安全可信智能的可能技术路径

崔　鹏[1] 邓　柯[1] 王国豫[2] 王禹皓[1]
张　江[3] 朱　军[1] 朱占星[4]
[1] 清华大学
[2] 复旦大学
[3] 北京师范大学
[4] 北京大学

关键词：安全可信智能　可解释性

当今，人工智能技术已经逐渐渗透于人们日常生活的各个环节中。从清晨唤醒的智能音箱，到智能导航等出行软件，到信息搜索和远程会议等工作系统，再到餐馆、视频推荐等休闲平台，都能或多或少地看到人工智能技术的身影。这些领域很大程度上受到互联网经济“以产出为导向”（outcome-oriented）理念的影响，对预测准确度和效率等性能指标的关注远大于对预测风险的关注。受此影响，人工智能技术也多以预测性能的优化为主要目标，以性能驱动的模式进行技术演进。

展望未来10~20年人工智能的发展，我们预期人工智能技术应用将进入“深水区”，在医疗、工业生产、金融甚至军事等更多与人类社会生产生活密切相关的领域进行渗透。可以预见，人工智能技术必然在进一步释放社会生产力方面发挥驱动性作用。但这些领域都是风险敏感型领域，人工智能技术产生的错误将会给人类带来巨大损失，事关人类的生命健康、社会正义甚至国家安全等，因此必然要求人工智能技术的发展从以往的“性能驱动模式”转变为“风险敏感模式”。能否有效降低人工智能的系统性风险，实现安全可信的人工智能，在一定程度上决定了人工智能应用的深度和广度。

毋庸置疑，当前的人工智能技术系统在安全可信层面还存在着许多隐患。虽然近些年深度学习在若干应用领域取得了突破性进展，但其将“黑盒模型”发挥到极致，导致其本质上“不可解释”；对于独立同分布假设的过度依赖，导致其在真实场景下的性能“不稳定”；由于对数据中所存在虚假关联不能有效区分，导致其在社会性问题方面的预测决策“不公平”。诸如此类问题如何解决，目前尚缺乏有效的理论体系与方法。

为了厘清安全可信智能的内涵和外延，以及可能的技术路径，“安全可信智能”讨论组进行了深入而激烈的思辨和探讨。

从“可解释性”说起

谈到安全可信智能，多数人会首先想到可解释性问题。一般认为，只有人工智能的过程和结果都符合人的逻辑，对人可解释，人类才能感觉到人工智能是安全可信的。

观点1：人工智能的可解释性应按照对象区分为三个层次（王国豫、崔鹏、邓柯）

人工智能的可解释性涉及到若干个对象。一种是对**人工智能领域专家**的可解释性，对专家可解释重要的是得到同行领域专家的理解、认可和背书。第二种是对**行业使用者**（例如医生、石油工人）的可解释性，需要按照行业的规范和逻辑体系保证技

术的可解释性。第三种是对**公众**的可解释性，公众对新技术的理解和接受往往建立在专家同行之间认可和第三方权威机构背书的基础上，不单纯是技术层面的可解释，而是一个复杂的过程和机制。这三种可解释性之间存在着一定的差别，在具体研究过程中需要加以区分。

观点 2：公众关心人工智能是否可靠可信，而不是可解释（王国豫）

人工智能技术的可解释性问题是回答人工智能计算中关乎“为什么”的专业问题。套用亨佩尔（C. Hempel）和奥本海默（P. Oppenheim）提出的科学性解释的结构，一个可以解释的人工智能技术需要回答“根据怎样的原始条件，按照何种定律，产生了最终的这一现象和事态”这一问题。通常，可解释才可追溯、可问责，因此可解释性是可靠性和可信性的基础。对公众来说，他们更关心的是人工智能是否可靠可信，而不是是否可解释。而这更需要公众和科学家、技术人员之间的沟通和互动。这绝不是一个单纯的技术问题，而是一个复杂的科学社会学问题。

观点 3：人工智能对应用行业人员的关键是指标可解释（王国豫）

行业人员一般需要理解人工智能技术应用的规范和原理，因此，需要人工智能发展出一套对行业人员可解释的规范和指标体系。比如医生采纳某种新的药物，并不需要知晓该药物的机理和原理细节，但需要药厂提供一系列药物临床实验的结果和安全性数据，以及使用条件、标准和规范。因此，发展和界定人工智能技术在各行业的应用标准和规范，包括人工智能应用的条件和边界，是影响其行业应用的关键，也是实现人工智能技术安全可信的重要环节。

观点 4：人工智能可解释性的关键是针对领域专家范畴的机理可解释（全体）

公众和行业人员的信任是有间接性的。其实，人工智能研究要重点考虑的还是领域专家内部以及同行之间对模型机理的可解释性。关于机理可解释，可以归纳为三个层面：第一是边界，即人工智能技术能否从原理上保证其所适用的场景以及不适用的场景；第二是可回溯，即当人工智能技术出现问题或错误的时候，能否对出错原因进行回溯；第三是可验证，即对于人工智能技术的边界性和可回溯性能够进行全面科学的测试和验证。

总而言之，从技术层面，人工智能研究不需要考虑对公众可解释；对行业人员也不需要可解释，但需要建立第三方验证和保障体系；人工智能可解释性问题更多是对领域专家和学者可解释，也就是机理可解释。由此引出安全可信智能的三个基本问题：边界性问题、可回溯问题和可验证问题。

安全可信智能需要解决的三个基本问题

边界性问题 边界问题的含义是智能模型可以自主判断自身对问题可解的范围与不可解范围。在明确了智能模型的边界的情况下，设计者才能真正地确定一个智能模型可以安全可信使用的场景。

可回溯问题 当智能模型在运行中出现错误和故障的时候，设计者可以定位错误的产生局部和成因。可回溯是一种追责的机制，能够帮助设计者设计约束智能模型应用的规范。

可验证问题 即第三方保障和背书问题。第三方需要评估智能模型安全可信性的衡量指标和评测标准，尤其是对边界性和可回溯性的验证和测试。相应地，设计者也需要一个能够验证智能模型安全可信性的平台，对模型的改进提供反馈指导。

边界性问题

针对边界性问题，我们在三条技术路径上形成了共识，包括不确定性统计学习、复杂任务驱动的机器学习以及因果启发的机器学习。

路径 1　不确定性统计学习

观点 1：边界性问题的本质是不确定性问题（朱军）

边界性的衡量是智能模型对自身解决问题能力的自主判断。理想情况下，智能模型可解能力的边界可表示为其对所给出预测和判断的置信区间。边界之所以难以刻画，根本挑战在于智能模型的实际

应用环境是开放和不确定的，同时，在给定的有限训练数据下，智能模型也存在认知上的不确定性。为了让智能模型具备“知道自己不知道”的能力，必须合理地刻画不确定性。概率和统计是一种引入不确定性的工具和路径，但是现今的学习模型（特别是深度神经网络）并没有充分、合理地利用或考虑不确定性问题。

观点 2：基于目标函数优化的智能模型偏离了统计学习的本质（邓柯）

当前智能模型的技术路线和真正的统计学习存在一些重要的区别。统计分析强调对事物的不确定性有定量的刻画和了解。但是当前智能模型通常的做法是设计一个损失函数，把不确定性通过积分等方式整合到损失函数里面。比如分类任务中追求分类的精确度，其精确度是损失了分类不确定性的结果。虽然这种做法在许多场景下很有效，但是从统计角度考察，这是以丢失分布中的复杂信息为代价的。这种做法把问题过分简化了。

因此，以实现不确定性推断为目标，需要把数据的内在分布和不确定性完整地表达出来，而不是简单地简化成损失函数。因为简化后的损失函数的适用场景存在一定局限性。一种简化方式下的损失函数在某些场景下可能是适用的，而在其他某些场景下可能不适用。这是一个重要的问题。过去的统计研究都在研究不确定性，但从过去的统计研究的范式来讲，是研究小数据的场景，所涉及的模型相对简单。而现如今，在与人工智能相关的问题里，数据结构十分复杂，规模也十分庞大，直接使用原来统计研究的方法，会面临很大的挑战。统计学界也对这个问题做了很多反思。如果可以真正把统计学习能够处理不确定性问题的优势与人工智能处理大规模复杂数据的优势相结合，对两个学科都是非常大的贡献。

观点 3：确定性的监督信息是实现不确定性学习的一个障碍（朱军、朱占星）

一张“狗”的图像，一般被确定性描述成“狗”作为监督信息，而其中只有一部分像素是对应于“狗”。在如此完全确定性的信息的监督条件下，很难产生一个包含不确定性的模型。因此，需要引入监督信息的不确定性，并在推理过程中保护和维持这种不确定性，才有可能在本质上实现不确定性模型。应该在更大的逻辑框架下重新定位不确定性的统计学习框架。而只有不确定性模型才能提供在各种情况下的置信区间。

路径 2　复杂任务驱动的学习

观点 1：无法得到边界的原因是学习任务过于简单（朱占星、邓柯）

现如今智能模型尚无法有效处理边界问题的原因之一是现在的学习任务过于简单。例如分类任务，仅仅学习一个分类界面就足以提高准确率，而对于数据的内部结构，包括数据的底层产生机制并没有充分的理解。在一个没有信息的二分类任务中，两种类别的比例不等，从统计的角度来看，应该学习出两个类别的真实比例来做随机分类的决策。而在现有机器学习的框架下，为了提高训练数据中分类的准确率，学习算法会对所有样例选择预测比例较高的类别。从优化的角度来看，这种做法做到了最优，但是模型没有理解到数据内在的结构和分布，因而难以对预测做出合理的边界性判定。

观点 2：学习任务的复杂性和可实现性需要平衡（邓柯）

在机器学习模式里，很早之前就有关于分类错误所产生的惩罚的启示。在一个分类任务中，不同的分类错误所带来的损失可能是不同的，因此可以优化的目标函数需要一些变化。损失函数需要凭借人类的经验提前设定，并且依赖于具体的应用场景。不同场景下的损失函数设计方法是不同的。而从统计的角度考虑，统计研究从根本上反对直接使用目标函数。统计研究希望模型能够准确刻画数据背后的分布规律和不确定性，而不引入具体应用中的代价和惩罚的信息。具体的代价和惩罚应该放在后端的使用阶段中考虑。

分类问题只关注分类结果正确与否，而把背后的分布规律函数刻画出来是一个描述问题。描述问题远比分类问题要难得多。很多场景下只需要解决简单的分类任务即可，但是也可能在一些更复杂任

务的情况下，有必要学习数据的分布。

如今人工智能已得到高度发展，想要进一步实现安全可信智能，寻找到模型的边界，理解数据的结构和分布可能是一个必要之路。也许是时候去解决这个挑战性问题了。

观点 3：复杂任务可以理解为多个任务或通用任务（朱军、朱占星）

机器学习从总体上来讲是从统计学出发产生的，但是它把统计里面的很多内容过分简化了。因此，机器学习虽然能够解决一些问题，但是整个系统过于集中在问题的局部。比如在分类问题中，系统只为了画一个分类界面，而忽视了数据内在的结构。因为这些内在的结构对于解决这个任务可能并不必要。而最近一些更复杂的任务（例如对比学习、自监督学习）开始使用一些更复杂的目标来探索数据的产生机制和本质，从而牵引整个学习的过程。只有学习出数据的分布结构和本质，才具备探讨边界问题的基础。

此处的复杂任务有两种理解方式。其一，该任务更具有通用意义，它和大多数高层任务不是很相关，但是这个任务的层面更本质。比如在自然语言处理领域，它学习了语言结构、语言模型；在视觉领域，它学习了图像的底层信息结构。其二，该任务是多个任务的综合，即希望学习得到的模型能够支撑比较多的任务，这就要求模型对底层的分布和结构要有充分的理解，才能以不变应万变，实现多任务通用。

路径 3　因果启发的机器学习

观点 1：噪声是产生不确定性的主要因素（崔鹏、邓柯）

边界问题可以归约为不确定性问题，而不确定性问题的产生与噪声有很大关系。假设数据由真实信号和噪声两部分构成。当前基于关联统计的学习模型以数据拟合为主要目标，而数据拟合中包含了信号和噪声两部分，最后导致模型不能有效区分信号和噪声。在这种情况下，模型很难实现对确定性和不确定性的有效区分。如果能够估计数据的真实产生机制（即 true model），就可更好地解决边界性问题。

观点 2：因果统计旨在探索数据的真实产生机制（崔鹏、王禹皓）

因果学习一直在探索数据的真实产生机制，也就是尝试从数据中识别真实信号。如果在学习模型的基础上引入因果，在一定程度上实现真实信号和噪声的识别和区分，在这个框架下或许可以更好地解决不确定性问题，从而解决边界问题。同时，在理想的情况下，明确因果机制可以帮助我们现有的模型体系和模型框架高度简化。现有的模型因为不具备对因果机制的了解，使用了非常庞大复杂的网络进行数据拟合。例如回归分析中，现有模型可能采用成千上万的因子进行回归。但是在掌握了背后因果机制的情况下，可以建立一个使用很少变量的回归模型。当前的人工智能模型“知其然，但不知其所以然”，即只求关联但不求因果，相当多的复杂计算旨在处理数据中的噪声。因果的引入，可以发现数据中所蕴含的本质结构和规律，或许可以赋予模型“以不变应万变”的能力。但在大数据环境下挖掘背后的因果机制是相当有挑战性的，是一个值得持续探索的方向。

可回溯问题

针对可回溯问题，我们主要提出了两条可能路径，即基于因果的反事实推理和自省学习系统。

路径 1　基于因果的反事实推理

观点 1：反事实推理是解决可回溯性的根本途径之一（王禹皓、崔鹏）

可回溯问题的本质是针对已发生的事实和结果，追溯导致其发生的原因。该问题可被描述为“what if”的问题，即对于既定事实，假如对前置条件进行一定修改，事实将会如何改变。因此，本质上这是一个反事实问题。而反事实推理则是以此类问题为对象所发展出来的一套方法体系。其核心是要厘清系统内所有变量之间的因果性依赖关系，当对系统某些变量施加一些干预或改变时，其他变量会依照因果依赖关系进行传导，直至对系统输出产生影响。而此过程的逆过程，即对某输出结果进行路径性回溯。

观点2：知识可以辅助可回溯，但和因果推理在不同层面（朱军、崔鹏）

引入知识图谱，可以帮助解决回溯路径的问题，在某种程度上是比较接近可解释的一条路径，即对模型的预测或决策给出一定的原因解释。但知识图谱本质上应该属于推理的载体层面，关乎推理的对象，而不是推理的过程和方法。知识图谱可以从推理对象层面提高可回溯和可解释性。而基于因果的反事实推理是从推理模型的层面保证可回溯，这属于推理过程层面。另外，因果不等于知识，它更像是一种推理手段和推理逻辑，可以从数据中发现变量之间的关系，但是这种关系不一定能沉淀为具有普世价值和具有传承意义的知识。因此，在解决可回溯性问题中，应将知识图谱和因果推理在不同层面区别对待。

路径2　自省学习系统

观点1：人类实现可回溯是靠“自省”机制（张江）

人类的大脑存在“双系统”机制。卡尼曼（Kahneman）等人的研究表明，人类有“快思考”和“慢思考”两种思考模式。人类在产生预测、判断或决策行为时往往效率优先，不一定是逻辑驱动，而且往往是不自知的，即“快思考”。但是当人类反思的时候，则会尝试用一套系统来解释过去的行为，从而弥补“效率优先”所带来的风险，即“慢思考”，这套系统游离于行为系统之外。因此，一个具有可回溯性的智能系统可能由两套系统构成。第一系统是产生行为的系统，第二系统是一个基于逻辑、基于可解释的反事实推理的自省学习系统。第二系统能够对系统自身行为进行反思，从而逐渐达到稳态，并实现知识的沉淀。

观点2：从控制论角度，自省学习系统是一个二阶系统（张江）

自省系统是一个二阶系统。比如因果推断会从一些数据行为上提炼出其中的因果结构，而自省系统相当于在一个推理系统之上，基于推理机器本身的行为再归纳出来一个系统。就像我们人类，除了从外界学习信息识别，还可以把这种归纳和学习的过程本身当成数据，在脑子里面进行反思，然后再经过提炼形成更高效的学习和归纳机制。系统科学中的二阶控制论的观点与此类似。在复杂系统方面已有一些相关研究，这些研究能够给予我们一些启发。无论模型是一个神经网络，一个元胞自动机，还是一个多主体系统，在追溯的时候，都需要有另外一套机制或者另外一个计算模型，能够从原系统中读取数据和信息，并重构一个类似于因果图的结构，而这就是自省学习系统的目标。

可验证问题

路径1　面向智能算法的测试验证平台

观点1：智能算法的测试验证远比软硬件测试挑战大（朱占星、朱军）

智能算法测试不同于传统的软件测试，使用软件测试的方法来验证智能算法是不合适的。智能算法不仅仅涉及逻辑正确性的问题，还有性能如何测定的问题。在普通的软件测试中，软件是固定的程序，但智能算法是可以调节的，新的训练数据会引起程序的参数变化，最后导致输出的变化。在这种情况下，即便智能算法的逻辑没问题，但是由于结果是不固定的，这种直接检查输出的软件测试方法并不适用。

同时，使用软件硬件测试的方法无法解决智能算法验证的可扩展性问题。神经网络原本是个连续的系统，需要将它转化为离散的基于规则的系统。为原始神经网络的整个推断过程寻找一个等价的形式，可以用基于逻辑的软件测试方法进行验证。现有基于软件硬件测试的方法往往只能验证最多上百个神经节点，尚无法处理大规模神经网络模型。

观点2：解决大型神经网络测试和验证的两个思路（朱占星）

目前利用软件测试理论和方法对智能算法进行测试验证的基本思路是将连续系统转换为一个离散系统，于是朴素的基于规则的验证近乎于对所有的可能组合进行穷举，出现组合爆炸的困境，只能处理输入维度较低以及非常小型的神经网络的验证。

为了应对这个挑战，两个可行的思路值得去尝试。第一个是对目前的神经网络的输入和计算逻辑进行合理的聚类划分，使得每一类具有类似的验证

特点，从而避免穷举带来的高计算复杂度问题；另一个思路是通过设计对验证友好的神经网络结构，同时兼顾智能决策和验证复杂度。

路径 2　模块化智能系统

观点:模块化既是功能需求，也是测试需求（朱军）

未来的智能系统很有可能是模块化的，而不是靠一个结构、一个算法解决所有的问题。很多智能系统的智能在模块层次上体现了出来，因此模块化设计在一定程度上是为了满足功能性需求。模块和模块之间如何融合以实现比较复杂的智能任务，是未来重要的考察方向。另一方面，模块化将为解决智能系统的验证问题提供便利。目前，智能系统验证的可扩展性和可验证性都是待解决的问题。特别是较复杂的智能系统，很难对其进行直接验证和测试。而对于模块化智能系统验证，如果能验证好每一个小组件，再以此为基础验证大的组件，则可以形成可实现、可扩展的一种验证方式。因此，在安全可信的意义上，模块化是为了解决验证测试，以及事后的归因定位。

小结

本文所呈现的安全可信智能的基本问题和可能技术路径，均围绕人工智能性能可保证、机理可解释和系统可验证等技术范畴展开讨论。但最终实现安全可信智能，还要考虑人们对数据隐私、算法公平性等方面的深度关切，这就需要人工智能理论方法、技术系统、行业应用，以及面向未来智能社会的规范伦理等多个层面相互协同。显然，实现安全可信人工智能之路尚未明晰，但我们希望本文的一些初步探讨能够为该领域更深入的思辨和研究起到抛砖引玉之用。 ■

崔　鹏

CCF 杰出会员、理事、YOCSEF 主席。清华大学长聘副教授。主要研究方向为因果推理与稳定学习、网络表征学习。

cuip@tsinghua.edu.cn

邓　柯

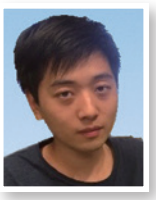
清华大学统计学研究中心长聘副教授、执行主任。主要研究方向为统计学、人工智能。

kdeng@tsinghua.edu.cn

王国豫

CCF 职业伦理与学术道德委员会主席。复旦大学哲学学院教授，复旦大学应用伦理学中心主任，博士生导师。主要研究方向为科学技术伦理学、技术哲学。

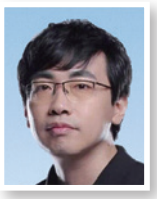
wguoyu@fudan.edu.cn



王禹皓

清华大学交叉信息研究院助理教授。主要研究方向为因果推断、高维统计。

yuhaow@tsinghua.edu.cn



张　江

北京师范大学系统科学学院教授。主要研究方向为复杂系统建模。

zhangjiang@bnu.edu.cn

朱　军

CCF 杰出会员、人工智能与模式识别专委会委员。清华大学教授。主要研究方向为机器学习。

dcszj@tsinghua.edu.cn

朱占星

北京大学助理教授。主要研究方向为机器学习、深度学习的理论及算法。

zhanxing.zhu@pku.edu.cn

（执笔人为第一作者，其他作者按照姓氏拼音排序）