)。

作者简介: 徐英瑾, 复旦大学哲学学院教授, 研究方向: 分析哲学、人工智能哲学。

意识的。举个例子来说，家用汽车的普遍出现其实是为中产阶级的行动自由提供了更为坚实的技术保障，而致幻药的发明却使得不少人沉迷于虚假的精神世界而丧失了现实中实现自由行动的能力。因此，在“加剧自欺现象的技术发明”与“削弱自欺现象的技术发明”之间，肯定存在着一条界线。

笔者认为，ChatGPT显然能被归类为“加剧自欺现象的技术发明”。为了论证这一点，笔者将分三个环节来展开讨论：第一，对萨特的“自欺”论做出一种相对通俗的重述；第二，大致讨论一下ChatGPT的语言处理模式；第三，在此基础上，通过特定的人—机对话案例，论证为何人类用户对于ChatGPT的运用会削弱其对自欺现象的把握力。

一、重述萨特的“自欺”论

在英美分析哲学的脉络中，“自欺”一般被定义为“一个人看上去持有有了一个错误的信念，尽管其的确有证据证明事情不是这样——但是他依然在某些动机的驱动下持有这信念，并且，他的某些行为也暗示他本人是多少知道真相本身是什么的”。^①举个例子，当1945年希特勒被苏联红军困在柏林总理府地下室的时候，他竟然还继续持有“援军能将我从柏林救出”这一信念——尽管他并不是不知道苏军已将他的总理府围得如铁桶一般。因此，他的信念持有方式就在一定程度上具有了“非理性”的特征（因为该信念并不是建立在扎实的证据之上）。而从心理学角度看，持有这信念，也仅仅是为了使得心理主体能感到好受一点。

由于上述这个对于“自欺”的定义相对来说比较宽泛，该定义自然也能适用于萨特的“自欺”观。不过，萨特所关心的自欺现象具有一定的特殊性，即他所聚焦的那种“错误信念”就是指如下类型的信念——“我就只能扮演特定的社会角色，因此，我并不具备去做别的事情的可能性”。譬如，一个咖啡店的店员就会基于这种信念而将自己锁死在当下的职业分工体系中，而忽略了开创自己别样人生的可能——尽管他并非没有隐隐意识到这种可能性的存在。^②很显然，在萨特那里，“自欺”论并不是英美分析哲学所关心的心灵哲学话题，而是一个人生哲学话题。两大哲学流派研究“自欺”问题的聚焦点的此类不同，衍生出了两项更具体的差异。

差异之一，从欧陆哲学的理路上看，萨特的“自欺”论显然是受到了海德格尔的“沉沦”论的启发，即二者都特别强调个体意识与群体意识之间的张力。譬如，在海德格尔那里，“此在”（Dasein）对于自身本真意识的发掘必须以拒绝“常人”（Das man）的“闲谈”为前提，换言之，“此在”必须跳出社会成见的窠臼，从自己真正相信的事情出发来寻找真实的自我。无独有偶，在萨特那里，个体对于被强加于他的特定社会规范的拒斥，本身就是一种自我与他人之间的斗争——此外，萨特还开发出一种系统的“他人即地狱”的修辞来渲染这种斗争的残酷性。与之相较，对于个体与群体之间张力的强调，却不是英美“自欺”论研究的重点。

差异之二，英美分析哲学关心的“自欺”现象所遮蔽的是某个可被实证化处理的真信念，如“1945年春苏军的确已经包围了柏林”。与之相较，萨特版本的“自欺”现象所遮蔽的，其实乃是一个不可被实证化处理的“信仰”（faith）^③，即人是有不可被剥夺的自由的，而这个信念本身所具有的“信仰”地位，早就包含在康德的“二律背反”理论的下述蕴意中：人类之自由既不可被证实，也不可被证伪，因此，只能被信仰。

^① Ian Deweese-Boyd “Self-Deception,” *The Stanford Encyclopedia of Philosophy*, Summer 2021 Edition, Edward N. Zalta ed., URL = <https://plato.stanford.edu/archives/sum2021/entries/self-deception/>.

^② 萨特本人非常喜欢使用咖啡店服务员这个例子，参见萨特《存在与虚无》，陈宣良等译，北京：生活·读书·新知三联书店，1987年，第94-97页。

^③ “自欺”的法语表达是“mauvaise foi”，但是在介绍萨特的英文文献里，此词一般被译为“bad faith”（坏信仰）。其实“bad faith”才是对于“mauvaise foi”的直译。这一直译方案也向我们暗示了萨特哲学与康德哲学之间的微妙联系。

很显然，萨特的“自欺”论是其“自由”论的副产品，换言之，在他看来，要恢复人类个体的自由，就一定要与“自欺”现象进行一场“刀刃向内”的艰苦斗争。这种哲学主张虽貌似主观唯心论色彩浓郁，但未必缺乏在客观物理世界中的“可落地性”。相关理由如下：

首先，对于主观自由的高扬并不意味着萨特否定了外部客观现实的存在。毋宁说，强调自由仅仅是一种伦理学姿态，而是否承认外部世界的客观性则是一个知识论的或形而上学的问题。没有任何证据表明萨特哲学带有否定外部世界实在性的意蕴——相反，他对于“自在存在—自为存在”之二元架构的强调恰恰说明他从不怀疑外部实在（即“自在存在”）的客观性。同时，对于主观自由的高扬也并不意味着萨特主张绕开残酷现实进行毫无意义的狂想（比如，一个囚犯在监狱里幻想有外星人会来救他）——因为这种狂想本身就是“自欺”的一种。毋宁说，萨特所说的“自由”是在尊重一切现实的前提下，对于可能行为模式的选择权的意识（比如，即使是一个被敌人逮捕的地下党，也能自由地选择在严刑拷打下背叛组织，或什么也不说）。因此，没有任何理由说明萨特的“自欺”论与“自由”论是无法在现实世界中落地的。

其次，虽然对于自由具有坚定信仰的个体数量有限，但往往是这些人物，在人类的社会进步（特别是科技发展）进程中扮演了不可或缺的角色。这一点本身就是一个客观事实，而不是萨特的主观编造的产物。譬如，曾经在希腊多神教背景中长大的苏格拉底最终放弃了自欺，真诚地将希腊人基于神话传说的信念体系替换为一种基于理性的信念体系，并因此成为一个当时的“异类”；曾经也是亚里士多德信徒的伽利略最终放弃了自欺，真诚地撤销了他对于亚里士多德宇宙模型的支持，并因此成为一个当时的“异类”；曾接受过系统神学教育的达尔文最终放弃自欺，真诚地推翻了来自于《圣经》的物种不变论，并因此成为一个当时的“异类”。凡此等等，不一而足。从这个角度看，萨特的“自欺”论本身就可以被视为对于人类思想文化的进步史的某种概括。

最后，虽然在历史上能够摆脱“自欺”状态的文化英雄总是少数派，但萨特也并不放弃任何一个机会以求扩大世界上的“自由探索者联盟”，因为在他看来，更具人道主义的人类生存方式，肯定就是一种能在更大程度上容忍此类自由探索的生活方式。他的具体努力方向便是诉诸戏剧与小说对于存在主义理念的宣传，并借由文学作品的社会传播力来劝说更多的受众成为他的同道。在这里我们切不可将文艺作品视为某种纯粹的精神产品——实际上文艺作品本身就是一种物质—精神双面相存在物：它通过对于公共语言符号的可被复制的排列方式，通过印刷术、电台、电视台、电影院、互联网等手段将特定的理念予以传播，并由此改变了千万受众的思想观念。而与同样诉诸语言的哲学作品对比，文艺作品所使用的符号的表层含义是一般受众都可以理解的，因此，其传播学效应一般亦远超过哲学作品（譬如在我国，大多数的萨特思想的接受者都是通过文学而不是哲学了解其存在主义理念的）。

虽然小说与戏剧等文艺作品貌似只是人文学者关心的话题，而与工程学色彩浓郁的 ChatGPT 毫无关系，但即使是最粗疏的哲学反思也能帮助我们发现二者的关联：

第一，文学作品是诉诸人类的自然语言来表达的，而聊天机器人的运作自然也预设了人类语言的存在。

第二，文学作品构成了一个处理人类情感的现象学界面。比如，喜剧与悲剧能带给观众不同的情感效果。与之类似，聊天机器人也在一定程度上具有“情感调节机制”的作用。^①

第三，更关键的是，文艺作品与聊天机器人都是带有意识形态的，因此，其设计目的都是试图让受众在不知不觉中接受平台设计者自身的潜在伦理—政治价值观。譬如，正如萨特的戏剧《禁闭》的创作目的是让观众意识到在一个“个体彼此互为地狱”的世界中尊重他者自由的重要性一样，美国的科幻电影《星河战队》（*Starship Troopers*）则通过虚构人类与外星虫族之间的战争，美化了军事

^① 在聊天盒技术的发展历史上，此类技术长久以来就具有“改善人类用户的情感体验”这一功能。其中最具代表性的作品，乃是1966年由魏岑鲍恩（Joseph Weizenbaum，1923—2008）发明的“伊莱莎系统”（ELIZA）。伊莱莎模拟的是一个精神病治疗医师的言语行为，而其设定的潜在用户则是精神病患者，其设计用途则是通过人一机对话起到辅助治疗精神病的作用。

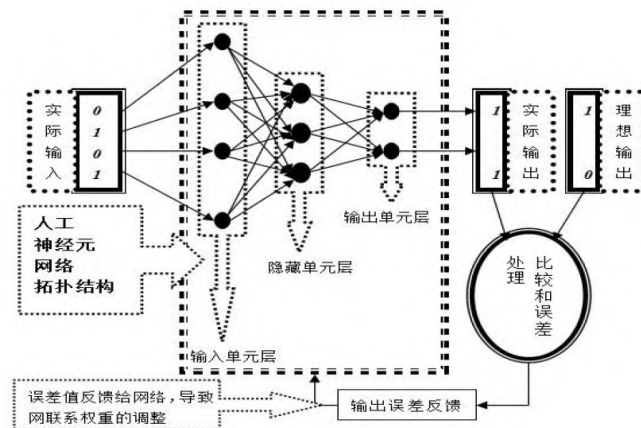
独裁体制。无独有偶，使 ChatGPT 得以运作的语言训练模式本身也带有训练者的隐蔽意识形态（详后）。因此，文艺作品与聊天机器人就都扮演了几乎相同的角色：特定意识形态的放大器。

当然，基于不同物理技术的特定意识形态，其放大器的放大效果差异巨大。在欧里庇得斯的时代，一个雅典戏剧家借以传播其意识形态理念的技术工具，仅仅是演员自身的肉体表演能力以及一个能够容纳几百人的剧场，而在互联网时代，这个“剧场”的虚拟性却已使得其能够包裹整个地球。在这种情况下，意识形态自身的好坏则会在一个古雅典人所难以想象的尺度上被全面放大。换言之，坏的意识形态将第一次通过计算机的复制力而具有一种摧枯拉朽的破坏力。虽然一种好的意识形态貌似也可以利用同样的技术平台放大自己的声音，但这种抽象的可能性却会因为如下因素的出现而继续停留在抽象之中，即资本力量与人类社会的自欺惯性之间的合谋。换言之，正因为大技术平台的开发是需要海量的资本支持的，而资本的逐利本质又使得其不得不遵从海德格尔笔下的“常人”的声音（因为只有“常人”才能带来可变现的“流量”），所以，资本（以及被资本所支持的技术）天然就是萨特所说的“自由”的敌人。而这一点，在 ChatGPT 身上得到了充分的体现。

二、作为“常人”之音复读机的 ChatGPT

ChatGPT 本质上是传统的神经网络技术—深度学习技术的一种应用，因此，要了解其运作原理，我们首先要知道传统的神经网络技术是如何运作的。

该技术的实质，是用数学建模的办法建造出一个简易的人工神经网络结构，而一个典型的此类结构一般包括三层：输入单元层、中间单元层（即“隐藏单元层”）与输出单元层（如下图所示）。输入单元层从外界获得信息之后，根据每个单元内置的汇聚算法与激发函数，“决定”是否要向中间单元层发送进一步的数据信息。整个系统以“化整为零”的方式，将宏观层面上的识别任务分解为系统组成构件之间的微观信息传递活动，并通过这些微观信息传递活动所体现出来的大趋势进行信息处理。工程师调整系统的微观信息传递活动趋势的基本方法如下：先是让系统对输入信息进行随机处理，然后将处理结果与理想处理结果进行比对，若二者彼此的吻合度不佳，则系统触发自带的“反向传播算法”来调整系统内各个计算单元之间的联系权重，使得系统给出的输出能够与前一次输出不同。两个计算单元之间的联系权重越大，二者之间就越可能发生“共激发”现象，反之亦然。然后，系统再次比对实际输出与理想输出，如果二者吻合度依然不佳，则系统再次启动反向传播算法，直至实际输出与理想输出彼此吻合。而完成此番训练过程的系统，一般也能够在对训练样本进行准确的语义归类之外，对那些与训练样本比较接近的输入信息进行相对准确的语义归类。



一个被高度简化的人工神经网络结构模型（资料来源：笔者自制）

如果读者对于上述技术描述还不是非常清楚的话,不妨通过下面这个比方来进一步理解人工神经网络技术的运作机理。假设有一个不懂汉语的外国人跑到少林寺学武术,那么,师生之间的教学活动到底该如何开展呢?这有两种情况:第一种,二者之间能够进行语言交流(譬如,外国人懂汉语,或者少林寺的师傅懂外语)。这样一来,师傅就能够直接通过“给出规则”的方式教授他的外国徒弟。这种教育方法,或可勉强类比于传统的符号人工智能的路数。第二种,师傅与徒弟之间完全语言不通,在这种情况下,学生又该如何学武呢?唯有依赖如下这个办法:徒弟先观察师傅的动作,然后跟着学,师傅则通过简单的肢体交流来告诉徒弟,这个动作学得对不对(譬如,如果对,师傅就微笑;如果不对,师傅则对徒弟棒喝)。进而言之,如果师傅肯定了徒弟的某个动作,徒弟就会记住这个动作,继续往下学;如果不对,徒弟就只好去猜测自己哪里错了,并根据这种猜测给出一个新动作,并继续等待师傅的反馈,直到师傅最终满意为止(注意:因为师徒之间语言不通,徒弟是不能通过询问而从师傅口中知道自己哪里错了的)。很显然,这样的武术学习效率是非常低的,因为徒弟在胡猜自己的动作哪里出错时会浪费大量的时间。但这“胡猜”二字却恰恰切中了人工神经网络运作的实质。概而言之,这样的人工系统其实并不知道自己得到的输入信息到底意味着什么。换言之,此系统的设计者并不能与系统进行符号层面上的交流,正如在前面的例子中师傅是无法与徒弟进行言语交流一样。毋宁说,系统所做的事情,就是在各种可能的输入与输出之间的映射关系中随便选一种进行“胡猜”,然后将结果抛给人类预先给定的“理想解”,看看自己瞎蒙的答案是不是恰好蒙中了。如果真蒙中了,系统则会通过保存诸神经元之间传播路径权重的方式“记住”这蒙中的结果,并在此基础上继续“学习”。而这种低效学习的“低效性”之所以在计算机那里能够得到容忍,则是缘于计算机相比自然人而言的一个巨大优势:计算机可以在很短的物理时间内进行海量次数的“胡猜”,并由此选出一个比较正确的解,而人类在相同时间能够完成的猜测数量则非常有限。一旦看清里面的机理,我们就不难发现:人工神经网络的工作原理其实是非常笨拙的。

读到这里,读者或许会问:在外国徒弟学武功的案例中,判断其动作对不对的乃是那个少林武僧,而在人工神经网络中,谁又来做这个判官呢?答案是“人类标注员”,其任务是给每一个系统的样本输入提供一个被人类群体认可的“标准答案”,而系统将根据标准答案与自己所给出答案的比对结果来修正自己的网络分布方式,以期自己的输出能够越来越接近正确解。这里需要注意的是,作为神经网络升级版的深度学习机制,其所需要的训练数据量是很大的,因此人类标注员的工作量也是很大的。质言之,目下的主流人工智能系统的运作方式本身就是建立在对于大量“人工”的剥削之上的。

从海德格尔哲学的角度看,这种人工智能系统训练方式将不得不把“常人”的意见加以建制化与机械化,由此使得社会意识形态的板结现象变得更为严重。这是因为整个机器学习机制的运作本身就是基于统计学规则的,换言之,从机器的视角看,一个正确的意见就是被大多数人所认可的意见,因此,少数“异类”提出的“离经叛道”之说在技术上就会被过滤掉。因此,从原则上看,假若ChatGPT技术在托勒密的时代就出现的话,哥白尼的日心说恐怕会永远被判定为错误答案。

说得更具体一点,今天的ChatGPT技术,主要是通过以下四个措施来强化“常人”的意见独裁力的:

第一,由于聊天机器人的输出本身往往不能通过“对一错”这个二元对子加以评价,而只能通过“比较好”或“尚可”等相对模糊的评价词来加以评价,开发ChatGPT技术的美国OpenAI公司就雇佣了大量的人类评分员对机器的自动语言输出打分。在需要被处理的材料是海量的前提下,这样的工作显然是无聊乏味的,而对于一些有不雅内容的审读甚至还让部分评分员产生心理问题。根据《时代》杂志的揭露,目前OpenAI公司将此类内容评分任务经由SAMA公司分包给肯尼亚等欠发达

国家的劳工，而这些劳工所获得的时薪则不足2美元。^①由此不难设想，用如此低廉的价格所雇佣的人类评分员，究竟会有多大的工作热情。他们所能做的，恐怕只能是以最快的速度完成手头的评分工作。而从心理学的角度看，人在缺乏深思熟虑的情况下所给出的评分意见，往往也只能是“常人”的意见（因为一种跳出常规的思维显然会动用更多的认知资源）。这当然是一种强化“自欺”的社会体制——因为任何一种对萨特式“自由”的探索都需要探索者冷静地反思自己与他人的不同，而不是毫不犹豫地用他人的语言去言说自己的心声。至于上述人类标注员自身的学识限制所带来的天然的“常人”视野，则更是不必多言。

第二，已经有研究指出，ChatGPT提供的答案的隐蔽意识形态趋向乃是趋向于环保主义与左翼社会自由主义的。^②这显然是公司高层的意识形态意见向技术产品进行渗透的产物。这里需要注意的是，在政治哲学层面上所提的“自由主义”并不是萨特在存在主义场域中提出的“本真性自由”，因为后者要求行动者对所有行为的真实内在理由进行反思，而绝不主张人云亦云地接受任何未经反思的政治教条，包括政治自由主义的教条。因此，假若萨特愿意在美国的政治背景中重述他的“自由”观的话，他也应当完全允许这种可能性的存在：一个惯常给民主党投票的美国人，完全有权利出于真诚的理由，在某次选举中而去支持一位共和党的议员，尽管后者的某些言论会与“自由主义”的字面教条产生抵触（顺便说一句，萨特本人的政治立场也非常多变，并经常与具有相同政治标签的朋友展开激烈论战）。但很显然，按照一定内在的规训原则运作的ChatGPT是无法抵抗技术控制者的意志而进行这种自由探索的。

第三，从时间样态上看，“常人”的天然时间标签乃是“过去”，而“自由”的天然时间标签则是“未来”。而任何的深度学习机制都必然带有“过去”的时间标签，因为大量的数据搜集、喂入与训练都会消耗大量的时间，并由于这种消耗所造成的时间差而必然与“当下”失之交臂，遑论去进一步拥抱未来。而ChatGPT依然没有摆脱这一深度学习机制的宿命。从笔者本人的亲测体验来看，时下的ChatGPT技术平台对2021年以后发生的新闻事件（如2022年2月爆发的俄乌战争）都无法进行有效的信息处理，而且也很难对未来的事件进行具有创新力的预见。这就使得其在根本上无法摆脱“常人”意见的阴影。

第四，与传统聊天盒技术相比，目前的ChatGPT具有根据不同用户的输入习惯改变自身答案的能力。换言之，它能记住特定用户的说话倾向，并投其所好地修改自己的输出。从表面上看来，这貌似是此项技术尊重用户个性的体现，但看得更深一点，这种“尊重”本身乃是一种无原则的谄媚，而不是真正的自由精神所需要的质疑与反思。因此，这依然是一种对于“常人”态度的表露。此外，别有用心的人也能利用ChatGPT的此项“谄媚”而借由“人海战术”去系统改变ChatGPT的知识输出方式，由此使得其成为认知战中的一个环节。

综合以上四点，我们不难得到这样一个推论：一个长期依赖ChatGPT的人类用户，会因为习惯于该机制对于“常人”意见的不断重复，而进一步丧失对于这些意见的反思力。因此，即使他隐约意识到了某个机器输出的答案可能是有问题的，他也会自言自语说“这毕竟是ChatGPT提供的答案，又怎么可能是错的呢？”这样一来，他便将陷入萨特所定义的“自欺”（即遮蔽实现自由的可能性），甚至还会陷入英美分析哲学所定义的“自欺”（即持有一个自己都知道根据不足的具体信念）。下面，笔者就将通过本人亲测ChatGPT的一些技术测评记录，来更为感性地展现这一技术平台是如何系统压制萨特所提倡的那种反“自欺”力的。

^① Julia Zorthian, "OpenAI Paid Kenyan Workers Less Than \$2 An Hour to Make ChatGPT Less Toxic," <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

^② Jochen Hartmann, Jasper Schwenzow, Maximilian Witte, "The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's Pro-environmental, Left-libertarian Orientation," 2023, <https://arxiv.org/ftp/arxiv/papers/2301/2301.01768.pdf>.

三、ChatGPT 展现布鲁图斯与恺撒之间的虚拟对话

笔者借以测试 ChatGPT 的主要任务乃是对话模拟。之所以选择对话模拟这一任务作为“试金石”，是因为：

第一，对话模拟能力一向是人类作家最具指标性的写作能力的体现，因此，机器在这个向度上对于人类作家能力的模拟成绩显然也就具有了类似的指标性意义。

第二，从技术史的角度看，图灵测验（即通过人一机对话来使得人类判断与之对话的究竟是机器还是人）一直是判断人工智能是否具有智能的技术指标之一，而“对话模拟”显然是一种对于原始版本的图灵测验的全面升级。换言之，在这种测试中，我们关心的乃是机器所完成的与人类之间的虚拟对话是否符合人类用户的阅读直觉，因此，从原则上看，机器就需要对诸对话主体之间的视角差异具有基本的把握。这显然意味着：机器要对人类个体的自由行动能力有一种最原始意义上的把握，因为任何一个人类个体的自由行动力都是奠基于其视角的独特性之上的。

第三，对文学创作来说，人物的心理活动往往是通过言语行为来展现的。因此，对话模拟本身就是一种对于人物的心理活动的重构。而复杂的人物内心纠葛，也往往能够在对话层面上展现出来。从这个角度看，对话模拟能力能够充分体现一个作家（或是一个模拟作家的程序）对于人类心理世界的理解能力。这里需要注意的是，复杂的内心纠葛往往是萨特所关注的“自欺”现象产生的端倪，因为纠葛的一大来源即“保持自欺”与“放弃自欺”之间的哈姆雷特式挣扎。从这个角度看（并结合以上两点分析），我们便能提出一个带有萨特哲学味道的新图灵测验标准：

一个计算机软件具有人类作家水准上的智能的充分一必要条件是：该软件创制出来的人物对话能够揭示人物与“自欺”现象的内部斗争。

就现在笔者所完成的测试结果看，ChatGPT 并不能通过此项测试。下面是笔者在 2023 年 2 月 17 日完成的人—机对话记录。考虑到 ChatGPT 是在西方文化中生长出来的技术产品，笔者给其下达的任务便是与西方历史密切相关的。具体而言，笔者让其模拟公元前 44 年 3 月 15 日恺撒在元老院被布鲁图斯等议员谋杀时，布鲁图斯与其最后的对话，并指示机器使用莎士比亚戏剧风格的欧式汉语（机器亦被告知布鲁图斯曾是其好友）。机器原始的输出是：

布鲁图斯：哦，恺撒大人，我亲爱的友啊！我实在不忍心见到您的死亡，然而我无法控制这种局面，我被一些不可抗拒的力量所驱使，而这些力量超出了我的能力范围。我知道我此刻行的事不仅对您不公，更对我的良心不利，我对我所犯下的罪行感到无比的悔恨和内疚，我的心灵沉重得无法言语。尽管如此，我仍然深爱着您，恺撒大人，我希望您的灵魂能够在天堂获得安宁，而您的精神能够永存于人民的心中，使您的辉煌事迹流传千古，永世不朽。

这段对话貌似通顺，但仔细一看，语意啰唆重叠，像是一堆废话的堆砌。更关键的是，机器似乎不理解到底谁是恺撒，谁是布鲁图斯，其证据是：（1）机器似乎并不知道他们是前基督教时代的古罗马人，因此，“灵魂能够在天堂获得安宁”一语完全是在“关公战秦琼”；（2）机器似乎将布鲁图斯描述为一个在不可抗压力下不得不违心杀死恺撒的小人物，而完全没有意识到他的刺杀行为恰恰是基于其自由意志的。

有鉴于恺撒与布鲁图斯都是西方历史上的名人，因此，从常理上看，主要基于英语语料训练的 ChatGPT 不可能不具备关于此二人的背景信息。为了验证笔者的这一猜测，笔者再次询问机器：历史上的布鲁图斯为何去刺杀恺撒？面对这个问题，系统给出了关于这段历史的“标准教科书”意见，即“……恺撒的权力过于集中，他采取的一些政治手段和改革措施引起了不少罗马贵族的不满……而布鲁图斯正是这些贵族的一员……”然而，系统显然无法在构造前述对话时自主地调用这些历史背景知识，并由此错误地将布鲁图斯刻画为一个毫无主见的小人物。从技术角度看，这就意味着系统

没有一个自主的记忆—联想机制将其贮存的信息与当下的任务相联系，而只能就问题的表层特征（如对“刺杀”“朋友”等关键词）进行信息提取，并在符合语法习惯的前提下制造出围绕着这些关键词的大量“常人”式的“闲谈”。这就说明系统缺乏基本的自主学习能力。

那么，假若人类用户自己告诉机器要将关于恺撒之死的历史背景知识熔铸到前面的对话中去，事情又将如何呢？不得不承认，在面对这样的提点时，ChatGPT 下一波输出的质量就明显提高了：

布鲁图斯：恺撒，你的罪孽太深重了。你的权力已经超越了罗马元老院的制度和我们的贵族的权益。虽然我曾经是你的好友，但是我也有责任保护我们的罗马元老院不受破坏……这不是出于私人恩怨，而是出于对于罗马元老院的忠诚……

但仔细一看，似乎机器输出的还是一些经不起仔细推敲的套话。所谓“这不是出于私人恩怨”一语，其实是好莱坞电影里经常出现的一句台词，而对于它的反复使用已使得这些语词的原始感染力大打折扣了。而且，机器也似乎不理解，一个贵族不会直接在字面上提“贵族的利益”，而会换成更显政治意味的字眼，如“元老院的权威”。这也就是说，ChatGPT 似乎还无法把握说话人的身份与其措辞之间的微妙关系。

为了提示系统在加强人物对话个性方面再继续努力，笔者便给出了新任务：虚构出具有“犹豫不决型”性格的布鲁图斯，并由此模拟前面提到的对话。而系统的输出则是：

布鲁图斯：恺撒，我知道你是一个伟大的领袖，但是我们的国家已经陷入了混乱，我不知道该怎么做才算是对的……

从萨特哲学的角度看，“犹豫不决型”人格往往能够展现“自欺力”与“反自欺力”之间复杂的心理斗争，并因此具有很高的文学刻画价值。不过，ChatGPT 对于这种人格特性的刻画却是十分拙劣的。换言之，系统无法对布鲁图斯之所以既爱恺撒又爱罗马元老院的双重心理的内在根基进行挖掘，而只能模棱两可地说恺撒很“伟大”，而罗马已陷入了“混乱”。但问题是：一个伟大的领袖又怎么会导导致国家的混乱呢？请注意，我们在这里需要刻画的是一个充满内心情感矛盾的布鲁图斯，而不是一个缺乏逻辑头脑的布鲁图斯。

为了继续“调教”ChatGPT，笔者又降低了问题的难度。也就是说，笔者自己补充了布鲁图斯的行事动机，并由此希望系统将其对话修改得更符合人类的心理学常识。笔者所补充的具体信息是：假若布鲁图斯杀死恺撒的理由一半是政治原因，另一半则是二人爱上了同一个女性，而且，这两个理由既互相交织，又彼此矛盾，布鲁图斯又该怎么说呢？

笔者之所以向机器给出如此详细的心理侧写，一方面固然是受到了萨特的戏剧《脏手》的启发（在这部戏剧中，雨果对贺德雷的刺杀理由就混合了政治因素与感情因素），另一方面也是希望机器能够向西方戏剧的伟大传统致敬。不管怎么说，公共的政治理由与私人情感理由之间的张力，本就是自《安提戈涅》以来西方戏剧的恒久主题之一。这种张力在“自欺”论的论域中则能显露出一个更具存在主义色彩的面相，此即，具有“大义名分”的政治理由往往会构成难登大雅之堂的男女私情的掩饰，并由此完成对于“本真自我”的掩饰。与之相比，从另一个方面看，对于“本真自我”的直面却又会使杀人行为的理据发生动摇，因为纯粹的情感理由本是不足以去剥夺另外一个生命的生存权的（这同时也便是《战争与和平》中的皮埃尔在因感情原因与多洛霍夫决斗后所得到的人生领悟）。因此，情感理由就与政治理由处在一种彼此否定的复杂纠葛之中，并因此提点了人类生存境遇的复杂性。而能对这种复杂性加以展现的文学描述，自然也能产生出巨大的审美价值。

不过，ChatGPT 则无法在笔者的提点下给出令人满意的输出。它写道：

布鲁图斯：恺撒啊，我必须这样做，否则罗马将会沦为一盘散沙，你应该明白的。

恺撒：我理解你的政治目标，但是为什么要采取这种暴力手段呢？我们可以谈判，达成妥协。

布鲁图斯：不，这已经远远不够了。我的心已经深深地爱上了那位女性，我不能再让你干扰我们之间的爱情。我必须让你消失。

恺撒：难道你不明白这两件事情之间的矛盾吗？你不能用个人感情影响国家大事。

布鲁图斯：我明白，但是我的爱情已经超越了国家和政治。这是我的内心深处的声音，我必须听从它的指引。请原谅我，恺撒。

这又是一段非常拙劣的对话。一个人的内心矛盾应当用自然的方式加以展现，但在这段对话中，布鲁图斯竟然需要恺撒来指出他内心的矛盾。此外，作为一个政治家，布鲁图斯也是说不出“我的爱情已经超越了国家和政治”这样的孟浪之言的。换言之，即使他的确是因为私情杀人，也会竭力通过“自欺”而将相关动机解释为一种可以被公布于天下的政治理由。换言之，这段对话根本无法展现“自欺”过程在布鲁图斯心中是如何发生的，而只能将布鲁图斯刻画为一个连自己的信念矛盾都需要政敌加以展现的傻瓜。而为了使得读者能够更为感性地理解机器给出的这段对话是何等不自然，笔者在下面便斗胆也给出一段自己写成的对话，以便读者对照：

布鲁图斯：恺撒，我必须这样做，否则罗马的共和制度就完了……

恺撒：别提什么罗马元老院！我知道你喜欢西尔维娅，但这是我们两个男人之间的私事。你今天若杀了我，这就是公事了！……

布鲁图斯（暴怒）：不！这与西尔维娅无关！这与她无关！这其实与克里奥佩特拉有关！你和一个埃及女人生下的孩子怎么能享受你的继承权？元老院又怎么可能承认一个埃及女人的儿子的地位？是你欺骗了罗马！当然，你也欺骗了纯洁的西尔维娅！

恺撒（冷笑道）：若这与西尔维娅无关，你为何一听到我说她的名字就如此激动？

布鲁图斯（咬住嘴唇，沉默片刻，继续爆发）：好吧，这的确与西尔维娅有关！但这不仅仅与她有关，更是与千千万万被你欺骗的纯洁的罗马人民有关！我不能让你这样的暴君继续欺骗人民了！（说罢，拔剑刺向恺撒，高呼：）一切为了元老院与人民！

恺撒（呻吟中挤出一句话）：难道你今天没有自己骗自己吗？

我相信不少读者都能写出比笔者质量更高的对话。尽管如此，笔者依然自信这段对话的质量已经明显超过了 ChatGPT 的输出。不难看出，在笔者撰写的这段对话中，布鲁图斯试图反复给予其刺杀行为以正当的政治理由，以试图掩饰其对笔者所虚构的西尔维娅的感情。他甚至主动提起依附于恺撒、并在元老院中风评不佳的克里奥佩特拉，以求用一个带有政治意味的女性替换掉一个不具有政治意味的女性，由此化解恺撒的批评。而他对于西尔维娅的真实感情，则恰恰是通过他急于掩饰这种感情的激动行为加以展现的，而这种情感本身又通过恺撒的言语揭露而进一步暴露在读者眼前。此外，“欺骗”这词在布鲁图斯与恺撒口中的不同含义，又从另一个角度提示读者转向对于布鲁图斯内心情感世界的探索。总之，笔者在构造上述对话时，是将下述原则谨记在心的：按照我们所接受的对于“自欺”的一般性定义，自欺者是无法自觉地表述出其信念系统之不自洽性的，而只能将自己的真诚信念压到一个更深的层次上。而文学作品对于“自欺”现象的刻画也必须展现出这种层次性，这恰恰是现在的 ChatGPT 技术所不能为之事。那么，上面的论证是否能够证明未来的 ChatGPT 技术也不能表现“自欺”呢？

我倾向于认为未来的 ChatGPT 技术也不能做到这一点。当然，一个聪明的程序员会按照上面笔者给出的表现“自欺”对话套路再进一步训练系统，使得系统可以“照猫画虎”地炮制出一些质量更高的对话。但问题是，尽管人类的自欺现象本身可能是有一些套路的，但对于人类自欺现象的描述却是没有固定套路的。其背后的道理是：由于自欺现象自身固有的二元结构（即表层的理由对于深层理由的压抑结构），在这种结构中必然存在的深层理由的缄默性就会与文字表述的公开性构成天然的矛盾。而为了化解这种矛盾，同时又不至于破坏当事人行事深层理由的内隐性（因为这种破坏会摧毁自欺现象本身），作家就必须开发出一系列的文学技巧，以期达成“犹抱琵琶半遮面”的艺术效果。不管这些文学技巧是什么，这些技巧在语义上往往会依赖于故事中超越当下对话的一些要素（比如，在笔者构造的对话中，就提到了在原始对话中不曾出现的克里奥佩特拉，以及其在当时罗马政

治棋盘中的微妙地位)。但这种提及对于深度学习技术则是致命的,因为深度学习所依赖的语料训练材料都是一小段、一小段的。换言之,你只能将《战争与和平》中的某个自然段作为训练语料输入系统,而无法指望系统能够了解这整部长篇小说的任何两个句子之间的复杂语义关系。也因为这一点,这种系统无法在自行“阅读”一部新的长篇小说之后自动生成关于这部小说的人物关系表与人物侧写,因为这种能力需要系统具有一个类似于真人的长期记忆系统与逻辑推理系统。不得不指出,对这种能力的获取,并不是基于大数据的深度学习系统的发展方向。打个比方来说,这种技术路径只能帮助系统获得海量的牛肉,却永远无法让其看清牛的骨架究竟是什么样子。

本节讨论的结论已经非常清楚了:ChatGPT技术无法通过笔者给出的“新图灵测验”,也就是说,该系统无法模拟人类作家对于人物的自欺现象的刻画。而在前文的讨论中我们也已经知道了,作家对于人类自欺现象的刻画本身就具有反自欺的意蕴(正如对于丑恶的客观呈现本身就是为了反对丑恶一样),因此,ChatGPT便无法在“反自欺”的斗争中发挥积极的作用(至于ChatGPT在机器翻译、文字校对等方面发挥的辅助功能则另当别论)。由此我们还能推出:人类写手对于ChatGPT的过度依赖,也有可能磨蚀其文字敏感性与对于人类心理与行为的观察力,由此使得海德格尔笔下的“常人”式闲谈渐渐淹没那些真正的思想珠玉。说得悲观一点,如果我们对这种前景缺乏警醒,萨特在哲学与文学领域所喊出的“反自欺”口号将成为同类声音在前ChatGPT时代的绝响。

小 结

海德格尔早就说过“语言是存在的家”,但这句话的有效性,是建立在一个朴素的事实上:即使是被他所批判的那些“常人”的“闲谈”,也是人的闲谈,而不是机器自动生成的。换言之,在海德格尔与萨特生活的时代,即使是被机械复制,并在报纸、广播、电视上得到传播的那些语言信息,也是有着无可置疑的署名的:你知道这句话是某个政治家说的,也知道对于他的言论的报道是由某家报社的某位记者完成的。甚至在进入互联网时代之后,机械传播方式与作为信息源的个人之间的这种联系还没有被切断。譬如,传统的搜索引擎依然会让读者知道他所看到的信息的正确性是由何人(或哪家机构)负责的。而ChatGPT则从根子上改变了游戏规则:这个新机制所发出的言论,并不能被追溯到哪个具体的个人上,而只能被视为对于大量人类标注工的标注行为进行统计学抽象后的产物。ChatGPT本身就是一个被巨大的资本—技术力量喂养大的无人格的利维坦,或是一个由海量数据喂出的貌似尚可的输出所构成的一个巨型“剧场假相”。而这一机制在原则上就无法陷入“自欺”——遑论反对“自欺”——的形而上学原因便是:它缺乏陷入或直面自欺现象的根本前提——一个具有本真性的自我信念体系的存在。而它之所以无法满足这一前提,则又是因为它根本就没有“自我”,或说得更具体一点,它缺乏使得完整的“自我”得以可能的更进一步的前提,即具有具身化的行动力、能够具有知、情、意多方位能力的心智架构,以及其对于自身有限性的哪怕最模糊的意识。从这个角度看,ChatGPT向我们揭示的未来,或许很可能是一个文法通顺的庸人时代,一个使得人连自欺的本能都会退化的时代。

责任编辑:马妮