人工智能道德决策设计的Top-down方法

姓名：刘沛雨  学号：2100012289  本科生导师/助教：张史梁

摘要：自德国哲学家莱布尼茨于1674年设计了可进行道德规则运算的机器以来，哲学上有关道德机器的讨论已经持续了数百年之久。20世纪以来，随着计算机科学技术的蓬勃发展，道德机器所必需的强大算力已不再是阻碍道德机器实现的难题。在这样的背景下，人工智能（尤其是强人工智能）的道德决策设计方法走出了哲学家的思维，成为了现实生活中亟待解决的问题。本文笔者将以【美】W.瓦拉赫 与【美】C.艾伦所著的《道德机器》为主要依据，讨论有关智能体道德决策设计的Top-down方法与Bottom-up方法的具体可行性，并提出笔者初步设想的可行的决策设计方案。

1. 概述

近期我阅读了【美】W.瓦拉赫 与【美】C.艾伦所著的《道德机器》一书，书中主要讨论了人工智能道德体的道德决策设计问题，令我收获颇丰，对人工智能在现实中的应用有了新的思考与体会，以下是对本书中关于人工智能道德决策设计的Top-down方法的一些介绍与总结。

1. 正文
2. Top-down方法
3. Top-down方法简介

Top-down方法，即自顶向下为AMA构建道德决策规则的方法。这种方法的特点是“自上而下、理论驱动”，设计者为AMA建立一个具有普适性的量化的道德原则或规范体系，使得AMA能够在任何情况下通过在该体系中进行“道德计算”来判别自身的行为是否由规则所允许。Top-down方法的本质是设计一套可以转换为算法（机器指令）的规则，使得有道德的行动在AMA中转化为遵守规范的可计算问题，从而使得AMA可以通过道德规则运算算出在各种情况下的最佳行动。

在计算科学迅速发展的今天，算力已不再是阻碍Top-down方法的主要障碍，这为Top-down方法的具体实现提供了基础。但总的来看，由于道德规则本身的模糊性与不完备性，建立自上而下实现人工道德的进路是十分困难的。后文将具体阐述Top-down方法的特点与不可行性。

1. Top-down方法的优势

在笔者看来，与Bottom-up方法相比Top-down方法最大特点在于可以充分运用机器的算力来进行算法地高准确度运行。一旦AMA的设计者定义好了道德规则向机器指令算法的转换方法，AMA便可以在任何情况下运用算力在极短时间内完成道德计算，进而做出在这种道德规则框架下最为“正确”（合乎规范）的行为。换句话说，Top-down方法的优势在于强可靠性，这种可靠性并不意味用这种方法设计出的AMA永远不会做“错事”，而是说其不会做出逾越设计者规定的道德体系的行为。

这种基于可靠性需要辩证地看待。一方面，这种可靠性保证了AMA行为与设计者设计初衷和意愿的强相关性，这种保守性确保了AMA不会主动做出违背设计者设计意愿的事。另一方面，这种人工设计主导下的AMA，其本身的智能性便有待考量。在笔者看来，这种自上而下设计出的AMA更像是设计者主观意志的执行机器而非具有独立思维能力的“智能体”。但从算法上看，AMA所做的仅仅是收集巨量的时空信息并根据设计者设定好的转化规则进行快速的道德量化计算，最终做出道德正确度最高的行为。而在该过程中AMA并没有完全体现出自身对于道德的理解与考量，只是作为一种高性能计算机器对复杂的输入进行运算并输出相应结果（做出行为），这在本质上与简单非智能化程序的运行并无太大区别，这种设计方法下AMA的智能性主要体现在其能够根据道德准则做出相应行为的能力上。

1. Top-down方法的可行性分析

正如前文所述，与莱布尼茨时代相比，今天算力已不再是阻碍Top-down方法实现的主要障碍，但从理论上看，完全通过这种自顶向下进路实现道德机器仍是不可行的。困难主要体现在道德规则-算法转换函数的设计上，原因主要有以下两方面：道德规则本身的不完备性、人类道德判断的固有歧义性。

在最一般的意义上，自上而下实现人工道德的进路涉及一套可以转换为算法的一组规则。这种规则只是对于需要明确规定或禁止的所有行为的任意集合，即这种规则只是明确了在某些或绝大多数情况下必须做或者必须不做的行为，即道德的“戒律”模式（如黄金法则、摩西十诫等）。显而易见，在现实生活中一些戒律并不适用于某些特定情况，甚至在某些情况下同一道德体系中不同的道德规则之间会发生冲突。概括地说，道德规则本身是不完备的，因为道德体系难以囊括任意时间任意空间中适用的所有道德准则，设计者更不可能将这些“枚举类型”的道德准则转化为算法输送给AMA。

为了解决同一道德体系中的规则冲突问题，哲学家们试图寻找更为普适和抽象的原则，希望AMA能够根据这些概括性原则在具体情况下推导出具体正确的道德原则。概括性原则的选取总体上可以分为功利主义学派和道义论学派。这两种学派的具体观点在此不做赘述，需要明确的是无论是功利主义还是道义论，在它们的概括性规则体系下设计AMA都会遇到难以解决的计算性问题——如何使得AMA在概括性规则的基础上通过算法对千变万化的情境进行道德计算处理。

功利主义主张效果论，认为道德的目的是实现世界“功利”总量的最大化，关注行动所带来的“功利”量。从AMA的角度看，困难在于如何判定不同行为方式的结果以得到某种功利测度结果的最大化——这需要智能体做到预测几乎所有行为可能带来的结果，这使得AMA的计算需求无法被满足。而道义论认为责任是道德的核心，重点关注道德行为的动机而非结果。从AMA的角度看，困难在于如何让AMA具有知道或分析自身行为动机的能力，即在具体情况下知道如何使用规则（或知道如何判定规则）。人类道德的固有歧义性可能使得AMA陷入两难的境地，这种情况下无论AMA做出什么样的行为对道德主体中的某一方都是不符合道义论的，这时AMA可能就无法在既定的道德规则体系下做出适当的行为。

参考文献：

[1]. 【美】W.瓦拉赫 与【美】C.艾伦《道德机器》