【大数据时代的信息技术与哲学思辨专题研究】

网络谣言与“认知安妥性”*

——一条非大数据化的智能鉴谣系统设计思路

徐英瑾

摘　要:站在当代认识论的立场上看,“网络谣言”的构成要件之一——“与事实不符的虚构”——可以被替换为“目标信念的认知欠安妥性”,即“目标信念太容易为假”。而一个自动避谣系统对于一个信念的“欠安妥性”的评估,则可以按照评测“特定背景信息下目标信念的‘使真者’的成真概率”的方式来进行。此外,由于上述背景知识的具体内容对于特定用户的价值观的依赖,这一方案在原则上就是与以“泯灭个体用户个性”为必然后果的大数据技术截然不同的,并因此足以构成对于大数据技术的某种对冲技术。站在更宏观的角度看,此项研究所具有的“打通人文思辨与工程学研究之间的壁垒”的意蕴,又包含了如下两个目标指向:其一,通过工程学问题自身自带的“跨文化性”,为英美认识论在华语世界的传播降低文化门槛;其二,通过上述这种工程学研究所自带的“可操作性”要求,倒逼学院哲学家对认识论领域内的“安妥性”概念进行“可编程性”检查。

关键词:网络谣言;通用人工智能;认知安妥性;概率论;背景知识

中图分类号:G20;B017　　**文献标识码**:A　　**文章编号**:1003-0751(2020)06-0111-10

一、导论

首先,笔者要解释本文标题中的“认知安妥性”(epistemic safety)究竟是何意。先请读者考虑如下案例:张三在化学测验之前复习不足,因此,在做选择题时,他只能通过抓阄来押答案。假设最后一题的答案是“B”,而且他也押中了,那么,这是否意味着他已经获取了囊括在该答案中的化学知识呢?从常识角度看,答案当然是否定的。说得具体一点,张三此次获取正确答案的过程包含了太多的运气成分,而只要幸运女神下次不再那么眷顾他,他就会马上得出错误的答案。此外,根据大多数人对于作为名词的“知识”(knowledge)与作为动词的“知道”(to know)的语义直觉,任何人要获取一项知识,或是要知道某事,其获知的方式都应当具有一定的公开可辩护性与可重复性,并由此在相当程度上排除掉“认知运气”(epistemic luck)的成分。因此,张三如果足够理智的话,他就不能在自我评测时认为自己已经掌握了囊括在答案B中的相关化学知识,因为他所蒙对的答案是缺乏“认知安妥性”的。由于“认知运气”本身就意味着产生错误的风险的存在,故此,我们就不妨先将“认知安妥性”(以下简称为“安妥性”)大致定义为一种能够在相当程度上豁免于错误风险的认知状态。在当代西方认识论(epistemology)[①]文献中,“认知安妥性”也被视为“知识之为知识”的一项必要条件。

——以上说的这些,与“网络谣言”又有什么关系呢?

由于“网络谣言”本身不是一个成熟的学术词汇,所以,在此我们只能满足于对于它的一个非常粗疏的常识定义。按照“百度百科”的说法,网络谣言是指通过网络介质(例如微博、国外网站、网络论坛、

收稿日期:2020-02-17

*基金项目:国家社会科学基金重大项目“基于信息技术哲学的当代认识论研究”(15ZDB020)。

作者简介:徐英瑾,男,复旦大学哲学学院教授,博士生导师,教育部长江青年学者(上海　200433)。

社交网站、聊天软件等)而传播的没有事实依据,带有攻击性、目的性的话语。[②]按照这个粗疏的定义,网络谣言的特征有:(甲)通过互联网传播;(乙)是与事实不符的虚构;(丙)带有一定的传播目的。很明显,在这三个特征之中,(乙)最为关键,因为不借助互联网传播的谣言依然还会是谣言,而带有一定的传播目的的真信息也依然不会成为谣言。那么,(乙)与"安妥性"的关系又是什么呢?

笔者认为,"安妥性"标准乃是我们用以判断一条信息是否触发了条件(乙)的简易标准。那为何我们需要一个简易标准呢?这是因为,对于一个知识与信息处理能力有限的智能体来说,要判断一条信息是不是与外部事实真正相符乃是非常困难的,因为这样的智能体往往缺乏独立调查真相的能力与精力。相对容易做的反而是判断一条信息的"安妥性",即判断这条信息是不是容易出错。譬如,若有条信息说有个数学功底很差的高中生通过在高考中一路抓阄而考上了清华,这样的信息虽然也可能是真的,但因为太欠缺"安妥性"(即太容易错),而依然会被任何一个心智正常的人判断为谣言。这里需要注意的是,从逻辑上看,当然存在着一条信息既欠安妥性,同时又为真的可能性,在这种情况下,按照条件(乙),该信息反而不算谣言。但考虑到这种"误杀"的可能性其实不高,将"欠安妥性"视为将某信息被判为谣言的"测试标准",依然是妥当的。

——以上说的这些,与人工智能又有什么关系呢?

假设有一台具有"通用人工智能"(Artificial General Intelligence,以下简称为 AGI)[③]特征的推理机器,设计者希望它能够自动地在海量的互联网信息中将网络谣言筛选出来。那么,我们该如何设计这样的系统呢?一种目下似乎很受青睐的做法,乃是基于大数据技术而设计的。譬如,根据所谓的"多能动者的信息主体信任识别机制",相关系统能够自动寻找每条信息背后的信息发布者的权威度或可信度(这一点又取决于发布者自身的互联网活动痕迹),然后延及对于信息传播者的权威度与可信度的计算,由此决定信息本身的可信度。[④]但这个做法本身却是有问题的,因为它没有办法应对如何揭穿"楚门的世界"这一难题,即,在网络本身充斥着大量虚假的网络意见领袖与"权威避谣人"的情况下,这样的避谣机制很容易被错误的网络风向带到错误的方向上去,而不能根据所传播的信息本身的质量给出独立的评断意见。换言之,这样的机制无法豁免于任何基于大数据的 AI 技术所无法避免的困难,即它只能根据既有的数据对未来的时态发展趋势进行推测,而无法对所涉及的信息内容本身进行真正的思考。

而要避免这个问题,笔者的建议是让 AI 能够对信息的内容进行独立的评断,特别是对于其"安妥性"的评估。由于"安妥性"本身意味着相关信息"不太容易为假",而对"不容易为假"的判断又牵涉到对于信息内容的内部特征——而不是信息发布者——的判断,所以,这样的系统的运作方式,显然就应当具备英语认识论文献中所经常提到的"内在主义理路"(the internalist approach)的某些特征,以便由此提供更多的理论空间来模拟人类的内部思维过程。

那么,我们又该使用怎样的技术手段,以一种亲和于"内在主义"的方式,来表征"安妥性"与"安妥性欠缺"之间的差别呢?一条很容易被想到的表征思路就是诉诸概率论,比如将一个更具"安妥性"的信念表征为一个具有更高成真概率的信念,并将一个不那么具"安妥性"的信念表征为一个成真概率较低的信念。虽然这个想法本身还是非常粗糙的,但考虑到基于贝叶斯式概率计算的 AI 研究早已通过珀尔(Judea Pearl)的工作[⑤]而成为 AI 学界的通识,因此我们不妨就认为:在概率论的大框架里去理解刻画"安妥性",至少在大方向上是具有可行性的。此外,特别需要指出的是,笔者所推崇的这种概率统计本身并不是基于那种泯灭用户个性的大数据处理技术的(尽管将概率论与大数据技术结合,其实并不困难),而恰恰是通过引入"概率值归赋者的背景知识",而与特定用户的个性画像发生了密切关联(详后)。因此,这种技术路径与其说是大数据技术的一种,还不如说是构成了大数据技术的对冲技术。

不过,有意思的是,在西方认识论讨论"安妥性"的典型文献中,特别是在英国哲学家普理查德(Ducan Prichard)的理论框架中,处理安妥性的技术工具却不是概率论,而是"可能世界语义学"(possible world semantics),即一种通过"可能世界"的理论假设来对"反事实条件句"(counterfactuals)的含义进行阐释的技术工具。而本文首先将论证,至少

站在 AGI 的立场上看,这一技术路径是根本无法被编程化的。与之相比较,笔者将要在本文中提出的基于概率论的“安妥性”概念,不仅不会引发“循环论证”的问题,而且它还将产生对以“斑马案例”为典型的诸认识论经典案例的足够解释力(此外,笔者还将提到“斑马案例”与一般谣言甄别过程之间的类似性)。从更宏大的视角去看,笔者所给出的技术路线,不仅对 AI 的研究具有意义,也将向哲学界提供一条经由经验科学改造传统的认识论理论的新道路,以便在 AI 时代以新的技术内容夯实奎因(Willard Van Orman Quine)所提出的“自然主义的认识论”(naturalized epistemology)概念的内涵。

二、基于“可能世界语义学”的“安妥性”概念是无法被编程的

用“可能世界”概念解释“安妥性”的学者很多,但最具代表性的乃是英国学者普理查德。他对于“安妥性”的定义,大致可以被概括如下。⑥

某主体所持有的某目标信念 P 是安妥的,当且仅当如下条件被满足:就与该主体所处的那个现实世界最为切近的大多数的可能世界而言,只要在那些可能世界中,主体继续以他在现实世界中持有 P 的方式持有 P,那么,P 依然是真的。

这个定义貌似还很抽象,我们不妨便用所谓的“谷仓案例”⑦来解释一下(顺便说一句,本文已经在“谣言甄别”的语境中对该案例进行了改写)。假设读者已经知道在横店影视城有大量的假谷仓(里面当然是没谷子的),以便为影视剧的拍摄提供布景。而且读者还知道,在这些假谷仓中,混有一个编号为“88 号”的真谷仓(里面藏有很多优质黑龙江产压缩真空大米),以便为广大群众演员提供后勤保障。⑧同时,读者还知道,关于谷仓之虚实,乃是包括读者在内的小圈子里的秘密知识,广大公众对此是一无所知的。这时候,假设读者再获得一条新信息:有个叫“张三”的窃粮贼(他不属于上述小圈子),骑了一辆自行车进入“谷仓阵”,而且,他在没有任何内部情报支援的情况下就随机找到了 88 号谷仓,并偷走了其中的一袋大米。你觉得这是谣言吗?

很多人恐怕会倾向于认为这是谣言,因为张三碰巧撞上 88 号谷仓这件事太容易成假了。譬如,只要张三稍微转转头,或者他的自行车骑得快一点或者慢一点,他就会看到那些假谷仓,并由此错过唯一藏有粮食的 88 号谷仓。若按照普理查德的“可能世界语义学”的话语方式来重述这个故事,那么在这个案例中,所谓的“现实世界”,就是指“张三撞上 88 号谷仓”这一状况,而所谓的“与现实世界非常切近的可能世界”,就是指“张三稍微转头并由此注意到假谷仓”的那些状况。由此看来,张三的信念(即“我看到的的确是真谷仓”)之所以是不安妥的,就是因为它仅仅在现实世界中为真,却在那些与现实世界非常接近的可能世界中为假。

上文提到的“非常切近”一词,显然涉及了某种空间性隐喻,而在威廉姆森(Timothy Williamson)的知识论模型中,关于“安妥性”的空间化隐喻还得到了一种更为淋漓尽致的表达。具体而言,在他的隐喻框架中,目标信念被看成了一个圆润的球体,而其真值则被视为球与所在平面之间的关系——如果球能够在平面上停稳,则代表“真”,若停不稳,则代表“假”。至于球在平面上的投影位置,则被视为“现实世界”,与该位置接近的那些位置,就代表那些“与现实世界非常接近的可能世界”。这样一来,如果一个信念被认为是“安妥的”,那么就可以通过这种空间隐喻而被理解为:即使球稍微离开其所在的位置,并在与之非常接近的位置出现,它也能停稳当。换言之,如果它一离开自己原来的位置就马上就无法停稳了,它就是不安妥的。⑨

然而,上述这种基于空间隐喻的理解方式虽然很直观,但从计算机编程的角度看,它却很难被恰当地程序化,因为上述描述方式中已经包含了一个难以被消除的循环论证。

很明显,任何人若要按照普理查德的要求去确定一个信念是否安妥的话,他就不得不先去依次做三件事情:

(1)分辨出哪种情况是现实世界,哪些情况是可能世界;

(2)根据每个可能世界与现实世界之间的举例,来判定哪些可能世界是与现实世界“切近”的,哪些则否;

(3)在那些与现实世界非常切近的可能世界中,我们再去判断目标信念在其中是否依然是真的——若是真的,则目标信念安妥,否则相关信念就是不安妥的。

然而,上述这个路线图的步骤(1)预设(2)已经被执行了,而(2)的执行却必须预设(1)已经被执

行。这种自相矛盾的状况就必然导致(1)与(2)都不能得到执行,由此导致整个路线图的崩塌。为何这么说呢?

为了理解这一点,我们不妨来考察一个被高度简化的谷仓案例。设想张三所面对的“谷仓”仅仅有 6 个(其中只有 1 个是真的),并均匀分布在张三周围。在这种情况下,面对那个真谷仓的张三,每将自己的头往旁边转 60°,就会看到一个假谷仓。由此,按照普理查德的理路,我们不难算出,在他看到真谷仓的现实世界之外,该案例涉及的可能世界一共有 5 个。但只要想得更仔细一点,我们就会发现上面这种关于可能世界的计算方式是有问题的。我们不妨问:为何张三不能将他的头仅仅转 20°呢?或者,为何我们不能设想张三干脆抬起头,不看谷仓,而注视天上飞过的麻雀呢?由此产生的“可能世界”,难道会与现实世界不够“接近”(或不够“相似”)吗?这样一来,可以被纳入考量的“可能世界”的数量就会暴增,由此超过任何智能体的信息处理上限。而对于这个问题,一种典型的普理查德式解释便是:这里所说的“可能世界”,还应当受到一个重要的限制条件的约束,即在这些可能世界之中,主体张三不仅继续持有信念 P,而且他产生与持有 P 的方式与其在现实世界中产生与持有 P 的方式是相同的。举例来说,如果在现实世界中主体是通过知觉来把握 P 的,他就不能在可能世界中通过回忆来把握 P。按此理,在谷仓案例中,那些在其中张三并没有真正知觉到一个“谷仓”的“可能世界”,就不应当被计算在内。由此,“可能世界”的数量“通货膨胀”的问题便得到了消除。

在这里,我们姑且先接受普理查德对于这个问题的解答。然而,这个解答实际上说的是,若我们要将“可能世界”识别出来,我们就要先确定:该可能世界至少在信念产生方式方面是与现实世界一样的。由于确定两个世界之间的信念生产方式之间的异同,本身就是评估这两个世界之间的距离(即相似性)的一个重要方式,所以,普理查德的上述答案无异于在说,我们要执行步骤(1)之前,必须先执行步骤(2)。很明显,这种要求就像“在造一层楼之前先造二层楼”一样荒谬。所以,普理查德的整个“安妥性”方案是缺乏编程指导价值的。

即使不谈上面这个麻烦,对于 AGI 的研究来说,普理查德的“安妥性”概念还会带来两个问题。

第一,他没有提供一个统一的算法化说明来告诉我们,在不同的认识论案例中,我们究竟该如何切分与识别可能世界,并度量其与现实世界之间的距离。换言之,用来处理“谷仓案例”的切分方案,可能并不适用于本文“导论”里提的“考试抓阄”案例,因为不同案例的不同细节,最终会导致可能世界切分方案自身的“特设性”(ad hocness)。但从 AGI 的角度看,这种“特设性”会倒逼认知架构的构架师去为不同的问题处理语境预先设计不同的可能世界划分方案,由此导致非常惊人的建模成本。同时,即使这种设计最后勉强完成,它也会因为无法应对构架师事先没有料到的那些新问题语境而失效,由此失去了 AGI 系统本该具有的通用性与灵活性。

第二,看得更深一点,普理查德并没有花费足够的篇幅来说明“可能世界”自身的本体论地位,而这个遗漏的理论缺口显然会让 AGI 的研究者很头疼,因为没有一个编程者能够在弄清楚“可能世界”的本体论意义之前去找到关于它的恰当建模方式。而更麻烦的是,在这个问题上,在形而上学领域内已存在的种种关于“可能世界”本性的讨论方案,也难以助普理查德一臂之力。为何这么说呢?首先可以肯定的是,若像“模态实在论者”刘易斯(David Lewis)所建议的那样,将每个可能世界视为一个与现实世界同具实在性(却与现实世界在时空与因果上脱钩)的巨型实在对象⑩,那么,由此导致的可能世界模型,无论对 AGI 来说,还是对普理查德本人的认识论来说,都是毫无用处的,因为此类可能世界既然与现实世界彻底隔绝,当然也就难以成为认知考量的对象。此外,即使我们按照认识论气味更浓郁的“语言代理论”(Linguistic Ersatzism)的要求,将一个“可能世界”视为一个内部不包含矛盾的巨型命题集合,这种观点也将在 AGI 的脉络中引发如下两个让人头疼的工程学难题:(甲)上述这个命题集合到底该有多大?(很显然,它不能无限大,否则相关的人工认知构架的信息库就会“爆仓”——但它又不能太小,否则很难说它构成了一个“世界”);(乙)一个智能体该如何发现该集合诸成员之间的逻辑不兼容性?(很显然,当一个信念系统足够庞大的时候,检验其内部兼容性的工作会带来巨量的工程学负担。)而在笔者看来,要避开上述这些难题的最干脆的办法,就是彻底抛弃可能世界语义学。

三、关于“安妥性”概念的概率论模型

在引入笔者自己的“安妥性”模型之前,先给出几点准备性说明。

第一,笔者有必要首先来澄清几个概率论符号的含义。符号“Pr(甲)”指“信念甲成真的概率”。由于“信念甲”在这里的出现是无条件的,所以,这样的概率又叫“无条件概率”或“先验概率”。与之对比,符号“Pr(甲|乙)”则指“在条件乙已经被满足的前提下,信念甲的成真概率”。由于“信念甲”在这里的出现是有条件的,所以这样的概率又叫“条件概率”或“后验概率”。这里需要注意的是,任何一个先验概率的赋值都预设了相关赋值者的相关背景知识,所以这些先验概率也可以被转化为一定背景知识下的后验概率,记作“Pr(甲|背)”。

第二,前面对概率符号的解释已经涉及“概率赋值者”这一概念,这也就是本理论框架中的“安妥性评估者”。那么,他到底是谁呢?以前面提到的谷仓案例为例,故事里的张三显然不是我们这里所说的“安妥性”的恰当评估者,因为他其实是不知道自己所面对的谷仓阵的虚实的,真所谓“不识庐山真面目,只缘身在此山中”是也。所以,做出“张三的信念不具有安妥性”这一评断的乃是故事的听众,也就是知道张三面对的谷仓大多数为假的“庐山外人”。就此,我们可以将一阶的“故事内主体”(简称为“主体”)与二阶的“故事外评价者”(简称为“评价者”)相互区分,并由此认为“安妥性”标签的颁发者(或“安妥性”概率的赋值者)就是二阶的评价者。需要注意的是,在经验的意义上,一个一阶的主体也可以在事后成为一个二阶的评价者。譬如,谷仓案例里的张三,也可以在事后知道他刚才所面对的谷仓大多数是假的,由此认定他刚才的信念是欠安妥的。但尽管如此,为了避免可能的混淆,在概念上区分二者,依然是非常必要的,因为一个处在“庐山”中的“张三”与跳出“庐山”的“张三”毕竟在信息上还是不对称的。另外,即使不从认识论的角度看,而从网络谣言鉴别机制设计的角度看,对于一阶主体与二阶评估者的区分也是非常重要的。具体而言,在该技术语境中,一阶的主体,就是“疑似谣言”的提供者(注意:在不少场合中,这一提供者可能是匿名的),而二阶的评估者,就是 AI 专家所试图设计的评估体系自身。由于评估体系自身的背景知识很可能是信息的提供者所不具备的,二者之间的信息不对称性就需要被某种形式加以刻画,而上述这种“一阶—二阶”二分法无疑就是一种简便的刻画方法。

第三,既然笔者引入了概率论,就自然会引入“成真概率”这一说法。但这一说法显然预设了“真”这个麻烦的逻辑语义学概念,并由此倒逼笔者给出相关的语义学解释。为了不陷入与之相关的复杂争辩,笔者就姑且先采用一种符合论的真理观,即认定一个信念是真的,当且仅当它与一个外部的“使真者”(truth-maker)是彼此符合的。但需要注意的是,在概率论的语境中,上述这种简朴的符合论观点却不得不经历两重非常关键的改造。

改造(甲):由于“符合”是一个“非黑即白”的二极化概念,而概率值却是允许处在[1,0]之间的任何一个梯度之上的,因此,我们就不得不将信念与使真者之间的关系改造为一种“间距关系”(因为间距关系本身也是允许被梯度化的)。而在概率论语境中,引入诸如“间距关系”之类的空间化隐喻,并不会导致此类隐喻在可能世界语义学框架中所引发的那些困惑,因为从概率论角度看,说“P”(指一个特定的目标信念)与“$P_{使真}$”(指 P 的使真者)之间的间距近,无非就是说二者联合发生的概率值高,或者说,$Pr(P\&P_{使真})$的值颇高。而这一联合概率的值,无疑是可计算的(详后)。

改造(乙):由于任何概率值的计算都是预设了二阶评估者的背景知识的,所以,对于 $Pr(P\&P_{使真})$的计算,实质上也就等于对于 $Pr(P\&P_{使真}|背)$的计算。这也就是说,信念的使真者对于新信念自身而言的外在性,仅仅是对一阶信念的持有者而言的,而对于二阶的评估者来说,甚至关于使真者的信息也是内在的。现在笔者再以“谷仓案例”为抓手,来说明这一点。在该案例中,目标信念当然是“张三撞上了一个真谷仓”(简称“真谷”),而该信念的使真者,则是同样为二阶评价者所知的“张三所撞上的谷仓,就是 88 号谷仓”(简称“谷 88”)这一点,因此,对于该信念的安妥性的计算,就是对于 Pr(真谷 & 谷 88|背)的值的计算。不过,经过一番概率推演以后,我们不难发现,对于 $Pr(P\&P_{使真}|背)$的计算,其实就是对于 $Pr(P_{使真}|背)$的计算。[11]因此,对于某个目标信念的安妥性的评估,就可以建立在对于 $Pr(P_{使真}|背)$的值的评估之上。

基于上述三重技术性说明，现在笔者就可以给出一个新的“安妥性”定义。

“安妥性”新定义：对于任何一个“安妥性”评估者的背景知识（“背”）来说，出现在一个目标信念 P 的安妥性程度，取决于 $Pr(P_{使真}|背)$ 的值。若该值高于一个内在于评估者的“安妥性心理阈值”，则该信念就会被打上一个“安妥性”的标签，否则，该信念则会被认为是欠安妥的。⑫

而建立在上述定义之上的“谣言”鉴别原则便可归纳如下：任何一个信念，只要按照上述“安妥性定义”被判断为欠安妥的，则可以被打上“谣言”的标签。

读者可能会问：上文所说的“背景知识”究竟是什么？我们又如何通过一个可以被程序化的算法，来精细地规定背景知识对于目标信念自身的概率值的影响机制呢？

对于这个问题，笔者的回应是：任何一个被系统解读的故事脚本，要么是“形式特征”凸显的故事脚本，要么是“经验特征”凸显的故事脚本。先来解释前者。比如，在简化版本的谷仓案例中，既然六个“谷仓”的外貌彼此之间是没有任何可知觉的区别的，那么，“谷仓”这一事项所勾起的经验记忆就与判定目标信念（“张三撞上了一个真谷仓”）的安妥性无关。换言之，对于该信念的安妥性的评判方式，与判断一个人随机投了一个六面骰子后，骰子的“六点”正好朝上的概率的方式，大致相同。由此，谷仓案例就是一个“形式特征”凸显的故事脚本。或说得更形式化一点，一个故事脚本是“形式特征”凸出的，当且仅当故事中的目标信念在被简化为“A 是 B”这样的主—谓关系之后，系统发现：该关系是诸如“1 比 6”这样的纯数量关系，或者是“P 或者非 P”这样的纯逻辑关系。在这种情况下，一个安妥性评估系统将按照“形式特征”的标准，来激发自身关于逻辑学与数学的背景知识，并由此来评估目标信念的成真概率。与之相对应，一个故事脚本是“经验特征”凸出的，当且仅当故事中的目标信念在被简化为“A 是 B”这样的主—谓关系之后，系统发现：A 与 B 之间的关系乃是经验对象（或其集合）之间的非形式关系。在这种情况下，一个安妥性评估系统将按照“形式特征”的标准，来激发自身关于与 A 或 B 相关的经验事例的经验性背景知识，并由此来评估目标信念的使真者的成真概率。⑬下一节所要提到的“斑马案例”，就是此类“经验特征”凸显的案例。

四、“斑马为骡”是谣言吗？

下面便是笔者在“网络谣言”甄别语境中，对于在认识论文献中经常被讨论的“斑马案例”的重述。⑭假设有一个叫李四的土豪在县城郊区开了一家私人野生动物园，并且在微信朋友圈里发广告说，他的野生动物园有大量非洲野生动物，包括斑马，欢迎大家来观摩。这时候又有某人在微博上发帖说，李四的动物园里的斑马都是假的，实际上它们都是用油漆抹上了斑马条纹的骡子。那么，这一条微博消息是不是谣言呢？

按照笔者提出的鉴谣机制设计思路，要判断这一条微博消息，我们首先要判断它是不是安妥的，而这一点又取决于：根据系统自身的背景知识，“动物园里的所谓‘斑马’都是被涂抹成斑马状的骡子”这个信念的使真者的成真概率是否足够高。但这样做的前提又是：我们应当知道该如何在概率论的话语框架中刻画这个信念。那么，这种刻画又当如何进行呢？

从语言分析的角度看，目标信念可以被展开为这个样子：“在看到的动物外形具有斑马形状的前提下，对于该前提的合理解释是：它们都是被涂上斑马纹路的骡子。”换言之，这是一个从特定知觉现象出发，获取特定解释的“溯因推理”（abduction）所具有的结论。从概率论角度看，这个展开式的表达就是“$Pr(涂了斑马纹的骡子_{使真}|斑马形状)$”，或简写为“$Pr(斑骡_{使真}|马状)$”。而对该概率式的值的计算，则可以按照贝叶斯定理来进行。按此定理，我们立即得出：

$$Pr(斑骡_{使真}|马状)=Pr(马状|斑骡_{使真})\times Pr(斑骡_{使真})/Pr(马状)$$

那么，我们又该怎么来计算上式右边的三个概率值呢？很显然，Pr（马状）= 1，因为根据题设，在动物园的观摩者面前出现了“貌似斑马的动物”这一点乃是毋庸置疑的。此外，我们也能肯定：$Pr(马状|斑骡_{使真})=1$，因为在“骡子被涂上斑马纹”这一点的确被一个使真者满足之后，这样的骡子的确在外形上是与斑马难以分辨的（在这里我们假定分辨者乃是一般的公众，而不是动物学家）。这样一来，$Pr(斑骡_{使真}|马状)$ 的值，也就等于 $Pr(斑骡_{使真})$ 的值

了。考虑到在笔者的理论模型中所有的概率赋值都预设了二阶评估者的背景知识的有效性，所以 Pr（斑骡$_{使真}$）的值实际上也就是 Pr（斑骡$_{使真}$|背）的值。

那么，系统又该如何评估 Pr（斑骡$_{使真}$|背）的值呢？很显然，被评估的信念，可以被转换为“骡子被涂抹上了斑马纹”这样的主—谓形式，而这一主—谓语句所刻画的乃是一个经验性事态，而不是某种纯粹的数量关系或逻辑关系。这样一来，按照前小节所给出的系统工作程序，系统的经验性背景知识就会被激活，以便测定该语句的成真概率。其具体测定方式如下。

（甲）系统会搜索其记忆库并检索互联网，以测定是否有关于“骡子被涂抹上斑马纹路”的使真者的案例。如果没有，则执行（乙）。如果有，则记录案例的次数，并由此输出 Pr（斑骡$_{使真}$|背）的值。该值再与特定的安妥性阈值比较，若低于阈值，则输出“欠安妥”标签，若高于阈值，则输出“安妥”标签。

（乙）系统将原有目标语句泛化为“动物 A 被涂上了另外一种动物 B 的颜色”，由此搜索其自身的记忆库或检索互联网，寻找类似案例。如果没有，则立即宣布目标概念欠安妥。如果有，则记录案例的次数，并由此输出 Pr（斑骡$_{使真}$|背）的值。该值再与特定的安妥性阈值比较，若低于阈值，则输出“欠安妥”标签，若高于阈值，则输出“安妥”标签。

针对上述程序，心细的读者或许会问：假若系统搜索了自己的记忆库与互联网后，仅仅找到一两个关于“骡子被扮成斑马”的使真者的例子，事情又当如何呢？很明显，按照对于概率本性的“频率主义”（frequentism）解释，这么少的案例是不足以构成合适的样本空间以构成概率赋值的基础的。而且更麻烦的是，如果系统以“世界上已知的骡子数量”为分母，以“被涂抹成斑马的骡子的数量”为分子，并由此来计算 Pr（斑骡$_{使真}$|背）的值的话，这个值将永远是很低的，因为世界上已知的骡子的数量肯定是海量的。由此，目标信念将永远因为“欠安妥”而被判为“谣言”。但这样的“一刀切”的评判方式会不会导致“谣言误杀率”太高的问题呢？

针对上述质疑，笔者的回复如下。

其实我们并不能以“世界上已知的骡子数量”为分母，以“被涂抹成斑马的骡子的数量”为分子，由此来计算 Pr（斑骡$_{使真}$|背）的值。其理由是：能够成为频率值计算的分母的，应当是相关假设的所有正面证据与负面证据的数量总和。具体而言，对于“出现被伪装成斑马的骡子”这一假设而言，构成该假设的负面证据的，并非“的确未曾被伪装成斑马的所有骡子”，而仅仅是“那些的确被证明无法被伪装成斑马的骡子”。因此，我们不能将“世界上已知的骡子数量”视为与“Pr（斑骡$_{使真}$|背）的值”对应的分子式的分母。

——但为什么这么说呢？从语言分析的角度看，“伪装成斑马的骡子”这一表达方式中的“伪装”一语，一方面既预设了“伪装物”与“被伪装的原物”的存在，另一方面又预设了将二者联系在一起的某种伪装目的的存在。从这个角度看，“伪装物”与“被伪装的原物”关系，或可类比于咳嗽药与咳嗽患者之间的关系：一方面，“咳嗽药”与“咳嗽患者”互相预设了对方的存在，而另一方面，将二者联系在一起肯定还有某种更深的目的，如“治疗咳嗽”。很显然，我们若要去判断“某型咳嗽药水具有疗效”这一假设自身的安妥性，我们所要考察的，便是被证明为可治病的药水的案例与被证明为不可治病的药水的案例的比例——而无论在此比例关系的分子项中，还是在其分母项之中，药水与病患之间的尝试性匹配总是存在的。换言之，刚刚走下流水线而尚未与任何病人发生关系的药水，并不构成“该型咳嗽药水具有疗效”这一假设的反例。同理，只有那些伪装用的油漆与被伪装的骡子彼此之间出现匹配尝试（而无论这些尝试成败与否）的案例，才能构成对于“存在着被伪装的骡子”这一假设的正面证据或反面证据，而那些与伪装用油漆毫无关系的自然状态的骡子，是不能被计入上述假设的证据集的。由此，我们也就不用担心由于世界上存在的海量骡子而导致的“目标信念永远欠安妥”的问题了。

有的读者或许会进一步提问：如若上述说法是正确的话，难道我们需要像检测药物一样，真的去找到相当数量的骡子，构成一个具有统计学意义的样本空间，并尝试着给这些骡子涂抹上斑马式迷彩色，由此计量“伪装成功”与“伪装失败”的比例吗？而由此导致的惊人的实验成本，又如何可能由一个安妥性评估机制来加以支付呢？

对于这个问题的解答乃是这样的：我们其实根本不必为此类“成本支付”的问题而担心，因为只要系统获取了少数的案例，以便证明“使用斑马迷彩

将骡子伪装为斑马”是可行的,并且,只要系统能够确定这些信息的信源具有一定的可靠性,那么,系统就立即可以输出一个关于目标信念的“安妥性”标签,并将其判断为“非谣言”。

不过,凭什么这里所说的安妥性评价机制能够以如此“节俭”的方式做出判断,而对于药物有效性的判断却不能如此“武断”呢?这是因为在伪装斑马的案例与试药的案例之间,还有一个非常重要的差异,此即,与药物相匹配的诸病患的内部生理条件之间的彼此差异,会使得单个的药物治愈案例难以具备普遍的推广意义,并由此倒逼假设验证者去建立一个具有统计学意义的样本空间,并在其中囊括不同种族、性别、年龄的病患。与之相比,诸骡子外形之间的差异,却不足以对某种伪装色的涂抹效果产生足够明显的影响。所以,如果有报道说仅仅有一头骡子被成功涂抹了斑马迷彩色,那么系统就应当可以立即推测出:这种情况可以被立即推广到所有的骡子的案例上——除非关于上述成功涂色的报道自身的信源是不可靠的。

——那么,倘若被二阶评估机制所评估的信念,恰恰是关于某种新药的药效的呢?很显然,在这种情况下,上面提到的这种规避样本空间构建的思路恐怕就难以成立了。这样一来,该系统又该如何在给出“目标信念安妥性与否”的判断过程中,继续规避巨大的假设验证成本呢?

对于这个问题,存在着一个简易的解决方案,同时还有一个更简易的解决方案。前者是这样的:如果系统本身就已经预存了关于医学的固定知识,特别是关于与此药类似的其他药物的疗效的知识,系统就可以通过类比推理来判断关于新药物的药效的信念是否具有安妥性。而关于此问题的更简易的解决方案是:系统在缺乏相关医学知识的前提下,将直接根据信念的信源可靠性来给出针对信念的安妥性判断。

那么,系统又当如何知道报道的信源本身是否可靠呢?按照“导论”所提及的“多能动者的信息主体信任识别机制”的研究思路,相关机制将根据各个信源之间的互联网活动轨迹,来给每个信源的权威性进行评估。但正如笔者所提到的,这一基于大数据技术的做法将无法避免“楚门的世界”的困境,因为大量的发言机器人的海量留言(或者是特定社会机制对于特定人类发言的删除)能够制造大量虚假的网络活动轨迹,由此制造虚假的意见领袖。而笔者对于该问题的解决办法则是基于与“大数据主义”截然不同的“无罪推定”(innocent until proven guilty)原则的。具体而言,除非如下条件被满足,否则,任何一个信源所给出的内容都会被认为是可信的。

条件(甲):该信源所给出的目标信念的内容,在逻辑上与评价系统内置的每条信念产生了严重的矛盾;

条件(乙):该信源历史上给出“谣言”的比例,高于平均水准,或是在与舆情热点所涉及的重要话题上曾经有过发布“谣言”的记录(这些记录的次数不必很多,有时候甚至一条就足以让信源被污染);

条件(丙):该信源被人类用户先验地判定为“不可信”的。

关于上述条件(丙),笔者还有进一步的说明。该限制条件的提出,体现了笔者心目中的鉴谣系统的“人类用户中心主义”色彩:任何本系统的人类用户,都可以根据他自己的教育背景与取信偏好,武断地规定特定类型的网络发言人是可靠的,或是不可靠的。当然,笔者也不否定,人类用户本身的这个决断,亦可能是某种范围更为宽广的“楚门的世界”精心诱导的产物——但尽管如此,由于个体人类用户之间自然存在的观点分歧,对于条件(丙)的吸纳,依然可以避免某种特定样式的“楚门的世界”的信息诱导机制在舆论场中占据统治地位,由此,真信息在舆论场中得以流通的概率,也可以得到提高。而这种安排本身,又是基于关于“合取概率”的基本概率学原理的:单一信源 A 出错的概率,先验地就要高于诸多彼此不可还原的信源全部出错的概率,因此,多个信源之间的彼此竞争,本来就是维护信息传统中的纠错机制的重要制度安排。从这个角度看,在 AI 的语境中对于这些信源之可靠性的自动化评估机制,也应当小心地维护人类文化与意见的多样性与丰富性,并在此过程中维护特定的人类用户自身的利益。与之相比,给出上述这种呵护的人本主义意图,却恰恰很难见容于以“赢家通吃”为自身运行原则,并以资本的抽象增值(或福柯式的全景式观察)为最重要目的的大数据技术。

五、总结与衍生型讨论

现在已经到对本文的讨论加以总结的时候了。

本文的宏观写作意图之一,就是在“网络鉴谣机制设计”的新技术语境中,重新激活某些学院意义上的认知论研究话题(特别是知识归赋中的“安妥性”问题),由此打通人文思辨与工程学研究之间的壁垒。这种打通至少还有两个具体目的:其一,通过工程学问题自带的“跨文化性”,来冲淡原始认识论话题中的“盎格鲁—撒克逊文化”痕迹,由此为英美认识论在华语世界的传播降低文化门槛;其二,也是更重要的是,上述这种工程学研究所自带的“可操作性”要求,将倒逼我们对认识论领域内的“安妥性”概念进行“可编程性”检查。至于这种倒逼性检查的具体结果,便是促使我们发现了:在可能世界语义学的框架中对于“安妥性”的处理方案实际上是缺乏编程指导价值的,因为这种处理方案包含了不可容忍的循环论证问题。而笔者给出的替代方案则是基于概率论的,并由此很自然地避免了对于“可能世界”这一神秘对象的建模难题。而在具体操作过程中,通过对于概率赋值的背景知识与目标信念之间关系的阐述,笔者的方案也尽量避免了对于大数据的依赖,这样一来,由此衍生的自动避谣机制也就能够更好地贴合特定人类用户的既有价值观。

但对于笔者所勾勒的这个技术路线图,读者们可能还会有进一步的疑问。第一个疑问是比较“务虚”的:如果真如笔者所言,每一个避谣机制都会成为特定人群的价值观的数码化增殖方式的话,那么每个机制所筛选出来的“谣言”就只可能是特定人群的偏见的产物了。而由此被筛选出来的“谣言”,又凭什么就一定是谣言呢?

对于这个问题,笔者的回答是:与人类用户一样,任何一种鉴谣程序都不可能是全知、全能者,所以我们在原则上就不可能制造出一台能够摆脱“视角主义”(perspectivism)之限制、并以“上帝之眼”来查看世间万物的“真理辨别机”。AI 所能做的,便是使得特定人群的鉴谣习惯能够得到机器的模拟,由此使得特定人群能够以更高的信息处理效率,最终看到他们所愿意看到的“真相”。至于特定人群的鉴谣用背景知识本身是不是有严重偏差,并会不会由此导致在另一群用户看来非常“荒腔走板”的鉴谣结果,则根本不是一个工程学问题,而是裁判话语权的建构问题。因此,这样的问题并不属于本文考量的范围,而应当留给政治哲学家去研究。需要指出的,大数据主义的支持者或许会通过海量数据的累积来达成一种对于所谓“上帝的视角”的模拟,并由此反对笔者提出上述技术进路的必要性。但为他们所忽略的一个事实却是:海量数据的累积,在原则上就会使得互联网舆论场中相对边缘的意见变得更加边缘化,由此导致评价背景的单一化。对于充满偶然性的高度复杂的现代社会来说,这种得到数码技术强化的“意见一致”,或许会给人类带来非常大的决策风险。

第二个疑问则比较“务实”:笔者所提议的这种鉴谣机制,是否具有立即投入使用的潜力呢?

对于这个问题,笔者的回答是:这一鉴谣机制的产品化,还面临着两个巨大的技术障碍,即“自然语言理解”(Natural Language Processing)与“常识表征”(Commonsense Representation)。具体而言,这样的机制应当有能力从不同的网络文本中提炼出各自的目标信念,并将这些信念改造为概率论可以理解的方式,以便成为相关概率值的合适承载者;同时,这样的机制还需要能够对特定用户的背景知识或常识进行计算化表征,以便利用这类表征结果来进一步计算“相关常识被给定的情况下目标信念的后验概率”。然而,对于 AI 研究的历史有所了解的人都知道,自然语言理解与常识表征,一向是 AI 研究中的两个高难话题,因此,从某种意义上说,笔者的鉴谣机制的正常运作,其实是预设了这些艰难的科研课题已经得到了全面的克服——但目前它们显然还没有被克服。从这个意义上说,笔者所提出的技术规划的确依然只是一个粗糙的方案,它还需要大量的工作对自身进行细化。但尽管如此,笔者依然要指出,如果我们要制造一台能够尽量像人类那样思考的谣言鉴别机器,对于上述两个技术困难的牵涉就是不可避免之事,因为作为“语言的动物”,并作为“一切社会关系的总和”,人类在本质上就是生活在自然语言与特定的常识所构成的“文化场”之中的,并因此天然地就是从事“自然语言处理”与“常识表征”的生物学机器。与之相比,那些试图仅仅从网络轨迹出发进行数据分析,而忽略所有这些重要事项的大数据鉴谣技术,却注定因为对于人性本质的忽略,而永远无法像人类那样思考。

注释

①本文所说的“认识论”乃是指现代分析哲学语境中的以“证成”(justification)与“知识”的本质为核心考量的研究分支,而与近代意

义上的认识论没有直接的关联。②《网络谣言》,百度百科,https://baike.baidu.com/item/%E7%BD%91%E7%BB%9C%E8%B0%A3%E8%A8%80/2425816?fr=aladdin.③“通用人工智能”研究与“人工智能”研究的区别是:前者将致力于设计一台能够通用于各种应用领域的智能机器,而后者则只致力于设计只能应用于特定领域的智能机器。④滕婕、夏志杰、罗梦莹、王筱莉:《基于 Multi-Agent 的网络谣言传播事件中信息主体信任识别研究》,《情报杂志》2020 年第 3 期。⑤Judea Pearl. Causality. *Models, Reasoning, and Inference* (*second edition*). New York: Cambridge University Press. 2009.⑥相关文献参见:Ducan Pritchard. *Knowledge*. London: Palgrave Macmillan, 2009; Ducan Pritchard. Safety-based Epistemology: Whither Now?. *Journal of Philosophical Research*. 2009, Vol. 34, pp.33-45. 正文中所给出的定义,根据汉语学术界的阅读习惯,做了适当简化。⑦该案例的原始版本见于:Alvin Goldman. Discrimination and Perceptual Knowledge. *Journal of Philosophy*. 1976, Vol.73, pp.771-791.⑧“88 号谷仓”只是为了讨论方便而采用的一个名字。取别的名字也不会影响讨论的结果。⑨这个隐喻的出处是:Timothy Williamson. *Knowledge and its Limits*. Oxford: Oxford University Press. 2000, p.124.⑩David Lewis. *On the Plurality of Worlds*. Oxford: Blackwell. 1986.⑪具体计算过程如下:根据合取概率的计算公式,$Pr(P\&P_{使真}) = Pr(P|P_{使真}) \times Pr(P_{使真})$。很显然,既然任何一个命题的使真者都能使得该命题为真,则 $Pr(P|P_{使真}) = 1$。这样一来,我们就得出:$Pr(P\&P_{使真}) = Pr(P_{使真})$。由于任何的概率估值都是在特定背景知识下进行的,因此,去计算 $Pr(P_{使真})$,也就等于去计算 $Pr(P_{使真}|背)$。⑫读者可能会问:为何在这里要强调“目标信念的使真者的概率”,而不干脆就是“目标信念的概率”呢?难道“桌上有水”这一信念的“使真者”,不就是“桌上有水”这一事实吗?对此,笔者的回应是:使得“桌上有水”这一信念成真的当然不是这个信念本身,而是与这个信念不同,却在内容上与之彼此符合的一个具体事态,如“张三的桌子上的确放着一杯水”。后者当然包含了比前者更多的信息。由此,在特定的背景知识中评估一个使真者的概率,自然会因为使真者自身的具体性而更容易调取评估者的相关记忆。⑬当然,会有不少信念的表面结构并不是“A 是 B”。如果遇到这种情况的话,系统会将这些信念先“翻译”为“A 是 B”的结构。比如,它会将“某病毒杀死了很多的民众”翻译为“某病毒是很多的民众的死因”。⑭该案例原始出处:Fred Dretske. Epistemic Operators. *Journal of Philosophy*. 1970, Vol. 67, pp.1007-1023.

责任编辑:涵　含

Internet Rumors and Epistemic Safety

—A Route-Map for Designing a Non-Big-Data-Based Automatic Rumor-Identifying System

Xu Yingjin

Abstract: From the perspective of contemporary epistemology, one of the necessary conditions for being "internet rumors", namely, "fiction which doesn't correspond to reality", can be replaced by the "epistemic unsafety of the truth-maker of the target belief", which literally means that it is very easy for the truth maker of the target belief to be false. The estimation of the degree of the safety/unsafety of the truth maker of a target belief, according to the blueprint of an Automatic Rumor-Identifying System (ARIS), is executed by virtue of the attribution of a proper value to the posterior probability of the truth maker of the target belief (with regard to certain background knowledge). Due to the completed interplay between the results of this probability-attribution and the foregoing type of background knowledge, which is in turn imbued with specific human users' social values, the desired type of ARIS is by nature not based on big-data technology, the digital tyranny induced by which could easily marginalize the voices of individuals of statistically insignificant personalities. More generally speaking, my pursuit of a "safety"-notion-based ARIS is motivated by an attempt to combine the insights both from epistemology and Artificial-General-Intelligence(AGI)-oriented engineering, and such attempt also provides a chance for us to re-check the feasibility of some typical epistemological proposals.

Key words: Internet rumor; Artificial General Intelligence (AGI); epistemic safety; probability theory; background knowledge