学界 | 法律框架下的人工智能问责：解释的角色

2017-11-07 机器海岸线

选自 arXiv

作者：Finale Doshi-Velez, Mason Kortz 等

机器海岸线编译

参与：方建勇

Accountability of AI Under the Law:
The Role of Explanation

Finale Doshi-Velez*, Mason Kortz*,
for the Berkman Klein Center Working Group on Explanation and the Law:

论文链接：https://arxiv.org/pdf/1711.01134

摘要：使用人工智能或“人工智能”系统的普遍性，已经引起人们越来越多的关注该如何管理这些系统。如何规范 AI 系统的选择将需要谨慎。人工智能系统有可能合成大量的数据，从而提供比以前更高级别的个性化和精确性应用，从临床决策支持到自动驾驶和预测性警务。也就是说，常识性推理仍然是人工智能的圣杯之一，对 AI 系统的有意和无意的消极后果存在合理的担忧。

让 AI 系统负责有很多方法。在这项工作中，我们专注于一个：解释。关于人工智能系统的合法解释权的问题，最近在欧盟通用数据保护条例中进行了讨论，因此仔细考虑 AI 系统何时以及如何解释可以改善问责制是及时的。何时需要解释的良好选择，可以帮助防止 AI 系统带来的负面影响，而糟糕的选择可能不仅不能使 AI 系统承担责任，而且还妨碍急需的有益的 AI 系统的开发。

下面，我们简要回顾一下当前的社会、道德和法律规范的解释，然后关注法律目前需要解释的不同背景。当需要解释时，我们会发现存在很大的变化，但是也存在重要的一致性：当需

要从人类角度进行解释时，我们通常想知道的是某些输入因素如何以及是否影响最终决策或结果。

这些一致性使我们能够列出必须考虑的技术因素，如果我们希望 AI 系统能够提供法律规定的人类目前需要的各种解释。与人工智能系统的流行智慧相反，我们认为这种解释水平通常应该是技术上可行的，但有时也可能是繁复的解释，某些方面对人类来说是简单的，但对 AI 系统有挑战，反之亦然。作为一个由法律学者、计算机科学家和认知科学家组成的跨学科小组，我们建议目前人工智能系统可以并且应该保持与人类目前相似的解释标准;将来我们可能希望把 AI 作为一个不同的标准。

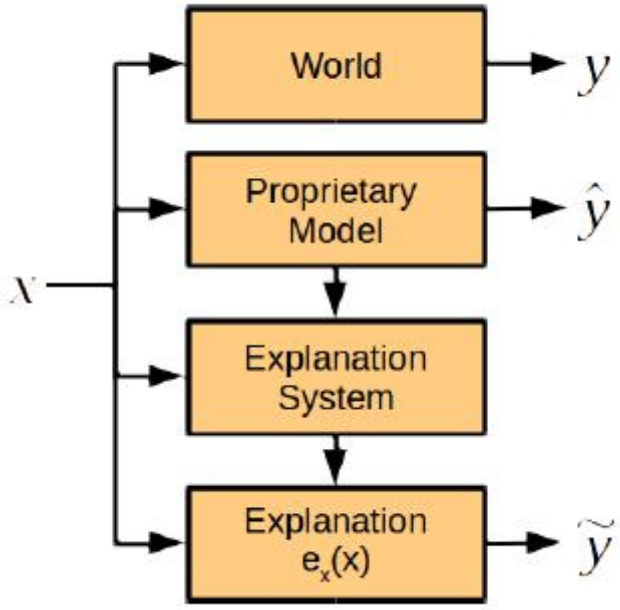

图 1：可解释的 AI 系统框架图。

	Human	AI
Strengths	Can provide explanation post-hoc	Reproducible, no social pressure
Weaknesses	May be inaccurate and irreliable, feel social pressure	Requires up-front engineering, explicit taxonomies and storage

表 1：人与人工智能的解释能力比较。

Approach	Well-suited Contexts	Poorly-suited Contexts
Theoretical Guarantees	Situations in which both the problem and the solution can be fully formalized (gold standard, for such cases)	Any situation that cannot be sufficiently formalized (most cases)
Statistical evidence	Problems in which outcomes can be completely formalized, and we take a strict liability view; problems where we can wait to see some negative outcomes happen so as to measure them	Situations where the objective cannot be fully formalized in advance
Explanation	Problems that are incompletely specified, where the objectives are not clear and inputs might be erroneous	Situations in which other forms of accountability are possible

表 2：对认定认可机构的方法的考虑。

✂----------