·人工智能的伦理学研究·

人工智能伦理建设的目标、任务与路径:六个议题及其依据*

陈　小　平

[摘　要]　人工智能伦理建设的必要性已形成全球共识，但建设目标、重点任务和实现路径仍存在较大分歧，概括为六个议题。本文首先介绍 AI 的两大类主要技术——强力法和训练法，在此基础上总结 AI 现有技术的三个特性，作为 AI 伦理的技术依据。同时，以全球公认的福祉原则作为 AI 伦理的根本依据。本文立足于这两个依据，阐述 AI 伦理建设应具有双重目标——同时回答应该和不应该让 AI 做什么，进而探讨另外五个重要议题: AI 的安全底线，AI 功能的评价原则，AI 治理责任的落实路径，AI 主体状况变迁的可能性，以及一种全新的创新模式——公义创新。

[关键词]　人工智能　伦理　评价　治理　公义创新　　[中图分类号]　N01/TP18

经过几年的广泛讨论，人工智能（Artificial Intelligence，简称 AI）伦理建设的必要性已形成全球共识。但是，关于 AI 伦理的建设目标、重点任务和落地路径，仍存在较大的分歧和争论，也有些关键问题尚未引起足够的重视，文本将这些内容概括为六个议题。显然，建设目标的定位将决定重点任务和落地路径的选择，从而决定 AI 伦理建设的发展大局。关于 AI 伦理建设目标的主要分歧是: AI 伦理应该是双重目标（即同时回答应该和不应该让 AI 做什么），还是单一目标（即主要回答不应该让 AI 做什么）？如果是单一目标，一些重大议题将被完全或部分地排除。引起分歧和争议的一个重要原因在于，对现阶段 AI 技术特性的认识存在巨大差异，从而导致对 AI 社会意义和伦理风险的截然不同甚至完全相反的判断。为此，有必要梳理七十年来 AI 研究的主要进展，澄清现阶段 AI 技术的主要特性，形成 AI 伦理的技术依据。同时，以全球公认的福祉原则作为 AI 伦理的根本依据。本文根据这两个依据讨论 AI 伦理的六个议题。

一、人工智能的强力法

AI 经过三次浪潮取得了大量进展，各种技术路线层出不穷，受到研究者较多关注的有两大类技术——强力法和训练法。强力法又包含推理法和搜索法两种主要类型，推理法是在知识库上进行推

* 本文根据作者在“第二届全球视野下的人工智能伦理论坛”（杭州，2020 年 7 月 25 日）上的演讲整理而成。作者在与赵汀阳、王蓉蓉关于 AI 伦理问题的讨论中受益良多。本文部分素材来自《人工智能伦理导引》（陈小平主编，中国科学技术大学出版社 2020 年），刘贵全、顾心怡、叶斌、汪琛、王娟、侯东德、苏成慧参与了该书编著。谨向以上诸位表示感谢。

理，搜索法是在状态空间中进行搜索。推理法通常由一个推理机和一个知识库组成，推理机是一个负责推理的计算机程序，往往由专业团队长期研发而成，而知识库则需要研发者针对不同应用自行开发。

一般来说，推理机的工作方式是：针对输入的提问，根据知识库里的知识进行推理，给出问题的回答。下面用一个简化的例子加以说明。假设我们要用推理法回答“就餐”这个应用场景的有关问题。为此需要编写一个关于“就餐”的知识库，其中部分知识如表 1 所示。表 1 中的第一条知识 $\forall x \forall y(dish(x) \rightarrow food(y) \rightarrow hold(x, y))$ 是一个逻辑公式，它的含义是：餐具可以盛食物；表中的第二条知识 $food(rice)$ 也是一个逻辑公式，它的含义是：米饭是食物；表中的其他知识类似。

表 1　一个知识库的例子

就餐知识的逻辑表达	含义
$\forall x \forall y(dish(x) \rightarrow food(y) \rightarrow hold(x, y))$	餐具可以盛食物
$food(rice)$	米饭是食物
$food(soup)$	汤是食物
$dish(bowl)$	碗是餐具

表 2　一些问答的例子

问题	问题的含义	回答
$hold(bowl, rice)$？	碗能盛米饭？	*yes*
$hold(bowl, soup)$？	碗能盛汤？	*yes*
$hold(bowl, x)$？	碗能盛什么？	*rice*，*soup*，……
……	……	……

表 2 列举了一些问题，比如第一个问题“$hold(bowl, rice)$？”问的是：碗能盛米饭吗？推理机利用知识库中的知识进行推理，可以给出此问题的回答 *yes*。表 2 中的第三个问题稍微复杂一点，它问的是：碗能盛什么？回答一般不是唯一的，但推理机仍然能够根据知识库中的知识，找出所有正确的答案：碗能盛米饭、能盛汤……。推理机还可以回答更复杂的问题。

值得注意的是，一般情况下，由推理机得到的回答，并不是知识库中存贮的知识。例如表 2 中的三个回答都是推导出来的，在知识库（表 1）中并没有直接保存“碗能盛米饭”“碗能盛汤”等答案。因此，知识库推理与数据库查询不同，不是提取事先保存的答案，而是推出知识库中没有保存的答案，可见知识库加推理机的组合能力之强大。知识库上的推理被认为是一种智能功能，是其他信息技术所不具备的。

目前强力法受到一个条件的限制——封闭性。[①] 封闭性在推理法上的具体表现是：要求存在一组固定、有限的知识，可以完全描述给定的应用场景。对于上面的“就餐”场景，如果存在着不可以盛汤的“破碗”（并且将“破碗”也当作“碗”），那么表 1 中的知识就不能完全描述这样的“就餐”

① 关于封闭性具体内涵的详细描述，通俗性介绍参见陈小平，2020 年 a；专业性介绍参见陈小平，2020 年 b。

场景，因为根据这些知识推出的某些回答（如“碗能盛汤”）在这个场景中是不正确的。

上述“就餐”场景是特意设计的一个小例子，而实际应用中的场景都很大、很复杂（否则就不必应用 AI 技术了），有时不满足封闭性条件。比如一个就餐场景中，一开始没有破碗，根据知识库推出的回答都是正确的；可是一段时间之后出现了破碗，根据知识库推出的某些回答就不正确了。这种情况也是不满足封闭性条件的。

关于推理法对于整个 AI 的重大意义，深度学习的三位领军学者 Geoffrey Hinton、Yann LeCun 和 Yoshua Bengio（他们共同获得 2018 年度图灵奖）在深度学习的总结性论文中指出：深度学习的根本性局限在于缺乏复杂推理能力。（cf. LeCun et al）而推理法代表着人类关于复杂推理能力的最高研究成果，所以推理法的局限性也代表着整个 AI 现有技术的局限性，封闭性对推理法的限制也是对整个 AI 现有技术的限制。

二、人工智能的训练法

训练法要求首先收集一组原始数据，并对其中的每一条数据都进行人工标准，做成训练数据集。然后用训练数据集训练一个人工神经元网络，用训练好的网络回答问题。

图 1 是一个人工神经网络的示意图。图中每一个圆圈代表一个“神经元”，每一个带箭头的线段代表神经元之间的一个“连接”。人工神经元网络就是由大量神经元和连接组成的网络。一个连接可理解为一条信息通道，并对通道中传递的信息进行加权运算；也就是说，一条连接首先从一个神经元接受输入数值，经过加权运算，再按照箭头的指向，向下一个神经元输出加权计算的结果。图 1 省略了所有连接上的权值。

如图 1 所示，一个神经元可以有多个输入连接，从而同时接受多个输入值。一个神经元也可以有多个输出连接，从而同时向多个神经元传递输出值。每个神经元能够独立地计算一个简单函数 f，即根据该神经元的所有输入值，计算得出函数 f 的值之后，作为输出值向所有输出通道同时发送，经过各条连接上的加权运算之后，传递给其他神经元。在图 1 中，x_0，x_1，……，x_n 是整个人工神经元网络的输入连接，具体输入值来自网络外部；y_0，y_1，……，y_m 是整个人工神经元网络的输出，具体的输出值就是网络的计算结果。

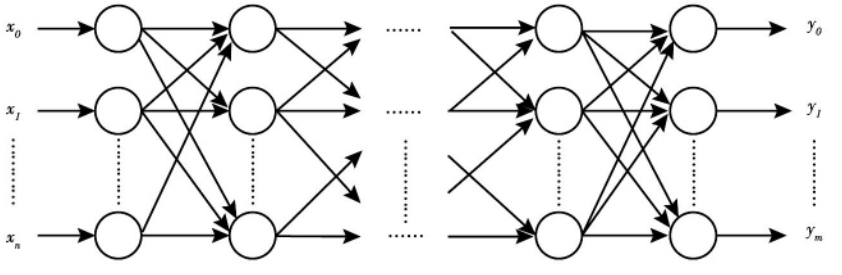

图 1　一个人工神经网络示意图

图 1 只画出了四列神经元，其他列被省略了。每一列神经元称为一个“网络层”。如果一个人工神经网络具有很多层，比如几十层、几百层甚至更多层，就称为“深层网络”，深层网络上的机器学习称为“深度学习”。

下面以著名的 ImageNet 图像分类比赛中的一个任务为例，说明训练法的工作过程。在比赛之前，组织者收集了一个大型图片库，包含 1400 多万张图片，并将其中一部分图片做了人工标注，这些带

人工标注的图片作为训练数据集，参赛队可以用这些图片训练他们的神经网络。图片库中没有标注的图片作为测试集。在比赛中，要求每一个参赛的图像分类软件，针对测试集中的大量图片，自动识别这些图片中动物或物品的种类，按识别正确率的高低决定比赛名次。

这个测试集中的图片被人工分为 1000 类，其中每一个类用 0 至 999 中的一个数字进行标注。一个类包含几十张到一百多张图片，这些图片中的动物或物品的种类相同，所以这些图片被标注为相同的数字。这 1000 个类包括 7 种鱼，第一种鱼的所有图片标注为 0，第二种鱼的所有图片标注为 1，……，第七种鱼的所有图片标注为 6；还包括公鸡和母鸡，公鸡和母鸡的图片分别标注为 7 和 8；还有 26 种鸟的图片分别标注为 9 至 34 等等；一直到最后一类——卫生纸图片，标注为 999。原始图片和人工标注的对照见表 3。采集好的原始图片经过人工标注，训练集就制作完毕，可以用于人工神经元网络的训练了。

表 3　原始图片与人工标注对照

原始图片	人工标注
7 种鱼的图片	0 – 6
公鸡、母鸡的图片	7、8
26 种鸟的图片	9 – 34
……	……
卫生纸的图片	999

如果训练之后，一个人工神经元网络的正确识别率达到了预定的要求（比如 95% 以上），就认为训练成功，可以应用了。正确识别指的是：对输入的任何一张图片，能够指认输出图片中动物或物品所对应的数字。比如输入公鸡的图片，人工神经元网络输出数字 7；输入卫生纸的图片，则输出数字 999。从实际效果来看，如果一个人工神经元网络达到了上述要求，就可以认为，该神经网络“学会”了识别图片中的 1000 类动物或物品。

训练法也受封闭性的限制，具体表现为：可以用一组固定、有限、带人工标注的代表性数据，完全描述给定的应用场景。（参见陈小平，2020 年 a，2020 年 b）所谓“代表性数据”，指的是能够代表所有其他数据的数据。例如，上面的图像分类比赛例子中，如果只用训练集中的图片训练神经网络，就可以训练出合格的网络，那么这个训练集就具有代表性，代表了图片库中所有 1400 多万张图片。反之，假如一个训练集不具有代表性，用它训练出的神经网络就不合格，比如正确识别率到不了预定的要求，不能实用。

三、人工智能现有技术的三个特性

普通算法通常直接计算一个函数。例如，图 2 中的算法计算一个自然数 x 是偶数还是奇数，算法规定了每一步计算过程，根据相关背景知识可以得知每一步计算的含义和作用是什么，进而判断这个算法是否正确。

通过“AI 算法”与普通算法的对比发现，它们是非常不同的。具体地说，强力法中的推理法是用知识和推理解答问题，要求针对一个应用场景编写相关的知识库，然后用推理机回答问题，而不是像普通算法那样直接计算结果。训练法则要求首先采集、制作训练数据集，训练出一个合格的神经网络，然后用该网络回答问题，而网络内部的运行一般是无法解释的（至少目前如此）。

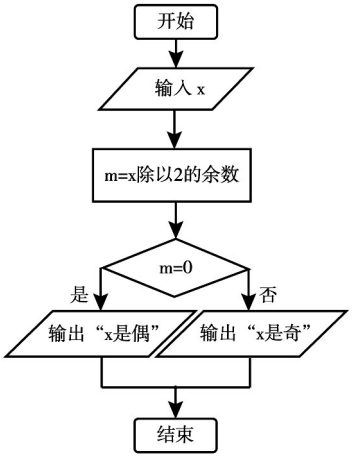

图 2　计算自然数奇偶性的普通算法

由此可见，AI 算法不仅更复杂，更重要的是原理不同，难以直接根据一个 AI 算法判断它能做什么、不能做什么、怎么做的、做得是否正确等等。为此，本文给出 AI 现有技术的三个特性，从而为分析 AI 伦理的六个议题提供技术依据。

AI 现有技术的第一个特性是封闭性（具体含义如上所述）。一个应用场景如果具有封闭性，则应用 AI 的强力法或训练法技术，可以保证应用成功；如果不具有封闭性，则不保证应用成功（但也不一定失败）。由于大量应用场景是封闭的，或者可以被封闭化，即改造为封闭的（参见陈小平，2020 年 a），所以封闭性条件对于大量实际应用成立，也为这些应用的研发提供了一个不可忽略的关键指标。

AI 现有技术的第二个特性是被动性。这些技术不具备主动应用的能力，只能被动地被人应用。有人认为，AI 可以自我学习，从而学会它原来不会做的事情。事实上，这样的技术确实在研究之中，但目前尚未成熟，无法投入实用，而且强行投入实用会带来极大风险。还有人认为，围棋 AI 程序“阿法狗”可以自学下围棋，而且通过自学战胜了人类。其实，围棋是一个封闭性问题，“阿法狗”技术只对封闭性场景有效（参见陈小平，2020 年 b），而且“阿法狗”的所谓“自学”完全是它的设计者事先安排好的，与通常人的自学不是一回事。

AI 现有技术的第三个特性是价值中性，也就是说，这些技术本身无所谓善恶，人对它们的应用方式决定其善恶。以推理法为例，推理机给出的回答会不会对人有害，完全取决于知识库是否包含可能隐含不良后果的知识。由于知识库是人编写的，所以是设计者决定了推理法的具体应用的善恶。也有研究者试图让 AI 自动寻找自己所需的知识，即具有自动获取知识的能力（例如 Chen et al，2012），但这些技术目前仍处于基础研究和实验测试阶段。

四、人工智能伦理的六个议题

议题 1：AI 伦理的建设目标——双重还是单一？

根据对伦理学的常识理解，伦理是人的行为准则，以及人与人之间和人对社会的义务。（参见《辞海》缩印本，第 221 页）因此，AI 伦理要回答两方面的问题：应该让 AI 做什么，不应该让 AI 做什么。同时回答两个问题是双重目标；只回答“不应该做什么”问题是单一目标。

鉴于世界各国都将“福祉”作为 AI 伦理的基本原则甚至第一原则，我们将福祉原则作为 AI 伦

理体系的指导性原则。显然，福祉的实现主要源于努力而非限制。由于 AI 具有被动性，AI 的发展必须经过人的努力，所以 AI 伦理应该引导和规范这种努力，这就是双重目标的根本依据。

在双重目标下，AI 伦理体系的基础架构（参见陈小平，2019 年）[①] 如图 3 所示。在此架构中，AI 伦理有三层结构：伦理使命（福祉）、伦理准则（如安全性、公平性等）和实施细则（详见议题 3）。其中，针对不同的应用场景，需要设立不同的实施细则，于是 AI 伦理与社会及经济发展相互紧密关联，不再是空中楼阁。在这个架构中，传统创新需要受到伦理准则的约束（这种约束过去没有充分建立起来），从而促使传统创新更好地服务于社会需求和重大社会问题的解决。

由于传统创新并不十分适合社会重大问题的解决，所以我们提出了一种新的创新模式——公义创新（详见议题 6）。公义创新和传统创新都要接受福祉原则的指导，这是不变的。同时，根据公义创新的成果可以改变现有伦理准则的内涵，也可以增加或减少伦理准则，以反映社会发展对 AI 伦理的反作用。在两种创新的促动下，社会需求和社会重大问题不断得到解决，推动社会进步，形成新的社会需求和重大社会问题，从而实现社会及经济的螺旋式发展。

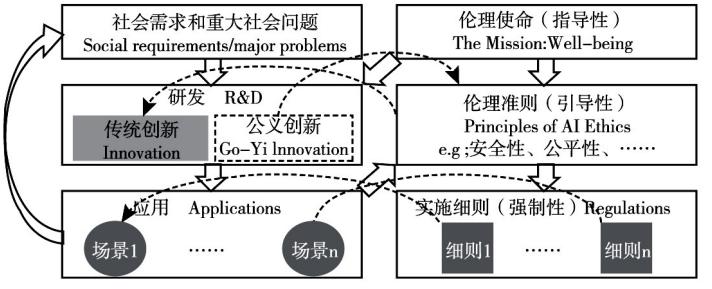

图 3　人工智能伦理体系架构

议题 2：AI 的安全底线——技术失控与技术误用？

在技术范围内，AI 的伦理风险主要有两类：技术失控和技术误用/滥用。技术失控指的是人类无法控制 AI 技术，反而被 AI 所控制，成为奴隶或宠物。技术误用/滥用指的是 AI 技术的非正当使用，由此带来对用户和社会的损害，但达不到失控的严重程度。技术误用/滥用是目前存在的现实伦理问题，亟需加强治理；而技术失控是人们的最大担忧，相关影视作品的流行大大增强了这种担忧。

对于 AI 技术失控的可能性而言，上文总结的 AI 三个特性具有关键性影响。人类对封闭性或封闭化场景具有根本性乃至完全的掌控力，因此这些场景中的应用不会出现技术失控。根据被动性，AI 技术应用都是由人类实施的，只要人类对不成熟、不安全的 AI 技术不实施应用，这些技术都无法进入应用空间，也就不会引起风险。根据价值中性，只要人类对 AI 技术的应用符合伦理准则，这些应用就不会对人类造成不可接受的损害。

因此，在 AI 三个特性成立，并且 AI 应用遵守伦理准则的情况下，不会出现技术失控，也不会对人类造成不可接受的损害。可是，在这三个特性不全成立，或者 AI 应用不遵守伦理准则的情况下，就可能出现伦理风险。例如，假如未来出现了可以在非封闭性场景中自主进化的 AI 技术，就无法排除各种伦理风险，甚至包括技术失控的可能性。（参见赵汀阳）再如，如果在 AI 技术应用中不遵守相关伦理准则，就会出现技术误用/滥用；数据安全问题、隐私问题、公平性问题等等，都属于这种

① 原文引入了“伦理创新”的术语，后经王蓉蓉建议，改为“公义创新”，但内涵保持不变。

情况，而且已经在一定范围内发生，亟需加强治理。这表明，针对技术误用/滥用的治理已经成为当务之急，而完整 AI 伦理体系的建设也必须提上议事日程。

议题 3：AI 功能的评价原则——“超越人”与“人接受”？

对 AI 技术的功能水平的传统评价原则是“超越人”，有时具体表现为“战胜人”，如阿法狗。不过在 AI 界，这个原则理解为 AI 与人的同类能力水平的对比，看谁的水平更高，而不是要在现实世界中用 AI 战胜人（虽然经常发生这种误解）。AI 研究界和产业界往往认为，当 AI 的某项能力超过了人，那么就可以在产业中实现该能力的产品化；如果尚未超过，则表示 AI 的该项能力还不够强，难以实用化。

不过，在上述传统评价原则之外，实际上还存在着另一种评价原则，这就是“人接受、人喜爱”。在一些应用场景中，AI 通过人－机器人交互提供服务，而且人－机器人交互以人机情感互动为基础，例如面向空巢群体的情感机器人、用于自闭症等人群心理干预的机器人、用于少儿娱乐教育的机器人等。在这些应用中，用户对机器人的接受度是第一重要的，否则产品的其他功能再好也难以被用户接受。

在接受原则下，相关 AI 产品的主要评价指标不是在某个方面比人强，而是人对 AI 的接受性和接受度是否满足用户的期望。例如，中国科学技术大学研发的情感交互机器人“佳佳”，其智能水平只是她的“姐姐”——“可佳”机器人（cf. Chen et al，2010，2012）的几分之一，但由于“佳佳”可以识别人（如用户）通过表情和话语呈现出的情绪，并通过机器人的表情和话语进行即时反馈，在一定程度上实现了机器人与人的情感互动，因而具有更高的用户接受度，在人机情感交互方面的性能远远超过“可佳”。

两种 AI 功能评价原则决定了人类对 AI 的两种观察角度和评判标准，所以它们绝不是单纯的技术问题，同时也决定了 AI 伦理对 AI 技术的观察角度和评判依据。因此，AI 伦理应该同时从这两个角度展开自己的研究和实践。目前对第一个角度的研究较多，而第二个角度的研究基本处于空白状态，亟待加强。

议题 4：AI 治理责任的落实——规范性与自主性？

目前法学界倾向于认为，AI 尚不具备法律主体地位。（参见刘洪华，2019 年）因此，与 AI 相关的法律责任的主体是人，比如产品的研发、运维机构。因此，与 AI 技术相关的主体责任和治理责任的落实，就成为 AI 伦理的一个重要议题。

我们认为，由于 AI 现有技术的三个特性，法学界的上述判断是符合现阶段实际情况的，AI 确实不应该、也不可能承担主体责任。另一方面，只要伦理规范足够具体化，以至于成为封闭性条件的一部分，那么在这种场景中，就可以利用 AI 现有技术，自主地执行这些规范，从而完成部分 AI 治理任务。对于非封闭的应用场景，或者伦理规范不能成为封闭性条件的一部分的情况下，则不能完全依靠 AI 技术的自主性，必须坚持人的管理和介入。总体上，人作为责任主体，绝不能放弃自己的职责。

根据以上分析可知，在伦理规范和管理体制下，让 AI 技术自主或半自主地实现其功能，是一种有效的责任落实方式。例如，利用 AI 技术，可以对消息的真伪性进行核查和推测，对通过核查的真实消息向目标用户进行分发推送，对敏感操作流程的合规性进行审核，等等。不过，由于这些应用的场景往往不是完全封闭的，所以仍然需要人工管理，但 AI 技术的应用能够大大减轻人工负担，显著提高工作效率，整体上明显改善管理水平。

产业部门的现行管理体制为主体责任的落实提供了一条可行路径，尤其其中的技术标准可以作为 AI 伦理准则的一种实施细则（见图 3）。对于 AI 相关产品，需要与其他工业品一样，设立四个层级的技术标准：国际标准、国家标准、行业标准和企业标准，其中企业标准和行业/团体/地方标准不得

与国家标准相抵触，而国家标准与国际标准之间，可以通过国际标准化合作达成协调一致。所有这些层级的技术标准都应符合 AI 伦理规范的要求。这样，伦理规范就通过技术标准及相关管理机制得到落实，不再是纸上谈兵的空中楼阁。

议题 5：AI 主体状况变迁的可能性——物、人还是“非人非物”？

上文已说明，目前 AI 在法律上是物，不是人。但是，由于大量应用需求的推动，以及“接受”评价原则的采纳及相关研究的深入和成果推广，AI 技术的发展已形成了一种新的可能性：在不远的将来，某些 AI 产品或技术载体如情感交互机器人，会被部分大众接受为“非人非物、亦人亦物”的第三种存在物。

在 AI 发展早期，曾出现少数用户将 AI 误认为人的情况，比如上世纪 60 年代有人将一个 AI 对话系统误认为人。不过，这是在人与物的二分法体系之中出现的混淆，没有突破二分法的边界。而现在出现的情况是，人在与某些机器人的交互中，一方面从理智上明确认识到和自己交互的机器人不是人，同时却在情感中不将机器人视为物，而更倾向于视为某种有情感能力的新型存在物。这种情况实际上比之前的要更复杂。

出现这种现象的原因在于：与科学和哲学中的默认假设不同，人们通常并不关心机器人表现出的情绪是不是真实的人的情绪，更不去仔细区分人的情绪和机器人的情绪有什么本质区别。（参见胡珉琦）

这种现象带来三方面的可能性。第一，有助于 AI 在某些领域的应用推广，满足用户的大量真实需求（尤其是情感交互方面的需求），从而带来 AI 研究和应用的新机遇；第二，为调整、拓展和改善人机关系开辟了新的探索空间；第三，带来一种新的伦理挑战——对自古以来从未受到怀疑的人－物二分法的挑战。虽然科学上可能不承认这种存在物的真实性，哲学上也不承认它的必要性，但如果越来越多的大众在认知和心理上接受这种存在物，就会形成一种普遍和重要的社会现象，甚至可能对人机关系和人际关系产生广泛的、震撼性的冲击和深远的影响。因此，忽视这些可能性将会造成 AI 伦理大厦的巨大缺口。AI 伦理的双重目标要求对正、反两方面的可能性展开积极探索。

议题 6：AI 时代的创新模式——传统创新与公义创新？

在图 3 所示的 AI 伦理体系架构中，一个核心部分是公义创新。与传统创新（参见黄阳华）相比，公义创新的主要内涵及特点如下。

第一，传统创新主要追求经济效益的显著增长，而公义创新追求经济效益和社会效益的协同提升。传统创新带来经济效益的显著增长是有目共睹的。与此同时，诸多重大社会问题不断积累和深化，包括气候变化、环境污染、人口老化、收入不均、大规模流行病等等。甚至有人认为，正是传统创新加剧了这些问题的恶化。作为对传统创新模式的反思和超越，公义创新将以经济效益和社会效益的协同提升为基本目标，以重大社会问题的解决为重点任务，改变经济效益和社会效益相互脱节的现象。在现代社会中，公益事业与商业创新是相互分离的，科技成果相对易于进入商业创新，不易进入公益事业，公益事业与商业创新的这种分立式组合，明显不利于重大社会问题的解决。

第二，传统创新的目标对象是满足用户需求的具体产品/服务，而公义创新的目标对象是符合社会发展需要的人工/人造系统。[①] 满用户需求、且具有显著经济利益的产品/服务这个目标对象贯穿于传统创新的全流程，是该流程一切环节的终极考核指标，因而难以避免各种损害社会效益的副作用。

① “人工”的例子如“人工降雨”，其结果（降下来的雨）是“真的”（自然的），而导致这个结果的过程是人为的（非自然的；“人造”的例子如“人造卫星”，其结果（卫星）及其过程都不是“真的”。AI 中的 Artificial 包含人工和人造两种类型，公义创新的目标对象也包括人工系统和人造系统。

因此，公益创新将不再以产品/服务本身作为目标对象，而是上升到人工/人造系统（参见司马贺，第30页）层面，并且全面重构人工/人造系统的设计－实施体系，将其改造为实现经济效益和社会效益综合提升的手段。

例如，很多高新技术的应用在提高经济效益的同时，也带来人工岗位的大量减少，[①] 并可能导致新的收入分化，这种情况在传统创新中比较普遍。而在公义创新的设计考虑中，一个人工/人造系统包含的要素有：产品/服务、制造方式、员工利益、用户利益、……。于是，设计方案的评价指标不仅反映经济效益，同时也反映社会效益。显然，这种人工/人造系统的设计和实施难度远远高于传统的产品设计和制造。为此，不仅需要将AI技术继续应用于产品设计环节（类似于传统创新），更需要将“规划”（参见李德毅，第216页）、机制设计、目标优化等AI技术应用于整个人工/人造系统的设计，从而使AI技术发挥更大的作用，帮助人类发现或创造社会经济发展的更多新机遇，如新的就业岗位、新的人机合作方式、新的生产－生活协同方式以及解决重大社会问题的新途径。

第三，传统创新延续、强化工业文明传统，而公义创新探索更具包容性的文明路径。除上面提到的问题之外，传统创新通过延续、强化工业文明传统，进一步加剧了人的异化、人机对立等长期存在的难题，甚至可能产生“无用阶层”（参见巩永丹）等文明层面的重大挑战。尤为重要的是，这些挑战性问题在工业文明传统下是无解的，因此有必要探索新的化解路径。公义创新的思想来源包括三个方面：历史观——道家哲学（特别是老子的“道”），文化观——儒家哲学（特别是孔子的“义”），社会观——希腊哲学，如梭伦的“正义”理论。（参见廖申白）这些不同文化传统的融合、发展将构成公义创新的理论基础，并在其上构建公义创新的方法论体系，最终形成可运行的公义创新模式。在这种新模式下，对人的关注将得到根本性加强，对人和机器的认识将大幅度更新，人与机器的关系将得到重新定义，并在福祉原则的指导下，推动人、机器和环境的更具包容性的一体化发展。

显然，在现行市场规则下，公义创新面临很多困难，因此公义创新的实行要求改变市场规则和管理方式。其次，公义创新也要求设计思维、教育理念及实践的彻底变革，并带来人的观念的重大变革。事实上，公义创新的实施将为社会经济发展带来大量新机遇。

为了实现其基本使命——增进人类福祉，AI伦理要能够同时解答两方面的问题：应该让AI做什么，不应该让AI做什么，所以AI伦理具有双重目标。根据双重目标，结合AI现有技术的特性，本文认为短期内AI的主要风险是技术误用/滥用，这应成为近期AI伦理治理的重点课题。同时，本文分析了AI功能评价的两种原则——超过人和人接受，需要同时从这两个角度展开AI伦理治理。针对以上任务，本文发现，在现行产业管理及技术标准体系的基础上加以扩展，在适当条件下将AI技术引入到管理过程中，可以更加有效地实施AI伦理治理，从而形成落实AI治理责任的一条切实可行的路径。一个较长期的挑战是AI主体状况的变迁，即某些类型的AI被部分人接受为“非人非物、亦人亦物”的可能性，由此带来从技术到人机关系再到AI法制的一系列新课题。另一个更大的挑战是面向重大社会问题，以经济效益和社会效益的协调统一为基本追求的公义创新，它在人类福祉原则的指导下，广泛深入地利用AI技术，将传统的产品设计和制造升级为人工/人造系统的设计和实施，这也是双重目标下AI伦理体系建设的最大特色和最终标志。

（下转第107页）

① 对此需要具体情况具体分析，比如目前在国内工业界，机器人替代的劳动岗位主要是工作环境恶劣、不适合人从事的工种，如喷漆、打磨等。值得重点关注的是经济效益与社会效益不一致的情况。

参考文献

海德格尔，2018 年 《谢林：论人类自由的本质》，王丁、李阳译，商务印书馆。

加布里埃尔，2018 年 《不可预思之在与本有——晚期谢林与后期海德格尔的存在概念》，王丁译，载《哲学分析》第 1 期。

王丁，2020 年 《对自由的诠释作为自由自身的实行——海德格尔、谢林与一种“自由诠释学”的可能》，载《哲学动态》第 3 期。

先刚，2008 年 《永恒与时间——谢林哲学研究》，商务印书馆。

谢林，2016 年 《近代哲学史》，先刚译，北京大学出版社。

2019 年 a 《论人类自由的本质及相关对象》，先刚译，北京大学出版社。

2019 年 b 《启示哲学导论》，王丁译，北京大学出版社。

Schelling，F. W. J.，1856 – 1861，*Sämmtliche Werke*（SW），14 Bände，hrsg. von K. F. A. Schelling，Stuttgart：Cotta.

1696，*Initia Philosophiae Universae*，hrsg. von Horst Fuhrmans，Bonn：H. Bouvier u. CO.

1972，*Grundlegung der positiven Philosophie*，hrsg. von Horst Fuhrmans，Torino：Bottega D'Erasmo.

1989，*Einleitung in die Philosophie*，hrsg. von Walter E. Ehrhardt，Stuttgart：Fromman-Holzboog.

1990，*System der Weltalter*，hrsg. von Siegbert Peetz，Frankfurt am Main：Klostermann.

（作者单位：华中科技大学哲学系）

责任编辑：陈德中

（上接第 87 页）

参考文献

陈小平，2019 年 《人工智能伦理体系：基础架构与关键问题》，载《智能系统学报》第 4 期。

2020 年 a 《封闭性场景：人工智能的产业化路径》，载《文化纵横》第 1 期。

2020 年 b 《人工智能中的封闭性和强封闭性——现有成果的能力边界、应用条件和伦理风险》，载《智能系统学报》第 1 期。

《辞海》（缩印本），1979 年，上海辞书出版社。

巩永丹，2019 年 《人工智能催生“无用阶级”吗？——赫拉利“无用阶级”断想引发的哲学审度》，载《国外理论动态》第 6 期。

胡珉琦，2020 年 《AI 与情感》，载《中国科学报》7 月 23 日。

黄阳华，2016 年 《熊彼特的“创新”理论》，载《光明日报》9 月 20 日。

李德毅主编，2018 年 《人工智能导论》，中国科学技术出版社。

廖申白，2002 年 《西方正义概念：嬗变中的综合》，载《哲学研究》第 1 期。

刘洪华，2019 年 《论人工智能的法律地位》，载《政治与法律》第 1 期。

司马贺（Herbert Simon），1987 年 《人工科学》，商务印书馆。

赵汀阳，2018 年 《人工智能会是一个要命的问题吗？》，载《开放时代》第 6 期。

Chen et al，2010，“Developing High-level Cognitive Functions for Service Robots”，in *Proc. of 9th Int. Conf. on Autonomous Agents and Multi-agent Systems*（AAMAS 2010），Toronto，Canada.

2012，Xiaoping Chen，Jiongkun Xie，Jianmin Ji，and Zhiqiang Sui，“Toward Open Knowledge Enabling for Human-Robot Interaction”，in *Journal of Human-Robot Interaction* 1（2）.

LeCun et al，2015，“Deep learning”，in *Nature* 521.

（作者单位：中国科学技术大学计算机学院）

责任编辑：孟宪清　周　丹

On the Meaning and Further Implications of Wang Yangming's "Nothing Exists Beyond the Mind": Re-examining the Dialogue "Nanzhen guanhua"

Qiao Qingju

In Wang Yangming's narrative of "Nanzhen guanhua(appreciating flowers at Nanzhen)," *mingbai* (brightness) (i. e. the color of the flower becoming vivid) refers not only to the object's manner of presenting itself, but also to the subject's comprehensive cognition. As the former, *mingbai* is the vigor and vitality of the object, and is ontological; as the latter, it is how human beings connect with the world, forming a consciousness of the "community of all lives." The concept of *Ji*("silence" or "absence of color") meanwhile is not referring to the non-existence of being. Instead, it emphasizes the breakage of links or loss of connections between human beings and the world. Following Wang's thought, the notion that "nothing exists beyond the mind" can be further illustrated through three other principles: "existence should be perceived," "existence should be given attention," and "existence and human beings are integrated." Essentially, "Nanzhen guanhua" refers to the existence of human beings, i. e., the way to become a sage, rather than the existence of things.

The Target, Tasks, and Implementation of Artificial Intelligence Ethics: Six Issues and the Rationale Behind Them

Chen Xiaoping

Despite worldwide consensus about the necessity of developing Artificial Intelligence (AI) Ethics, there exist serious disagreements regarding its target, major tasks, and implementation, which are presented and clarified as six issues in this article. The rationale for resolving the disagreements presented here stems from the principle of well-being and the characteristics of existing AI technology, which are identified in this article from a summary of AI achievements so far. On the basis of this rationale, we argue that the target of AI Ethics is twofold: to answer what AI should do and what AI should not do. We then investigate five other issues: the AI safety bottom-line, evaluation principles for AI functions, the implementation of AI governance, the changing subjectivity of AI, and innovation in terms of public justice.

Being, History, and Freedom: The Basic Problem of Schelling's Late Philosophy

Wang Ding

After Kant, the all-encompassing task of German idealism was to justify reason itself as the substance and principle of the world, and to justify the unfolding of the world as the consequence of the activity of reason itself; that is, to construct a rational, immanent, and complete science. But the ultimate question begged by this proposed solution is, what motivates the fact that the world and reason exist rather than do not? That is, "why is there something rather than nothing?" This problem not only points to a higher understanding of existence and freedom from the perspective of meta-idealism, but also boosted Schelling beyond the rational internalism of German idealism, letting him put forward the distinction between "the logical – negative philosophy" and "the historic – positive philosophy", thereby opening up another path for German idealism.