DOI: 10.3724/SP.J.1224.2020.00252

“新兴医学工程伦理与治理”专刊

基于“医患信任”理论的医疗 AI 可信度问题的探讨

徐汉辉

（南开大学 医学院，天津 300071）

摘 要: 21 世纪以来，人工智能技术在医疗领域中得到了广泛的应用，其在疾病预测方面显示出超出人类医生的精准和高效。然而，由于尚不清楚人工智能系统内部的工作原理，这种精准预测是如何实现的仍然无法解释。这被称为人工智能的“黑箱问题”。“黑箱问题”引发了人们对于人工智能“不透明性”的担忧，进而对其可信度提出了质疑。面对质疑，一种观点认为，对于“黑箱”的担忧并无必要。这种观点的持有者往往强调，很多时候，人类医生是凭着直觉和经验在为病人做诊断，而这些直觉和经验连他们自己也很难解释清楚。既然我们能够信任人类医生，也应该信任人工智能。基于信任理论，从医患信任的角度来回应上述观点，尝试论证，医患信任和患者对于医疗人工智能的信任，这两种信任的基础完全不同，不具有可比性。

关键词: 医疗人工智能；黑箱；医患关系；信任

中图分类号: R-02 ;TP39 **文献标识码**: A **文章编号**: 1674-4969(2020)03-0252-08

基于深度学习（Deep Learning）的人工智能，其工作机理是模拟人脑神经网络，对输入的数据进行“自主”学习，进而输出精准的结果。正如人类尚不清楚人脑的工作原理一样，应用于医疗诊断的人工智能系统对于患者、医生甚至程序设计者来说都是不折不扣的“黑箱（Black Box）”。即，虽然这些人工智能系统能够对某些疾病进行精准预测，但研究人员并不知道这种精准预测是如何实现的。由此引发了人们对于其“不透明性”的担忧，进而对其可信度提出了质疑。国内围绕人工智能中伦理问题的探讨更多地集中在数据隐私保护[1]、社会公平公正等当面[2]。对于“黑箱”的不透明性的讨论也多是指出这一问题的存在[3]。相比较而言，本文将从信任这一概念的本质出发，基于医患信任的理论来展开讨论，并对医疗人工智能黑箱问题的已有辩护予以回应。

1 医疗 AI 的发展及其可信度问题

21 世纪以来，人工智能技术（Artificial Intelligence，AI）在医疗领域中得到了广泛的应用，其在疾病预测方面显示出超出人类医生的精准和高效。首先，在乳腺癌的预测和诊断方面，人工智能有着不俗的表现[4]。在大量学习乳腺癌患者的乳房 X 光片后，对于新的乳房 X 光片，人工智能系统能够准确预测出就诊者是否有得乳腺癌的风险。对于其他疾病的预测，人工智能也有着骄人的“成绩”。2015 年，位于纽约西奈山医院（Mount Sinai Hospital）的一个研究团队研发出了一套名为“深度患者（Deep Patient）”的人工智能（诊断）系统。通过对该医院历年来总计 70 余万份电子病例的“深度学习”，“深度患者”系统可以准确地预测出新病例中“隐藏”的疾病[5]。不

收稿日期: 2020–02–26; 修回日期: 2020–06–02

作者简介: 徐汉辉（1988–），男，博士，讲师，研究方向为生命伦理学、医学伦理学。E-mail：018205@nankai.edu.cn

仅如此，“深度患者”系统对于精神分裂症这类精神疾病的预测有着难以想象的精准度[5]。相比较而言，对人类医生来说，由于病症的复杂，病因的不确定性，精神类疾病一直以来都是最难诊断的。

然而，这些出色表现并未打消人们对于人工智能“不透明性”的担忧。一种观点认为，将诊断结果寄希望于一个无法解释工作原理和内部机制的“黑箱机器”，是对患者的不负责任，因此，考虑到人工智能的不透明性，不能盲目地信任其所给出的诊断结论，以免造成严重的后果。与之相对的另一种观点则认为，人工智能优于人类医生的高精准度就是其可信度的保障，对于“黑箱”的担忧，尽管可以理解（人们总是会对某些未知的不确定的事物感到担心），但其实并无必要。这种观点的持有者往往强调，就所谓“黑箱”而言，人类医生并不比人工智能好多少。很多时候，人类医生是凭着直觉和经验在为病人做诊断，而这些直觉和经验连他们自己也很难解释清楚。如韦德·潘杰（Vijay Pande）表示：“人类智能本身就是，而且一直总是，一个黑箱。例如，当一个人类医生做诊断的时候，患者可能会问这位医生，她是如何给出这个诊断结论的？这位医生大概率会告诉患者她是如何从已有的数据中得出的结论。然而，这位医生真的可以清楚地解释出她是从什么样的研究中怎样得到支撑她结论的数据吗？当然，她一定会给出自己做诊断的理由，但这其中难免有猜测或者凭借直觉的成分。”[6]如果我们能够信任医生给出的“难以解释”的诊断，也就没有理由排斥人工智能系统给出的结论。

上文中对于医疗 AI 可信度的第二种观点就涉及了人们对医生和对医疗 AI 的信任的比较，为回应这一比较，必须厘清使用与未使用医疗 AI 时的“信任”情况。不同的医患关系模式下，医患信任的依据和表现有所不同；现代医患关系中，医患信任的基础是患者知情同意权的有效保障。从这一点说来，患者对于“黑箱”模式下人工智能系统的诊断结论缺乏医患信任中的必要因素。换句话说，患者对于医生诊断信任的基础是自己可以充分知情及在知情情况下的自主决定，哪怕只是部分知情（如上文潘杰所言，很多时候，医生提供的信息也有基于自身揣测的成分），而这种充分知情的缺失恰恰是患者对于医疗人工智能“黑箱”不信任的原因。因此，认为“人类医生的诊断很多时候也难以解释，患者既然信任人类医生也应该信任医疗人工智能”的观点并不如其所言的那样可靠。基于此，笔者将尝试从哲学角度厘清“信任”这一概念的本质及不同种类的信任，进而梳理医患关系模式的发展历程，并分析不同医患关系模式下“医患信任”的基础和特点。为区分使用与未使用医疗 AI 时的“信任”情况，需要论证在现代医患关系模式下，医患信任的基础之一是患者知情同意权的保障，也就是说，尽管在一些个案中，医生可能会给出“难以解释”的诊断，但患者的信任并不是所谓的“盲目信任（Blind Trust）”。而这将使得医患信任区别于患者对医疗人工智能的信任，因为，在不理解人工智能系统内部工作原理的情况下，对其诊断结论的信任必然是一种盲目信任。

2 “信任”与“医患信任”

分析医患信任这一特殊关系中的信任之前，需要厘清更一般意义上的信任概念。一个有意思的现象是，很多有关信任的哲学讨论[7, 8]都选择从著名的“囚徒困境（Prisoner’s Dilemma）”开始。“囚徒困境”是这样一种假设：两名银行劫匪 A 和 B 被捕入狱，警方对单独关押的 A 和 B 说了同样的话：“如果你招供而你的同伙保持沉默，我们将无罪释放你，并用你的证词指控你的同伙，他将因此被判刑 5 年；反之，则是你的同伙被释放而你被判 5 年。如果你和你的同伙同时招供，由于证据确凿，你们都将被判 3 年。而如果你和你的同伙都保持沉默，那么，由于证据不充分，你们都将只被判 1 年。”面对这样的境况，两名劫匪该如何选择？如果我们把这两名劫匪看作是一个

利益共同体，那么，显然，两人相互配合的利益最大化结果是都保持沉默，均被判 1 年。然而，要实现这一结果，双方都要（1）信任对方不会招供，且（2）确信对方也信任自己不会招供。

从上述分析中，我们能够对信任做出一个初步的概述：信任是一种态度和倾向——将自己置身于一种对他人的依赖之中并持有一种信念，即所依赖之人有能力且出于好意（或至少没有恶意）完成（某事）[9]。首先，信任需要寄予信任的一方对于被给予信任的一方有所依赖。如上文中的囚徒困境，对于两名劫匪而言，利益最大化的结果依赖于对方的决定。这种依赖也就意味着，寄予信任的一方对于被给予信任的一方来说是处于一种脆弱的或者易受伤害的（Vulnerable）境地。这种境地体现在，寄予信任的一方有可能遭到背叛。在囚徒困境中，假设劫匪 A 信任其同伙 B 不会招供，为求二人利益最大化，劫匪 A 最终选择保持沉默。然而，劫匪 B 却并未如 A 所料想的那样守口如瓶，而是选择了招供。这样一来，A 对 B 的信任便遭到了背叛。信任的这一特征表明信任本身是有风险的。其次，信任需要寄予信任的一方对被给予信任的另一方（完成某事）的能力持乐观态度。仍以囚徒困境为例，当劫匪 A 决定保持沉默以求实现二人利益最大化时，他一定对于同伙 B 抵御警方“威逼利诱”有所信心。最后，信任需要寄予信任的一方认定被给予信任的一方对其有善意或者至少不会有恶意。在囚徒困境中，我们也能想象到，当劫匪 A 决定保持沉默时，他一定相信其同伙 B 不会有意出卖自己。如果劫匪 A 了解到，其同伙 B 早有致自己于死地以独吞赃款之意，那么，A 对 B 的信任一定会大打折扣。这里需要指出的是，信任并不要求寄予信任一方排除被给予信任一方出于自利动机行动的可能。也就是说，当我们信任他人时，并不一定要求对方一定对我们有善意（Good Will），有些时候，即使当他们出于自利动机行动时，我们也会信任他，只要其行动动机不是恶意的。比如，囚徒困境中，劫匪 A 可以信任劫匪 B，相信即使 B 依据自利动机行动，也应该会选择保持沉默。

就信任的分类而言，朱莉·安妮·罗斯坦（Julie Anne Rothstein）[9]56-72 将信任分为自我信任（Self-Trust）、婴儿期信任（Infantile Trust）、人际信任（Interpersonal Trust）、角色信任（Role Trust）以及机构信任（Intuitional Trust）。简单说来，自我信任和婴儿期信任是一个人成长工程中所具有的建立其他信任的能力和基础。这里不做展开。人际信任是指人与人之间所建立的信任，如与亲人、朋友、同事、合作伙伴之间形成的信任，可以表达为：P 信任 Q 做 X。囚徒困境中的信任就是这样一种信任。影响人际信任的主因来自两方面，一是 P 对 Q 的了解，如基于 Q 之前的表现来判断 Q 是不是一个值得信任的人，或者基于对 Q 的了解来判断 Q 有没有能力完成 X；二是 P 和 Q 之间的亲密程度，即 P 和 Q 的关系是否亲密到了足以让 P 放心 Q 不会辜负 P 的信任去完成 X。回到囚徒困境中，当劫匪 A 决定保持沉默的时候，一定是他对同伙 B 有信心，知道以 B 的为人不会轻易出卖朋友也能够抵挡住警方的“攻势”，同时，A 也相信，其和 B 之间的关系紧密到 B 不会出卖自己。

与之相比，角色信任弱化了个体之间的直接信任。所谓角色信任是指对于处在某个特定角色中的个体的信任。这种信任最直接的体现在职业信任，即对于从事某种职业的人的信任。例如，在当前的医疗模式下，很多患者去医院看病，对于给自己看病的医生也许并不认识，患者不了解他们面前的这位医生是不是一个值得信赖的人，但是，基于对医生这一职业的信任，患者仍然会选择与其合作并相信这位医生给出的治疗建议。类似的信任还包括，对于某位（陌生的）警察的信任，相信他会为市民提供保护，对于某位（陌生的）老师的信任，相信他能够传道授业。不同

于人际信任，角色信任并非直接基于对一个人的了解和彼此之间的亲密程度，而是基于对一个特点群体（如医生）的信任。罗斯坦将角色信任进一步细化为广义和狭义。他说：“广义的（角色信任）更加社会性和角色性。从这个意义上说，一个人相信飞行员可以安全地驾驶飞机从一个地方到另一个地方。狭义的（角色信任）则加入对具体情境和个体的考虑。从这个意义上说，一个人可能不那么信任一个一身酒气踉踉跄跄进入驾驶舱的飞行员。同样的，一个人可能也会不那么相信一身酒气上手术台的医生。”[9]65-66 也就是说，在一个具体的医疗环境中，人们对于一位素不相识的医生的信任，来自两个方面，一方面是人们对医生群体的信任，另一方面则是基于该医生的具体表现而产生的信任。前者的信任程度受医生群体公信力的影响。如果整个医生群体的公信力很高，那么患者基于此对于该医生也就更信任；反之，患者则会更加谨慎。20 世纪以来的众多违反伦理的医疗事件的出现，如纳粹医生活体实验、塔斯基吉梅毒实验等，在一定程度上削弱了医生共同体的公信力。这也部分解释了为什么在现代医患关系中，患者对于医生的信任更加谨慎。后者的信任程度受医生个人表现的影响。问题在于，作为一名医生，他的哪些表现更能够获得患者的信任？针对这一问题，笔者将在下文详细阐述。这里只是说明了狭义的角色信任的影响因素。

机构信任是指个人对于机构或者公共服务部门的信任。此类信任同样受到机构公信力的影响。但机构公信力的表现不同于上文中提到的某个群体的公信力的表现。机构公信力主要体现在制度建设上。换句话说，要想让个人信任某个机构，要从制度建设上保障个人权利，使得人们相信，在于该机构打交道的过程中，自己的利益可以得到有效的保障。那么，具体到医疗机构，如医院，赢得患者信任的制度建设又是如何体现的呢？这一问题，笔者也将在下文中加以论述。

3 “医患信任”及其与医疗 AI 可信度的关系

上文中，笔者将信任这一概念的含义及其分类进行了简要的阐释。那么，具体到医患关系中，医患信任又有什么样的特点呢？要弄清这一问题，首先需要对医患关系的不同模式有所了解，因为不同的医患关系模式所形成的医患信任有所不同。

从发展历史来看，医患关系模式经历了从“医生主导型”向“医患共同决定型（Shared Decision Making）”的转变[10]。“医生主导型”的医患关系模式即常被批评的“家长主义”医患关系模式。在这种模式下，医生基于自己的专业知识和技能为患者做出“最好的”安排和治疗，无需过多征求患者的意愿。这种模式下的医患关系也因此被类比为“家长-婴儿（Parent-infant）”关系。在这种模式中，患者对于医生往往是一种绝对的信任，甚至是盲目的信任（Blind Faith）。按照上文中的划分，这种信任更多地表现为一种人际信任，即医生和患者直接地建立一种人际关系，并在相处的过程中产生对彼此的信任。实际上，这与古代行医模式有着密切关系。近代之前，无论中外，医生大多以自我雇佣（Self Employed）的形式出现。他们生活在固定的地方，有自己的诊所（很多时候医生的家就是诊所），前来看病的多是周围的邻居和乡亲。这些人对于这位医生的信任首先是基于对他这个人的信任。比如，某人的发小会些医术，去找他看病时对他的信任自然而然的是基于彼此所形成的关系的一种信任，也就是上文中的人际信任。关系密切的人际信任恰恰是盲目信任的基础，这好像，孩子应该充分信任父母为他们做的安排和规划，如果有质疑，反而是不合适的。同样的，去找发小看病，却问东问西，总是担心他开的药是不是对症，这反而伤了他的心，他会想，“我们的关系这么好，难道我会害你吗？”因此，在这种基于人际信任的医患信任中，“盲目

信任”其实并不盲目，也可以解释得通。而对于那些走街串巷的游医，患者的盲目信任则是基于医生共同体的公信力。这种公信力的形成来自两个方面，一方面是相比之患者，医生对于疾病的诊疗有绝对的权威；另一方面，医生共同体有自己的行业规范和职业伦理，如《希波克拉底誓言》，以保证大众相信其医疗行为是为患者利益着想的。换句话说，患者相信医生共同体能够遵守诸如《希波克拉底誓言》，加之医生在疾病治疗方面的专业权威，绝对信任也就顺理成章。

近代医患关系向着“医患共同决定型”转变。这种医患关系模式的转变与行医模式的改变有关。传统的自我雇佣式的行医模式逐渐被医疗机构所取代，大多数医生在医院工作且流动性增大。这样一来，传统“熟人模式”的医患关系变为患者去医院找“陌生的”医生就诊。因此，传统医患关系模式下双方的人际信任也就消失了。在转型后的医患关系中，尽管是医患共同决定，仍然需要患者对于医生一定程度的信任，如对于医生医术的信任，对于医生医德的信任等。转型后的医患信任更多地体现为患者对于医生的（狭义的）角色信任加上患者对于医院的机构信任。

如上所述，影响患者对（某位）医生的角色信任的因素包括，医生群体的公信力和具体这位医生的表现。首先，受到 20 世纪以来一系列不良事件的影响，加之现代媒体常常对于医疗事故进行放大报道，总的说来，医生群体的公信力有所下降。这种下降还表现在随着医学知识的普及以及信息技术的发展，医生在医疗诊断方面的权威也在受到挑战。例如一些患者经过自学医学知识，常会对医生的专业判断提出质疑。那么，具体到某个医生，他的哪些表现更能够获得患者信任呢？在这里，笔者尝试归纳为以下四个方面的表现。第一，医生的能力。医生的能力可以由其学历、职称、从业年资以及过往救治病人的表现所体现，这些都可以增加患者的信任。第二，医生和患者良好的沟通。这种沟通包括及时地告知患者病情、可选的治疗方案、治疗所涉及的风险以及其他患者希望了解到的情况等。第三，医生的热情和耐心。第四，没有利益冲突。利益冲突可以体现在医生和某种药品/器材有利益关联，或者医生的收入与患者医药费挂钩等。避免这些利益冲突，更能让患者相信，医生的诊断、治疗建议等是出于患者利益最大化的考量，也更能获得患者信任。

至于患者对于医院的信任，如上所述，机构信任靠机构公信力加以实现，主要体现在机构的制度建设上。具体到医疗机构，最重要的公信力保障之一便是对患者知情同意权的制度化保障。知情同意是“医患共同决定”模式最直接的体现。在现代医疗实践中，知情同意权被看作是对保护患者利益最有效最重要的措施[11]。所谓知情同意权，简单说来，是指患者对于自己的病情有充分知情权，对于医生提供的多种治疗方案所依据的理由和涉及的风险有充分的知情权，同时，在患者充分知情的基础上，需征得患者同意，某种治疗才可以被实施[12]。从这一定义中，我们能够看出，充分知情是患者知情同意的前提。问题在于，如何才算是（使患者）充分知情？一般而言，患者应该被告知的信息包括：（1）患者病情；（2）治疗方案及替代方案；（3）治疗方案的依据、涉及的风险、常见的或者可预见的副作用；（4）其他患者希望了解的与其病情及诊断相关的信息。医疗机构对患者知情同意权的制度化的保障是赢得患者信任的关键。正如哲学家欧若拉·奥尼尔（Onora O’Neill）所认为的，患者知情同意权的有效保障不仅是出于对患者自主性的保护，同时也是现代医患信任重塑所必须的[13]。也就是说，正因为知情同意权的确定，患者才相信，医生的诊疗是符合自己的最大利益的，才相信不会出现类似纳粹医生的行为或者塔斯基吉梅毒实验事件。

综上，现代医患信任不是之前医患关系模式下那种人际信任，而是患者对医生（狭义）角色信任及对医院机构信任的结合。影响这种信任的因素，有医生群体的公信力、某个医生的个人表现以及医院的公信力（主要靠对患者知情同意权的制度化保障来体现）。由此可见，在具体的医疗实践中，患者信任一名医生给出的诊断，并不仅仅因为其过往诊断的高精确性，而是诸多因素的综合。这其中，医生对于患者病情的充分告知是不可缺少的要素，一方面，这是医生获得患者（狭义）角色信任的重要组成部分；另一方面，也是医院对于患者知情同意权制度化保障的直接体现。

让我们回到文章开头中关于医疗人工智能黑箱问题的关切。那种认为“很多时候，人类医生是凭着直觉和经验在为病人做诊断，而这些直觉和经验连他们自己也很难解释清楚；如果我们能够信任医生给出的‘难以解释’的诊断，也应该信任人工智能系统给出的结论”的观点很难站得住脚。因为，人工智能由于黑箱问题，对于其做出的诊断结论无法给出任何可能的解释，让患者信任这样的诊断结论实质仍是一种盲目信任。然而，如上所述，现代医患关系模式下，患者对医生的盲目信任不复存在，让患者充分知情才是获得信任的前提和基础。

4 高精确性为医疗 AI 可信度带来的风险

医疗 AI 的优势之一是其具有的高精确性，医疗人工智能在疾病诊断方面表现出了高于人类医生的精准性，这成为其优于人类医生的重要理由。但同时这一优势也带来了可信度方面所存在的风险，即，诊断的高准确度或许意味着患者被误诊的风险降低，但由于人工智能系统的不透明性，一旦被误诊，对患者的伤害可能更大。换句话说，高准确性并不意味着高安全性。在这里，笔者尝试通过一个思想实验来论证，即使人类医生和人工智能都具有一定程度的不可解释性，人类医生给出的诊断和人工智能系统给出的结论却有可能有着本质的差异。

假设有四种疾病 A、B、C、D。其中，A 通过几天的休息就能痊愈；疾病 B 需要服用止痛片外加充足睡眠；C 需要使用抗生素；疾病 D 则必须截肢。假设当具有特殊症状的病人前来就诊的时候，某位人类医生首先排除 C 和 D，而后在 A 和 B 中做判断。其实，根据病人的症状，这位医生并不十分清楚为什么是 A 而不是 B。但是根据他多年行医的经验和直觉，大多数情况下，他会建议病人回去休息。另一边，假设人工智能系统面对相同症状的病人前来就诊时，首先排除的是 B 和 C，而后在 A 和 D 之间做选择，并在绝大多数情况下将此类症状诊断为疾病 A。然而，人工智能系统的诊断精确度高于人类医生，但并非百分之百精准。在这种情况下，很多人就医的时候可能仍然会选择人类医生而非人工智能。理由也很明显：虽然人工智能系统的诊断准确性更高，但一旦出现误诊（将疾病 A 误诊为疾病 D），其给患者造成的伤害更大。相比较而言，人类医生即使误诊，也最多是将疾病 A 误诊为疾病 B，而不会将其误诊为疾病 D。这里需要指出的是，笔者并不是认定所有的应用于医疗诊断的人工智能系统都存在类似风险。但是由于人工智能系统的不透明性，我们并不清楚其内部工作机制，因此，上述思想实验中的假设就有可能为真。正是这种可能性使得人们对于人工智能系统不透明性的担忧具有合理性和必要性。换句话说，尽管具有高精确性，人工智能系统的不透明性也会使其可信度大打折扣。

回到文章开头提到的“深度患者（Deep Patient）”系统。通过对总计 70 余万份电子病例的“深度学习”，该系统可以准确地预测出新病例中“隐藏”的疾病。然而，“黑箱”模式下的“深度患者”系统同样受到了可信度的质疑。这种质疑和担心并非杞人忧天。患者有理由而且应该对于不透明的高准确度保持谨慎的态度，毕竟，高精

确性并不意味着绝对的安全。而且，由于其不透明性，很难预料一旦出现误诊，会给患者带来怎样的后果。如上所述，“不透明性”恰恰是现代医患关系，即“医患共同决定型”的医患模式，所要尽力避免的。因此，现代医患关系中的信任是基于患者的充分知情，而对于“黑箱”模式下的医疗人工智能的信任却要求患者放弃知情而仅仅依据其准确性去盲目信任。那种认为“人类医生也是‘黑箱’，患者既然信任人类医生，也应该信任医疗人工智能”的观点正是没有意识到这两种信任的前提根本不同。对于医疗人工智能的黑箱问题，笔者更倾向于一种谨慎的态度，即医疗人工智能系统应该更加透明，即使这种透明化有可能降低其效率或者增加社会成本。因为，患者对于医疗人工智能不具备信任的基础，且任何技术的应用都应以保障患者的安全为前提。

5 结论

综上所述，对于“黑箱”的医疗人工智能的盲目信任难以得到辩护。一方面，这种盲目信任与现代医患关系模式中，对患者充分知情权的尊重和保障有巨大的价值冲突；另一方面，由于人工智能系统的不透明性，其诊断的高精确性并不等于高安全性，一旦出现误诊，对患者造成的伤害可能更大。因此，解决医疗人工智能黑箱问题的可行之路是使其更加透明化，让诊断结果更加的“可解释”，而不是要求患者基于其高精确性而盲目信任其诊断结论。这里，一种可能的透明化方案是，通过对人工智能系统的设置，使得当医疗人工智能的诊断结论与医生的诊断出现偏差的时候，或者当不同的医疗人工智能给出不同的诊断意见的时候，医生能够知道造成这种不一致的原因，以便向患者解释相关情况。

参考文献

[1] 潘若琳, 郑秋莹. 人工智能在健康产业应用中的伦理问题分析[J]. 中国医学伦理学, 2019, 12: 1541-1546.

[2] 杨庆峰. 从人工智能难题反思 AI 伦理原则[J]. 哲学分析, 2020, 2: 137-150.

[3] 徐 凤. 人工智能算法黑箱的法律规制——以智能投顾为例展开[J]. 东方法学, 2019, 6: 78-86.https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html. 2020.

[4] Alakwaa F M, Chaudhary K, Garmire L X. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data[J]. Journal of Proteome Research, 2017, 17(1): 337-347.

[5] Miotto R, Li L, Dudley J T M. Deep learning to predict patient future diseases from the electronic health records[C]. European Conference on Information Retrieval, 2016: 768-774.

[6] Pande V. Artificial intelligence's 'black box' is nothing to fear[EB/OL]. (2018-01-25)[2020-01-15]. https://scholar.harvard.edu/people_analytics/publications/artificial-intelligences-black-box-nothing-fear.

[7] Annette B. Trust and antitrust[J]. Ethics, 1986, 96(2): 231-260.

[8] Russell H. Trust[M]. Oxford: Polity Press, 2006.

[9] Rothstein J A. Reconsidering trust in the physician-patient relationship[D]. Yale: Yale University, 1996.

[10] Kaba R, Sooriakumaran P. The evolution of the doctor-patient relationship[J]. International Journal of Surgery, 2007, 5(1): 57-65.

[11] Del Carmen M G, Joffe S. Informed consent for medical treatment and research: A review[J]. The Oncologist, 2005, 10(8): 636-641.

[12] Faden R R, Beauchamp T L. A history and theory of informed consent [M]. Oxford: Oxford University Press, 1986.

[13] O'neill O. Autonomy and trust in bioethics[M]. Cambridge: Cambridge University Press, 2002.

On Credibility of Medical Artificial Intelligence Technologies based on Physician-Patient Trust

Xu Hanhui

(*School of Medicine*, *Nankai University*, *Tianjin* 300071, *China*)

Abstract: Since the 21st century, artificial intelligence (AI) technology has been widely used in the medical field, and has shown more accuracy and efficiency than human doctors for diagnosis. However, as the internal mechanisms are often unclear, an AI system is usually regarded as a “black box.” This “black box” problem has raised concerns regarding the credibility of AI. However, one view claims that the concern regarding the “black box” is unnecessary, as human doctors do not seem not much better than AI in terms of providing diagnoses based on intuition and experience. In many cases, the human intuitions and/or experiences are also difficult to explain. Thus, if we can trust human doctors, it seems reasonable to trust medical AI. In this study, I attempt to respond this view, based on a theory of physician-patient trust.

Key Words: medical AI; Black Box; physician-patient relationships; trust