# 可信智能与法则问题

## ——智能的决策性问题

刘炼

更新：
时间：*May 20, 2019*
学号：*1711361*

**摘　要**

人工智能的发展，很大程度上是基于计算力的提升和数据量的大量扩展。在多噪声的数据时代，构建可信的人工智能，摆脱人对于数据的封闭，做出更优的决策，存在极大的意义。但同时，在此过程中，要实现维护社会的法则以及伦理道德的自由，需要更多的“人机融合”。通过把握“人”“机”的整体观念，来重新探讨智能决策下的法则问题，能使得问题得到简化。

**关键词：** 可信智能；智能决策；自然法则；伦理道德；“人机交融”

近年来，人工智能已经越来越多地融入到人的生活当中，在提供便利的同时，这也带来了新的问题。智能化的工具无法确保百分之百的正确和可信，然而，却没有道德与法律上的约束限制其作用，此问题突出体现在利用人工智能进行决策这一方面。因此，如何构建可信智能，并且实现道法权上的管理，成为新的讨论的问题。

在第三届世界智能大会中，华为，联想，阿里巴巴等各大企业都在利用 5G 技术构建智慧城市，智慧全球进行了深入的技术研究，其中一个重要的问题在于，如何保证人工智能决策的可信以及如何实现问责。要实现人工智能新时代的发展，则无法逃避对此问题的探讨。

# 1　人工智能发展的双重困境

今天的人工智能的发展，在本文所提及的可信智能和自然法则问题中都陷入了困境之中，其使得人工智能的技术无法得以大规模的应用，阻碍了人工智能的发展。

## 1.1　可信智能

对于可信智能，还没有一个完全的定义。我们必须从可信算力，可信算法，可信算子三个方面去探讨可信智能的概念。对于无穷的计算问题与有限的计算力而言，实现可信算力成为了一个十分

巨大的挑战。从 [1] 中做出的一个预测来看，计算力仍无法满足计算问题本身的需求。同时，可信算力又是必要的，2019 年阿里云数据中心发生故障，影响了几乎整个华北地区的数据计算，造成上十亿的损失。与此同时，传统的人工智能算法是不可信的，人不仅仅是无法预测智能机器的行为，甚至无法深入解释算法中的理论，因此，对于人工智能所进行的决策行为，人完全处于一种不可见的黑箱之外。所谓可信算子，即可信的数据支持。对于我们采集的数据而言，永远是有噪声的，因此，无法保证数据本身的真实有效，使得可信算子也陷入了发展的困境中。早在 2012 年，Wang et al.[2] 就提出了关于云计算中的数据可信的重要意义，也提出了一种带有启发式的解决方案，利用动态调度实现数据的可信存储，然而，实际上，其仍然存在着多重问题。

### 1.2 自然法则

从很早开始，人类就开始了对于法与道德的探究。在《圣经》中，有“十诫”这样的戒律来规定人的道德；在中国战国时期，有法家主张以法治治理国家，实现国家意志上的道德；还有例如希腊神话中的宙斯，即掌管律法之神。

到了康德之后，其将“自然法”区分为“自然法则”与“自然法权”，本文主要讨论的是“自然法则”的问题，这是由于现在实际上还没有一个整体的法则规定，更无法实现从“法则”到“法权”的过渡；因此，首先必须要有“法则”，并使得其能维持当前的社会道德秩序。康德对于自然法则的规定如下：对于赋予责任的（verbindenden）法则而言，一种外在的立法是可能的，一般而言这些法则就叫做外在的法则（leges externae）。在这些法则中间，有一些法则，对它们的责任即便没有外在的立法也能被理性先天地认识，它们虽然是外在的法则，但却是自然的法则（natürliche Gesetze）……[3]

这样一种自然法则是理性的他律。它实际上涉及到责任问题，而这在产生决策的智能中变得困难，这与人工智能的技术状况息息相关。一方面而言，人工智能无法承担起相应的责任，正如其没有先天的理性认识以及自然的法权一样，其对于责任依然是不具有相应的承担意识与能力的。邓晓芒提出，这里，自然法则是“外在的法则”，它本身虽然不是外在立法的产物，但却是一切外在立法的“实证的法则”（positive Gesetze）之所以可能的前提，因为它是能够“被理性先天地认识的”。[4] 人工智能不具有这种先天理性的认识。在另一方面，正如在可信算法中所提到的，人（人工智能的创造者）无法解释人工智能的行为，缺少一种科学的可控性，使得如果让人（创造者本身）来承担责任的话，实际上是破坏了人的自由的法权。

### 1.3 困境的科学问题与划界问题

这一双重的困境可以从两个层面得到划分，即科学问题和划界问题。

对于科学问题而言，实际上，这一困境的基本来源就是人工智能只有实践上的意义，而在理论上仍然十分匮乏。当我们无法用数学理论解释人工智能的行为与现象的时候，我们必然处于一种无知的状态，而这样一种无知导致了无法赋予人工智能准确的定位，即无法给予它相应的法规与法权。

而在另一方面，当讨论这一责任的问题时，探讨的就是人的法则与法权问题，这是由于最初的

划界导致的。传统的划分方式使得研究本身无法广延到人以外的任何类之外，使得讨论本身不具有普遍性或者全局性，因而导致在人工智能的时代，没有办法将人工智能本身所该具有的责任考虑到。

# 2 决策问题的“人机交融”

在一个经典的问题——决策问题上，人机之间的交互实际上收到了很好的效果，这在另一方面激励了人们对于实现可信智能的人机交融的尝试，同时，也带来了新的启示。

## 2.1 人机交互实现决策的经典案例

### 2.1.1 国际象棋中的人机大战与人机合作

远在 Alpha go 之前，深蓝就在国际象棋这一棋类项目中战胜了当时的世界冠军卡斯帕罗夫。所以，在人类引以为自豪的很多方面，机器似乎都已经做得足够好，甚至好过人类的表现。但是，一个更有趣的比赛结果诞生在 2005 年，一个被称为自由式象棋竞标赛的比赛中。这个比赛有许多顶级国际象棋高手和超级电脑参赛，但是最终的冠军属于三个利用了普通个人电脑的业余棋手。合理地结合人机之间的差异与不同的能力，使得决策能力得到了极大的发展，而这也是真正的“人机交融”吸引人的原因。

### 2.1.2 人机交融实现的“艺术创作”

（本文暂且不去讨论其实现的结果是否为真正的艺术作品，只是对此实现的结果进行探讨。）Gatys et al.[5] 和 Johnson et al.[6] 所提出了基于深度学习的风格迁移技术，使得我们能利用不同的绘画作品的风格来修改照片本身的风格，这看似只是机器本身的工作，但实际上，一个好的风格迁移的结果，离不开人类对于原内容图片与风格图片之间的选择。

同样的，包括近年来令人惊奇的小冰，九歌的诗歌创作系统，其在实现深度学习之前需要大量的人类为其选择及标注的数据，因而，最终创作的结果，实际上是人机交互融合的结果。因此，实际的一个好的决策中，人与人工智能都不是单独出现的，而是以一种人机交融的方式出现。

## 2.2 基于“人机交融”的可信智能和法则问题

在前述的两重困境中，我们将人工智能本身视作独立的个体，这在现世仍然是缺少意义的，考虑决策中的问责问题，实际上是考虑人机交融过程中的法则问题。

因此，更多的考虑应该放在如何将人机交互的整体纳入到我们整个法则法权的责任承担机制中，而不是简单地进行人与人工智能的划分来讨论这一问题。

在这种转化之下，我们的问题变为，如何在现有的法则法权基础之上纳入人机交互的整体。这使得问责问题本身可以继续被探索，而不致陷入前述的双重困境中。由于智能机器本身是与人相互

关联的，这使得问责问题本身可以在现有基础上归约为人的法则与法权问题；并在此基础之上，在法权与法则的范围之中扩充人机的概念，将其纳入到这一讨论范围之中。

在另一方面，可信智能的困境解决也难以依靠机器本身，而仍然需要人自身的帮助。极为重要的是在算子方面的可信智能，这是实现人机决策的关键。一个正确的决策和错误的决策之间的差距，主要的根源在于算子的正确性与可信性。

## 3 小结

使用人工智能，实现人机交融的新的决策方式已经在许多决策问题中取得了卓有成效的结果；然而，现世的人工智能应用在多个方面对传统的社会结构造成了较大程度的影响。因此，在我们讨论人工智能应用的过程中，我们不得不考虑其出现错误（尤其针对于越来越依赖人工智能的决策问题所犯的错误）之后，如何对人工智能进行问责。不同专业的学者主要已经从两个方面——可信智能与新的人工智能下的法则，着手解决这一问题，但现在的讨论仍然存在许多困境。结合人工智能与人在决策问题过程中扮演的角色，这样的讨论更应立足于人与智能机器交融的整体来进行考虑；简单的二分将导致问题变得难以解决。

## 参考文献


[1]Malek M . Predictive Analytics: A Shortcut to Dependable Computing[C]// International Workshop on Software Engineering for Resilient Systems. Springer, Cham, 2017.

[2]Wang C , Wang Q , Ren K , et al. Toward Secure and Dependable Storage Services in Cloud Computing[J]. IEEE Transactions on Services Computing, 2012, 5(2):0-0.

[3] 李秋零主编：《康德著作全集》第 6 卷，〔北京〕中国人民大学出版社 2007 年，第 232 页

[4] 邓晓芒. 康德论道德与法的关系 [J]. 江苏社会科学, 2009(4).

[5]Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style[J]. arXiv preprint arXiv:1508.06576, 2015.

[6]Johnson J , Alahi A , Fei-Fei L . Perceptual Losses for Real-Time Style Transfer and Super-Resolution[J]. 2016.