

结合道义逻辑的人工智能论辩系统法律推理

余喆

卢一苇

伯克哈特·谢弗

摘要: 形式论辩是一种可以在知识不一致情境下实现推理的非单调推理手段。在底层知识的表示上, 通常可以根据需求灵活采用不同的逻辑形式。在已有的法律推理系统, 例如法律本体、法律知识图谱或法律专家系统中, 智能化的法律推理往往依托于经典逻辑语言实现。这导致这些法律知识库的推理无法反映出法律推理中的不确定性、不一致性、辩证性等特征, 并限制了它们在法律任务中的实际应用。因此, 本文结合了道义逻辑与形式论辩, 刻画基于法律语义下的规范推理。这一尝试旨在通过引入新的推理功能与表达方式更好地包含法律推理的底层原则, 并为之后更复杂的法律应用打下基础。考虑到 AI 可解释问题, 本文还基于所提出的论辩系统给出了一个关于解释的形式定义。

关键词: 形式论辩; 道义逻辑; 规范推理; 可解释人工智能; 非单调推理

中图分类号: B81

文献标识码: A

1 引言

形式论辩旨在以计算机可理解的方式, 运用符号逻辑形式化地捕捉人类推理和论辩的本质, 建模以论辩为代表的人类推理, 并为论证的可计算化分析和评价提供可行的框架, 是当前人工智能与逻辑领域颇受关注的一个研究分支。形式论辩系统的优点包括可以灵活应用于认知推理、实践推理等多种不一致推理情境, 可以高效地处理系统动态性并降低计算复杂度, 易于嵌入各种逻辑推理手段, 以及在推理过程的可解释和易理解上具有优势等。([17])

规范推理是一种处理规范性陈述和原则的推理, 主要基于规范性考虑(例如伦理原则、道德价值观、社会或法律规范等)做出决定或判断。因此, 这类推理

收稿日期: 2024-10-24

作者信息: 余喆

中山大学逻辑与认知研究所暨哲学系

yuzh28@mail.sysu.edu.cn

卢一苇

爱丁堡大学法学院

Y.Lu-104@sms.ed.ac.uk

伯克哈特·谢弗

爱丁堡大学法学院

B.Schafer@ed.ac.uk

基金项目: 国家社会科学基金重大项目“汉译逻辑术语本土化与中国逻辑学话语体系建设研究”(21&ZD065)。

致谢: 感谢第一届全国人工智能逻辑大会审稿人对本文提出的修改意见。

常用于伦理、法律、政策制定和决策领域,以指导道德判断、法律解释和基于价值的决策。其中的论证通常是为了证明或支持规范性主张而构建的。尽管整体自洽是大部分规范系统的基本要求,由于社会环境和社会需求的变化、规范解释的主观性、价值观的多元化、道德困境等原因,规范之间的不一致并不鲜见。人们常常面临着需要在规范存在冲突的实际情况下进行合理推理决策的情境。

以具有代表性的法律领域为例,人工智能技术的发展,尤其是以人工智能系统为核心机制的产品,例如自动驾驶汽车的出现,给法律系统的适应与调整带来了极大的压力。这种压力不仅体现在法律自身的调整上,也体现于各追求法律合规性的主体——例如人工智能系统设计者——更难以进行大规模、快节奏的法律推理上。这主要是因为现有的法律规范在实际处理新的法律情境中的推理时,往往会产生更多冲突的、不确定的,甚至是缺失的法律推理依据。尤其是当基于逻辑推理的法律技术系统,例如专家系统或法律检测系统试图为大范围的人工智能产品提供智能化的法律协助时,这种不一致性与不确定性带来的困难更加明显。造成这些困难的原因主要包括以下几个:

1. 法律文本的模糊性表达。出于长久以来默认人类会是法律文本的唯一阅读者,现有的法律系统主要建立在自然语言之上。因此,文本中不可避免地蕴含了大量模糊的、语义不固定的表达。这一方面是因为自然语言自身的特点,另一方面也是法律刻意的保留。因为解释与论辩是法律中核心的价值观之一,而这种语义模糊性是对法律功能的一种保护。然而,对于一个基于逻辑的法律推理系统而言,这很有可能造成在一个既定情境下选择不同语义时的推理规则冲突等问题。

2. 法律推理对不一致与不确定性推理的包含。法律推理中的不一致推理与不确定推理并非由人工智能产品引发,后者仅仅是加剧了这种情况。事实上,不一致与不确定推理可以被视作法律推理的本质体现之一。法律推理与逻辑推理不同的是,尽管从形式上它也包含了由法律事实经由法律规则,推理出法律结果的过程,但是这个过程中各元素的确定包含了大量的法律解释。法律解释是在诸多冲突与模糊的法律可能中选择而来的。例如,对GPS数据的追踪侵犯的是个人财产权还是隐私权?某个存在争论的证据能否被法庭接受?如果为了维护某一法律规范而做出的行为违反了另一种法律规范,是否应当负有法律责任,或应以哪个规范为优先?不同情境下在众多不一致的法律解释中选择最适用的那一种,正是法律在当今社会中发挥作用的核心推理方式。然而,这种推理是基于人类的常识或专家对法律抽象原则的理解的。当它被嵌入到基于逻辑推理的智能系统之中,就很容易变成导致推理停滞的冲突。尽管也有一部分学术理论认为法律应当是一个完美的固定的推理机器,但德沃金式的“唯一正确答案”情况几乎不会在现实的法律推理中出现。([11])

3. 现存法律的语义和惯例对人工智能情境的解释能力有限。如上文所述,尽

管法律解释的过程充满了冲突处理与主观选择,无法选择或冲突无法处理的情况仍是少数。但人工智能产品的涌现显著地加剧了这一情况,这是因为面向人类的法律在直接引用于人工智能产品时,包含了许多不能完全解释的情况。例如,自动驾驶系统中的司机是乘客还是驾驶系统?必须有成年人监护乘车的儿童是否可以独自乘坐自动驾驶汽车?对于这些情况,无论选择何种法律解释,都会引发缺乏进一步法律支持的冲突。这一方面是因为人工智能产品带来了当今法律系统不理解的新概念,另一方面是因为法律系统是一个追求自洽的复杂系统,对一个子结构的调整很可能带来其他结构的不兼容。例如,如果因为法律责任的不确定,在自动驾驶汽车上安装酒精探测器并阻止醉酒乘客上车,虽然符合交通法或刑法要求,却可能造成违背隐私法和消费者权益保护法的问题。

我们也可以从以下例子中看出为何如此。

例 1. 根据目前大多数国家的现行交通法规,当人和货物同时涉及交通事故或其他危险情况时,应优先保护驾驶员的生命安全。这一法规基于下述法律原则:

1. 驾驶员安全优先于货物安全。

上述法律原则旨在确保在危险情况下必须保护驾驶员的安全。然而,自动驾驶的出现可能会改变这一意图。这是因为在自动驾驶情况下,驾驶员不再是人,而是系统或汽车本身,这可能使之不如其货物重要。在一个更极端的情景中,我们可以设想一辆载有狗的自动驾驶汽车——在危险中我们应该保护谁?基于上述法律原则,汽车会做出牺牲狗以保护自身安全的决定。这是因为它符合驾驶员的概念,在法律语境中,宠物或牲畜被视为财产,即一种货物。但显然,会有很多人不希望汽车做出这种选择,此时多数人认为应该应用的法律原则为:

2. 生命优先于无生命物体。

在这种情况下,我们希望确保自动驾驶汽车所采取的行动已经改变为生命安全(或重要货物)被优先保护。因此,应该保护狗的安全。相应地,法律规定需要被修改。

从例 1 中,首先,我们可以明显看出,人工智能产品带来了法律推理中的冲突。这种冲突是由于情景变化,最佳的法律解释相应发生了变化。因此,所需要的法律推理系统,应当能够表达并处理这种冲突。更重要的是,能够反映出背后法律推理原则的偏好顺序发生的变化,也就是反映出法律解释的来源,从而具备基本的法律可解释性。其次,在这个例子中,无论是法律调整还是设计者的调整,其核心问题事实上都在于行为层面,即车辆究竟应该如何做。这一层面也是法律的使用和服从者与法律制定者实际上发生交际的层面。因此,作为法律推理系统

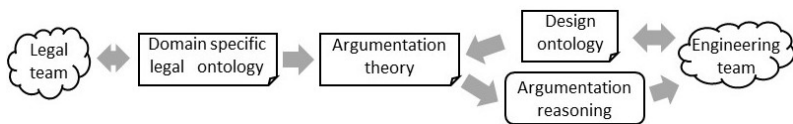


图 1: 基于论辩的法律本体推理过程

的基础, 责任与行为义务应当被明确且着重地表达。这也是为何我们选择道义逻辑, 这一已被证明非常适合表达义务与责任的逻辑作为基础语言。

可见, 使基于逻辑的推理系统具备处理不确定与不一致情境下推理的能力, 对于法律应用来说是至关重要的。并且, 相关法律推理系统应当具备如下几个基本的特性, 才能更好地反映出法律推理的内核与特征:

- 1) 对冲突情境下推理的表达与处理;
- 2) 对责任归属与行为指导的着重描述;
- 3) 对推理原则的偏好变动的描述。

图 1 展示了基于形式论辩系统处理 (可能包含不一致的) 法律本体推理的过程和作用方式。([7, 8])

下文第二节将讨论和比较与本文相关的国内外研究工作; 第三节给出基于道义逻辑语言和形式论辩的规范推理系统基本定义; 最后, 第四节总结全文并指出未来的研究方向。

2 相关工作

从中文文献来看, 目前较多的工作专注于道义逻辑以及规范的不一致性本身的思考和探讨, 例如文献 [16, 20], 对不同的非单调道义逻辑研究方法给出分析和比较, 但不倾向于推理系统本身的构建。本文的侧重点则在于希望通过道义逻辑方面研究的融入, 提升作为非单调推理手段的形式论辩系统针对不一致规范推理刻画和处理的能力, 使之能够更好地将自身所长应用于人们所需要的规范推理情境。

在与论辩系统结合的道义推理方面, 较有代表性的是文献 [14], 结合输入/输出逻辑 (Input/Output Logic) 扩展了基于结构化论辩系统 ASPIC 类的论辩系统。([9, 10]) 此外, 一些研究将讨论的重点落在规范之间的优先关系上, 例如文献 [6]。本文更倾向于将道义算子用于知识表示; 在结构化论辩系统方面, 本文借鉴几种结构化论辩系统设定的基本思路, 但重点并不在于论辩系统的扩展, 而是希望使之更适用于相关具体推理情境, 因此本文在相关论辩理论的基础上做了适当简化。

除了上述文献, 在使用论辩系统处理不一致规范推理时, 目前比较常见的仍

是基于命题逻辑或谓词逻辑作为知识表示的方法。同时,推理中的不确定性可能通过前提中的假设,也可能通过可废止推理规则来刻画,前者的代表是基于假设的论辩系统([13]),后者的代表是基于规则的论辩系统。([3, 9])基于假设的论辩系统所使用的规则与演绎推理规则类似,基于规则的论辩系统则通常既包含严格的(strict)推理规则,也包含可废止的(defeasible)推理规则。基于假设的方法在刻画自然语言论证时常常显得不够自然,而基于规则的方法的一个问题在于难以解释为何有些规则被归为严格的,有些规则被归为可废止的。我们希望基于道义逻辑作为底层逻辑的方法可以借助道义算子和逻辑对应关系,自然地得到大部分严格的规则。

3 结合道义逻辑的论辩理论

依据上文的分析,本节主要以下述自动驾驶汽车与交通法规情境为例,对相关定义进行阐释。

例 2. 面对自动驾驶汽车技术的发展,传统法律法规尚有许多需要进一步完善和厘清的细节。假设在一辆自动驾驶汽车上,约翰是唯一的乘客。如果仅将约翰视为乘客,那么按照现有的交通法规,尚未有明确规定指出他必须处于清醒的状态;而如果将约翰视为司机,那么他应当保持清醒。

在例2中,如果此时约翰处于醉酒状态,而汽车造成了交通事故,那么约翰是否应当保持清醒对于是否判定他为“醉驾”,乃至于是否应该承担法律责任而言至关重要,可见上述案例中的问题可能导致法律判决上的不一致。然而这种规范之间的冲突并不容易通过一阶逻辑很好地反映。

为了刻画规范概念和规范之间的不一致,本节尝试以道义逻辑为底层逻辑,并以形式论辩为推理工具,处理基于不一致规范的推理。

根据一些学者的分析(例如[4, 5, 12]),法律规范之间可能存在不同意义上的不一致,仅用一个初始算子“应当”难以进行刻画。因此,本文引入下述两个道义算子:“ O ”(应当)“ P ”(允许),并使用“ $\Diamond$ ”表示一种一般的可能性。我们将在常见意义上使用逻辑符号“ $\neg$ ”“ $\vee$ ”“ $\wedge$ ”“ $\rightarrow$ ”“(,)”等,并基于刻画一种(在任意对应的情境下)不可能使得两种行为或决策同时被接受的情况,即“ $\neg\Diamond(\varphi \wedge \psi)$ ”,用集合“ $\perp$ ”表示。除此之外,为了区别于确定性的推出关系,本文使用“ $\Rightarrow$ ”

表示（规范中的）不确定性推出关系。¹ 由此，例2中的规范可以分别表示为：

$$Passenger(John) \implies P(\neg Sober(John))$$

$$Diver(John) \implies O(Sober(John))$$

鉴于本文的重点聚焦于法律推理的应用情境，下文中的定义主要针对法律规范的特点给出，相关定义中所指的规范也主要是法律规范。但是法律规范作为一种典型代表，与实践推理可应用的其他领域的规范之间有天然的相通之处，我们相信相关设定也可以推广应用于其他类型的规范推理。

3.1 基本定义

我们将基于道义逻辑的逻辑语言的集合记为 $\mathcal{L}^D$ ，构造论证的规则集（即规范的集合）记为 $\mathcal{N}$ ，知识库（即论证的前提集合）记为 $\mathcal{K}$ ，一个形式论辩理论（Argumentation Theory，简称 AT ）可定义如下。

定义 3.1. 一个论辩理论可以表示为四元组 $AT = (\mathcal{L}^D, \mathcal{N}, \mathcal{K}, \perp)$ ，其中

- $(\mathcal{L}^D, \mathcal{N})$ 构成一个演绎系统， $\mathcal{L}^D$ 是基于道义逻辑的逻辑语言， $\mathcal{N}$ 是规范的集合，其中的规范形式为： $\varphi_1, \dots, \varphi_n \implies \varphi$ ($\varphi_i, \varphi \in \mathcal{L}^D$)；
- $\mathcal{K} \subseteq \mathcal{L}^D$ 是一个非空的知识库（前提集）；
- $\perp$ 是冲突关系集合，任意一组 $\varphi \wedge \psi \in \perp$ ，当且仅当 $\neg \diamond(\varphi \wedge \psi)$ 。²

基于前提集和规范集，可以构造论证。参考结构化论辩框架 [1, 9]，我们引入几个函数：令 $Prem$ 返回一个论证所使用的所有中元素的集合； $Conc$ 返回一个论证的最末结论； Sub 返回一个论证的所有子论证的集合； $Norms$ 返回一个论证中所使用的所有中的规范的集合； $TopRule$ 返回一个论证中所使用的最后一条推理规则。

基于一个论辩理论的论证可以定义如下。

定义 3.2. 设 $AT = (\mathcal{L}^D, \mathcal{N}, \mathcal{K}, \perp)$ 是一个论辩理论，基于 AT 所构造的论证是通过有限次应用以下步骤所获得的，其形式为：

1. φ ，如果 $\varphi \in \mathcal{K}$ ，并且 $Prem(A) = \{\varphi\}$ ， $Conc(A) = \varphi$ ， $Sub(A) = \{\varphi\}$ ， $Norms(A) = \emptyset$ ；或者

¹本文默认，尽管可以赋予较高的权重或优先级，但法律规范都可归为“可废止”（defeasible）的推理类型。因而不像如同“约翰是醉酒状态，那么约翰不是清醒的”（ $Drunk(John) \rightarrow \neg Sober(John)$ ）这样的推理那么确定和严格（strict）。这也正与法律推理的非单调性和可废止性特点相契合。（[18]）此外，由于本文专注于实践/规范推理情境，我们暂不考虑如“Tweety 是鸟，那么 Tweety 会飞”（ $Bird(Tweety) \Rightarrow Fly(Tweety)$ ）这样的认知方面的可废止推理，也就是说，定义中涉及的所有可废止规则都属于规范。

²其中自然也包括 φ 和 $\neg\varphi$ 这样的最基本的逻辑冲突，即 $\neg \diamond(\varphi \wedge \neg\varphi)$ 。

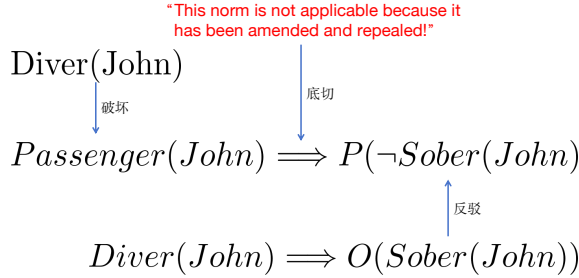


图 2: 攻击关系类型示意图

2. $A_1, \dots, A_n \Rightarrow \psi$, 如果 $A_1, \dots, A_n$ 都是论证, 并且 $\mathcal{N}$ 中存在一条规范 $\text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow \psi$, 使得 $\text{Prem}(A) = \text{Prem}(A_1) \cup \dots \cup \text{Prem}(A_n)$, $\text{Conc}(A) = \psi$, $\text{Sub}(A) = \text{Sub}(A_1) \cup \dots \cup \text{Sub}(A_n) \cup \{A\}$, $\text{Norms}(A) = \text{Norms}(A_1) \cup \dots \cup \text{Norms}(A_n) \cup \{\text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow \psi\}$.

接下来我们需要定义论证之间的冲突关系, 由此获得可以进行论证状态评价的抽象论辩框架。([2]) 关于论证之间的攻击关系, 最基本的思路之一是基于“矛盾”(contradictory)或者“反对”(contrary)关系, 定义针对一个论证前提的“破坏”(undermining), 针对一个论证推理过程的“底切”(undercutting), 以及针对一个论证结论的“反驳”(rebutting)。([1, 9, 13]) 上述几种攻击形式如图2所示。

从逻辑合理性和应用的角度上来说, 现有的关于攻击关系定义的选择是否令人满意依然存在一些争议。([19]) 本文从实践推理的特点出发, 基于定义3.1中所给出的集合, 定义一种更“广义”意义上的破坏或反驳攻击。

定义 3.3. 设 $AT = (\mathcal{L}^D, \mathcal{N}, \mathcal{K}, \perp)$ 是一个论辩理论, A 、 B 和 B' 是基于 AT 所构造的论证。 A 攻击 B , 当且仅当 A 广义地破坏或反驳 B 于其子论证 B' , 使得 A 和 B' 的结论不可能同时成立, 即:

- A 破坏 B 于 B' , 当且仅当对于形式为“ φ ”的 $B' \in \text{Sub}(B)$, $\text{Conc}(A) \wedge \varphi \in \perp$;
- A 反驳 B 于 B' , 当且仅当对于形式为“ $B'_1, \dots, B'_n \Rightarrow \varphi$ ”的 $B' \in \text{Sub}(B)$, $\text{Conc}(A) \wedge \varphi \in \perp$;
- A 底切 B 于 B' , 当且仅当对于 $B' \in \text{Sub}(B)$, $\text{TopRule}(B) = n$ 且 $n \in \mathcal{N}$, $\text{Conc}(A) = -n$.³

根据定义3.3, 但凡一个论证中所包含的两条规范所导向的行动、决策或法律判决等不可能同时采取或者被接受, 就都会导致论证之间产生攻击关系。例如“在

³ “ $-n$ ”表示“ n 不可应用”。

2024年8月8日上午10点乘飞机去甘肃”和“在2024年8月8日上午10点乘火车去甘肃”，或者“认定约翰违反交通规则”和“认定约翰未违反交通规则”，都是不可能同时被采取的行动。

当考虑论证偏好时，有些攻击可能不会成功，因而不会保留到最终所得的抽象论辩框架（简称 AF ）中。参考例3.1，法律语境下相关论证的偏好可以基于规范所依据的法律原则之间的优先关系而获得。因此，当我们认为例3.1中的原则2相较于原则1而言优先级更高时，根据定义3.3，推出“应当保护自动驾驶汽车”的论证 A 和推出“应当保护狗”的论证 B 互相冲突，但结合偏好获取的合理定义，只有后者对前者的攻击会被保留到论证评价的步骤。

在下述定义中，我们暂时假设论证之间的攻击关系都被保留到抽象论辩框架中，记为集合 $\mathcal{D}$ ， $(A, B) \in \mathcal{D}$ 代表基于一个论辩理论所构造的论证 A 攻击论证 B 。下述关于论辩语义的定义主要引自关于抽象论辩框架的经典文献 [2]，其中一组同时可接受的论证的集合被称为一个“外延”（Extension）。

定义 3.4. 设 $AF = \langle \mathcal{A}, \mathcal{D} \rangle$ 是基于论辩理论 $AT = (\mathcal{L}^D, \mathcal{N}, \mathcal{K}, \perp)$ 所得的一个抽象论辩框架，其中 $\mathcal{A}$ 是所有论证的集合， $\mathcal{D}$ 是所有论证间攻击关系的集合。对于一组论证的集合 $E \subseteq \mathcal{A}$ ，如果对于所有 $A, B \in E$ ，不存在 $(A, B) \in \mathcal{D}$ ，则称 E 无冲突；如果对于所有 $A \in E$ ， $(B, A) \in \mathcal{D}$ ，存在 $C \in E$ ， $(C, B) \in \mathcal{D}$ ，则称 E 可防御论证 A ，从而：

- E 是可相容的，当且仅当 E 无冲突，且 E 中的所有论证都可被 E 防御；
- E 是一个完全外延，当且仅当 E 是可相容的，且 $\mathcal{A}$ 中所有可以被 E 防御的论证都在 E 中；
- E 是一个优先外延，当且仅当 E 是（集合包含意义上）极大的完全外延；
- E 是一个基外延，当且仅当 E 是（集合包含意义上）最小的完全外延。

根据上述任何一种论辩语义的定义，都可以得到结论 $Passenger(John)$ 和 $P(\neg Sober(John))$ 的论证应该出现在某个语义下的同一外延集合中，而结论为 $Driver(John)$ 和 $O(Sober(John))$ 应该出现在某个语义下的同一外延集合中。

3.2 基于论辩理论的解释

形式论辩系统对于论证的计算评估，根本目的是决定哪些论证是“可辩护的”（justified）。根据论辩语义所得到的外延，可以判定论证的状态：一个论证被某外延接受，当且仅当它是外延中的元素；一个论证对于某外延而言被拒绝，当且仅当它受到该外延中的一个（对于该外延而言可接受的）论证的攻击。进而，根据一个论证对该语义下所有的外延而言都可接受，还是对该语义下的至少一个

外延而言可接受, 可以判断一个论证是“怀疑”(skeptical)的可辩护, 还是“轻信”(credulous)的可辩护。

2019年4月欧盟出台的《人工智能道德准则》(*Ethics guidelines for trustworthy AI*)中提出: AI发展方向应该是“可信赖的”, 包含安全、隐私和透明、可解释等方面。而文献[15]通过实证调研表明, “可辩护性”是在几种解释策略中对人类用户而言最高效的。

受此启发, 我们给出下述基于论辩系统的解释定义。

定义 3.5 (基于论辩的解释). 设 A 是在特定论辩语义 $\mathcal{S}$ 下一个怀疑或轻信可接受的论证, $\text{Conc}(A) = \varphi$, 那么

- 关于 φ 这个断言是如何得到的一个解释是 $\text{Prem}(A) \cup \text{Norms}(A)$;
- 关于 φ 这个断言为何可辩护的一个解释是 $\text{Prem}(B) \cup \text{Norms}(B)$, 其中对于所有 $C \in \mathcal{A}$, 如果 $(C, A) \in \mathcal{D}$, 那么 $(B, C) \in \mathcal{D}$.

这一解释旨在为决策提供清晰的推理链, 并在存在与此决策冲突的其他决策的情况下, 解释为何与之冲突的决策不可辩护。

根据定义3.5, 考虑例2, 假设加入一条规范“自动驾驶汽车上唯一的乘客必须承担作为司机的责任”, 那么根据定义3.3, 由此规范构造的论证结论与“将约翰视为乘客”不能同时成立, 论证之间产生攻击, 即使假设论证偏好相同, 该论证也通过与结论为 $\text{Passenger}(\text{John})$ 和 $P(\neg \text{Sober}(\text{John}))$ 的论证互相攻击, 为结论为 $\text{Driver}(\text{John})$ 和 $O(\text{Sober}(\text{John}))$ 的论证提供了辩护。因此, 根据上述定义, 关于 $O(\text{Sober}(\text{John}))$ 如何得到的解释包括 $\{\text{Driver}(\text{John})\}$ 和 $\{\text{Driver}(\text{John}) \implies O(\text{Sober}(\text{John}))\}$; 而关于 $O(\text{Sober}(\text{John}))$ 为何可辩护的解释可能包括作为事实的 $\{\text{OnlyPassenger}(\text{John})\}$ 和 $\{\text{OnlyPassenger}(\text{John}) \implies O(\text{Driver}(\text{John}))\}$ 。

在考虑论证偏好时, 我们可以基于价值或法律原则, 得到相应的规范之间的优先关系, 从而获得论证之间的偏好, 这时根据需要, 相应的优先排序也可以在解释中进行演示。

4 结语

本文旨在基于结构化论辩理论, 引入道义逻辑在语言层面上更好地描述权利和义务等问题, 并使用论辩系统作为推理和解释的工具。本文的另一个重点是关注到了基于形式论辩的人工智能可解释问题, 并给出了一种可能的解释方案。由于人工智能理论和技术的飞速发展和广泛的影响力, 确保系统的可解释从伦理、法律、实践和技术层面来说都至关重要。可解释人工智能有助于相关的前沿技术和人类用户之间的沟通, 保证这些工具能够被有效应用且值得信赖。然而, 当前

基于深度学习等方法的人工智能算法和模型常被描述为“黑箱”，难以向人类用户诠释其决策过程和方法，从而面临着解释困难的挑战。有鉴于此，开发能够增进系统透明度的人工智能理论和技术手段是当前研究的关键问题之一。本文主张引入道义算子等元素，使论辩理论能够在知识表示上比使用更常见的命题逻辑或一阶逻辑更好地捕捉法律上什么是必要的、可能的或不可能的，以及处理义务和许可的概念。因此，在法律情境下，不仅可以使针对规范推理的论辩理论更好地刻画相关概念和规范之间的关系，更能够在提升表达能力的同时让论辩系统可以在自身具备的易解释性优势的基础上，给出更精准的描述和解释。

作为一项尝试性工作，本文主要给出了一种结合道义逻辑、基于论辩系统处理不一致规范推理方法的基本思路，尚有许多细节有待探究和完善。

首先，在论辩系统的合理性要求方面〔9〕，本文所主要参考的结构化论辩框架 *ASPIC+* 已被证明能够满足“直接一致性”“间接一致性”等基本的理性假设要求。〔1〕而在规范推理的应用语境中，例如法律规范语境下，整体的自洽性是系统的核心要求之一。法律的一个基本目标是，在确定的情境下，公民知道自己应当如何行动并会引发什么明确的法律后果。如果法律系统本身包含不可解释或不可调和的不一致，会造成行为指导的混乱与法律责任的难以划分，从而丧失最基本的社会功能。这种法律系统的自洽性是诸多专家经过长时间的调整以及在大量的实际案例中给出不同的法律解释得到的，因此具备了相当程度的广度与深度。有鉴于此，本文也以维护这种法律自身蕴含的自洽性为优先。基于上述考虑，一方面，规范推理的应用情境通常不会违背基本的理性，另一方面，我们不希望使得一个形式推理系统的合理性优先于应用情境本身，即让规范去适应推理系统的合理性要求。因此，我们认为这部分工作的主要落脚点应该在于定义上的细化，以及如何保证论辩理论中的矛盾集合“ $\perp$ ”的合理获得，由于篇幅有限，我们会在未来工作中继续完善这部分内容，并探索其他的系统特性。

除了系统性质方面的研究之外，未来可以考虑的进一步工作还包括关于规范之间的优先关系如何获得和论证偏好方面的讨论，以及如何结合具体的应用情境，给出人们所需要且易理解的解释等。将道义逻辑与其他人工智能技术相结合，可能为增强决策系统提供更多的途径，特别是在需要严格遵守规则和规范的领域，如自动驾驶汽车、人工智能伦理等方面。例如，通过深度学习算法训练自动驾驶汽车以识别和应对各种交通场景，而基于道义逻辑的推理系统可以确保车辆的决策遵守交通法规和道德准则。将道义逻辑纳入深度学习模型，有助于构建能够自动理解和应用道德规范的人工智能系统，同时，在形式论辩系统的易解释性优势基础之上，结合道义逻辑可以进一步提供规范性解释。因此，未来我们将继续探索如何通过解决这些挑战，使结合道义逻辑的论辩系统能和深度学习等技术协同作用，为各种应用中更可靠、更合乎道德、更易于理解的人工智能系统铺平道路。

参考文献

- [1] M. Caminada and L. Amgoud, 2007, “On the evaluation of argumentation formalisms”, *Artificial Intelligence*, **171(5)**: 286–310.
- [2] P. M. Dung, 1995, “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games”, *Artificial Intelligence*, **77(2)**: 321–357.
- [3] A. J. García and G. R. Simari, 2014, “Defeasible logic programming: delp-servers, contextual queries, and explanations for answers”, *Argument & Computation*, **5**: 63–88.
- [4] L. Goble, 2013, “*Prima Facie Norms, Normative Conflicts, and Dilemmas*”, *Handbook of Deontic Logic and Normative Systems*, 241–352, London: College Publications.
- [5] J. F. Horty, 1993, “Deontic logic as founded on nonmonotonic logic”, *Annals of Mathematics and Artificial Intelligence*, **9**: 69–91.
- [6] B. Liao, N. Oren, L. van der Torre and S. Villata, 2019, “Prioritized norms in formal argumentation”, *Journal of Logic and Computation*, **29(2)**: 215–240.
- [7] Y. Lu, Z. Yu, Y. Lin, B. Schafer, A. Ireland and L. Urquhart, 2022, “A legal support system based on legal interpretation schemes for ai vehicle designing”, *Proceedings of the 35th International Conference on Legal Knowledge and Information Systems (JURIX 2022)*, pp. 213–218.
- [8] Y. Lu, Z. Yu, Y. Lin, B. Schafer, A. Ireland and L. Urquhart, 2022, “Handling inconsistent and uncertain legal reasoning for ai vehicles design”, *Proceedings of Workshop on Methodologies for Translating Legal Norms into Formal Representations (LN2FR 2022)*, pp. 76–89.
- [9] S. Modgil and H. Prakken, 2013, “A general account of argumentation with preferences”, *Artificial Intelligence*, **195**: 361–397.
- [10] H. Prakken, 2010, “An abstract framework for argumentation with structured arguments”, *Argument & Computation*, **1(2)**: 93–124.
- [11] M. Rosenfeld, 2005, “Dworkin and the one law principle: a pluralist critique”, *Revue Internationale de Philosophie*, **59(233 (3))**: 363–392.
- [12] L. Royakkers and F. Dignum, 1997, “*Defeasible Reasoning with Legal Rules*”, in D. Nute(ed.), *Defeasible Deontic Logic*, 263–286, Dordrecht: Springer Netherlands.
- [13] F. Toni, 2014, “A tutorial on assumption-based argumentation”, *Argument & Computation*, **5(1)**: 89–117.
- [14] L. van der Torre and S. Villata, 2014, “An ASPIC-based legal argumentation framework for deontic reasoning”, *Computational Models of Argument*, **Vol. 266**, pp. 421–432.
- [15] L. R. Ye and P. E. Johnson, 1995, “The impact of explanation facilities on user acceptance of expert systems advice”, *MIS Quarterly*, **19(2)**: 157–172.
- [16] 孔红, “规范冲突问题及其非单调逻辑研究”, 湖北大学学报(哲学社会科学版), 2016年第9期, 第33–39页。
- [17] 廖备水, “论新一代人工智能与逻辑学的交叉研究”, 中国社会科学, 2022年第3期, 第37–54+204–205页。

- [18] 熊明辉, “法律人工智能的推理建模路径”, 求是学刊, 2020 年第 06 期, 第 89–100 页。
- [19] 余喆, 徐康, 廖备水, “结构化论辩系统——限制性反驳和非限制性反驳”, 逻辑学研究, 2018 年第 3 期, 第 3–14 页。
- [20] 张莉敏, “国外非单调道义逻辑研究探析”, 学术论坛, 2008 年第 9 期, 第 16–19 页。

(责任编辑: 袁之)

An Argumentation System for Legal Reasoning Combined with Deontic Logic

Zhe Yu Yiwei Lu Burkhard Schafer

Abstract

Formal argumentation is a non-monotonic reasoning approach that enables reasoning in contexts where knowledge is uncertain and inconsistent. Typically, the representation of underlying knowledge can flexibly adopt various logical forms within a structured argumentation framework as needed. This paper primarily focuses on applications of formal argumentation systems in AI and Law. In the domain of normative reasoning—where legal contexts serve as a prime example—conflicts between legal norms frequently arise. Existing legal reasoning systems, such as legal ontologies, legal knowledge graphs, and legal expert systems, typically rely on classical logical languages to implement intelligent legal reasoning. However, this reliance prevents these systems from effectively capturing the uncertainty, inconsistency, and dialectical nature of legal reasoning, thereby limiting their practical applicability in legal tasks. To address this limitation, this paper integrates deontic logic with structured argumentation theory to model normative reasoning within legal semantics. This integration enhances the foundational principles of legal reasoning by introducing novel reasoning functionalities and expressive capabilities, laying the groundwork for more advanced legal applications in the future. Furthermore, considering the growing emphasis on explainable AI, we provide a formal definition of explanation within the proposed argumentation system.

Zhe Yu	Institute of Logic and Cognition, Department of Philosophy, Sun Yat-sen University yuzh28@mail.sysu.edu.cn
Yiwei Lu	School of Law, University of Edinburgh Y.Lu-104@sms.ed.ac.uk
Burkhard Schafer	School of Law, University of Edinburgh B.Schafer@ed.ac.uk