从算法发展研究人工智能应用问题与对策*

郭金朋[1,2],韩敏捷[3],许正荣[2]
(1.浙江宇视科技有限公司,浙江 杭州 310051;2.安徽农业大学 信息与计算机学院,安徽 合肥 230036;
3.中国科学技术大学 生命科学学院,安徽 合肥 230022)

摘 要:基础硬件的发展带来了算力的提高,深度学习算法的出现和应用奠定了人工智能技术理论基础,Google、百度等高科技公司将人工智能应用的作为重点领域而加大投入,从而加速了人工智能的落地。当前人工智能已成为引领信息产业革命的最重要的技术手段。与此同时,人工智能对人类的潜在威胁也在加剧。对人工智能算法进行分析,归纳算法发展过程及缺陷,并总结这些缺陷可能引发的技术问题和伦理问题。最后,对人工智能的应用,从算法和伦理角度给出一定的对策。

关键词:人工智能;深度学习;算法;技术问题;伦理问题
中图分类号:TP391.4 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.201115

中文引用格式:郭金朋,韩敏捷,许正荣. 从算法发展研究人工智能应用问题与对策[J].电子技术应用,2021,47(7):21-28.
英文引用格式:Guo Jinpeng,Han Minjie,Xu Zhengrong. Research on artificial intelligence application based on algorithm analysis[J]. Application of Electronic Technique,2021,47(7):21-28.

Research on artificial intelligence application based on algorithm analysis

Guo Jinpeng[1,2],Han Minjie[3],Xu Zhengrong[2]
(1.Zhejiang Uniview Technologies Co.,Ltd.,Hangzhou 310051,China;
2.School of Information and Computer,Anhui Agricultural University,Hefei 230036,China;
3.School of Life Sciences,University of Science and Technology of China,Hefei 230022,China)

Abstract:AI (Artificial Intelligence) makes significant improvements with fast development of hardware improvement and algorithm iteration. Google, Baidu and other high-tech company increase investment in AI applications year by year, considering AI as a key areas. At present, artificial intelligence has become the most important technological means to lead the revolution of the information industry. However, the threat of artificial intelligence to humans is also increasing. In this paper, we analize the existing AI algorithm, explain their defects, and discuss what kinds of risks would be caused. Finally, we propose some approaches to generate more explainable, more robust, and more ethical AI system.

Key words:artificial intelligence;deep learning;algorithm;technical risk;ethical risk

0 引言

人工智能(Artificial Intelligence)是一门边缘学科,一般被认为是自然科学和社会科学的交叉,涉及计算机、逻辑学、生物学、物理学、社会学、认知学、数学及统计学、心理学和犯罪学等。关于人工智能的定义,不同的学者有不同的见解。McCarthy[1]在1955年将人工智能定义为“制造智能机器的科学与工程”。Andreas Kaplan和Michael Haenlein[2]将人工智能定义为“系统正确解释外部数据,从这些数据中学习,并利用这些知识通过灵活适应实现特定目标和任务的能力。美国麻省理工学院Winston教授在《人工智能》中认为:“人工智能是研究如何让计算机去实现人类做的智能工作”。美国斯坦福大学Nilson教授指出:“人工智能是关于知识的学科,是怎样表示知识、获得知识并且使用知识的学科”。当前,比较普遍的认识是,人工智能是人通过创造机器来实现人类智能的技术,它的任务是研究与设计智能主体(Intelligent Agents)。智能主体指一个可以观察周围环境(perception)并作出行动(action),以达致目标的系统。由于人工智能主要通过计算机程序来实现智能技术,因此在计算机科学领域持续发展。目前人工智能的基本研究范畴有计算机视觉、自然语言处理、语音识别、知识获取、感知识别、模糊控制等。近年随着计算机领域基础硬件的发展和软件算法的优化,人工智能发展迅速,并得到了广泛的应用。其主要应用有智能控制、机器人、自动化

*基金项目:安徽省高等学校省级自然科学研究重点项目(KJ2018A0145)

技术等。

人工智能扩展了人类的智慧，比人类思考更快、计算更快、输出结果更快。以人工智能为代表的信息产业被迅速“赋能”，从早期的算法研究到现在的产品化，科技进步推动产业变革，人工智能结合现有基础设备得以合理部署，前沿技术不断落地，人工智能正一步一步地改变人类的生活。然而，人类在享受人工智能的智慧成果的同时，也面临着人工智能的威胁。2014 年起，Amazon 使用人工智能招聘软件，该软件对男性求职者有倾向性，被认定为性别歧视。2018 年 3 月，Uber 自动驾驶汽车在美国 Temper 市测试时，撞到一名女子，致使其死亡。Deep Fakes 可以支持面部交换。2019 年，美国 Kneron 测试团队在荷兰机场用照片骗过智能登机终端。Facebook 曾收集超过 100 万张图片(约 10 万张人脸图片)构建 MS Celeb 用于商业公司的算法训练和军事面部识别研究等。人工智能是否能稳定运行，是否安全可靠，其开发和使用过程会不会侵犯人类的权利，如生命权和隐私权？人工智能的使用者是否有某些倾向而赋予人工智能倾向性等？人工智能的风险正逐渐受到人类的关注。

1 人工智能发展历程

1943 年，Warren McCulloch 和 Walter Pitts[3]提出人工神经元(Artificial Neuron)的概念，并给出其逻辑演算的数学模型，奠定了人工神经网络(Artificial Neural Networks)的基础。1950 年，Alan Turing[4-5]设计了一套机器智能测试方法，对机器表现智能行为的能力进行测试，提出了最早的机器智能概念“Baby Machine”(图灵模型)。图灵准则成为人工智能领域评价机器智能的重要标准。1956 年，计算机学家、认知学家、神经学家等在 Dartmouth 会议上正式提出“Artificial Intelligence”概念，一致认为，学习的任一方面及智能的任一特征都能够被准确表达，因而可以制造机器来实现这些智能。1957 年，心理学家 Rosenblatt[7]提出感知器(Perceptron)的概念，即具有简单配置的单层神经网络(One-Layer Network)。单层神经网络可以实现基本形状的区分(如三角形等基本图形)，使人们认识到人工智能的发展是有可能的。但是，囿于当时的计算机硬件，Rosenblatt F 的概念无法被广泛应用。1960 年，Widrow 和 Hoff[8]提出了一种在没有隐藏层的两层神经网络中根据输出误差产生的随机梯度下降法。1967 年，AMARI S[9]提出了一种在没有隐藏层的两层神经网络中根据输出误差产生的随机梯度下降法。1969 年，MINSKY M L[10]等人发表关于神经网络的缺陷的论文，指出，基于单层神经网络的感知器无法解决异或问题，因为人为设计出的特征层是固定的，不符合人工智能对智能机器的定义。至此，人工智能研究一度陷入困境。RUMELHART D E[11] 等人在 1986 年提出了第二代神经网络(BP 神经网络)，是一种与最优化方法(如梯度下降法)结合使用的，用来训练人工神经网络的常见方法，推动了人工智能算法的进一步发展，同时，该算法存在梯度下降问题，这一缺陷限制了 BP 神经网络的快速发展。1989 年，LECUN Y[11]开发了用于使用卷积神经网络(Convolutional Neural Networks，CNN)识别手写数字的神经网络体系结构，对于提高 DL 的效率至关重要，他们提出了使用反向传播算法直接从手写数字图像中学习卷积核系数的概念。手动设计参数依赖于研究者的经验，而自动学习可以自动从数据中提取特征，代入系统进行运算，在效率和性能上表现更为优异，因此更受研究者的青睐。2006 年，HINTON G E[13]提出基于自适应编码方法(Auto-encoder)的图模型，并提出深度信念网改善神经网络的梯度消失缺陷，由此深度学习的概念被提出，深度学习使机器能够深层次提取特征，同时抑制人为因素对神经网络模型好坏的影响，促进人工智能系统向更准确、更稳定、更智能的方向快速发展。2011 年起，微软将深度学习应用于语音识别研究，2012 年，Jeff 建立 Google Brain，运用神经网络进行猫识别的研究，建立高达 10 亿神经元的深度学习网络，并将同样的方法用于语音识别研究，将语音识别的误报率降至 20%，至此，人工智能应用在语音识别领域率先取得重大突破。此后，Google、科大讯飞、百度和 Apple 等相继进行语音识别的研究，并开发出 Deep Voice、微软小冰和苹果 Siri 等产品。2015 年，微软计算机视觉团队在 ImageNet 比赛中，其算法的误报率为 4.94%，而同等测试条件下，人类的误报率为 5.1%。这一结果表明人工智能的准确度超过了人类。2017 年，Demis 团队设计的 AlphaGo 打败了世界围棋排名第一的柯洁[15-16]。AlphaGo 的成功表明大数据、高算力、深度算法正在将人工智能推向新的高峰，这一阶段人工智能正在接近人类智慧。

2 算法发展与缺陷

2.1 信息表示的不确定性

以牛顿为代表人物的确定性科学人物认为世界是有序的、确定的、和谐的，只要设定初始条件，就可以计算出未来的一切[19]。对于不确定性，他们的观点是，不确定出现是因为人类先验知识的局限性，或者对初始条件的计算有误差。

麦克斯韦认为概率演算才是表达世界的真正逻辑，玻尔兹曼则在物理学中引入速记行，建立统计力学。当前，普遍被认可的观点是，不确定性是客观世界的基本特征，客观世界的绝大部分现象都是不确定的，而确定的现象，都要在特定的边界条件约束性才能发生[20-21]。

机器体获取信息的不确定通常与以下方面有关：客观信息存在和表达都具有随机性、模糊性，基因突变，遗传特征显现，海水走向，甚至地震等，客观世界充满了随机性。炊烟、云彩等常见的现象又无法通过准确的、规则的模型去反映，这都是模糊性的表现。随机性可以“概率”来表达，并借助随机变量的分布函数进行研究，例

如，人工智能界应用比较广泛的是以贝叶斯公式为基础的贝叶斯理论，贝叶斯理论利用先验知识和样本数据来获得未知样本的估计，从而使原本不确定的信息能够逻辑化地表达出来。1965 年，美国学者 ZADEH L A[22]提出模糊集合论，对模糊信息的表达做出了突破性贡献。在人工智能研究中，学者们通常通过模糊规则、模糊语义、模糊逻辑等方法来处理[22]。

2.2 模拟算法的不确定性

人工智能在发展过程中大致经过了符号主义(Symbolism)和连接主义(Connectionism)。符号主义(Symbolism)是由认知学家在人类脑科学等先验知识的背景下提出的，侧重逻辑演算，以先验知识和推理能力建立起模拟人类行为的符号模型，以 Warren McCulloch、McCarthy J 和 Simon H[17]的理论为代表。这一时期符号模型的应用要求模拟的知识没有模糊性和二义性，即必须是准确的，但是推理模型研究者的经验毕竟有限，模拟对象的知识存在极大的不确定性，因此符号模型面临模型难以构建等困境。

连接主义(Connectionism)是由数学家和神经学家提出的，模仿人类中枢神经系统的神经元，构建具有类神经元的网络模型(神经网络)，来实现机器的智能。这一理论是建立在如下假设之上：心理现象可以用统一的、简单的相互连接的网络来表示。因此，连接主义学派又称为仿生学派。连接主义学派的神经网络算法分为两个阶段，第一阶段，即人工智能发展早期(1956－1986)，Widrow 提出了有效的人工神经网络设计方法，Minsky 论证了人工神经网络的局限性，Rumelhart 提出了 BP 神经网络算法，使人工设计神经网络方法得到了广泛的应用。第二阶段，即人工智能快速发展时期(1989 至今)，1989 年 LeCun 基于 BP 神经网络设计了自动学习的网络算法，2006 年，Hinton 的自适应编码方法(auto－encoder)奠定了深度学习的基础，机器所认知的内容也从经验知识迈向常识知识，从此人工智能进入深度学习时代。连接主义同样面临着不确定性带来的挑战[18]。人工设计神经网络阶段，神经网络模型的好坏往往依赖设计者的先验水平，同时，人类的认知也具有不确定性和局限性，对人工智能系统的优劣有直接的影响。

2.3 深度学习算法的缺陷

深度学习阶段，当前深度学习算法在特定的数据库(图片、视频等)进行演算，其准确率可以超过人类。但同样囿于客观世界的不确定性，深度学习算法同样表现出鲁棒性差的特点。

2.3.1 易被攻击

SZEGEDY C[24]等提出深度学习具有反直觉性(counter－intuitive)。第一，在高维神经网络中，信息由空间携带，而非由单个的神经元携带。第二，深度神经网络的输入－输出是不连续的，实验表明，在图像分类应用上，可以通过某些扰动欺骗神经网络，导致其对图像分类错误。对于这些扰动，可称之为对抗样本(adversarial examples)，如图 1 所示(对抗样本图片链接 http://goo.gl/huaGPb)。Sezegedy 在同时提出，对抗样本具有迁移性，即，通过一个网络模型训练出来的对抗样本，不仅对该网络模型具备攻击性和欺骗性，它对其他的网络模型同样具有攻击性和欺骗性。

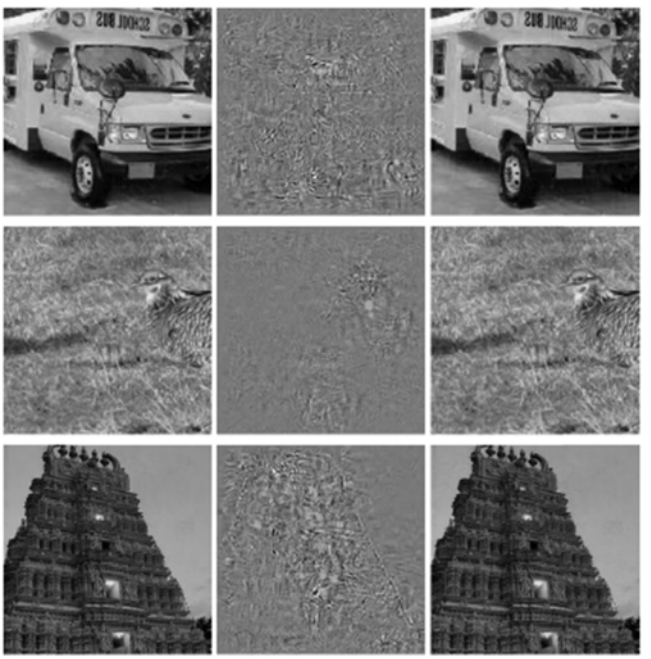

图 1 对抗样本产生

GOODFELLOW I J[25]则认为对抗样本来源是高维空间的线性性质。对抗样本的存在极大地威胁人工智能系统的安全性，此后，研究者们针对对抗样本防御进行了一系列研究[26]。包括图片预处理(pre－processing)、精馏法(defensive distillation)、正则化表达(regularize)、对抗性训练(adversarial training)等。对抗样本防御研究如表 1 所示。JPG 算法处理对抗样本如图 2 所示，精馏法原理如图 3 所示。

2.3.2 不可解释性

浅层的神经网络发展到多层的神经网络，有两个重要的变化，一是神经网络输入的信息不同，浅层的神经网络需要人工提取特征并输入，而深层的神经网络不需要，只需要开发者输入原始数据即可。二是多层的神经网络层数明显增多，且含有隐藏层，同时其性能有了较大的提升(2015 年的 ResNet[33]已达到 152 层)。深度学习模型的不可解释性是由其学习本质决定的，是典型的黑箱算法。第一，深度学习模型较为复杂，含有大量的超参数。如数据采集、预处理、特征提取、特征选择、神经网络层数、单层神经元梳理、关联性权重、线性和非线性表达等。不可解释性给深度学习的应用带来了潜在的安全威胁。在应用过程，在神经网络的快速计算和收敛，一旦

表 1 对抗样本防御研究

方法	原理	缺陷	研究者
图片预处理	用 JPG 等预处理方法对图片进行处理	如果对抗样本较大，预处理算法作用不明显，即使对抗样本的扰动程度远在分类器由于归纳偏差，同样会发生分类错误	Dziugate G K[27]
精馏法	通过精馏进行知识转移，即从较复杂的神经网络中提取知识，转移到较小的神经网络中进行计算，以此降低计算的复杂度，减小对抗样本的扰动的影响	对标准攻击进行略微修改，认为见效 softmax 的输入度，即可在防御精馏网络上发现对抗样本	PSPERNOT N[28]、CARLINI N[29]等
正则化表达	提出 DCN(Deep Contractive Network)弱化对抗样本的干扰	容易产生新的对抗样本，无法消除对抗样本	Luca[30]、ROSS A S[31]等
对抗样本训练	将对抗样本加入训练集中，对算法进行主动训练，增强算法的适应能力	无法解决对抗样本迁移导致的黑盒攻击	Kurakin A[32]等

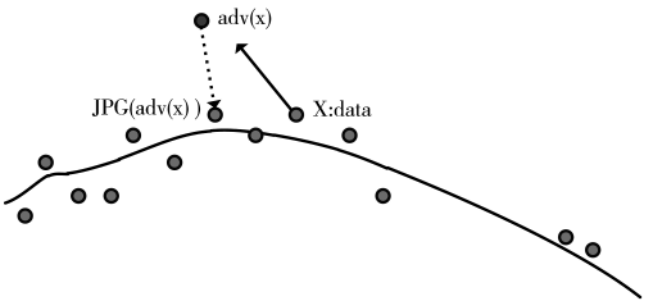

图 2 JPG 算法处理对抗样本

某个参数出现错误，在快速的收敛过程中，错误则被无限放大，同时由于大量的超参数存在，无法有效地排查是哪个参数出了问题，从而无法对相应的参数进行优化以规避或解决问题。这一障碍限制了其在医疗、无人驾驶、军事航天等领域的广泛应用。研究者们有的使用精馏、抗样本训练、梯度优化等方法来试图消除扰动，在完成这一动作后再单独处理可解释性。Mike[34]等人将多个正则化的深度神经网络与易于解释的区域决策树结合，在一定程度上提高了算法的可解释性。

2.3.3 鲁棒性

深度学习算法的准确率相对达到了很高的水平。在计算机视觉方面，2014 年，汤晓鸥[35]等在 LFW 数据集上取得了 97.35%的识别率，次年，Google[36]的 FaceNet 在 LFW 数据集上取得了 99.63%的识别率。语音识别方面，微软[37]在日常对话数据上的准确率达到了 5.9%。然而，这一切都是有条件的，比如，在特定的数据集上、语义没有二义性、确定信息以及限定领域。离开了特定的条件，表现往往不佳。比如在训练集上可以达到较高准确度，在测试集上准确度往往较低。人脸识别中姿态、年龄、光照等问题没有得到很好的解决。在生物特征识别、卫星、军事等依赖高可靠系统的领域。深度学习的鲁棒性亟待提高。

3 人工智能应用的问题

随着人工智能基础硬件[38-39]的飞速发展，基于深度学习的人工智能产业同样快速发展，并在城市智能交通、智能安防、智能机器人、智能医疗诊断、智能无人机、智能医疗系统、智能网联汽车等众多领域应用广泛[40]。人工智能产品研究和应用面临着安全性和伦理性等诸多风险。

3.1 安全问题

3.1.1 不确定性

以随机性和模糊性为基础的不确定性给决策过程带来了极大的困难。致使人工智能系统的决策具有不稳定性，不犯错则已，一犯错就是大错。人工智能系统在决策过程中将不确定性以概率的形式表达出来，在决策过程中，同一时刻、相同环境也可能做出差异较大的决策；人类囿于自身大脑的思考水平，在进行决策时，往往倾向选择自己思考的最优解，而人工智能系统拥有更加强大的“思考”能力，能在短时间进行大量计算并给出最优解，这一决策结果超过了人类先验知识的决策结果，同

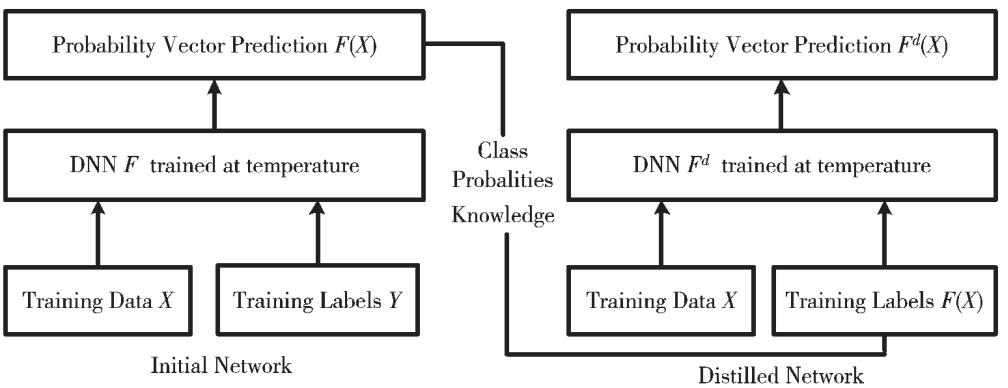

图 3 精馏法原理

人类的预期结果迥异。Kurzweil R[41]在《The singularity is near：When humans transcend biology》一书中表达了对“不可控”的担忧。在应用领域，如生物信息学领域，研究者们用深度学习进行药物分子适应性选择和预测[42-43]，Michael[44]等人训练基于基因特征的深度神经网络模型，来研究遗传变异，其他方面如医疗智能诊断[45]、远程外科手术、智能机器人、无人机和自动驾驶影响最大。

3.1.2 易被攻击

对抗样本的存在决定了基于深度学习的人工智能系统会受到来自外界信息的“欺骗”，这些“欺骗”会同真实信息一样，触发人工智能系统，输出一个正确的结果，这一问题对人工智能系统应用来讲是致命的，是不可接受的。以生物特征识别[46]领域为例：生物特征识别(biometrics)是指利用生物个体本身的特征来区分生物个体的技术。主要研究领域有脸、声音、指纹[47]、掌纹[48]、虹膜[49]等。近年来尤以人脸识别[50]较为火热。生物特征识别被广泛应用于公共安全[51-53]、智能支付[54]等领域。在实际使用中，由于对抗样本的存在，在一些条件下，可以用静态指纹或者人脸照片等轻而易举地骗取人工智能系统的信任，从而获取人工智能系统的授权[55]。近年来，研究者们尝试用多种方法提高人工智能系统的辨别能力[56-60]，但依然无法消除影响。这些缺点限制了人工智能系统在银行、公安等行业的使用。

3.1.3 不可解释性

人工智能的核心深度学习算法拥有大量的超参数，且模型复杂，层级众多，这些特点决定了人工智能系统的解释性较差，造成算法黑箱效应[61]。算法黑箱带来的影响有：(1)人工智能技术开发门槛降低，开发者只需要输入初始化参数，算法即可自己进行多次迭代运算，输出一个云端结果，但同样的，一旦结果有误，无从查起，更严重的是，算法在某个环节发生微小错误，这一错误随着算法多次迭代后，会输出不可控的、与预期大相径庭的结果。在人工智能系统中，这类错误一旦出现，必将是致命的[62]。(2)交互和共享难度大，开发人员无法进行模型共享和平台共享。(3)人工智能产品设计和开发人员的恶意使用会带来灾难性后果。黑箱具有极强的隐蔽性，在应用过程中容易被恶意利用，且难以预防。Armstrong S[63]在阐述人工智能军备竞赛模型中提到，拥有额外的开发团队和团队之间的敌意可能会增加 AI 灾难的危险，尤其是在承担风险比开发 AI 的技能更重要的情况下。

3.1.4 鲁棒性

当前深度学习的应用属于一种封闭式的、静态的、简单环境的应用，研究者们做了大量的假设和条件设定，以保证深度学习算法获得很好的效果。随着人工智能的快速落地，研究者们面临的环境将不再限定为某些特定场景，而是要面向开放的、动态的、复杂的场景[64]。人工智能系统越来越多地面对高风险的应用，面对多元化未知的场景，如复杂环境的自动驾驶[65]、智能电网管理[66-67]、证券交易预测[68]、控制性武器使用[69]，人工智能系统要做得更精准，同时差的时候也不能太差，可以归纳成对“鲁棒性”的要求更高[70-71]。

3.2 伦理问题

人工智能的伦理讨论伴随着人工智能发展的全过程[72]，一方面，人工智能的应用给人类生活带来了极大的便利，扩展了人类的智慧，能做到人类做不到的工作。另一方面，人工智能是否会侵犯人类的权益以及如何保障人工智能系统(如智能机器人)的权益一直处于争论中。由算法引发的人工智能伦理主要有：(1)责任界定问题。基于深度学习的人工智能系统具有超过人类的自我学习和计算能力，人工智能系统输出的结果可能是好的，也可能是坏的。例如，在自动驾驶应用中，算法的不确定性会导致决策失误，造成车祸；在生物特征识别中，人类可以使用欺骗样本获取生物识别系统的信任，从而完成包括鉴权、支付等一系列动作。(2)人工智能产品开发依赖于开发人员的道德约束。这一风险主要来自于人工智能算法的不可解释性。人工智能算法具备自动化的特征，一旦开发人员输入不当操作，无法及时停止，会致使不可预期的后果发生。

4 人工智能应用的思考

要创造和使用人工智能，至关重要的就是创造一个可信的人工智能，可信包括可靠、安全、符合伦理等。人工智能最终会以各种形态出现在我们的家中、道路上、医生办公室和医院、企业以及社区中。因此增强人工智能的可信度极为重要。换言之，只有可信的人工智能才能为人类所接受。

4.1 算法层面

如果人工智能不能可靠地表现，那么就不能信任它。不可解释、鲁棒性差等从人工智能发展伊始便一直存在且没有被完全解决。通常会看到，一个在特定环境下训练有素的算法可以轻松地识别出一辆校车，而当这辆校车处于翻车状态时，算法立即失去了作用。自动驾驶汽车行驶在复杂的、未知的、不可预测的环境中，医疗诊断系统在罕见的疾病分析上进行工作。世界是开放的、动态的，提高人工智能的可靠性是推广其应用的关键。第一，建立可靠模型。具体方法如引入变量、先验知识、逻辑推理和丰富认知模型，通过大量的现实场景获取变量、知识，从而通过推理和训练构造可靠性高的模型。第二，增强算法的可解释性。深度学习的最大挑战是黑箱效应。最大限度减小黑箱效应，为用户提供可靠的、安全的模型，帮助用户了解人工智能输入结果的原理和过程，才能使人类信任人工智能。如近年来已开展的基于特征的解释、模型逼近和解释模型等。

4.2 伦理层面

第一，规范人工智能开发过程。人工智能开发由多个环节组成，每个环节承担相应的工作，并实际对开发的结果负有责任。对于每个环节，要制定明确的开发任务，且这些任务是可被检查的，尽量消除开发过程的隐蔽性，防止开发者将个人偏见植入人工智能中，影响人工智能的决策进而影响到执行结果（如不可预期的结果、坏的结果等）

第二，制定道德规范。人工智能首先便是其开发者的价值体现，开发者的道德管理是首要任务。可以从以下几个方面入手：(1)谨慎选择开发者；(2)对开发者进行必要的道德训练；(3)制定明确的人工智能开发道德规范。

第三，建立人工智能应用监管准则。(1)限制人工智能的使用场景；(2)监管人工智能的使用过程；(3)培训合格的人工智能操作者；(4)明确人工智能的道德主体。

5 结论

当人类认识到人工智能的缺陷时，正是人工智能得到发展的时候，相对于对人工智能的缺陷表现出极大的恐惧，人类应该正视这些缺陷并不断努力去改善它。当前人工智能已经具备一定的学习能力，但是远远不够，人工智能像小孩子一样获取到了一定的知识，但对于这些知识并没有深入的理解和转化，也就是说，人工智能还没有拥有完整的认知过程。在未来的时间里，人工智能需要学习大量的知识，如政治、经济、社会等，提高同人类世界的交互能力，变得更加可靠。

“Rome won′t built in a day”，同样，完善的人工智能不是一蹴而就的，而是动态的、长期的。在人工智能开发过程中，不断地提高人工智能可信程度，约束开发人员和开发过程管理，制定相应的人工智能道德规范和行为准则，并且在应用过程中向人类证明，人工智能不具备或者具备较低的威胁（如不会完全取代人类或者造成较大伦理纠纷），建立起人与人工智能的信任关系。同时，人工智能的研究者需要注意到，任何信任关系都需要长时间的努力，从基础的信任到持续的信任，最终形成稳定的信任关系，要经过数年甚至数十年的努力，才能维护持续的信任，才能推动人工智能的进一步发展。

参考文献

[1] MCCARTHY J.What is artificial intelligence[R].1995.
[2] KAPLAN A，HAENLEIN M.Siri，Siri，in my hand：who′s the fairest in the land? on the interpretations，illustrations，and implications of artificial intelligence[J].Business Horizons，2019，62(1)：15-25.
[3] MCCULLOCH W，PITTS W.A logical calculus of ideas immanent in nervous activity[J].Bull Math Biophys，1943(5)：115-133.
[4] STRONG A I.Applications of artificial intelligence & associated technologies[J].Science[ETEBMS-2016]，2016，5(6).
[5] TURING A M.Computing machinery and intelligence[J].Mind.2018，LIX(236)：433-460.
[6] MCCARTHY J，MINSKY M L，ROCHESTER N，et al.A proposal for the Dartmouth summer research project on artificial intelligence(1955)[DB/OL].[2020-11-15].http：//www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html，2018.
[7] ROSENBLATT F.The perceptron，a perceiving and recognizing automaton：(Project Para)[Z].Buffalo(NY)：Cornell Aeronautical Laboratory，85-460-1，1957.
[8] HANSON S J.A stochastic version of the delta rule[J].Physica D：Alonlinear Pbenomena，1990，42(1-3)：265-272.
[9] AMARI S.A theory of adaptive pattern classifiers.IEEE Transactions on Electronic Computing，1967，EC-16(3)：299-307.
[10] MINSKY M L，PAPERT S A.Perceptrons：an introduction to computational geometry.Cambridge(MA)：Institute of Technology，1969.
[11] RUMELHART D E，HINTON G E，WILLIAMS R J.Learning representations by back-propagating errors[J].Nature，1986，323(6088)：533-536.
[12] LECUN Y，BOSER B，DENKER J S，et al.Back-propagation applied to handwritten zip-code recognition[J].Neural Comput，1989，1：541-551.
[13] HINTON G E，SALAKHUTDINOV R R.Reducing the dimensionality of data with neural networks[J].Science，2006，313：504-507.
[14] HE K，ZHANG X，REN S，et al.Delving deep into rectifiers：surpassing human-level performance on imagenet classification[C].2015 IEEE International Conference on Computer Vision(ICCV).IEEE，2016.
[15] SILVER D，HUANG A，MADDISON C J，et al.Mastering the game of Go with deep neural networks and tree search[J].Nature，2016，529：484-489.
[16] CHOI J，SHIN K，JUNG J，et al.Convolutional neural network technology in endoscopic imaging：artificial intelligence for endoscopy[J].Clinical Endoscopy，2020，53(2)：117-126.
[17] NEWELL A，SIMON H A.Computer science as empirical inquiry：symbols and search[J].Communications of the ACM，1976，19(3)：113-126.
[18] DONALD H.The organization of behavior[M].New York：Wiley，1949.
[19] 牛顿.自然哲学之数学原理[M].王克迪，袁江洋，译.北京：北京大学出版社，2006.
[20] 李德毅，杜鹢.不确定性人工智能[M].北京：国防工业出版社，2005.
[21] ROBLEY W.A brief history of time[M].The Cambridge Economic History of Europe/.Cambridge University Press，2013.
[22] ZADEH L A.Fuzzy sets[J].Information and Control，1965，

8(3):338-353.
[23] GAU W L,BUEHRER D J.Vague sets[J].IEEE Transactions on Systems,Man and Cybernetics,1993,23(2):610-614.
[24] SZEGEDY C,ZAREMBA W,SUTSKEVER I,et al.Intriguing properties of neural networks[J].arXiv preprint arXiv:1312.6199,2013.
[25] GOODFELLOW I J,SHLENS J,SZEGEDY C.Explaining and harnessing adversarial examples[J].arXiv preprint arXiv:1412.6572,2014.
[26] 张思思,左信,刘建伟.深度学习中的对抗样本问题[J].计算机学报,2019(8):1886-1904.
[27] DZIUGAITE G K,GHAHRAMANI Z,ROY D M.A study of the effect of jpg compression on adversarial images[J].arXiv preprint arXiv:1608.00853,2016.
[28] PAPERNOT N,MCDANIEL P,WU X,et al.Distillation as a defense to adversarial perturbations against deep neural networks[C].2016 IEEE Symposium on Security and Privacy(SP),IEEE.2016:582-597.
[29] CARLINI N,WAGNER D.Defensive distillation is not robust to adversarial examples[J].arXiv preprint arXiv:1607.04311,2016.
[30] GU S,RIGAZIO L.Towards deep neural network architectures robust to adversarial examples[J].arXiv preprint arXiv:1412.5068,2014.
[31] ROSS A S,DOSHI-VELEZ F.Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C].Thirty-Second AAAI Conference on Artificial Intelligence,2018.
[32] TRAMÈR F,KURAKIN A,PAPERNOT N,et al.Ensemble adversarial training:attacks and defenses[J].arXiv preprint arXiv:1705.07204,2017.
[33] SZEGEDY C,VANHOUCKE V,IOFFE S,et al.Rethinking the inception architecture for computer vision[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:2818-2826.
[34] WU M,PARBHOO S,HUGHES M,et al.Regional tree regularization for interpretability in black box models[J].arXiv preprint arXiv:1908.04494,2019.
[35] SUN Y,WANG X,TANG X.Deep learning face representation from predicting 10,000 classes[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2014:1891-1898.
[36] SCHROFF F,KALENICHENKO D,PHILBIN J.Facenet:a unified embedding for face recognition and clustering[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015:815-823.
[37] XIONG W,DROPPO J,HUANG X,et al.Achieving human parity in conversational speech recognition[J].arXiv preprint arXiv:1610.05256,2016.
[38] BENNIS M.Smartphones will get even smarter with on-device machine learning[J].IEEE Spectrum:Technology,Engineering,and Science News,Accessed,2019,25.
[39] JIANG Y,HUANG P,ZHU D,et al.Design and hardware implementation of neuromorphic systems with RRAM synapses and threshold-controlled neurons for pattern recognition[J].IEEE Transactions on Circuits and Systems I:Regular Papers,2018,65(9):2726-2738.
[40] SILVER D,HUANG A,MADDISON C J,et al.Mastering the game of Go with deep neural networks and tree search[J].Nature,2016,529(7587):484-489.
[41] KURZWEIL R.The singularity is near:when humans transcend biology[M].Penguin,2005.
[42] MA J,SHERIDAN R P,LIAW A,et al.Deep neural nets as a method for quantitative structure-activity relationships[J].Journal of Chemical Information and Modeling,2015,55(2):263-274.
[43] LIU R,WANG H,GLOVER K P,et al.Dissecting machine-learning prediction of molecular activity:is an applicability domain needed for quantitative structure-activity relationship models based on deep neural networks?[J].Journal of Chemical Information and Modeling,2018,59(1):117-126.
[44] LEUNG M K K,XIONG H Y,LEE L J,et al.Deep learning of the tissue-regulated splicing code[J].Bioinformatics,2014,30(12):i121-i129.
[45] YU K H,ZHANG C,BERRY G J,et al.Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features[J].Nature Communications,2016,7(1):1-10.
[46] JAIN A K,ROSS A A,NANDAKUMAR K.Introduction to biometrics[M].Springer Science & Business Media,2011.
[47] AHMAD S,LU Z M.A joint biometrics and watermarking based framework for fingerprinting,copyright protection,proof of ownership,and security applications[C].2007 International Conference on Computational Intelligence and Security Workshops(CISW 2007),IEEE.2007:676-679.
[48] LU G,ZHANG D,WANG K.Palmprint recognition using eigenpalms features[J].Pattern Recognition Letters,2003,24(9-10):1463-1467.
[49] WILDES R P.Iris recognition:an emerging biometric technology[J].Proceedings of the IEEE,1997,85(9):1348-1363.
[50] AMOS B,LUDWICZUK B,SATYANARAYANAN M.Openface:a general-purpose face recognition library with mobile applications[R].CMU School of Computer Science,2016,6(2).
[51] WEITZBERG K.Biometrics,race making,and white exceptionalism:the controversy over universal fingerprinting in Kenya[J].The Journal of African History,2020,61(1):

23-43.
[52] TAMISIER L,THIBURCE N,PRAT L,et al.The fingerprint disaster victim identification toolkit:from powder to biometrics[J].Journal of Forensic Identification,2019,69(4):413.
[53] TOOLEY J.The future of biometrics in policing worldwide[J].Biometric Technology Today,2020,2020(1):5-7.
[54] PIRZADEH K,KEKICHEFF M B.Mobile payment application architecture:U.S.Patent 10,454,693[P].2019-10-22.
[55] XU Y,PRICE T,FRAHM J M,et al.Virtual u:defeating face liveness detection by building virtual models from your public photos[C].Proceedings of the 25th USENIX Conference on Security Symposium,2016:497-512.
[56] REHMAN Y A U,PO L M,LIU M.LiveNet:improving features generalization for face liveness detection using convolution neural networks[J].Expert Systems with Applications,2018,108:159-169.
[57] ALOTAIBI A,MAHMOOD A.Deep face liveness detection based on nonlinear diffusion using convolution neural network[J].Signal,Image and Video Processing,2017,11(4):713-720.
[58] KOSHY R,MAHMOOD A.Optimizing deep CNN architectures for face liveness detection[J].Entropy,2019,21(4):423.
[59] REHMAN Y A U,PO L M,LIU M.SLNet:stereo face liveness detection via dynamic disparity-maps and convolutional neural network[J].Expert Systems with Applications,2020,142:113002.
[60] CORSETTI B,SANCHEZ-REILLO R,GUEST R M.Ergonomics in mobile fingerprint recognition systems:a user interaction evaluation[C].International Conference on Applied Human Factors and Ergonomics,2020:382-389.
[61] KUMAR D,TAYLOR G W,WONG A.Opening the black box of financial ai with clear-trade:a class-enhanced attentive response approach for explaining and visualizing deep learning-driven stock market prediction[J].arXiv preprint arXiv:1709.01574,2017.
[62] YUDKOWSKY E. Artificial intelligence as a positive and negative factor in global risk[J].Global Catastrophic Risks,2008,1(303):184.
[63] ARMSTRONG S,BOSTROM N,SHULMAN C.Racing to the precipice:a model of artificial intelligence development[J].AI & Society,2016,31(2):201-206.
[64] SARMA G P,HAY N J,SAFRON A.AI safety and reproducibility:establishing robust foundations for the neuropsychology of human values[C].International Conference on Computer Safety,Reliability,and Security,2018:507-512.
[65] JIN I G,SCH?RMANN B,MURRAY R M,et al.Risk-aware motion planning for automated vehicle among human-driven cars[C].2019 American Control Conference(ACC).IEEE,2019:3987-3993.
[66] HAN F,TAYLOR G,LI M.Towards a data driven robust event detection technique for smart grids[C].2018 IEEE Power & Energy Society General Meeting(PESGM).IEEE,2018:1-5.
[67] BERRIEL R F,LOPES A T,RODRIGUES A,et al.Monthly energy consumption forecast: a deep learning approach[C].2017 International Joint Conference on Neural Networks(IJCNN).IEEE,2017:4283-4290.
[68] KITAMORI S,SAKAI H,SAKAJI H.Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning[C].2017 IEEE Symposium Series on Computational Intelligence(SSCI).IEEE,2017:1-7.
[69] TUTTY M G,WHITE T.Unlocking the future:decision making in complex military and safety critical systems[C].Systems Engineering Test and Evaluation Conference 2018:Unlocking the Future Through Systems Engineering:SETE 2018.Engineers Australia,2018:557.
[70] DIETTERICH T G.Steps toward robust artificial intelligence[J].AI Magazine,2017,38(3):3-24.
[71] MARCUS G.The next decade in ai:four steps towards robust artificial intelligence[J].arXiv preprint arXiv:2002.06177,2020.
[72] ANDERSON M,ANDERSON S,ARMEN C.Towards machine ethics:implementing two action-based ethical theories[C].Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics,2005:1-7.

(收稿日期:2020-11-15)

作者简介:
郭金朋(1992-),男,硕士,工程师,主要研究方向:人工智能、模式识别。
韩敏捷(1992-),女,硕士,主要研究方向:生物信息学。
许正荣(1974-),女,硕士,副教授,主要研究方向:通信系统。

扫码下载电子文档

版权声明