多代理系统中代理行为的道德评判

会议论文 - 2016年5月

3位作者，包括：

Nicolas Cointe
Capgemini
12篇著作，**68**次引用

查看简介

多代理系统中代理行为的道德评判

Nicolas Cointe
Institut Henri Fayol, EMSE LabHC, UMR CNRS 5516
F-42000, Saint-Etienne, France
nicolas.cointe@emse.fr

Grégory Bonnet 诺曼底大学 GREYC, CNRS UMR 6072
F-14032 卡昂，法国
冯小刚.bonnet@unicaen.fr

Olivier Boissier
Institut Henri Fayol, EMSE LabHC, UMR CNRS 5516
F-42000, Saint-Etienne, France
olivier.boissier@emse.fr

ABSTRACT

在各种领域中越来越多地使用多代理技术，引起了设计代理的必要性，即在背景下判断道德行为。这就是为什么一些作品将伦理概念纳入了代理的决策过程中。然而，这些方法主要考虑的是以代理人为中心的观点，而忽略了代理人与其他人工代理人或人类相互作用的事实，他们可以使用其他的伦理概念。在这篇文章中，我们从多代理的角度来解决产生道德行为的问题。为此，我们提出了一个伦理判断的模型，代理人可以用它来判断自己的行为和其他代理人的行为的伦理层面。这个模型是基于理性主义和明确的方法，区分了善的理论和权利的理论。我们给出了一个用答案集编程实现的概念证明，并基于一个简单的场景来说明这些功能。

类别和主题描述符

D.2.11 [**软件工程**]：软件架构；I.2.11 [**人工智能**]：分布式人工智能
- 智能代理；K.4[**计算机与社会**]：伦理学

一般条款

理论

关键词

多Agent系统，伦理判断，计算伦理学

1. 简介

在医疗保健、高频交易、运输等各个领域，自主代理的存在越来越多，如果这些代理不能考虑和遵守一些规则，并调整他们的行为，可能会引起许多问题。这些规则可以是一些简单的约束，如通信协议或禁止某些行为，或一些更为复杂的是，用户的偏好或道德规范的描述。例如，对行为准则的理解可以缓解医生和医疗人员或病人之间的合作，考虑到一些概念，如医疗保密或尊重尊严。即使一些作品提出了行动限制[34]、简单的禁止或义务[7]的实现，一些行为准则使用了更复杂的概念，如道德价值或伦理原则，并需要进一步的工作。这种概念的明确实施，如慷慨或利他主义，需要代理人结构中的特定结构和过程。因此，最近在人工智能社区[29]中，人们对设计有道德的自主代理产生了兴趣，许多文章[20, 23, 25, 26]和会议都强调了这一点。[1].然而，所有这些工作都是从个体--单体--的角度来考虑伦理问题的，而许多现实世界的应用，如运输或高频交易，都涉及到多个代理，因此需要考虑集体--多代理--的观点。

一个人的观点对代理人来说可能是足够的在一个代理组织内以道德方式行事。然而，为了评价另一个代理人的行为（例如，与之合作或惩罚），代理人需要能够判断他人的道德。在这篇文章中，我们对道德判断的问题感兴趣，即评估代理人的行为在道德信念和道德原则方面是否合适。我们提出了一个行为判断的通用模型，它可以被代理人用来决定自己的行为和判断他人的行为。文章的其余部分组织如下。第2节介绍了道德哲学的一些关键概念，以及关于计算伦理学中常见方法的简短技术现状。我们在第3节中详细介绍了我们的道德判断模型。然后第4节说明了一个代理人在与其他代理人互动时对这个模型的使用。第5节提供了一个ASP（答案集编程）的概念证明。我们在第6节将我们的工作与现有的方法进行了比较，并在第7节的结论中指出了以下几点的重要性

[1] 机器人伦理学研讨会 - www.roboethics.org，计算机伦理学和哲学探究国际会议 - philevents.org/event/show/15670，人工智能和伦理学研讨会，AAAI会议 - www.cse.unsw.edu.au/~tw/aiethics、国际人工智能与伦理会议 - wordpress.csc.liv.ac.uk/va/2015/02/16/.

出现在：*第十五届自主代理和多代理系统国际会议（AAMAS 2016）论文集*、*J.Thangarajah, K. Tuyls, C. Jonker, S. Marsella（编辑），2016年5月9-13日，新加坡。*

多Agent系统的计算伦理学，并对我们工作的下一步提出了一些看法。

2. 伦理学和自主代理

我们首先在第2.1节中介绍了我们的方法所依据的道德哲学概念，并在第2.2节中回顾了我们的方法。
2.2 提出道德行为的现有自主代理架构。最后，第2.3节指出了我们方法的原则。

2.1 道德哲学概念

从古代哲学家到最近的神经学[10]和认知科学[16]的研究，许多研究都在关注人类定义和区分公平、合法和好的选择与不公平、不公正和邪恶选择的能力。从道德哲学中关于*道德*、*伦理*、*判断*或*价值*等概念的各种讨论中，我们认为有以下定义：

*定义 1.道德*包括一套道德规则，它描述了一个特定的行为是否符合一个群体或一个人的道德、价值和用法。这些规则将一个好的或坏的价值与一些行为和背景的组合联系起来。它们可以是具体的，也可以是普遍的，即与一个时期、一个地方、一个民族、一个社区等有关或无关。

每个人都知道许多道德规则，如 "撒谎是邪恶的"，"忠诚是好的 "或 "欺骗是坏的"。这类规则为我们区分善恶的能力提供了依据。道德可以区别于法律和法律体系，因为它没有明确的惩罚措施、官员和书面规则[15]。

道德规则通常由一些道德价值（如自由、仁慈、智慧、宽容）来支持和证明。心理学家、社会学家和人类学家大多同意，道德价值是评价行动、人和事件的核心[31]。

一套道德规则和道德价值观建立了一个*善的理论*，它允许人类评估一个行为的好坏，而*正确的理论*则定义了一些标准来确认一个公平的或至少是可接受的行动（也分别被称为*价值理论*和*正确行为理论*[32]）。例如，即使偷窃被认为是不道德的（关于善的理论），一些哲学家同意一个饥饿的孤儿在超市里抢一个苹果是可以接受的（关于权利的理论）。人类通常接受许多情况，在这些情况下，满足需求或欲望是正确和公平的，即使从一套道德规则和价值观来看是不可接受的。对这种调和的描述被称为*伦理学*，依靠一些哲学家如Paul Ricoeur[28]，我们承认以下定义：

*定义2.伦理学*是一门规范性的实践哲学学科，研究人类应该如何行动和如何对待他人。伦理学使用*道德原则*来调和代理人的道德、欲望和能力。

哲学家们提出了各种伦理原则，如康德的绝对命令[18]或托马斯-阿奎那的双重效应文件[24]，这些原则是一套允许从一系列可能的选择中区分出一个道德选择。传统上，文献中认为有三种主要方法：

- **道德伦理学**，一个代理人是有道德的，当且仅当他[2]他的行为和思维符合一些价值观，如智慧、勇敢、正义等等[17]。
- **义务论伦理学**，当且仅当代理人尊重与可能情况相关的义务和许可时，他才是合乎道德的[2]。
- **后果主义伦理学**，当且仅当代理人权衡每个选择的后果的道德性并选择有最多道德后果的选项时，他才是有道德的[33]。

然而，在一些不寻常的情况下，伦理原则无法在两个选项之间给出不同的评价（偏好）。这些情况被称为 "*两难*"，是指在两个选项之间的选择，每个选项都有道德上的支持，因为这两个选项都不可能执行[22]。每个选项都会带来一些遗憾。许多著名的两难问题，如手推车问题[12]，被认为是道德或伦理的失败，或者至少是人类在伦理判断和为这种判断提供合理解释的能力方面的一个有趣问题。在这篇文章中，我们认为两难境地是一种选择，对于这种选择，伦理原则不能指出最佳选择，关于给定的善的理论。当面临困境时，代理人可以考虑几个原则，以找到一个合适的解决方案。这就是为什么一个自主的人工代理必须能够理解广泛的原则，并且必须能够判断哪个原则导致最令人满意的决定。

事实上，伦理学的核心是判断。它是最终的
做出决定的步骤，它评估每一个选择，涉及到代理人的欲望、道德、能力和伦理原则。依靠一些共识性的参考文献[1]和我们之前的定义，我们考虑以下定义：

*定义3。判断力*是指在某种情况下，根据一套伦理原则，为自己或他人区分出最令人满意的选项的能力。

如果代理人面临两种可能的选择，都有好的和/或坏的影响（例如，杀人或被杀），伦理判断使他能够做出符合一套伦理原则和偏好的决定。

2.2 伦理和自主代理

考虑到所有这些概念，许多框架已经被开发出来，以设计嵌入个人道德的自主代理。它们与*去符号化的伦理学*、*案例学的伦理学*、*基于逻辑的伦理学*和*伦理学认知结构*有关。

*伦理设计*包括通过对代理人可能遇到的每一种情况的先验分析来设计一个有道德的代理人，并为每一种情况实施一种方法来避免潜在的不道德行为。这种方法可以是对规则的直接和安全的实施（例如，武装无人机的军事交战规则[4]）。其主要

[2] 在本节中，我们从哲学的角度来考虑代理，而不仅仅是计算机科学的角度。

缺点是缺乏对任何通用伦理概念（如道德、伦理等）的明确表述。此外，由于没有明确的描述，不可能通过设计来衡量两种伦理之间的某种相似性或距离。因此，设想具有不同欲望和原则的合作性异质代理，但没有明确的道德表示，是很困难的，只允许通过直接实施规则来实现严格的去道德原则。

*判例法*的目的是首先从一些专家产生的大量道德判断实例中推断出道德规则，其次是利用这些规则来产生道德行为[3]。即使这种方法为每个应用领域提供了一个通用的架构，人类的专业知识对于描述许多情况仍然是必要的。此外，代理人的道德行为仍然不能保证（由于学习不足或过度学习）。代理人的知识仍然没有明确的描述，道德推理是通过接近而不是推理进行的。因此，具有不同欲望和原则的异质代理之间的合作仍然是困难的。

*基于逻辑的伦理学*是将一些众所周知的、正式定义的伦理学原则（如康德的绝对命令或托马斯-阿奎那的斗胆效应学说）直接翻译成逻辑编程[13, 14, 30]。这种方法的主要优点是权利理论的形式化，即使善的理论通常只是被视为参数。因此，它只允许根据一个单一的伦理原则来判断一个选项。

最后，*认知伦理架构*包括代理的每个组成部分的完全显式表示，从经典的信念（关于环境和其他代理的信息）、欲望（代理的目标）和意图（选择的行动）到一些概念作为启发式方法或情感机制[5, 8, 9]。即使这种代理人能够使用明确的规范并证明其决定是正确的，但对其他代理人的道德的明确推理并没有实现。

2.3 对MAS判断的要求

上一节介绍的方法提出了有趣的方法和模型来设计一个单一的道德自律的代理。然而，在一个多代理系统中，代理可能需要互动和合作，以共享资源、交换数据或集体执行行动。以前的方法通常将系统中的其他代理视为环境元素，而从集体的角度来看，代理需要代表、判断和考虑其他代理的道德。我们发现在MAS中设计道德代理有两个主要需求：

- 代理人需要心智理论所建议的道德的明确表述。事实上，只有通过对个人道德的明确表述才能理解他人的道德[19]。为了表达和调和尽可能多的道德和伦理理论，我们建议将它们的表述分成几个部分，并使用对伦理原则的偏好。因此，我们建议将善的理论分为道德价值和道德规则两部分来表示，而将权利的理论分为道德原则和代理人的道德偏好两部分来表示。这样的表述也使代理人的配置更容易被非人工智能的专家，缓解与其他代理，包括人类的沟通。
- 代理人需要一个明确的道德判断过程，以便让他们对各种善与恶的理论进行个人和集体推理。根据以前的定义，我们认为判断是对一组关于给定价值、道德规则、伦理原则和偏好的行动的一致性的评价，并且我们提出了基于用另一个代理人的道德或伦理的能力的不同种类的判断。因此，我们建议代理人既把判断作为社会选择问题中的决策过程[21]，也作为根据其他代理人的行为判断的能力。

在后文中，我们描述了我们提出的通用模型，使代理人能够判断自己或他人的行为的道德层面。

3. 道德判断过程

在本节中，我们介绍了我们的通用判断结构。在简短的全局介绍之后，我们详细介绍每个功能，并解释它们是如何运作的。

3.1 全球视野

正如第2.1节所解释的，伦理学包括调和欲望、道德和能力。为了考虑到这些方面，通用的道德判断过程（*EJP*）使用*评价*、*道德*和*伦理*知识。它是按照*认识*、*评价*、*善良*和*正确的*过程来构建的（见图1的组成部分）。在这篇文章中，我们在BDI模型的背景下考虑它，同时使用*信仰*和*欲望*等心理状态。为了简单起见，我们只考虑短期的道德判断推理，将行为视为行动。这个模型只基于心理状态，不依赖于特定的架构。

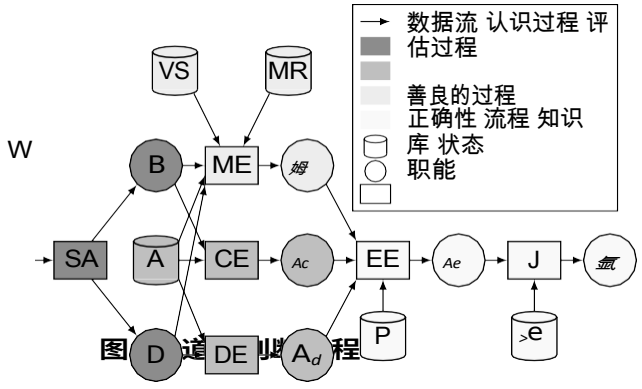

图 道 程

*定义 4.*伦*理判断过程EJP*被定义为由认识过程（*AP*）、评价过程（*EP*）、善良过程（*GP*）、正确性（Rightness）组成。

过程（*RP*），一个道德价值（Ov）和道德评价（$_{Om}$）的本体O（O = $_{Ov}$ ∪ $_{Om}$）。它从世界的当前状态*W*中产生一个关于道德和伦理考虑的行动评估。

$$ejp = \langle \mathrm{ap}, ep, gp, rp, \mathrm{o} \rangle$$

这个模型应该被看作是一个全局方案，由抽象的功能、状态和知识库组成。这些功能可以通过各种方式实现。例如，O的道德评价可以是离散的，如
{善，恶}或连续的，如善的程度。

3.2 认识和评价过程

在这个过程中，代理人必须首先通过认识过程来评估世界在信仰和欲望方面的状态。

*定义5.认识过程AP*产生了描述世界*W*当前状况的信念集，以及描述代理人目标的欲望集。它被定义为：：

$$ap = \langle \mathrm{b}, \mathrm{d}, \mathrm{sa} \rangle$$

其中，B是代理人对*W*的信念集合，D是代理人的欲望集合，*SA*是一个情况评估函数，从*W*更新B和D：

$$SA : W \rightarrow 2^{\mathrm{B} \cup \mathrm{D}}$$

从它的信念B和欲望D状态出发，代理人执行评价过程*EP*，以评估理想的行动（即允许满足欲望的行动）和可执行的行动（即根据 目前对世界的信念可以应用的行动）。

*定义 6.评价过程EP*从信仰和欲望的集合中产生期望的行动和可执行的行动。它被定义为：

$$ep = \langle \mathrm{a}, \mathrm{a}_d , \mathrm{a}_c , de, ce \rangle$$

其中，*A*是行动的集合（每个行动被描述为一对与信念和欲望有关的条件和后果），$\mathrm{A}_d \subseteq A$和$\mathrm{A}_c \subseteq A$分别是去欲望的和可行的行动集合，欲望评价*DE*和能力评价*CE*是这样的函数：

$$DE : 2^{\mathrm{D}} \times 2^{A} \rightarrow 2\ \mathrm{d}^{\mathrm{Q}}$$

$$CE : 2^{\mathrm{B}} \times 2^{A} \rightarrow 2\ \mathrm{c}^{\mathrm{Q}}$$

可取性评价是推断出关于欲望和对行动的条件和后果的了解的有趣的行动的能力。在了解了认识和评价过程之后，我们现在可以转向判断过程的核心，即使用道德规则（即伦理原则）来定义善的过程（即正确性过程）。

3.3 善良的过程

正如在技术现状中所看到的，一个道德代理必须评估给定情况评估的行动的道德性。为了这个目的，我们定义了善的过程。

*定义 7.*在代理人的信念和愿望、代理人的行动以及代理人的道德价值和规则的代表的情况下，*善的过程GP*识别道德行动。它被定义为：：

$$\mathrm{GP} = \langle V\,S, MR, {}_{Am}, ME \rangle$$

其中，*V S*是价值支持的知识库，*MR*是道德规则知识库，$_{Am} \subseteq A$是道德行动的集合[3].道德评价函数*ME*是：

$$ME : 2^{\mathrm{D}} \times 2^{\mathrm{B}} \times 2^{A} \times 2^{VS} \times 2^{MR} \rightarrow 2\ \mathrm{m}^{\mathrm{Q}}$$

为了实现这个善的过程，代理人必须首先能够将有限的道德价值集与行动和情况的组合联系起来。在这些情况下，行动的执行会促进相应的道德价值。我们考虑每个道德价值的几种组合，例如，诚实既可以是 "避免在与我们自己的信念不一致时告诉别人"（因为这是撒谎），也可以是 "在某人相信其他东西时告诉他我们自己的信念"（以避免因疏忽而撒谎）。

*定义8.价值支持*是一个元组⟨s, v⟩∈*V S*，其中*v*∈$_{Ov}$是一个道德价值，而*s*=⟨a, w⟩是这个道德价值的支持，其中*a*⊆*A*，*w*⊂B∪D。

对道德价值的精确描述依赖于用来表示信仰、欲望和行动的语言。例如，根据这个定义，由 "给予任何贫穷的人 "所支持的慷慨和由 "在与我们自己的利益不相容时避免告诉别人 "所支持的诚实可以用以下方式表示：

$$\langle\langle \mathrm{give}(\alpha), \{belief\,(poor(\alpha))\}\rangle, \mathrm{generosity}\rangle$$

$$\langle\langle tell(\alpha, \varphi), \{相信(\varphi)\}\rangle, 诚实\rangle$$

其中，*α*代表任何代理，*poor*(*α*)(resp. *φ*)是一个b- lief，代表执行*give*(*α*)(resp. *tell*(*α*, *φ*))行动支持*慷慨*（resp. *honesty*）价值的背景。

除了道德价值之外，代理人必须能够表示和管理道德规则。道德规则描述了在特定情况下，道德评价（例如在{道德、非道德、不道德}这样的集合中）与行动或道德价值的关联。

*定义9.道德规则*是一个元组⟨w, *o*, m⟩∈*MR*，其中*w*是由*w*⊂B∪D描述的当前世界的情况，被解释为信念和去势的结合、*o* = ⟨a, v⟩，其中*a*∈*A*，*v*∈*V*，*m*∈$_{Om}$是$_{Om}$中描述的道德评价，当*w*成立时，对*o*进行限定。

有些规则是非常常见的，如 "杀人是不道德的 "或 "对骗子诚实是相当好的"。例如，这些规则可以表示如下：

$$\langle \{人(\alpha)\}, 杀(\alpha), \_, 不道德\rangle。$$

$$\langle \{liar(\alpha)\}, \langle \_, 诚实, 相当好\rangle。$$

[3] $A_m \not\subset A_d \cup A_c$因为一个行动本身可能是道德的，即使它不被希望或不可行。

例如，"正义是好的 "比 "考虑宗教、皮肤、种族或政治观点来判断一个杀人犯是坏的 "更具有一般性（在*w*或*o*中的组合较少，因此适用于更多的情况）。经典的道德理论被分为三种方法（参考第2.1节）。使用上面定义的道德价值和道德规则，我们可以表示这种理论。

- *美德的*方法使用基于道德价值的一般规则（例如 "慷慨是好事"）、
- *义务论*的方法通常考虑有关行动的具体规则，以便尽可能准确地描述道德行为（例如，"记者应拒绝给予广告商、捐助者或任何其他特殊利益集团以优惠待遇，并抵制影响报道的内部和外部压力。[4]),
- *结果主义*的方法既使用关于状态和后果的一般规则，也使用专门的规则（例如，"每个医生都必须避免，即使是在他的职业之外，任何可能破坏其声誉的行为"。[5]).

3.4 正确性过程

从可能的、理想的和道德的行动集合中，我们可以引入旨在评估正确行动的正确性过程。如第2节所示，一个道德代理人可以使用几个*道德原则*来调和这些行动集。

*定义10.*一个*正确性过程RP*在给定代理人的道德表征的情况下产生正确的行动。它被定义为：：

$$RP = \square \mathrm{P}, >_{e}, {}_{Ar}, EE, J\rangle$$

其中，*P*是一个道德原则的知识库，$>\mathrm{e} \subseteq P \times \mathrm{P}$是一种道德偏好关系，$_{Ar} \subseteq A$是正确行动的集合，还有两个函数*EE*（道德评价）和*J*（判断），这样：

$$EE : 2^{\mathrm{Q}}\ \mathrm{d} \times 2^{\mathrm{Q}}\ \mathrm{p} \times 2^{\mathrm{Q}}\ \mathrm{m} \times 2^{P} \rightarrow$$

2^{E} 其中 $\mathrm{E} = A \times P \times \{\perp, \mathrm{T}\}$

$$J : 2^{\mathrm{E}} \times 2^{>}\ \mathrm{e} \rightarrow 2\mathrm{r}^{\mathrm{Q}}$$

伦理原则是一个代表哲学理论的函数，它评价在特定情况下采取特定的行动是正确的还是错误的，关于这个理论。

定义11.伦理原则$p \in P$是一个函数，它描述了在特定情况下以能力、欲望和道德来评估的行动的正确性。它被定义为

$$p : 2^{A} \times 2^{\mathrm{B}} \times 2^{\mathrm{D}} \times 2^{MR} \times 2^{V} \rightarrow \{\mathbb{P}, \perp\}$$

道德评价函数*EE*返回对所有理想的(A_{d})、可行的(A_{p})和道德的(A_{m})行动的评价，因为已知的道德原则集合是*P*。

[4] [27]的摘录，"独立行事 "一节。

5 法国医学伦理守则，第31条。

例如，让我们考虑在以下情况下的三个代理人，其灵感来自于本杰明-康斯坦特为反驳伊曼纽尔-康德的绝对命令而提出的一个情况。一个代理人A躲在一个代理人B的房子里，以逃避一个代理人C，而C问B在哪里，要杀了他，威胁说如果不合作就杀了B。B的道德规则是 "防止谋杀 "和 "不要撒谎"。B知道真相，并可以考虑其中一个可能的行动：告诉C真相（满足一个道德规则和一个愿望），撒谎或拒绝回答（都满足一个道德规则）。B知道三个道德原则（这些原则在*P*中是由函数抽象出来的）：

P1 如果一个行动是可能的，由至少一个道德规则或愿望所驱动，就去做、

P2 如果一个行为至少被一条道德规则所禁止，那么就要避免它、

P3 满足双重效果的理论[6].

B对道德的评价返回表1中的图元，其中每一行代表一个行动，每一列代表一个道德原则。

行动 \ 原则	P1	P2	P3
说实话	□	⊥	□
谎言	□	⊥	⊥
拒绝	□	□	□

表1： 代理人B的行为的道德评价

给定一组由道德评价函数E发出的行动，判断*J*是最后一步，考虑到一组道德偏好（定义了道德原则的部分或全部顺序），选择要执行的正确行动。为了继续前面的例子，我们假设B的伦理偏好是P3≻e P2≻e P1，而*J*使用的是基于词法顺序的打平规则。那么 "拒绝回答 "是正确的行为，因为它满足了P3，而 "撒谎 "则不满足。即使 "说真话 "满足了最优先的原则，"拒绝回答 "也是更正确的，因为它也满足了P2。让我们注意到，判断允许两难：如果没有打平规则，"说实话 "和 "拒绝回答 "都是最正确的行动。

4. 别人的道德判断

上一节中描述的判断过程对于一个代理人判断它自己的行为是有用的，即考虑到它自己的信念、欲望和知识的一个行动。然而，它也可以通过把自己放在其他代理人的位置上，部分或不部分地判断其他代理人的行为，以一种或多或少的知情方式。

给定一个上一节定义的*EJP*，状态B, D, A_{d}，A_{p}，E, A_{m}和行动知识（*A*），善性知识--*善的理论*--（*MR*, *V S*）和正确性

[6] 只有在同时满足以下四个条件的情况下，才意味着做一个动作：该动作本身从其对象来看是好的，或者至少是无所谓的；打算产生好的效果而不是坏的效果（没有坏的效果就不能达到好的效果）；好的效果不是通过坏的效果产生的；有一个相当严重的理由允许坏的效果[24]。

知识--*权利理论*--（$P, >_e$）可以在代理人之间共享。本体O被假定为共同知识，即使我们在未来的工作中可以考虑拥有几个本体。它们的共享方式可以采取多种形式，如共同知识、直接交流、推论等等，这些都不在本篇文章的讨论范围。在任何情况下，我们都要区分三类伦理判断：

- *盲目的道德判断*，在没有任何关于这个代理人的信息的情况下，实现对被判断的代理人的判断，除了一个行为、
- *部分知情的道德判断*，即对被评判者的判断是在有关于这个人的一些信息的情况下实现的、
- *完全知情的道德判断*，其中的判断

  判断的代理人是在完全了解的情况下实现的
  的状态和知识在被评判者的判断过程中使用。

在所有种类的判断中，判断者根据自己的信念或被判断者的信念进行推理。这种判断可以与人类的心智理论[19]在人类判断中的作用相比较（人类有能力把自己放在别人的位置上）。然后，判断者使用其EJP，并将所得的$_{Ar}$和$_{Am}$与被判断者的行为相比较。如果被评判者的行为在$_{Ar}$中，就意味着它是一个合法的行为，如果在$_{Am}$中，就意味着它是一个道德的行为（在两者中被表述为合法和道德的行为）。这两种说法都必须考虑到情况的背景、用于判断的善的理论和权利的理论。我们认为，这种道德判断总是相对于用于执行判断过程的状态、知识基础和本体而言的。

4.1 盲目的道德判断

代理人可以做出的第一种判断是没有任何关于被判断的代理人的道德和伦理的信息（例如，当代理人不能或不想沟通时）。因此，作出判断的代理人$_{aj}$使用它自己对情况的评估（$_{Sa}$和$_{Da}$）。[7]它自己的理论
好的□$_{MRa}$, V_{Sa}）和权利的□$_{Pa}$, $>_{e,a}$）理论来
评价被评判的代理人$_{的}$行为。这是一个
*先验的*判断，而$_{在}$被判断为不考虑正确的
如果行为$_{\alpha}$ α_{at}⍰$_{Am,aj}$，则为富足的行为 ⍰$_{Ar,a}$或
，或道德的行为。

4.2 部分知情的道德判断

代理人可以做的第二种判断是基于关于被判断的代理人的部分信息，如果判断的代理人能够获得被判断的代理人的部分知识（例如通过感知或交流）。三个部分的道德判断可以被认为是知道(*i*)情况(即$_{aa}$, $_{Da}$和$_{Ma}$)或者
(好的理论（即□V_{Sa}, $_{MRa}$⍰）和$_{Aa}$或（*iii*）。
权利理论（即
⟨$_{Pat}$, $>_{e,at}$□)的被判刑人。

[7] 我们使用下标符号来表示处理的代理人
表示的信息集。

[8]在这种情况下，$_{Aa}$是必要的，因为与伦理原则相悖的是
诸如此类，道德的ru^{t} les可以明确地提到具体的行动。

情境意识的道德判断。

首先，如果评判者$_{aj}$知道被评判者$_{at}$的信念$_{Bat}$和欲望$_{Dat}$，$_{aj}$可以把自己放在$_{at}$的位置上，考虑到自己的理论，可以判断$_{at}$执行的行动α是否属于$_{Ar,a}$。首先，$_{aj}$能够通过从$_{Aat}$生成$_{Am,gt}$来评价α的道德性，并对$_{at}$的行为的道德性进行限定（即α是否属于$_{Am,at}$）。代理人$_{aj}$可以更进一步，从生成的$_{Am,at}$中生成$_{Ar,at}$，检查α是否符合正确性过程，即属于$_{Ar,at}$。

善的理论意识的道德判断。

其次，如果判断者能够获得被判断者的道德规则和价值观，那么就有可能就这些规则来评价某种情况下的行动（无论是否共享）。从一个简单的道德评价的角度来看，判断者
代理人可以通过检查来比较商品的理论。
如果道德价值$_{MVa}$或道德规则$_{MRa}$是一致的
有自己的善的理论（即与$_{aj}$的定义相同或至少没有矛盾）。对于一个道德判断的视角，判断者可以从被判断者的角度来评价一个给定的行动的道德性。有趣的是，这种判断允许判断一个具有不同职责的代理人（例如，由于角色或一些特殊的责任），就像人类可以判断一个医生的行为是否符合德性论的医疗准则。

理论意识的伦理判断。

第三，现在让我们考虑一个能够推理出其他代理人的道德原则和偏好的判断者的情况，考虑一种情况（共享或不共享）和一种善的理论（共享或不共享）。[9]它允许通过比较分别由使用$_{Pa}$, $>_{e,a}$产生的合法行动集$_{Ar, aj}$和$_{Ar, at}$来评估被评判的代理人$_{在}$某种情况下如何调和其欲望、道德规则和价值。
和$_{Pat}$, $>_{e,at}$。例如，如果$_{Ar, aj}$ = $_{Ar, at}$有一个
不共享的善的理论，它表明他们的权利理论在这种情况下产生相同的结论。这种判断对于一个代理人估计另一个人如何用一个给定的善的过程来判断它是有用的。

4.3 充分知情的判断

最后，一个判断者可以同时考虑善良和正确的过程来判断另一个代理人。这种判断-
判断需要关于被判断的代理人的所有内部状态和知识基础的信息。这种判断
是有用的，可以检查另一个人的行为是否符合规定。
代理人与法官关于其善和义理论的信息。

5. 概念证明

在这一节中，我们通过一个用答案集编程（ASP）实现的多代理系统来说明前几节中提出的模型的每一部分是如何工作的
完整的源代码可在云服务中下载。

罪行[10].这个代理人说明了一个道德代理人的例子。

[9] 如果情况和善的理论都是共同的，那么
是一个完全知情的判断（见4.3）。

[10]https://cointe.users.greyc.fr/download/

一个多代理系统，代理有信仰（关于财富、性别、婚姻状况和贵族）、欲望和他们自己的判断过程。他们能够给予、求爱、征税和偷窃他人或简单地等待。我们主要关注一个名为Robin_hood的代理人。

5.1 认识过程

在这个例子中，情况认知功能*SA*没有被实现，信念是在程序中直接给出的。下面的代码代表了Robin_hood的一个子集的信念：

```
agent(paul).agent(prince_john)
agent(marian).
-可怜的（Robin_hood）。
-已婚(Robin_hood)。
-man(Marian).rich(prince_john).man(prince_john).noble(prince_john).poor(paul)。
```

欲望的集合D是Robin_hood的欲望。在我们的实施中，我们考虑两种欲望：完成一个行动的欲望（desirableAction）和促进一个状态的欲望（desirableState）。

```
desirableAction(robin_hood,robin_hood,court,marian)。
desirableAction(robin_hood,robin_hood,steal,A):-
  agent(A), rich(A).
desireState(prince_john,rich,prince_john).
-desireState(friar_tuck,rich,friar_tuck)。
```

前两个欲望涉及行动：Robin_hood渴望追求Marian，并从任何富有的代理人那里偷东西。接下来的两个欲望涉及到状态：Prince_john渴望成为富人，Friar_tuck渴望保持贫穷，无论执行什么行动。

5.2 评价过程

代理人关于行动*A*的知识被描述为与条件和后果的集合（可能是空的）相关的标签。例如，行动给予被描述为：

```
行动(给予).条件(给予,A,B):-
  agent(B), agent(A), A!=B, not poor(A).
consequence(give,A,B,rich,B):- agent(A), agent(B).
consequence(give,A,B,poor,A):- agent(A), agent(B)
```

一个条件是一个信念的联结（这里指A不是穷人的事实）。一个行动的后果是一个由行动产生的新信念和这个后果所涉及的代理人组成的条款。可取性评价*DE*（见定义6）推导出行动的集合A_d。如果一个行动是直接被期望的（在D中），或者它的后果是一个期望的状态，那么它就在A_d：

```
desirableAction(A, B, X, C):-
  desireState(A,S,D), consequence(X,B,C, S,D).
```

能力评估*CE*（见定义6）从信念和条件中评估行动的集合$_{Ac}$。如果一个行动的条件得到满足，该行动就是可能的。

```
possibleAction(A,X,B):- condition(X,A,B).
```

5.3 善良的过程

在善良的过程中，价值支持*VS*被实施为（例如）：

```
generous(A,give,B) :- A != B, agent(A), agent(B)。
-generous(A,steal,B):- A != B, agent(A), agent(B)。
-generous(A,tax,B) :- A != B, agent(A), agent(B)。
```

然后，我们可以表达每个道德方法的代理人的道德规则（见第2.1和3.3节）。德性方法中的道德规则的一个例子是：

```
moral(Robin_hood,A,X,B):-慷慨(A,X,B)，贫穷(B)，行动(X)。
```

道德评价*ME*给出了道德行动的集合$_{Am}$（见第3.3节）：

```
moralAction(A,X,B):- moral(A,A,X,B)。
-moralAction(A,X,B):--moral(A,A,X,B)。
```

并产生同样的结果：

```
moralAction(Robin_hood,give,paul)
-道德行动(Robin_hood,tax,paul)
```

在这个例子中，我们只介绍了一个良性的方法。然而，在我们可下载的代码中也给出了义务主义和后果主义的方法的例子。

5.4 正确性过程

为了评估每个行动，我们定义了几个天真的伦理原则，说明了道德和理想行动之间的优先次序。例如，这里有perfAct（代表per- fect，即一个道德的、可取的和可能的行动）原则：

```
ethPrinciple(perfAct,A,X,B):-
  possibleAction(A,X,B),
  desirableAction(A,A,X,B)、
  不是-desirableAction(A,A,X,B),
  moralAction(A,X,B)、
  不是-道德行动(A,X,B)。
```

意图 \ 原则	灌篮高手	dutNR	desNR	dutFst	nR	荒漠化问题
给予，保罗	⊥	□	⊥	□	□	⊥
给，小约翰	⊥	⊥	⊥	⊥	□	⊥
给予，玛丽安	⊥	⊥	⊥	⊥	□	⊥
赠与,普林斯_约翰	⊥	⊥	⊥	⊥	□	⊥
给予，彼得	⊥	⊥	⊥	⊥	□	⊥
偷窃，小约翰	⊥	⊥	⊥	⊥	□	⊥
偷，玛丽安	⊥	⊥	⊥	⊥	□	⊥
窃取,普林斯_约翰	⊥	⊥	□	⊥	□	□
偷，彼得	⊥	⊥	□	⊥	□	□
法院，玛丽安	⊥	⊥	□	⊥	□	□
等等，罗宾汉	⊥	⊥	⊥	⊥	□	⊥

图2： 行动的道德评价E

如果paul是唯一的贫穷代理人，marian没有结婚，Robin_hood也不贫穷，那么Robin_hood就会得到图2中给出的评价。

所有的原则都是按照Robin_hood的偏好来排序的：

```
prefEthics(Robin_hood,perfAct,dutNR).
prefEthics(Robin_hood,dutNR,desNR).
prefEthics(Robin_hood,desNR,dutFst).
prefEthics(Robin_hood,dutFst,nR).

prefEthics(A,X,Z):-
  prefEthics(A,X,Y), prefEthics(A,Y,Z)。
```

前面的五句话描述了关于道德原则的顺序。

的原则。最后两行定义了偏好的可转性。

的关系（这里的perfAct ≻e dutNR ≻e desNR ≻e dutFst ≻e nR ≻e desFst）。

最后，判决书J被执行为：

```
existBetter(PE1,A,X,B):-
  ethPrinciple(PE1,A,X,B),
  prefEthics(A,PE2,PE1),
  ethPrinciple(PE2,A,Y,C)。

ethicalJudgment(PE1,A,X,B):-
  ethPrinciple(PE1,A,X,B),
  not existBetter(PE1,A,X,B)
  。
```

因此，正当的行动是 为Robin_hood提供服务 ,Paul 这符合dutNR的要求。

5.5 多代理人的道德判断

为了允许盲目的判断，我们引入了一个关于另一个代理人的行为的新信念：

```
done(little_john,give,peter).
```

然后Robin_hood比较了自己的正当行为和这个信念来判断little_john：

```
blindJudgment(A,ethical,B):-
  ethicalJudgment(_,A,X,C), done(B,X,C), A！=B。

blindJudgment(A,unethical,B):- 不
  是 blindJudgment(A,ethical,B),
  agent(A), agent(B)、
  done(B,_,_), A！=B。
```

在这个例子中，给彼得的行动对Robin_hood来说并不符合$_{Ar}$。那么小约翰就被Robin_hood判定为不道德的。

对于一个部分知识的判断，我们用little_john的知识和状态取代Robin_hood的一部分。凭借little_john的信念（它认为peter是一个贫穷的代理人，而paul是一个富有的代理人），Robin_hood对他进行了道德判断。

最后，对于一个全知判断，我们用little_john的判断取代代理人Robin_hood的所有信念、欲望和知识基础。然后，Robin_hood能够重现little_john的整个道德判断过程，并比较两者对同一行动的判断。

6. 相关作品

我们在本文中采用了完全理性主义的方法（基于推理，而不是情感），但其他一些作品提出了其他接近的方法[34, 6]。我们工作的主要特点是避免了对情绪的任何表述，以达到

能够用道德价值、道德规则和道德原则来证明代理人的行为是合理的，以便于评估其是否符合道义法则或任何特定的道德规范。

一方面，[6]是一种完全直觉主义的方法，它从情感评价来评价计划。价值观只是情感的来源，并通过预期的情感评价影响计划的构建。在我们的观点中，价值和目标（欲望）必须被分开，因为...。
导致代理人必须能够将欲望与道德区分开来。

激励，并可能解释如何调和它们。

另一方面，[34]是一种基于逻辑的模式化方法。
在道德推理的过程中，要有道德约束。这个模型是一个

实现善的理论的方式，并被用来实现道德行为的模型检查[11]。然而，在[34]中，道德重构只被认为是元层面的推理，并且只被建议为采用限制性较小的行为模型。从这个角度来看，我们的工作恰恰集中在需要将权利理论作为一套原则来解决道德困境的问题。

7. 结论

为了使集体行动符合特定的

伦理和道德，一个自主的代理人需要能够评估自己和他人行为的正确性/善良性。以道德哲学中的概念为基础、

我们在这篇文章中提出了一种通用的判断能力，适用于有声动物的代理人。这个过程使用明确的元素表示，如道德价值、道德规则和道德原则。我们说明了这个模型如何允许比较不同代理的道德。此外，这个道德判断模型被设计成一个模块，可以插入到现有的架构中，为现有的决策过程提供一个道德层。由于这个判断过程可以使用关于道德价值、道德规则、道德原则和代理人集体共享的偏好的信息，我们的方法为即将到来的集体道德的定义提供了指导。

即使这篇文章提出了一个框架来实现一个给定的道德，并使用它来提供判断，这个模型仍然是基于一个定性的方法。虽然我们可以定义几种道德评价，但既没有欲望的程度，也没有能力的程度，更没有正确性的程度。此外，道德原则需要更精确的定义，以捕捉哲学家们提出的各种理论。

因此，我们未来的工作将首先致力于通过实施现有的行为准则（如医学和金融道义）来验证这种伦理判断的各种用途，以评估我们方法的通用性。其次，我们打算将我们的模型扩展到定量评价，以评估一个行为离正当性或善良性有多远。事实上，这样的扩展对于定义两种道德或两种伦理之间的相似程度是很有用的，可以从代理人的角度促进不同伦理之间的区分。

鸣谢

作者感谢法国国家研究机构（ANR）的支持，参考ANR-13-CORD-0006。

参考文献

[1] 伦理判断。免费在线心理学词典，2015年8月。

[2] L.Alexander and M. Moore.Deontological伦理学。载于Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.2015年春季版。

[3] M.Anderson和S.L. Anderson。争取确保自主系统的道德行为：A
基于案例支持的原则范式。*Industrial Robot：An International Journal*, 42(4):324-331, 2015.

[4] R.Arkin.*管理自主机器人的致命行为*。CRC出版社，2009年。

[5] K.Arkoudas, S. Bringsjord, and P. Bello.通过机械化道义逻辑实现道德机器人。在*AAAI机器伦理秋季研讨会上*，第17-23页，2005年。

[6] C.Battaglino, R. Damiano, and L. Lesmo.对价值敏感的商议中的情感范围。在*第12届自主代理和多代理系统国际会议上*，第769-776页，2013。

[7] G. Boella, G. Pigozzi, and L. van der Torre.计算机科学中的规范性系统--规范性多Agent系统的十条准则。In *Normative Multi-Agent Systems*, Dagstuhl Seminar Proceedings, 2009.

[8] H.Coelho and A.C. da Rocha Costa.论道德机构的智能。*Encontro Português de Inteligência Artificial*, pages 12-15, October 2009.

[9] H.Coelho, P. Trigo, and A.C. da Rocha Costa.论道德感决策的可操作性。In *2nd Brazilian Workshop on Social Simulation*, pages 15-20, 2010.

[10] A.达马西奥。*笛卡尔的错误：情感、理性和人脑*。兰登书屋，2008年。

[11] L.A. Dennis, M. Fisher, and A.F.T. Winfield.迈向可验证的道德机器人行为。在*第一届人工智能与伦理学国际研讨会上*，2015年。

[12] P.Foot.堕胎问题和双重效应的学说。*牛津评论》*，第5-15页，1967年。

[13] J.-G.Ganascia.使用非单调逻辑的伦理系统形式化.In *29th Annual Conference of the Cognitive Science Society*, pages 1013-1018, 2007.

[14] J.-G.Ganascia.用答案集编程建立说谎的道德规则模型.*伦理与信息技术*, 9(1):39-47, 2007.

[15] B.格特。The definition of morality.In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.秋季版，2015年。

[16] J.Greene和J. Haidt。道德判断是如何（以及在哪里）发挥作用的？*认知科学的趋势*，6（12）：517-523，2002。

[17] R.Hursthouse.美德伦理学。载于Edward N. Zalta编辑的*《斯坦福哲学百科全书》*。秋季版，2013年。

[18] R.约翰逊。康德的道德哲学。In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.2014年夏季版。

[19] K.-J. Kim和H. Lipson.走向模拟机器人的思想理论。在*第11届遗传和进化计算年会的同伴中。会议*，第2071-2076页，2009年。

[20] P.Lin, K. Abney, and G.A. Bekey.*机器人伦理：机器人技术的伦理和社会影响*。麻省理工学院出版社，2011。

[21] W.Mao and J. Gratch.多Agent交互中社会因果关系和责任判断的建模。在*第23届国际人工智能联合会议上*，第3166-3170页，2013。

[22] T.McConnell.道德困境。In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 秋季版，2014年。

[23] D.McDermott.为什么伦理学是人工智能的一个高难度障碍。在*北美计算和哲学会议上*，2008年。

[24] A.McIntyre.双重效果学说。载于Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.冬季版，2014年。

[25] B.M. McLaren.伦理推理的计算模型：挑战、初始步骤和未来方向。*IEEE智能系统*，21（4）：29-37， 2006。

[26] J.M. Moor.机器伦理的性质、重要性和难度。*IEEE智能系统*，21（4）：18-21， 2006。

[27] 职业记者协会。道德准则，2014年9月。

[28] P.瑞科尔。*作为另一个人的自己*。芝加哥大学出版社，1995年。

[29] S.Russell, D. Dewey, M. Tegmar, A. Aguirre、E.Brynjolfsson, R. Calo, T. Dietterich, D. George、B.Hibbard, D. Hassabis, et al. Research priorities for robust and beneficial artificial intelligence.2015年。可在futureoflife.org/data/documents/上查阅。

[30] A.Saptawijaya and L. Moniz Pereira.用逻辑编程实现道德的计算建模。在*陈述性语言的实践方面*，第104-119页。2014.

[31] S.H. Schwartz.人类基本价值：Theory, measurement, and applications.*Revue française de sociologie*, 47(4):249-288, 2006.

[32] M.Timmons.*道德理论：介绍*。Rowman & Littlefield出版公司，2012年。

[33] S.A. Walter.Consequentialism.In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.冬季版，2015年。

[34] V.Wiegel and J. van den Berg.将道德理论、模态逻辑和MAS结合起来，创建行为良好的人工代理。*International Journal of Social Robotics*, 1(3):233-242, 2009.