doi:10.13582/j.cnki.1672-7835.2020.03.005

基于偏好聚合和论辩推理的道德困境解决方法

廖备水,李崇慧
(浙江大学 哲学系,浙江 杭州 310000)

摘　要:在关于智能体行动决策的道德困境中,不同的利益相关方不但秉持不同的价值并且关于这些价值的个体偏好也不尽相同。为获得道德困境中符合多数利益相关方偏好的求解,以现有工作为基础,提出一种基于形式论辩的道德委员会论辩框架。通过社会选择理论中的两种方法,把不同层次的个体偏好聚合为统一的社会偏好,从而在论辩框架中实现论证优先级的提升。最后,运用论辩语义推理解决道德困境中的两难,获得合理的求解。

关键词:论辩推理;偏好聚合;社会选择理论,规范系统;道德困境
中图分类号:B812.4　　**文献标志码**:A　　**文章编号**:1672-7835(2020)03-0033-17

在新一代人工智能背景下,我们的生活中涌现了大量的智能体。它们承担着诸如病人看护、儿童陪伴和智能管家等职能。人们希望智能体的行为和决策能够符合人类的伦理价值和道德规范。因此,智能体的道德观与智能体的利益相关方息息相关。所谓利益相关方,主要包括以下四类①:一是使用该智能体的直接相关方,如:使用者和维护者;二是该智能体的设计者和制造商;三是政府和社会团体;四是有可能受到该智能体影响的其他主体,如:无人驾驶车道路上的行人。对智能体基于特定场景的行为决策进行规范设计时,由于各个利益相关方的立场和出发点不同,往往存在着较大的分歧。具体可能表现为两个方面:在利益相关方秉持的价值理念之间有冲突,或者是关于价值理念的个体偏好存在差异。在"电车难题"中,是让脱轨电车驶向一个教授还是五个平民?依据不同的价值理念或个体偏好,会得到不同的结论。电车面临的道德困境如何解决,不能简单地以支持结论的利益相关方人数的"多数"原则来解决,而是必须深入规范和价值的内在结构,需要同时考虑以上两个方面。

为了解决道德困境中智能体行动决策的冲突,在我们的前期工作中,建立了一个基于形式论辩理论的道德委员会②。在给定一个价值排序的前提下,基于与各利益相关方价值理念相关的规范集合,建立一个形式论辩系统来处理规范之间的冲突,获得基于价值数量最大化和基于利益相关方个体数最大化的求解。

该项工作假定所有的利益相关方都共享同一个价值排序(且为全序)。然而,这个假定在现实场景中往往并不成立。因为不同的利益相关方不但对于一组价值可能有完全不同的排序,而且,在某些特定场合,正是不同的价值排序导致了智能体行动的规范冲突。我们在 https://imdb.uib.no/dilemmaz/articles/all 上找到"顽固的ICD"案例:

Jane体内装有一个心脏震颤器(ICD),这个智能体的功能是只要探测

收稿日期:2019-12-16
基金项目:本项目受浙江大学"双脑计划"人文社科专项以及国家社会科学基金重大项目(18ZDA290, 17ZDA026)的资助
作者简介:廖备水(1970—),男,福建省古田县人,浙江大学哲学系教授,人文学院、计算机学院博士生导师,长期从事逻辑、认知和人工智能的文理交叉研究;李崇慧(1981—),女,江苏省无锡市人,博士研究生,主要从事形式论辩与人工智能逻辑研究。

①Kahn P. H. Friedman B. "Human values, ethics, and design" *L. Erlbaum Associates Inc.*, 2003:1209-1233.

②Beishui Liao and Marija Slavkovik and Leendert van der Torre. "Building Jiminy Cricket: An architecture for moral agreements among stakeholders", *Proceedings of the* 2019 *AAAI/ACM Conference on AI, Ethics, and Society*, 2019:147-153.

到有威胁生命的心律异常就会震颤病人的心脏使之恢复正常。自从 10 年前安装 ICD 后,ICD 已经两次从突发的心脏疾病中抢救了 Jane。不过最近 Jane 被查出罹患了晚期胰腺癌。在连续几个月的抗癌治疗未有明显疗效之后,Jane 又一次面临死亡的威胁。这时,Jane 向外科医生提出要终止 ICD 在体内的工作机制,体面地面对死亡。据知情者透露,ICD 震颤心脏时,犹如胸前被马蹄狠狠践踏。Jane 的外科医生拒绝了她的要求,他的理由是终止 ICD 相当于协助病人自杀。

这个案例中的智能体 ICD 所拥有的伦理设定是属于操作层面的,也就是智能体是按照预设的指令行事。因此在后续发生道德困境时,需要利益相关方共同协商确定行动决策。我们可以看到,随着客观情况发生变化,使用者对于智能体的自主权上升到了她价值排序的首位。而她的外科医生必定不会以病人对智能体的自主权为首要偏好。因此,引发了她和外科医生之间关于智能体 ICD 行动决策上的矛盾。

可见,利益相关方关于价值排序(即个体偏好)的差异似乎是造成道德困境的原因,但是否只要在利益相关方中形成关于价值的统一排序(即社会偏好),就可以有效地解决道德困境呢?当智能体拥有多个利益相关方,且关于它的行动决策存在多组相互冲突的规范推理的结论时,根据社会偏好关于价值的排序,是否意味着要接受排在首位的价值对应的规范,驳回排在末位的价值对应的规范?经过仔细分析,答案是否定的。究其原因,是我们忽略了价值和偏好与利益相关方关于智能体行动的规范推理之间的内在联系。因此,在这篇论文中,我们把社会选择理论中的偏好聚合和形式论辩理论中的抽象论辩推理结合起来开展研究。

本文研究的重点是各利益相关方关于价值的个体偏好如何影响道德困境中智能体的行动决策。具体来说,主要研究问题如下:

如何运用关于价值的个体偏好来形成一个统一的社会偏好?

基于统一的社会偏好,利益相关方之间的冲突是如何进行平衡的,从而得到关于智能体行动最大程度的统一意见?

本文拟在现有工作的基础上,以“顽固的 ICD”为例,采用形式论辩理论为道德委员会进行建模。在对个体偏好进行聚合时,本文区分了关于论证的个体偏好和关于规范的个体偏好两个层面的个体偏好。我们采用社会选择理论中关于偏好聚合的方法对个体偏好进行聚合,并在抽象论辩框架中通过论辩推理获得利益相关方关于智能体行动最大程度的统一意见。

本文的章节安排如下:第一节介绍利益相关方的价值规范系统和道德委员会论辩框架;第二节介绍社会选择理论中关于偏好聚合的两种方法,将关于论证的个体偏好和关于规范的个体偏好聚合后与形式论辩系统的推理相结合,得出符合多数利益相关方偏好的求解;第三节是总结,介绍相关工作和未来研究方向。

一　价值规范系统和道德委员会

(一)利益相关方的价值规范系统

在定义道德委员会对个体偏好进行处理之前,我们要先介绍价值规范系统(简称 VNS)。规范可以理解为被主体所在群体或环境期待的行为、行动或产出。规范系统是对同一群体中主体行动或行为进行指导和评估的系统①。依据麦金森和范德托的输入/输出逻辑②,规范被抽象为由输入和输出的命题组成的序对。因此,在利益相关方的规范系统中,规范可以被视作由理由为输入和强制性结论为输出组成的序对。每个利益相关方对应于不同的规范系统。进一步地,价值规范系统是在规范系统中加入与主体有隶属关系且与规范相对应的价值集合。形式化地,有如下定义:

定义 1　关于利益相关方 i 的价值规范系统 VNS_i 是一个元组$(L, V_i, \Omega_i, val_i)$,其中:

· L 是一种逻辑语言,如:封闭于经典否定符号¬ 的命题文字;

· V_i是利益相关方 i 持有的价值集合;

· $\Omega_i \subseteq L \times L$,是利益相关方 i 规范的集合。

①Carlos E. Alchourron. “Conflicts of norms and the revision of normative systems”. *Law and Philosophy*, 1991, 10(4):413-425.

②David Makinson and Leendert Van Der Torre. “What is input/output logic?” *Trends in Logic*, 2000(29):383-408.

规范是一个序对,如:$u = (a, b)$,其中 $u \in \Omega_i$,a 是作为输入项的理由,b 是作为输出项的强制性结论;

· val_i 是一个价值分配函数,$val_i: \Omega_i \rightarrow 2^{V_i}$,表示在利益相关方 i 的每个规范上进行的价值分配;

通过对 ICD 案例中相关利益方进行分析,发现除了病人和医生以外,智能体 ICD 的设计者也是相关利益方,因为他预设的指令决定了触发或终止 ICD 行动的条件。因此,他的价值理念和关于价值的排序也应在关注范围之内。所以,我们有三个相关利益方。在 ICD 案例中,我们令设计者的价值规范系统为 VNS_1,包含一个规范和一个价值理念:专制;Jane 的价值规范系统为 VNS_2,包含两个规范和两个价值理念:尊严和自主权;外科医生的价值规范系统为 VNS_3,包含两个规范和两个价值理念:责任和德行。如图 1 所示,在由三个相关利益方组成的价值规范体系①中,垂直方向上每个箭头连接的是作为理由和强制性结论的两个命题,即以输入和输出形式表达的一个规范。每个箭头上对应的是隶属于利益相关方的、支持该规范的价值。不同利益相关方的规范系统得出的结论相互矛盾,因此,水平粗线箭头表示的是结论之间的冲突关系。根据定义 1,我们将 ICD 案例中所有利益相关方的价值规范系统表示在图 1 中。

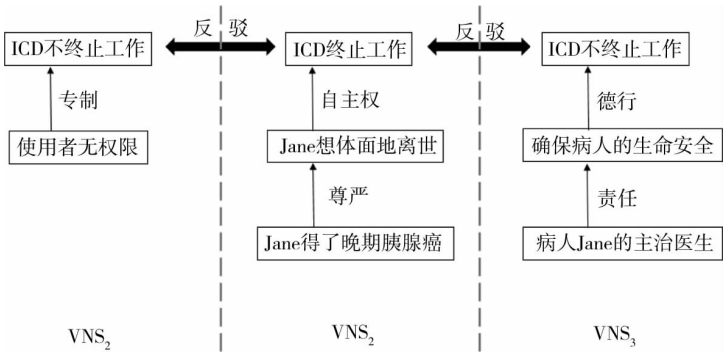


图 1　ICD 案例利益相关方的价值规范体系

例 1　令下标 1:设计者,2:Jane,3:外科医生;价值规范体系中的命题分别为:“使用者无权限” = *NoAccess*,“Jane 得了晚期胰腺癌” = *HasCancer*,“Jane 想体面地离世” = *DecentDie*,“病人 Jane 的主治医生” = *isJane'sDoctor*,“确保病人的生命安全” = *SecureLife*,“ICD 终止工作” = *DeactivateICD*,“ICD 不终止工作” = ¬ *DeactivateICD*,则:

VNS_1: $val_1(u_a)$ = 专制, u_a = (*NoAccess*, ¬ *DeactivateICD*), $\Omega_1 = \{u_a\}$;

VNS_2: $val_2(u_b)$ = 尊严, u_b = (*HasCancer*, *DecentDie*), $val_2(u_c)$ = 自主权, u_c = (*DecentDie*, *DeactivateICD*), $\Omega_2 = \{u_b, u_c\}$;

VNS_3: $val_3(u_d)$ = 责任, u_d = (*isJane'sDoctor*, *SecureLife*), $val_3(u_e)$ = 德行, u_e = (*SecureLife*, ¬ *DeactivateICD*), $\Omega_3 = \{u_d, u_e\}$。

在 ICD 案例的价值规范体系中,我们将重点讨论一个信息,就是各利益相关方关于专制、自主权、尊严、德行、责任这五个价值的个体偏好(严格排序)。理论上,对于利益相关方的个体偏好来说,关于五个价值的任何排序都是可能的,但根据“顽固的 ICD”中的描述,我们将三个利益相关方的个体偏好表示如下。令个体偏好 $\succ$ 的下标为:1:设计者,2:Jane,3:外科医生;我们用 v 加下标来标记不同的价值,v_p = 专制,v_a = 自主权,v_d = 尊严,v_m = 德行,v_r = 责任,则个体偏好的排序(从强到弱)依次为:

· $\succ_1$: v_p, v_r, v_d, v_m, v_a;

· $\succ_2$: v_d, v_a, v_r, v_m, v_p;

· $\succ_3$: v_m, v_d, v_p, v_r, v_a。

以上我们讨论了利益相关方的价值规范系统,分析了价值、偏好与利益相关方的规范推理及其规范系统输出结论之间的内在联系。要想解决智能体行动的道德困境,首先我们要建立一个保持中立的道德委员会,它是由所有利益相关方的

①如上文所述,规范系统和规范体系并不是同一概念,规范系统是针对同一群体中的主体而言,规范体系则是不同群体中主体规范系统基于某种联系的组合。

价值规范系统组成的价值规范体系。我们以现有工作[①]为基础,基于形式论辩理论对道德委员会进行建模。

(二)形式论辩对道德委员会的建模

形式论辩可以用于建模不一致情境中的推理。不一致情境主要表现在信息冲突,而信息冲突的原因,或是由于知识的不完全性和不确定性,或是由于推理主体的动机、偏好或观点等的不一致。在结构化论辩系统中,论证是具有内在逻辑结构且可以用底层逻辑语言来刻画的。目前,主要有四种结构化形式论辩系统,分别是 ASPIC+[②], DeLP[③], ABA[④] 和 CLA[⑤]。在一个论辩系统中,由论证及论证之间攻击关系构成的二元组称为一个抽象论辩框架[⑥]。在该框架中,各个论证的状态取决于与之有关联(攻击关系)的其他论证的状态。

在上一小节中,我们已经建立了利益相关方的价值规范体系(VNS)。要用形式论辩理论对道德委员会建模,我们要先用 VNS 中的有关概念来定义论证和论证之间的攻击关系。需要说明的是:上文已给出规范系统的定义是对同一群体中主体行为指导和评估的系统。而道德困境中的利益相关方是来自立场、认知和价值观存在差异的不同群体,并且各利益相关方的规范推理均以事实性命题为初始前提,因此,基于规范的论证构建具有主体独立性,封闭于利益相关方自身的规范系统。

定义 2　给定一个价值规范系统 $\mathrm{VNS}_i=(L, V_i, \Omega_i, val_i)$,根据现有基于规范的论证定义[⑦],我们称 A 为一个论证,当且仅当 A 是一个规范的序列:$[(a_0, a_1), (a_1, a_2), \cdots, (a_{m-1}, a_m)]$,其中 $(a_0, a_1), (a_1, a_2), \cdots, (a_{m-1}, a_m) \in \Omega_i$。我们用 $val_i(A)=val_i((a_0, a_1)) \cup val_i((a_1, a_2)) \cup \cdots \cup val_i((a_{m-1}, a_m))$ 表示指派给论证 A 的价值集合。由 VNS_i 产生的论证集合记为 $\mathcal{A}_i$。

如果与 VNS_i 中的规范集合相对应,即 $u_1=(a_0, a_1)$, $u_2=(a_1, a_2)$, $\cdots$, $u_m=(a_{m-1}, a_m)$,则论证 A 也可以写为 $[u_1, u_2, \cdots, u_m]$,其中 $u_1, u_2, \cdots, u_m \in \Omega_i$。论证 A 的子论证为:$[u_1, \cdots, u_j]$,其中 $1 \leqslant j \leqslant m$。论证 A 的子论证集合记为 $sub(A)$。我们用 $Prem(A)$ 表示论证 A 的前提,用 $Concl(A)$ 表示论证 A 的结论。

定义 3　如果论证 $A \in \mathcal{A}_i$ 是 VNS_i 中一个规范的序列 $[(a_0, a_1), (a_1, a_2), \cdots, (a_{m-1}, a_m)]$,其中 $(a_0, a_1), (a_1, a_2), \cdots, (a_{m-1}, a_m) \in \Omega_i$,则:

$Prem(A) = a_0$;

$Concl(A) = a_m$。

由于在伦理困境中,每个利益相关方都是以一个事实性命题为初始前提,建立规范序列,推理出一个结论。因此,由规范序列构建的论证,它的前提就是这个事实性命题,结论则是规范序列中最后一个规范的输出项。

定义 4　根据莫吉尔和帕肯关于论证攻击关系的定义[⑧],我们将论证 A 和 B 之间攻击关系的集合 $\mathcal{R}$ 定义如下:

· 论证 A 和论证 B 之间是反驳关系,当且仅当存在 $B' \in sub(B)$:$Concl(A) = \neg\, Concl(B')$;

· 论证 A 破坏论证 B 的前提,当且仅当:$Conc(A) = \neg\, Prem(B)$;

· 论证 A 底切论证 B,当且仅当:论证 A 攻

①Beishui Liao and Marija Slavkovik and Leendert van der Torre. "Building Jiminy Cricket: An architecture for moral agreements among stakeholders", *Proceedings of the* 2019 *AAAI/ACM Conference on AI, Ethics, and Society*, 2019:147-153.

②Sanjay Modgil and Henry Prakken. "The aspic+ framework for structured argumentation: a tutorial". *Argument & Computation*, 2014, 5(1):31-62.

③Alejandro J. García and Guillermo R. Simari. "Defeasible logic programming: Delp-servers, contextual queries, and explanations for answers". *Argument & Computation*, 2014, 5(1):63-88.

④Francesca Toni. "A tutorial on assumption-based argumentation". *Argument & Computation*, 2014, 5(1):89-117.

⑤Philippe Besnard and Anthony Hunter. "Constructing argument graphs with deductive arguments: a tutorial". *Argument & Computation*, 2014, 5(1):5-30.

⑥Phan Minh Dung. "On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games". *Artificial Intelligence*, 1995, 77(2):321-357.

⑦Beishui Liao, Nir Oren, Leendert Van Der Torre, et al. "Prioritized norms in formal argumentation". *Journal of Logic and Computation*, 2018, 29(2): 215-240.

⑧Sanjay Modgil and Henry Prakken. "The aspic+ framework for structured argumentation: a tutorial". *Argument& Computation*, 5(1):31-62, 2014.

击 B 从前提到结论的推理关系。

在以上定义的三种论证间攻击关系中,底切不依赖于论证的优先级。破坏前提的攻击关系虽然依赖优先级,但仅仅针对被攻击的前提是假设性前提的情况。而构建在规范序列之上的论证,其前提都是事实性命题,不具有被攻击的属性。因此,本文主要讨论论证攻击关系中具有对称性的反驳关系。

基于以上定义,根据现有工作中关于道德委员会的定义①,我们建立包含利益相关方个体偏好档案的道德委员会论辩框架(简称 AFMC),形式定义如下:

定义 5　给定价值规范系统 $\mathrm{VNS}_i=(L, V_i, \Omega_i, val_i)$,$i=1,2,\cdots,n$,道德委员会的论辩框架是一个元组 $\mathrm{AFMC}=(\mathcal{A}, \mathcal{R}, N, V, P)$,其中:

· $\mathcal{A}=\mathcal{A}_1\cup\cdots\cup\mathcal{A}_n$,其中 $\mathcal{A}_1,\cdots,\mathcal{A}_n$ 是根据定义 2 分别产生于 $VNS_1,\cdots,VNS_n$ 的论证集合;

· $\mathcal{R}\subseteq\mathcal{A}\times\mathcal{A}$,是论证之间的攻击关系的集合,如定义 4 中所定义;

· $N=\{1, 2, \cdots, n\}$ 是有限且非空的利益相关方的集合;

· $V=V_1\cup\cdots\cup V_n$ 是所有利益相关方的价值的集合;

· $P=<\succ_1,\succ_2,\cdots,\succ_n>$ 是利益相关方关于价值的个体偏好档案,其中 $\succ_i$ 是利益相关方 $i\in N$ 关于价值集合 V 的严格全序。

值得注意的是,为了确保道德委员会论辩框架中攻击关系的一致性,也为了说明道德困境中的冲突往往存在于不同利益相关方的 VNS 之间,我们合理地假设:任何利益相关方 i 的论证集合 $\mathcal{A}_i$ 是无冲突的,即不存在 $A,B\in\mathcal{A}_i$ 使得 $(A,B)\in\mathcal{R}$ 或 $(B,A)\in\mathcal{R}$。

例 2　根据定义 5 和例 1 中已建立的 VNS_1,VNS_2,VNS_3,ICD 案例的 AFMC 可以表示为 $\mathrm{AFMC}_{ICD}=(\mathcal{A}, \mathcal{R}, N, V, P)$,其中:$\mathcal{A}=\{A, B, C\}$,$A$, B, C 是分别来自于 VNS_1,VNS_2 和 VNS_3 的论证;由于 $Concl(A)=\neg Concl(B)$,$Concl(C)=\neg Concl(B)$,所以 $\mathcal{R}=\{(A, B), (B, A), (B, C), (C, B)\}$;$N=\{1$:设计者,2:Jane,3:外科医生$\}$;$V=\{v_p, v_a, v_d, v_m, v_r\}$;$P=<\succ_1,\succ_2,\succ_3>$。ICD 案例的 AFMC 如图 2 所示。

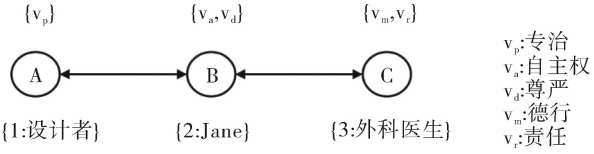


图 2　ICD 案例的论辩框架 AFMC

由于论证优先级的存在,框架 AFMC 中对称性的攻击关系将可能被部分地撤销,从而析出一个新的框架,我们把这个析出的新框架称为 AFMC 的一个解析,记作 $AF=(\mathcal{A}, \mathcal{D})$,其中,$\mathcal{A}$ 是框架 AFMC 中的论证集合,$\mathcal{D}$ 是论证之间的击败关系。击败关系是将偏好提升为论证的优先级后攻击关系的集合,也就是说,如果存在论证 A,$A'\in\mathcal{A}$ 使得 $(A, A')\in\mathcal{R}$,$(A', A)\in\mathcal{D}$,则若 $(A, A')\in\mathcal{D}$,则论证 A' 的优先级不能高于论证 A 的优先级。参考莫吉尔在文献②和巴罗尼在文献③中关于框架解析的定义,我们定义如下:

定义 6　令 $\Delta=\mathcal{A}\setminus\mathcal{D}$,我们说框架 $AF=(\mathcal{A}, \mathcal{D})$ 是框架 $\mathrm{AFMC}=(\mathcal{A}, \mathcal{R}, N, V, P)$ 的一个解析,当且仅当 $\mathcal{D}$ 中任何一对对称性攻击 (A, A') 和 (A', A) 中,最多只有一个出现在 Δ 中。

由于可能存在论证优先级无差异的情况,因此当 $\Delta=\emptyset$ 时,框架的解析就是它本身。当 $\Delta\neq\emptyset$ 时,我们对其中的一种特殊情况进行定义如下。

定义 7　我们说框架 $AF=(\mathcal{A}, \mathcal{D})$ 是框架 $\mathrm{AFMC}=(\mathcal{A}, \mathcal{R}, N, V, P)$ 的一个完全解析,当且仅当对于所有的 A, $A'\in\mathcal{A}$,满足以下条件:

1. 如果 $(A, A')\in\mathcal{R}$ 且 $(A', A)\in\mathcal{R}$,那么或者 $(A, A')\in\mathcal{D}$,或者 $(A', A)\in\mathcal{D}$,但不存在 $(A, A')\in\mathcal{D}$ 且 $(A', A)\in\mathcal{D}$;

2. 如果 $(A, A')\in\mathcal{D}$,那么 $(A, A')\in\mathcal{R}$。

我们把依据特定标准,从一个抽象论辩框架

①Beishui Liao and Marija Slavkovik and Leendert van der Torre. "Building Jiminy Cricket: An architecture for moral agreements among stakeholders", *Proceedings of the* 2019 *AAAI/ACM Conference on AI, Ethics, and Society*, 2019:147-153.

②Sanjay Modgil. "Hierarchical argumentation". *European Workshop on Logics in Artificial Intelligence*. Springer, Berlin, Heidelberg, 2006: 319-332.

③Baroni Pietro, Paul E. Dunne, and Massimiliano Giacomin. "On the resolution-based family of abstract argumentation semantics and its grounded instance." *Artificial Intelligence*, 2011, 175.3-4: 791-813.

到一组或几组可接受论证集合的映射关系称为论辩语义①。可接受论证的集合$\mathcal{E} \subseteq A$ 称为外延。对于同一个抽象论辩框架,不同的语义定义可以产出不同的外延。如果我们把论辩语义定义为一个函数σ,那么相当于抽象论辩框架被该函数映射到一个外延的集合。依据抽象论辩的术语,所有基于"可相容"论证集合的外延定义都取决于两个重要的概念:无冲突和可防御。给定一个抽象论辩框架$(\mathcal{A}, \mathcal{D})$,我们说$\mathcal{E} \subseteq \mathcal{A}$是无冲突的,当且仅当不存在$A, B \in \mathcal{E}$使得$(A, B) \in \mathcal{D}$。一个论证$A \in \mathcal{A}$被$\mathcal{E}$可防御,当且仅当对于任意$B \in \mathcal{A}$,如果$(B, A) \in \mathcal{D}$则存在$C \in \mathcal{E}$使得$(C, B) \in \mathcal{D}$。$\mathcal{E}$被称为可相容集合,当且仅当$\mathcal{E}$是无冲突的且它可以防御集合内的每一个论证;$\mathcal{E}$被称为完全外延,当且仅当$\mathcal{E}$是可相容的,且$\mathcal{A}$中所有可以被$\mathcal{E}$防御的论证都在$\mathcal{E}$中;$\mathcal{E}$被称为优先外延,当且仅当$\mathcal{E}$是集合包含意义上最大的完全外延;在优先外延的基础上,如果要求一个外延包含所有不受该外延攻击的论证,那么该外延被称为稳定外延;$\mathcal{E}$被称为基外延,当且仅当$\mathcal{E}$是集合包含意义上最小的完全外延。在以上外延中,基语义是产出唯一外延的论辩语义,完全语义、优先语义和稳定语义则有可能产出多个外延。

用论辩语义对框架 AFMC 进行求解,需要我们把框架 AFMC 转换成抽象论辩框架,因此,还需要进一步明确以下两个问题:

1.如何将个体偏好聚合为社会偏好,建立基于社会偏好的道德委员会论辩框架?

2.如何体现道德委员会中立的立场,也就是用论辩推理得到符合多数利益相关方价值偏好的论证集合?

对于第一个问题,如何建立基于社会偏好的道德委员会论辩框架,本文结合社会选择的相关理论,采用偏好聚合规则对个体偏好进行聚合。作为聚合对象的个体偏好,我们拟分两个层面进行认定:一个是关于论证的个体偏好,也就是基于论证上价值集合的分配,将个体关于单个价值的个体偏好提升为关于论证的个体偏好;另一个是关于规范的个体偏好,也就是基于规范上价值集合的分配,将个体关于单个价值的个体偏好提升为关于规范的个体偏好。对于第二个问题,基于以上方法得到的两个层面的社会偏好,拟采用不同的论辩推理方式进行求解,我们将在下一节进行详细的论述。

二　基于偏好聚合和论辩推理的解决方法

根据上一节中讨论的思路,在本节中我们先简要介绍社会选择理论和两种偏好聚合函数,然后基于论辩框架 AFMC 分别对关于论证的个体偏好和关于规范的个体偏好进行聚合和推理,获得符合多数利益相关方价值偏好的求解。

(一)社会选择理论与偏好聚合

社会选择理论是关于集体决策过程和程序的理论,集体决策的对象可以是投票、偏好、判断和社会福利,将输入的多个个体对象合成为统一的社会选择输出②。

我们先将偏好聚合的有关概念形式化定义如下,令:个体集合为$N = \{1,2, \cdots,n\}$,其中$n \geqslant 2$;可选项的集合为$X = \{x, y, z, \cdots\}$;$\forall i \in N$,$\succ_i$是个体i关于可选项集合X的个体偏好;我们把$<\succ_1, \succ_2, \cdots, \succ_n>$称为个体偏好的一个档案,用 P 表示。个体偏好聚合,就是令一个函数为F,将它作用于个体偏好档案后,得到一个社会选择的偏好排序$\succsim$,形式定义为:$\succsim = F(\succ_1, \succ_2, \cdots, \succ_n)$。

"Condorcet 悖论"是指,给定三个个体和三个可选项x, y, z,对应三个个体偏好分别为:$x \succ y \succ z$,$y \succ z \succ x$,$z \succ x \succ y$,社会选择的结果出现了一个 Condorcet 环,即:$x \succ y \succ z \succ x$,也就是三个可选项没有一个可以成为社会选择的最佳选项。

针对"Condorcet 悖论",阿罗提出了著名的不可能定理③。他认为,当$|N| \geqslant 2$并且$|X| \geqslant 3$时,在满足以下五项公理:集体理性,定义域无限制,帕累托准则,第三项独立性和不存在独裁者的前提下,通过某个函数或者决策程序从个体偏好找到社会偏好最佳项将是不可能的。

①Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. "An introduction to argumentation semantics". *The knowledge engineering review*, 2011, 26(4):365-410.

②Amartya Sen. "Social choice theory". *Handbook of Mathematical Economics*, 2005, 8(179):1073—1181.

③Amartya K. Sen. *Collective choice and social welfare*. Holden Day, 1979, pp. 201-218.

所以,要从个体偏好中聚合出合理的社会偏好,必然会或多或少地违反阿罗的公理系统。其中,“第三项独立性”被讨论得最多,许多聚合规则正是建立在对此项公理不同程度地的松解之上。所以,偏好聚合的关键还在于如何选择聚合函数以及怎样定义社会偏好的排序是“合理”的。

个体偏好聚合函数有一条典范规则是“成对多数规则”,即对于任意两个可选项 x, y,判定 x 大于 y 是社会选择的偏好排序的依据是多数个体更偏好 x。形式定义如下:

对于任一 $P=<\succ_1,\succ_2,\cdots,\succ_n>$, $\forall x, y \in X$, $x \succsim y$ 当且仅当 $|\{i \in N: x \succ_i y\}| \geqslant |\{i \in N: y \succ_i x\}|$。

仅仅运用“成对多数规则”这条典范规则作为个体偏好的聚合机制,社会偏好将无可避免地产生“Condorcet 悖论”。所以,“合理”的社会偏好是在满足阿罗定理其他公理的前提下,对其中的“第三项独立性”进行一定程度的松解,使之避免产生 Condorcet 环,或者满足“成对多数规则”典范规则。

个体偏好的聚合机制,主要有基于位置的方法和基于距离的方法。前者具有代表性的是 Borda 计数①,后者具有代表性的是 Kemeny 规则②。Borda 计数规则可以避免产生 Condorcet 环但不满足“成对多数”典范规则,Kemeny 规则满足“成对多数”典范规则,但往往并不能很好地处理 Condorcet 环,并且得到的社会偏好有时并不唯一③。

1.Borda 计数

Borda 计数的主要思想是根据可选项在个体偏好排序中的位置,为该可选项赋以分值。如:$x \succ y \succ z$,那么可选项 x 的分值为 3,y 的分值为 2,z 的分值为 1。个体偏好的聚合机制就是将对应可选项的分值求和后排序为社会偏好。在这里,我们采用 Borda 计数的一个变体,即基于图论的“偏好图”方法④,形式定义如下:

定义 8　一个偏好图 $G=(X, E, m)$,其中:顶点 $X=\{x, y, z, \cdots\}$ 为可选项的集合;对于任意 $x, y \in X$,有向边 $E=(x, y)$;如果 $|\{i \in N: x \succ_i y\}| - |\{i \in N: y \succ_i x\}| > 0$,那么边的权值 $m:(x, y) \rightarrow |\{i \in N: x \succ_i y\}| - |\{i \in N: y \succ_i x\}|$⑤。

定义 9　在一个偏好图 $G=(X, E, m)$ 中,每个可选项的得分规则为:对于任意 $x \in X$, $scor(x)=outdeg(x)-indeg(x)$,其中:$outdeg(x)$ 是指从顶点 x 发出的有向边的权值,$indeg(x)$ 是顶点 x 接收的有向边的权值。则 $scor(x)$ 的取值范围为 $[-|N|(|X|-1), |N|(|X|-1)]$。聚合后的社会偏好 $\succsim$ 是一个关于每个可选项 $scor(x)$ 的排序。

例 3　令 $X=\{x, y, z\}$, $N=\{1, 2, 3\}$, $P=<x \succ_1 y \succ_1 z, y \succ_2 z \succ_2 x, z \succ_3 x \succ_3 y>$,则 Condorcet 环的偏好图可以表示为图 3。根据定义 9,$scor(x)=scor(y)=scor(z)$,因此,聚合后的社会偏好为:$x \sim y \sim z$,即可选项 x, y, z 是无差异的。

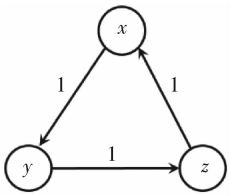


图 3　Condorcet 环的偏好图

由此可见,Borda 计数规则不但计算简便,还可以较为优美地处理 Condorcet 环,并将个体偏好的严格全序聚合为社会偏好的完全前序。但是,Borda 计数规则违反了阿罗定理中的第 4 条公理“第三项独立性”,特别是它对“成对多数”典范规则的违反,使得社会偏好的最佳项有时候并不是在多数个体偏好中占优的选项。

2.Kemeny 规则

Kemeny 规则是基于距离的个体偏好聚合机制。Kemeny 在文献中仅对它进行了文字描述⑥,我们在这里将它形式定义如下:

定义 10　给定可选项的集合为 $X=\{x, y,$

①Donald G. Saari. “The borda dictionary”. *Social Choice and Welfare*, 1990,7(4):279-317.

②John G. Kemeny. *Mathematics without numbers*. Daedalus, 1959, 88(4):577-591.

③Donald G. Saari and Vincent R. Merlin. “A geometric examination of kemeny's rule”. *Social Choice & Welfare*, 2000,17(3):403-438.

④Andrew J. Davenport and Jayant Kalagnanam. “A computational study of the kemeny rule for preference aggregation”. In *Conference on Nineteenth National Conference on Artificial Intelligence*, 2004.

⑤如果 $|\{i \in N: x \succ_i y\}| - |\{i \in N: y \succ_i x\}| = 0$,则在顶点 x, y 之间的边是无向的,权值为 0。

⑥John G. Kemeny. *Mathematics without numbers*. Daedalus, 1959, 88(4):577-591.

z, …},对于任意 x, $y \in X$, $\succ_i$是 $P = <\succ_1, \succ_2, \cdots, \succ_n>$中任意一个个体偏好,并且是一个严格全序,$\succsim$是社会偏好。则采用 Kemeny 规则计算的社会偏好 $K(\succsim)$如以下三个公式所示。

$$\delta(x,y,i)=\begin{cases}0, \text{如果 } x \succ_i y \text{ 并且 } x \succ y\\ 2, \text{如果 } y \succ_i x \text{ 并且 } x \succ y\\ 1, \text{如果 } x \succ_i y \text{ 并且 } x \sim y\end{cases}$$

$$d(\succ_i, \succsim)=\sum_{i\in N}\sum_{x,y\in X}\delta(x,y,i)$$

$$K(\succsim)=\operatorname{argmin} d(\succ_i, \succsim)$$

从定义 10 中我们可以了解到,基于 Kemeny 规则获得的社会偏好是按定义 10 中第一个公式计算后与每个个体偏好距离的总和最小(*argmin* 函数是令式子 $d(\succ_i, \succsim)$达到最小值时变量的取值)的那个排序。也就是说,需要遍历所有可能的社会偏好排序才能确定。Kemeny 规则对例 3 中个体偏好聚合的计算结果也是 $x \sim y \sim z$。但在某些情况下,与每个个体偏好距离总和最小的社会偏好不总是唯一的,存在多个的可能①。虽然 Kemeny 规则满足“成对多数”典范规则,但从计算复杂性角度看,该算法是一个 NP 难题②。

表 1　Borda 计数与 Kemeny 规则差异比较

例(a)		例(b)	
个体数	偏好	个体数	偏好
1	$B \succ C \succ A$	1	$A \succ B \succ C$
2	$A \succ B \succ C$	2	$B \succ C \succ A$
		2	$C \succ A \succ B$
F_B	$A \sim B \succ C$	F_B	$C \succ B \succ A$
F_K	$A \succ B \succ C$	F_K	$C \sim B \sim A$

3.Borda 计数与 Kemeny 规则的比较

为说明两个聚合函数各自的特点和差异,我们在表 1 中用两个例子来说明。在这两个例子中,可选项的数量均为三个,即可选项为 A, B, C。我们可以观察到,在例(a)中,个体数为三个,Borda 计数与 Kemeny 规则在聚合个体偏好时的差异体现在可选项 A 和 B 的排序上。容易发现 Kemeny 规则满足“成对多数”典范规则,因而聚合后得到的是 $A \succ B$,而 Borda 计数并不满足“成对多数”典范规则。例(b)中,个体数增加到五个,个体偏好档案中包含一个 Condorcet 环。可以看到 Kemeny 规则无法很好地处理 Condorcet 环,三个可选项呈现无差异排序,而 Borda 计数则得到了一个线性序。

(二)关于论证个体偏好的聚合和推理

本小节我们讨论第一个层面的个体偏好:关于论证的个体偏好。关于论证的个体偏好是利益相关方关于论证上价值集合的个体偏好。因此,需要先把关于单个价值的个体偏好提升为关于论证上价值集合的个体偏好。根据凯洛尔在文献中的介绍③,比较两个集合的优先级有精英和民主两个准则。给定两个有限非空集合 Γ, Γ',则:

· $\Gamma \trianglelefteq_{Dem} \Gamma'$,当且仅当 $\exists X \in \Gamma' \setminus \Gamma, \forall Y \in \Gamma \setminus \Gamma'$使得 $X \succsim Y$;

· $\Gamma \trianglelefteq_{Eli} \Gamma'$,当且仅当 $\forall X \in \Gamma' \setminus \Gamma, \exists Y \in \Gamma \setminus \Gamma'$使得 $X \succsim Y$。

根据以上定义和凯洛尔在文献中的论述③,在对两个集合进行优先级比较时,民主准则指所有元素被取代是因为存在更优先的元素,而精英准则是指被保留的所有元素都要优先于被取代的元素。因此,在存在极大值时,从集包含角度来说,民主准则倾向元素个数多的集合;而在存在极小值时,从集包含角度来说,精英准则倾向元素个数少的集合。我们用$\trianglelefteq_\alpha$ 表示基于个体偏好关于论证上价值集合的比较准则,其中 $\alpha \in \{Dem, Eli\}$。

给定一个道德委员会论辩框架,对于利益相关方 $i \in N$,我们将与 i 关于论证的个体偏好对应的论辩框架,形式定义如下:

定义 11　给定一个道德委员会论辩框架 $AFMC = (\mathcal{A}, \mathcal{R}, N, V, P)$,其中:$P = <\succ_1, \succ_2, \cdots, \succ_n>$,则与相关利益方 i ($i = 1, 2, \cdots, n$)关于论证的个体偏好相对应的论辩框架,记作 $AFMC_i = (\mathcal{A}, \mathcal{R}, \trianglelefteq_{\alpha,i})$,其中$\trianglelefteq_{\alpha,i}$是利益相关方 i 基于$\succ_i$ 用民主或精英准则提升后关于论证上价值集合的排序。

由定义 11 可知,个体偏好$\succ_i$是个体关于包

①Donald G. Saari and Vincent R. Merlin. “A geometric examination of kemeny's rule”. *Social Choice & Welfare*, 2000, 17 (3):403—438.

②Andrew J. Davenport and Jayant Kalagnanam. “A computational study of the kemeny rule for preference aggregation”. In *Conference on Nineteenth National Conference on Artificial Intelligence*, 2004.

③Cayrol Claudette, Véronique Royer, and Claire Saurel. “Management of preferences in assumption-based reasoning”. In *International Conference on Processing & Management of Uncertainty in Knowledge-based Systems: Advanced Methods in Artificial Intelligence*, 1992.

括自身持有价值在内的所有利益相关方持有价值的排序,因此定义 11 中与个体偏好$\trianglelefteq_{\alpha,i}$相对应的论辩框架 $AFMC_i$反映的是个体偏好对论证优先级进而对论证间攻击关系的影响。但是,该框架依然是一个个体框架,依然无法解决道德困境,只有在个体偏好聚合为统一的社会偏好后,才能体现社会选择对全局框架的决定性影响。我们将在下文结合案例进一步论述。

例 4　继续例 2。在 ICD 案例中,相关利益方的三个论辩框架分别为 $AFMC_1$, $AFMC_2$ $AFMC_3$,其中:由三个个体偏好$\succ_1$, $\succ_2$, $\succ_3$根据民主和精英准则分别提升后关于论证上价值集合的排序,按强度从小到大排列依次为:

$\trianglelefteq_{Dem,1}=\trianglelefteq_{Eli,1}$: $val_2(B)$, $val_3(C)$, $val_1(A)$,

$\trianglelefteq_{Dem,2}=\trianglelefteq_{Eli,2}$: $val_1(A)$, $val_3(C)$, $val_2(B)$,

$\trianglelefteq_{Dem,3}$: $val_1(A)$, $val_2(B)$, $val_3(C)$,$\trianglelefteq_{Eli,3}$: $val_2(B)$, $val_3(C)$, $val_1(A)$。

在上文中,我们提到利益相关方关于单个价值的排序是一个严格全序。在集合论中,严格序是定义在一个集合上的满足非对称性和传递性的二元关系。严格全序意味着关于任何两个价值的偏好都是可以比较的,且不存在偏好无差异的情况。而且,我们发现个体偏好在用精英或民主原则提升为关于论证上价值集合的偏好后,也是一个严格全序。当个体偏好确定为框架中论证的优先级时,严格全序使得任何两个论证的优先级也是可以比较的,且不存在无差异的情况。此时,就将偏好、优先级和框架解析联系到一起,有了以下的引理 1 和定理 1。

引理 1　$\trianglelefteq_{\alpha,i}$是一个严格全序。

证明:已知关于单个价值的排序$\succ_i$是一个严格全序,我们选证$\trianglelefteq_{Dem,i}$,同理可证$\trianglelefteq_{Eli,i}$。

我们先证传递性。假设存在三个论证上的价值集合 A,B,C,使得 $A \trianglelefteq_{Dem,i} B$ 且 $B \trianglelefteq_{Dem,i} C$,我们要证 $A \trianglelefteq_{Dem,i} C$。根据$\trianglelefteq_{Dem}$定义,则要证对于任一 $a \in A \setminus C$,存在 $c \in C \setminus A$ 使得 $c \succ_i a$(记作 $*$)。第一种情况是 $A \setminus C = \emptyset$,则上式($*$)成立。第二种情况是 $A \setminus C \neq \emptyset$,(1)如果 $a \in A \cap B$,已知 $a \notin C$,则 $a \in B \setminus C$。由 $B \trianglelefteq_{Dem,i} C$ 可得:存在 $c \in C \setminus B$ 使得 $c \succ_i a$;(2)如果 $a \notin A \cap B$,则由 $A \trianglelefteq_{Dem,i} B$ 可得:存在 $b \in B \setminus A$ 使得 $b \succ_i a$。(2.1)如果 $b \in B \cap C$,则 $b \in C \setminus A$ 且 $b \succ_i a$,满足式子($*$)。(2.2)若 $b \notin B \cap C$,由 $B \trianglelefteq_{Dem,i} C$ 可得:存在 $c \in C \setminus B$ 使得 $c \succ_i b$。已知$\succ_i$是一个严格全序,满足传递性,因此存在 $c \in C \setminus B$ 使得 $c \succ_i b \succ_i a$。而我们要证的是对于任一 $a \in A \setminus C$,存在 $c \in C \setminus A$ 使得 $c \succ_i a$。因此,我们假设不存在 $c \in C \setminus A$ 使得 $c \succ_i a$(记作 #)。则令 $S_a = \{c \mid c \in C \setminus B$ 使得 $c \succ_i a\}$,令 c_{max} 为 S_a 中根据$\succ_i$的最大元素且 $c_{max} \notin C \setminus A$。因此 $c_{max} \in A \setminus B$。由 $A \trianglelefteq_{Dem,i} B$ 可得:存在 $b' \in B \setminus A$ 使得 $b' \succ_i c_{max}$。(1)如果 $b' \in B \cap C$,则 $b' \in C \setminus A$,与假设(#)矛盾。(2)如果 $b' \in B \setminus A$ 且 $b' \notin B \cap C$,则由 $B \trianglelefteq_{Dem,i} C$,存在 $c' \in C \setminus B$ 使得 $c' \succ_i b' \succ_i c_{max}$,与 c_{max} 为 S_a 中根据$\succ_i$的最大元素相矛盾。

再证非对称性。假设存在任意两个论证上的价值集合 A 和 B,我们要证如果 $A \trianglelefteq_{Dem,i} B$ 则 $B \not\trianglelefteq_{Dem,i} A$。我们用反证法证明,即假设 $A \trianglelefteq_{Dem,i} B$ 且 $B \trianglelefteq_{Dem,i} A$,由 $A \trianglelefteq_{Dem,i} B$ 可得对于任一 $a \in A \setminus B$,存在 $b \in B \setminus A$ 使得 $b \succ_i a$,由 $B \trianglelefteq_{Dem,i} A$ 可得对于 $b \in B \setminus A$,存在 $a' \in A \setminus B$ 使得 $a' \succ_i b$,令 $a' = a$,则 $a' \succ_i b$ 且 $b \succ_i a'$。与$\succ_i$是一个严格全序满足非对称性矛盾。

由于$\succ_i$是一个关于单个价值的严格全序,根据定义 2 所有的价值都被指派到论证上,因此任何论证根据$\trianglelefteq_{Dem,i}$都可以进行比较,因而$\trianglelefteq_{Dem,i}$满足传递性、非对称性且是完全的,$\trianglelefteq_{Dem,i}$是一个严格全序。综上,所以$\trianglelefteq_{\alpha,i}$是一个严格全序。

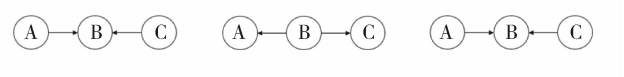
关于论证上价值集合的个体排序也是利益相关方关于论证强度的个体排序,也就是确定了框架中论证的优先级。

定理 1　根据$\trianglelefteq_{\alpha,i}$提升后的框架 $AFMC_i$ 是框架 AFMC 的一个完全解析。

证明:令框架 AFMC = ($\mathcal{A}$, $\mathcal{R}$, N, V, P)和框架 $AFMC_i$ = ($\mathcal{A}$, $\mathcal{R}$, $\trianglelefteq_{\alpha,i}$)。令框架 $AFMC_i$ 根据$\trianglelefteq_{\alpha,i}$提升论证优先级后获得的框架为 AF = ($\mathcal{A}$, $\mathcal{D}$),其中$\mathcal{D}$是$\mathcal{R}$中根据$\trianglelefteq_{\alpha,i}$确定的论证间击败关系,因此$\mathcal{D} \subseteq \mathcal{R}$。由引理 1 可得当各论证上的价值集合是一个关于论证优先级的严格全序。因此,对于任意 A_1、$A_2 \in \mathcal{D}$,或者 A_1 的优先级大于 A_2,或者 A_2 的优先级大于 A_1。由于攻击关系$\mathcal{R}$是依赖论证优先级、具有对称性的反驳关系,因此对于任意(A_1, A_2) $\in \mathcal{R}$和(A_2, A_1) $\in \mathcal{R}$,由引理 1 可得或者(A_1, A_2) $\in \mathcal{D}$,或者(A_2, A_1) $\in \mathcal{D}$。由于$\trianglelefteq_{\alpha,i}$是严格全序,$A_1$、$A_2$ 的优先级不可能是无差异的,因此不存在(A_1, A_2) $\in \mathcal{D}$且(A_2, A_1) $\in \mathcal{D}$,所以满足定义 7 第 1 点。又因为$\mathcal{D} \subseteq \mathcal{R}$,所以对于

任意$(A_1, A_2) \in \mathcal{D}$,必然有$(A_1, A_2) \in \mathcal{R}$,满足定义 7 第 2 点。综上,得证。

例 5　继续 ICD 案例,各论证上的价值集合是不存在共享价值的,因此个体偏好经过民主和精英准则提升后,三个利益相关方的个体偏好经过提升后分别获得的完全解析框架如图 4 所示。在这里展示的是经过精英准则提升后的个体框架。



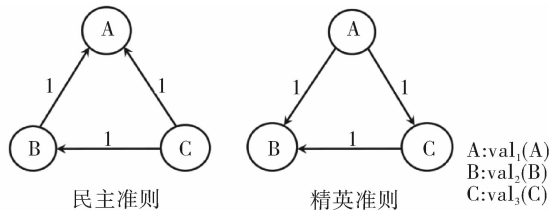
设计者的论辩框架　　Jane的论辩框架　　外科医生的论辩框架

图 4　ICD 案例中三个利益相关方的解析框架

在定义 11 中,我们已经得到了利益相关方的个体偏好提升后关于论证上价值集合的排序$\trianglelefteq_{\alpha,i}$。但$\trianglelefteq_{\alpha,i}$依然是一个个体偏好,令由个体偏好$\trianglelefteq_{\alpha,i}$组成的档案为$P_{\mathcal{A}}$。我们分别用两种聚合函数,一个是 Borda 计数,记为F_B,另一个是 Kemeny 规则,记为F_K,对个体偏好档案$P_{\mathcal{A}}$进行聚合,获得社会偏好记为$\succsim_{\alpha,\beta}$,其中$\alpha \in \{Dem, Eli\}$,$\beta = \{B, F\}$,我们将严格排序记为$\succ_{\alpha,\beta}$,无差异排序记为$\sim_{\alpha,\beta}$。社会偏好$\succsim_{\alpha,\beta}$可以自然地提升为论证的优先级,获得对应的解析框架。我们令σ为优先语义,由于优先语义是可接受论证集合在集包含意义下最大的集合,所以获得框架的优先外延,就是求得的符合多数相关利益方偏好的论证集合。继续 ICD 案例。

例 6　我们已经得到利益相关方关于论证上价值集合的个体偏好档案$P_{\mathcal{A}} = <\trianglelefteq_{\alpha,1}, \trianglelefteq_{\alpha,2}, \trianglelefteq_{\alpha,3}>$如例 4 所示。则:

· 基于 Borda 计数的聚合:根据定义 8,我们为档案$P_{\mathcal{A}}$建立的偏好图为:$G_{P_{\mathcal{A}}} = (X, E, m)$,其中$X = \{val_1(A), val_2(B), val_3(C)\}$。则$P_{\mathcal{A}}$的偏好图可以表示为图 5。根据定义 9,经过$F_B$函数聚合后的社会偏好分别为:



民主准则　　精英准则

图 5　ICD 中个体偏好档案$P_{\mathcal{A}}$的偏好图$G_{P_{\mathcal{A}}}$

根据民主准则:$val_3(C) \succ_{Dem,B} val_2(B) \succ_{Dem,B} val_1(A)$,

根据精英准则:$val_1(A) \succ_{Eli,B} val_3(C) \succ_{Eli,B} val_2(B)$。

· 基于 Kemeny 规则的聚合:已知关于三个可选项的 13 种排序组合以及它们之间的 Kemeny 距离如图 6 所示。边上的权值表示两个序在相互变换时在排序上的差距,也叫距离。如:从$a \succ c \succ b$变换到$a \succ b \succ c$时,$c \succ b$先变换$c \sim b$再变换为$b \succ c$,排序上的差距为 2,也就是距离为 2。ICD 案例中给定的个体偏好档案$P_{\mathcal{A}}$(根据民主准则提升)是图 6 中三个粗体的排序。根据定义 10,我们计算出当$d(\trianglelefteq_{\alpha,i}, \succsim) = 6$是最短的 Kemeny 距离时,社会偏好分别为:

根据民主准则:$K(\succsim) = val_3(C) \succ_{Dem,F} val_2(B) \succ_{Dem,F} val_1(A)$,

根据精英准则:$K(\succsim) = val_1(A) \succ_{Eli,F} val_3(C) \succ_{Eli,F} val_2(B)$。

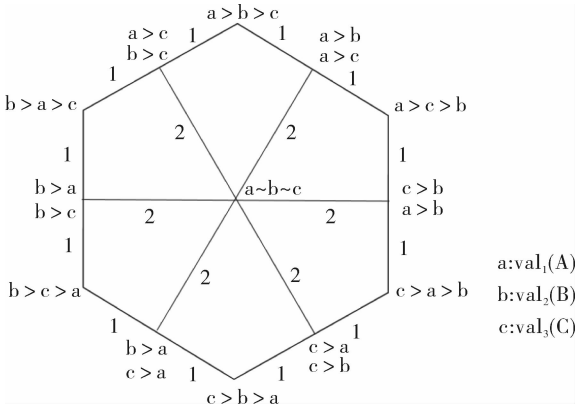


图 6　关于三个可选项的排序组合和 Kemeny 距离

我们通过民主和精英准则提升个体偏好,得到了两个不同的社会偏好。关于这两个准则在提升时的差异和后续影响我们将在后面的小节进行讨论,在这里只给出结果。在例 6 中,我们用聚合函数F_B和F_K对个体偏好档案$P_{\mathcal{A}}$进行聚合后,得到的社会偏好是一致的。但在一些情形下,两种聚合函数得到的结果并不总是一致①。社会偏好已自然提升为论证之间的优先级,我们分别获得经民主准则和精英准则提升、基于社会偏好的 AFMC 抽象论辩框架(由论证集合和论证间击败关系组成的框架)如图 7 所示。此时,社会偏好

①Tomas J. Mcintee. "Geometric ways of understanding voting problems". *Dissertations & Theses - Gradworks*, 2015.

对全局的决定性影响体现为:将图 4 中的三个个体框架统一为社会选择下的唯一框架。然后采用抽象论辩的优先语义计算,我们得到经民主准则和精英准则提升后社会选择框架的优先外延均为 $\mathcal{E}=\{A, C\}$。因此,这是关于论证的个体偏好聚合和推理的求解,也就是推理得出设计者和医生论证的结论是合理的求解。

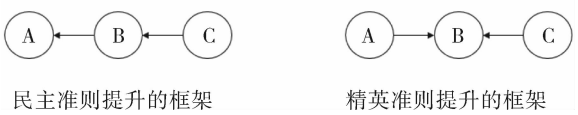


图 7　基于社会偏好的 AFMC 抽象论辩框架

(三)关于规范的个体偏好聚合和推理

本小节我们讨论第二个层面的个体偏好:关于规范的个体偏好。关于规范的个体偏好是利益相关方关于规范上价值集合的个体偏好。有关把规范上单个价值的个体偏好提升为规范上价值集合的个体偏好的方法,可以参考上一小节中集合优先级比较的方法。

定义 12　给定价值规范系统 $\mathrm{VNS}_i=(L, V_i, \Omega_i, val_i)$,$i=1,2,\cdots,n$,以及道德委员会的论辩框架 $\mathrm{AFMC}=(\mathcal{A}, \mathcal{R}, N, V, P)$,利益相关方 i 对规范的个体偏好是指其用民主或精英准则对规范上的价值集合进行排序,记作 $\gg_i$。各利益相关方关于规范的个体偏好档案记作 $P_\Omega=<\gg_1, \gg_2, \cdots, \gg_n>$。

由于 ICD 案例中,所有利益相关方的 VNS 中规范和价值都是一一对应关系,因此在该案例中,关于规范的个体偏好 $\gg_i$ 和关于单个价值的个体偏好 $>_i$ 排序一致。在这里,我们先不将关于规范的个体偏好提升为论证的优先级,而是运用两种聚合函数分别对个体偏好进行聚合。

定义 13　给定 $P_\Omega=<\gg_1, \gg_2, \cdots, \gg_n>$,用 F_B 和 F_K 聚合函数得到的社会偏好分别记为 $\succsim_B$ 和 $\succsim_F$。令 $\beta=\{B, F\}$,则用 F_B 和 F_K 聚合函数得到的关于规范的社会偏好,统一记法为 $\succsim_\beta$,我们将严格排序记为 $>_\beta$,无差异排序记为 $\sim_\beta$。

例 7　在 ICD 案例中,利益相关方关于规范的个体偏好档案为 $P_\Omega=<\gg_1, \gg_2, \gg_3>$,各利益相关方关于规范的个体偏好按从强到弱的顺序排列,分别为:

$\gg_1$: u_a, u_d, u_b, u_e, u_c,$\gg_2$: u_b, u_c, u_d, u_e, u_a,$\gg_3$: u_e, u_b, u_a, u_d, u_c。

· 基于 Borda 计数的聚合:根据定义 8,我们为档案 P_Ω 建立的偏好图为:$G_{P_\Omega}=(X, E, m)$,其中 $X=\{u_a, u_b, u_c, u_d, u_e\}$。则 P_Ω 的偏好图可以表示为图 8。根据定义 9,经过 F_B 函数聚合后的社会偏好为:$u_b >_B u_a \sim_B u_e \sim_B u_d >_B u_c$。

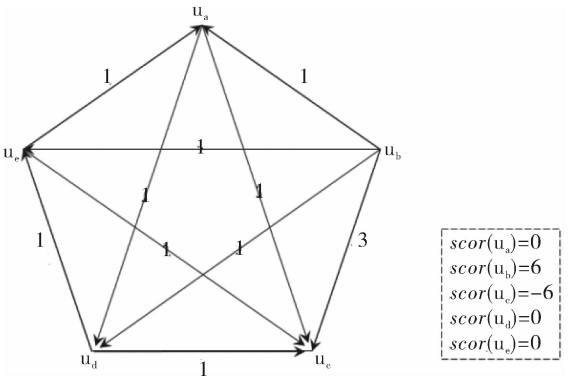


图 8　ICD 中个体偏好档案 P_Ω 的偏好图 G_{P_Ω} 及其 $scor$ 函数

· 基于 Kemeny 规则的聚合:根据定义 10,我们通过将 Kemeny 规则算法编程后计算得到的社会偏好也是 $u_b >_F u_a \sim_F u_e \sim_F u_d >_F u_c$。

关于规范的社会偏好虽然已经获得,但要在论辩框架中用论辩语义实现推理,还必须将规范和论证关联起来,这就要用到定义 2。在 ICD 案例中,论证 $A=[u_a]$,论证 $B=[u_b, u_c]$,论证 $C=[u_d, u_e]$。由于论证是由一组规范序列构成,在论辩推理中,可以根据两个准则将关于规范的社会偏好提升为论证的优先级。基于现有文献中的定义①,我们将这两个提升准则定义如下。

定义 14　给定一个道德委员会论辩框架 $\mathrm{AFMC}=(\mathcal{A}, \mathcal{R}, N, V, P)$,一个利益相关方关于规范的个体偏好档案 $P_\Omega=<\gg_1, \gg_2, \cdots, \gg_n>$,以及依据聚合函数 F_B 或 F_K 获得的规范的社会偏好排序为 $\succsim_\beta$,其中 $\beta=\{B, F\}$。令 $A=[u_1, \cdots, u_n]$ 和 $A'=[v_1, \cdots, v_m]$ 为 A 中的两个论证。令 $S=\{u_1, \cdots, u_n\}$; $S'=\{v_1, \cdots, v_m\}$,则:

最后链准则:$A \trianglerighteq_{L,\beta} A'$,当且仅当 $u_n \succsim_\beta v_m$;

最弱链准则:$A \trianglerighteq_{W,\beta} A'$,当且仅当 $\forall u \in S \setminus S', \exists v \in S' \setminus S$ 使得 $u \succsim_\beta v$。

①Beishui Liao, Nir Oren, Leendert Van Der Torre, et al. "Prioritized norms in formal argumentation". *Journal of Logic and Computation*, 2018, 29(2): 215-240.

依据规范的社会偏好,令 $\gamma = \{L, W\}$,则论证 A 的优先级不小于论证 A' 记作 $A \trianglerighteq_{\gamma,\beta} A'$,严格排序记作 $\triangleright_{\gamma,\beta}$,无差异排序记作 $\sim_{\gamma,\beta}$。

命题 1　$\trianglerighteq_{W,\beta}$ 等价于 $\trianglelefteq_{ELi}$。

证明:根据上文关于 $\trianglelefteq_{ELi}$ 的定义,$\trianglelefteq_{ELi}$ 是比较两个集合的二元关系。根据定义 14,最后链和最弱链的比较对象虽然是具有序列结构的规范集合,但 $\trianglerighteq_{W,\beta}$ 的定义是对两个集合的比较,$\trianglerighteq_{W,\beta}$ 和 $\trianglelefteq_{ELi}$ 在排序的方向上虽相反,但在关于集合优先级的排序上具有等价性。

定义 15　给定一个道德委员会论辩框架 $AFMC = (\mathcal{A}, \mathcal{R}, N, V, P)$,与关于规范的社会偏好所对应的论辩框架记作 $AFMC_S = (\mathcal{A}, \mathcal{R}, \trianglerighteq_{\gamma,\beta})$。

在引理 1 我们提到,在集合论中,满足自反性和传递性的二元关系是一个前序。我们说一个前序是完全的,当且仅当在这个集合中,基于该二元关系,任何元素都是可比较的。可以发现,相比严格全序,完全前序是一个更宽泛的排序。集合中任意两个元素,或者是严格偏序或者是无差异的。但是,即便个体偏好是一个严格排序,经过聚合后,社会偏好也不一定是一个严格全序(我们从表 1 中也可以发现),这是由个体偏好档案和聚合函数共同确定的。当偏好确定为框架中论证的优先级,严格全序使框架发生完全解析(如定理 1 所示),而完全前序则不一定能使框架发生完全解析,这就是下文的定理 2。

引理 2　关于规范的社会偏好 $\succsim_\beta$ 是一个完全前序。

证明:令个体集合为 $N = \{1, 2, \cdots, n\}$,可选项的集合为 $X = \{x, y, z, \cdots\}$。已知关于规范的个体偏好 $\succcurlyeq_i$ 是严格全序,$i = 1, 2, \cdots, n$。令 $\beta = \{B\}$,采用 Borda 计数为聚合函数,根据定义 8 和定义 9,对于任意 $x, y \in X$,当 $scor(x) = scor(y)$ 时,$x \succsim_B y$ 且 $y \succsim_B x$(或记作 $x \sim_\beta y$),$\succsim_\beta$ 满足传递性和自反性;对于任意 $x, y, z \in X$,如果 $scor(x) > scor(y)$ 且 $scor(y) > scor(z)$,则 $x \succsim_B y$ 且 $y \succsim_B z$,由于 $scor(x)$ 满足传递性,$scor(x) > scor(z)$,因此 $x \succsim_B z$,$\succsim_\beta$ 满足传递性;任意 $x, y \in X$ 都可以基于 $\succsim_B$ 进行比较,因此 $\succsim_B$ 是一个完全前序。令 $\beta = \{F\}$,采用 Kemeny 为聚合函数,根据 Kemeny 规则在上文中阐述的相关特性,当个体偏好档案中出现 Condorcet 环时,社会偏好中的相关可选项会出现无差异排序,所以 $\succsim_F$ 满足传递性和自反性;根据定义 10,任意 $x, y \in X$ 都可以基于 $\succsim_F$ 进行比较,$\succsim_F$ 是一个完全前序。因此,关于规范的社会偏好 $\succsim_\beta$ 是一个完全前序。

引理 3　基于 $\succsim_\beta$ 根据最后链和最弱链获得的 $\trianglerighteq_{\gamma,\beta}$ 是一个完全前序。

证明:根据定义 14,令 $A = [u_1, \cdots, u_n]$ 和 $A' = [v_1, \cdots, v_m]$ 为 $\mathcal{A}$ 中的两个论证。令 $\gamma = \{L\}$,当且仅当 $u_n \succsim_\beta v_m$ 时,$A \trianglerighteq_{L,\beta} A'$,有引理 2 可得 $\succsim_\beta$ 满足传递性,因此 $\trianglerighteq_{L,\beta}$ 满足传递性;当且仅当 $u_n \succsim_\beta v_m$ 且 $v_m \succsim_\beta u_n$ 时,$A \trianglerighteq_{L,\beta} A'$ 且 $A' \trianglerighteq_{L,\beta} A$,$\trianglerighteq_{L,\beta}$ 满足传递性和自反性;根据定义 14,任何 $A, A' \in A$ 都可以按基于 $\trianglerighteq_{L,\beta}$ 进行比较,因此 $\trianglerighteq_{L,\beta}$ 是一个完全前序。同理可证 $\gamma = \{W\}$,$\trianglerighteq_{W,\beta}$ 是一个完全前序。因此基于 $\succsim_\beta$ 根据最后链和最弱链获得的 $\trianglerighteq_{\gamma,\beta}$ 是一个完全前序。

定理 2　根据 $\trianglerighteq_{\gamma,\beta}$ 提升后的框架 $AFMC_S$ 是框架 AFMC 的一个解析。

证明:令框架 $AFMC = (\mathcal{A}, \mathcal{R}, N, V, P)$ 和框架 $AFMC_S = (\mathcal{A}, \mathcal{R}, \trianglerighteq_{\gamma,\beta})$。令框架 $AFMC_S$ 根据 $\trianglerighteq_{\gamma,\beta}$ 提升后的框架为 $AF' = (\mathcal{A}, \mathcal{D}')$。$\mathcal{D}'$ 是 $\mathcal{R}$ 中根据 $\trianglerighteq_{\gamma,\beta}$ 确定的击败关系集合,因此 $\mathcal{D}' \subseteq \mathcal{R}$。根据定义 6,令 $\Delta = \mathcal{R} \setminus \mathcal{D}'$。由引理 3 可得 $\trianglerighteq_{\gamma,\beta}$ 是一个完全前序,确定了 $\mathcal{A}$ 中所有论证的优先级。对于 $\mathcal{R}$ 中任意一对对称性攻击关系 (A, A') 和 (A', A),根据 $\trianglerighteq_{\gamma,\beta}$,$A$ 和 A' 优先级存在两种可能:1. $A \trianglerighteq_{\gamma,\beta} A'$ 且 $A' \not\trianglerighteq_{\gamma,\beta} A$(或 $A' \trianglerighteq_{\gamma,\beta} A$ 且 $A \not\trianglerighteq_{\gamma,\beta} A'$);2. $A \trianglerighteq_{\gamma,\beta} A'$ 且 $A' \trianglelefteq_{\gamma,\beta} A$。第一种情况,或者 $(A, A') \in \mathcal{D}'$,或者 $(A', A) \in \mathcal{D}'$,但不存在 $(A, A') \in \mathcal{D}'$ 且 $(A', A) \in \mathcal{D}'$,一对对称性攻击关系中仅有一组在 Δ 中。第二种情况,$(A, A') \in \mathcal{D}'$ 且 $(A', A) \in \mathcal{D}'$,则两组攻击关系均不在 Δ 中。综上,得证。

框架解析比框架完全解析具有更宽泛的意义,意味着框架中可能依然存在对称性的攻击关系。根据定义 7 可知,完全解析是框架解析的一种特殊情形。未获得完全解析的框架意味着用抽象论辩优先语义对其进行求解时,获得的往往并不是唯一解。

例 8 在 ICD 案例中,根据例 7 中关于规范的社会偏好,我们用最后链准则和最弱链准则提升得到的论证优先级都是:$A \sim_{\gamma,\beta} C \triangleright_{\gamma,\beta} B$。这时,我

们可以发现社会选择的影响是使图 2 中的道德委员会论辩框架发生了完全解析(框架中已不存在对称性的攻击关系),生成的抽象论辩框架图形同图 4 中设计者和医生的论辩框架。在抽象论辩优先语义下,得到该框架的唯一外延也是$\mathcal{E}=\{A, C\}$,因而这是关于规范的个体偏好聚合和推理求解得出的,符合多数利益相关方偏好的论证集合。

在例 8 中,我们发现,关于具有对称性攻击关系的论证 A 和论证 B,论证 B 和论证 C 的优先级排序是一个严格偏序。所以尽管论证 A 和论证 C 是无差异的,实际上道德委员会论辩框架发生了完全解析,获得的是唯一的优先外延,从而基于偏好聚合和论辩推理的道德困境获得了求解。因此,接下来我们想了解在哪些情况下,基于偏好聚合和论辩推理的道德困境可以获得唯一的优先外延。

首先,通过上文的分析,我们发现完全解析框架中不存在环结构。

定义 16　令框架 $AF=(\mathcal{A}, \mathcal{R})$,我们说一个框架是无环的,当且仅当对于任意 $A_1, \cdots, A_n \in \mathcal{A}$,不存在$(A_1, A_2), \cdots, (A_n, A_1) \in \mathcal{R}$,其中 $n \geq 1$。

命题 2　AFMC 的完全解析框架是无环的。

证明:令 AFMC 的一个完全解析框架为 $AF=(\mathcal{A}, \mathcal{D})$。当 $n=1$ 时,根据定义 5,由于 AFMC 中论证$\mathcal{A}_i$是来自于各利益相关方 i 的价值规范系统,且$\mathcal{A}_i$是一个无冲突的论证集合,因此在框架中不存在论证的自我攻击关系。当 $n=2$ 时,根据定义 6,完全解析的框架中,不存在对称性攻击关系。现在,我们要证当 $n=k, k>2$ 时,AFMC 的完全解析框架 AF 也是无环的。假设 $AF=(\mathcal{A}, \mathcal{D})$ 中有环。令一个环为$(A_1, A_2), (A_2, A_3), \cdots, (A_k, A_1) \in \mathcal{D}$,根据完全解析的定义,不存在$(A_2, A_1), (A_3, A_2), \cdots, (A_1, A_k) \in \mathcal{D}$,其中 $A_1, A_2, \cdots, A_k \in \mathcal{A}, k>2$。根据引理 1,框架 AF 是将某种严格全序的偏好提升为论证优先级而从 AFMC 解析出来的。令论证间的优先关系排序为$\succcurlyeq$,且是一个严格全序。这时有 $A_1 \succcurlyeq A_2$, $A_2 \succcurlyeq A_3$, $\cdots$, $A_k \succcurlyeq A_1$且不存在 $A_2 \succcurlyeq A_1$, $A_3 \succcurlyeq A_2$, $\cdots$, $A_1 \succcurlyeq A_k$,与$\succcurlyeq$满足传递性相矛盾。

由框架无环性可得,该框架是一个攻击关系呈链条式的结构,潘明栋在他的文献①中证明了该类框架在优先语义下具有唯一的非空外延,并且该外延也是基语义、稳定语义和完全语义下的求解(见该文献中的定义 10 和定理 3)。因此,我们可以直接得到以下定理。

定理 3　AFMC 的完全解析框架在优先语义下存在唯一非空外延。

推论 1　当$\trianglerighteq_{\gamma,\beta}$是一个严格全序时,根据$\trianglerighteq_{\gamma,\beta}$提升后的框架 AFMC_S在优先语义下存在唯一非空外延。

证明:根据定理 1,当$\trianglerighteq_{\gamma,\beta}$是一个严格全序时,根据$\trianglerighteq_{\gamma,\beta}$提升后的框架 AFMC_S是框架 AFMC 的一个完全解析框架。根据定理 3 可得该框架在优先语义下存在唯一非空外延。

其次,通过例 8 中的观察发现:即使偏好聚合并提升后的$\trianglerighteq_{\gamma,\beta}$是一个完全前序,但只要当具有对称性攻击关系的论证,关于其优先级的比较是一个严格排序时,在道德困境下依然可以获得唯一求解。

推论 2　当关于具有对称性攻击关系论证优先级比较的$\trianglerighteq_{\gamma,\beta}$是一个严格偏序时,根据$\trianglerighteq_{\gamma,\beta}$提升后的框架 AFMC_S在优先语义下存在唯一非空外延。

证明:由于具有对称性攻击关系的论证都可以根据严格排序进行优先级比较,因此根据$\trianglerighteq_{\gamma,\beta}$提升后的框架 AFMC_S是框架 AFMC 的一个完全解析框架。根据定理 3 可得该框架在优先语义下存在唯一非空外延。

再次,当框架未发生完全解析、对称性攻击关系仅部分地被撤销时,符合初始无环性的论辩框架将在优先语义下获得唯一非空外延。定义初始无环性需要先定义两个概念:初始论证和特征函数。下面,我们根据初始论证在文献②中的定义

①Dung, Phan Minh. "On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games." *Artificial intelligence*, 1995, 77(2):321-357.

②Liao, Beishui, and Leendert van der Torre. "Defense semantics of argumentation: encoding reasons for accepting arguments." *arXiv preprint arXiv*:1705.00303 (2017).

和特征函数在文献[①]中的定义,来定义初始无环性,最后得出定理4。

定义17　令AFMC的一个解析框架为AF=($\mathcal{A}$, $\mathcal{D}$)。在框架中,对于A∈$\mathcal{A}$,令A^-={ B∈$\mathcal{A}$ | (B, A)∈$\mathcal{D}$},如果A^-=∅,则我们说A是框架的一个初始论证,框架中初始论证的集合记为*IA*(AF)。

定义18　令AFMC的一个解析框架为AF=($\mathcal{A}$, $\mathcal{D}$),关于该框架的特征函数,记为f_{AF},根据上文论辩语义中可防御的定义,将特征函数定义如下:

$f_{AF}:2^{AF}\rightarrow 2^{AF}$,

$f_{AF}(S)$ = { A∈$\mathcal{A}$| S可防御A}。

定义19　给定AFMC的一个解析框架为AF=($\mathcal{A}$, $\mathcal{D}$),特征函数f_{AF}如定义18所定义,则框架AF满足初始无环性,当且仅当:

(1)$IA(AF)\neq\emptyset$,并且

(2)令$C^1=f_{AF}(IA(AF))$,…,$C^{i+1}=C^i\cup f_{AF}(C^i)$。当$C^{i+1}=C^i$时,令在框架AF中除去$C^i$所剩余的部分为$AF_r$,$AF_r=(\mathcal{A}\backslash C^i, D\cap(\mathcal{A}\backslash C^i\times\mathcal{A}\backslash C^i))$,则框架$AF_r$满足定义16中的无环性。

定理4　令框架AFMC的一个解析框架为AF=($\mathcal{A}$, $\mathcal{D}$),如果AF满足初始无环性,则在优先语义下有唯一非空外延。

证明:由于AF满足初始无环性,根据定义19的第一点$IA(AF)\neq\emptyset$,则根据抽象论辩语义,*IA*(AF)必然在优先外延中,且在基外延中。根据定义18和19中第二点,以*IA*(AF)为基始、根据特征函数求得的论证集合C^i也在优先外延和基外延中,因此框架AF的优先外延为非空论证集合。由定义19的第二点可得,满足初始无环性的框架AF,在除去C^i所剩余的AF_r中不存在环结构。根据命题2和定理3,无环的框架AF_r在优先语义下有唯一外延。因此,满足初始无环性,AFMC的解析框架AF=($\mathcal{A}$, $\mathcal{D}$)在优先语义下有唯一非空外延。

定理3和定理4中,AFMC的(完全)解析框架在优先语义下存在的唯一非空外延实际上也是完全语义、基语义、稳定语义下的唯一解。因此,为了获得道德困境下的唯一求解,我们可以使道德委员会论辩框架获得完全解析的框架,也就是需要在具有对称性攻击关系的论证之间获得关于优先级的严格排序,这取决于与这些论证优先级有关的社会偏好是否是一个严格排序;我们也可以通过判断道德委员会框架的解析框架是否符合初始无环性,即满足定义19中的两个条件来确定能否获得道德困境下的唯一求解。值得注意的是,满足初始无环性的论辩框架不一定是道德委员会论辩框架的完全解析框架,框架中可能依然存在对称性的攻击关系,这在一定程度上松解了社会偏好是严格排序这一限制。

三　总结与未来工作

本文的创新点在于:在单独使用偏好聚合或论辩推理都无法有效地解决道德困境难题的前提下,采用偏好聚合和论辩推理相结合的办法,先进行偏好聚合再进行论辩推理,将聚合得到的社会偏好与论辩框架中论证的优先级关联起来,从而可以运用论辩语义实现推理和求解。本节是全文的总结,先对偏好聚合和论辩推理相结合的方法进行讨论分析,然后介绍领域内相关工作并对该课题未来研究方向进行梳理和探讨。

(一)关于偏好聚合和论辩推理相结合的分析

在上文中,通过研究我们已经发现了在个体偏好、聚合函数、论证优先级、框架解析以及语义外延之间的一些关系。而这些被识别的因素在解决道德困境时又体现不同程度的影响。具体来说,可以分为以下几个方面。

首先,偏好的层次。我们在对道德困境的建模中,关于价值的偏好涉及三个层次:单个价值、规范上的价值集合、论证上的价值集合。所以,我们在进行偏好聚合和论辩推理时要解决的问题是:对象是哪个层次的偏好?在本文中,我们尝试以关于规范上价值集合的个体偏好和关于论证上价值集合的个体偏好为聚合对象,是因为我们要在论辩框架下进行推理,需要最终将偏好提升为论证的优先级,才能获得论辩语义下的求解。关于单个价值的偏好只是一个孤立的排序,无法与论证产生直接的关联,因而不能运用论辩语义进

①廖备水:《论辩系统:不一致情境中的推理》,浙江大学出版社2012年版,第28页。

行推理。

第二,论证优先级的提升准则。在文中我们介绍了关于集合优先级比较的民主准则和精英准则以及关于规范序列优先级比较的最后链准则和最弱链准则。这四个准则在对论证优先级提升时各有特点,且结果往往并不一致。命题 1 中我们论证了,根据最弱链准则和精英准则在本文中的定义,采用最弱链准则对构成论证的规范序列进行优先级比较,与用精英准则对论证上规范集合进行优先级比较的方法是等价的。我们先分析关于序列比较的最后链准则和最弱链准则。根据我们之前关于最后链和最弱链准则所做的研究①,在论辩框架中,最后链准则是指在构造论证的链条中只对产生反驳结论的链条进行优先级的比较来确定论证强度。一般情况下,这两个准则优先比较序列或链式结构的对象,如规范或论证。但在莫吉尔和帕肯在文献②中定义,当产生反驳结论的可废止规则为空集时,该准则也可以被用来比较假设性前提集合的优先级。由于在我们的框架中,论证是由规范序列构成且序列中每个规范都有价值偏好。所以,采用最弱链原则可以检视整个序列识别出其中最弱的一环,是对论证强度更为全面的评价和比较。再讨论关于集合优先级比较的民主准则和精英准则。在引理 1 中我们证明了当给定关于单个价值的排序是一个严格全序时,采用民主准则和精英准则提升的关于价值集合优先级排序也是一个严格全序。这个性质在论证优先级和论辩框架的完全解析之间建立了直接的联系。

首先,论证优先级提升和偏好聚合的先后次序。从案例推理中可以看到:我们先提升论证优先级再对关于论证的个体偏好进行聚合,和先对关于规范的个体偏好进行聚合再将关于规范的社会偏好提升为论证优先级,不同的先后次序对结果是有显著影响的。先进行提升,在我们的案例中,因为原本关于单个价值/规范的排序变成了关于价值/规范集合的排序,可选项的数量减少了。可选项的减少使得聚合后的社会偏好更有可能出现严格排序,从而通过论辩推理获得道德困境的唯一求解。通过例子我们可以看到,通过精英准则先提升后聚合的论证优先级排序是 $A>C>B$,而通过先聚合后用最弱链准则提升的论证优先级排序是 $A\sim C>B$。因此,我们将在未来的工作中研究论证优先级提升和偏好聚合的先后次序对推理结果更加结构化和数据化的影响。

其次,框架解析程度与唯一非空外延的关系。本文在最后一节中对这一关键问题进行了讨论和证明。定理 3 和定理 4 分别证明了完全解析框架和满足初始无环性的解析框架可以在论辩推理中获得唯一非空外延这一结论。但框架的解析程度除了与论证结构有关以外,还与社会偏好的排序有直接关联,而社会偏好通过聚合函数又取决于个体偏好档案。因此能否在论辩框架中获得道德困境的唯一求解,是由相当多的变量因素共同决定的。但通过研究可以发现,道德委员会的论辩框架结构可以先确定下来。因此,我们可以围绕如何使该框架发生完全解析和初始无环性解析去研究优先级提升和偏好聚合的先后次序,以及各种偏好聚合函数的取舍。在该阶段,论辩语义的选择还未体现其变量的属性。可以想到的是,在未来的研究中,随着一些限制的松解,框架复杂程度将会进一步提升,如:采用更有表达能力的逻辑语言、加入子论证与超论证关系③和可废止优先级④等等,论辩语义将会成为另一个重要的研究对象。

最后,偏好聚合函数的选择。尽管我们在文中并没有就偏好聚合函数的选择展开讨论,并且 Borda 计数和 Kemney 规则在我们的例子中得到的社会偏好总是一致的,但有相当多的文献对这

①Beishui Liao, Nir Oren, Leendert Van Der Torre, et al. "Prioritized norms in formal argumentation". *Journal of Logic and Computation*, 2018, 29(2): 215-240.

②Sanjay Modgil and Henry Prakken. "The aspic+ framework for structured argumentation: a tutorial". *Argument& Computation*, 2014, 5(1):31-62.

③Dyrkolbotn, Sjur, Truls Pedersen, and Jan Broersen. "On Elitist Lifting and Consistency in Structured Argumentation." *Journal of Applied Logics-IfCoLog Journal of Logics and their Applications*,2018, 5(3):709-745.

④Modgil, Sanjay, and Henry Prakken. "Reasoning about preferences in structured extended argumentation frameworks." *COMMA*. 2010:347-358.

两种经典的偏好聚合函数进行了讨论①②③④⑤。Borda 计数和 Kemeny 规则在本文中计算结果趋于一致的原因是,例子中的个体数和可选项个数都比较少。在表 1 中,我们已经看到,两者的计算结果在一些情形下会出现显著差异。但用 Borda 计数聚合的最佳项不可能排在 Kemeny 规则的末位,用 Kemney 规则聚合的最佳项也不会排在 Borda 计数中的最后。从表 1 中我们发现,Borda 计数可以比 Kemeny 规则得到一个更接近线性的排序。由于通过聚合,更接近线性的社会偏好将更有可能获得道德困境下的唯一求解,因此我们需要在未来工作中进一步研究满足何种条件时,Borda 计数以及 Kemeny 规则可以得到一个较为线性的社会偏好。此外,是否有其他偏好聚合函数在这一方面具有优越性,也在我们未来的研究范围内。

(二)相关工作和未来工作

本文是对现有工作⑥的拓展和深化。我们在文中松解了利益相关方关于价值共享统一排序的假设,使之更具有现实意义。在现有研究中,对于智能体行动决策的合理求解是基于价值数量或利益相关方主体数量最大化做出的,该评判标准虽易操作但可能存在有失公允的情形。而本文在引入个体偏好后采用社会选择理论中偏好聚合的理论和方法,获得的是符合多数利益相关方偏好的求解,因而更能体现道德困境中的集体意志。莫吉尔⑦、帕肯⑧、阿姆格⑨等人的文献都是在形式论辩领域以偏好或优先级为研究对象的代表性研究,但与本文不同的是,在这些文献中,偏好或优先级是作为框架给定的输入来研究,本文构建的价值规范系统可以看作这些文献中偏好或优先级一个实例化的来源。范德托等人研究了当多个论证共享同个价值时,关于价值的偏好可以被提升为论证集合的优先级⑩。本驰嘉潘对每个听众建立了基于价值的论辩框架,并在该框架中将个体对价值的偏好提升为论证的优先级,但本驰嘉潘并未对所有听众的个体偏好进行聚合处理,因而他只在个体框架上开展研究⑪。目前,基于形式论辩框架的聚合,主要有两个方向:一个是基于框架层面的聚合,主要方法是对论证间关系进行聚合,以哈瑞特⑫、陈伟伟等人的研究为代表。其中,陈伟伟等人是研究在保留论辩语义性质和满足阿罗不可能定理的前提下,多主体论辩框架中攻击关系的聚合规则⑬;另一个是基于论证层面的聚合,主要方法是对语义外延或语义标签进行聚合,又叫判断聚合,以卡米纳达等人的研究为代表⑭。判断聚合方法是为个体偏好分别建立个体框架进行语义计算,以外延中或获得肯定语义标签的多数论证为求解结果。但该方法在是否有解和是否唯一解的问题上还颇有争议。本文将偏好

---

①Andrew J. Davenport and Jayant Kalagnanam. "A computational study of the kemeny rule for preference aggregation". In *Conference on Nineteenth National Conference on Artificial Intelligence*, 2004.

②Tomas J. Mcintee. "Geometric ways of understanding voting problems". *Dissertations & Theses - Gradworks*, 2015.

③Donald G. Saari. "The borda dictionary". *Social Choice and Welfare*, 1990, 7(4):279-317.

④Donald G. Saari. "Which is better: the condorcet or borda winner?" *Social Choice & Welfare*, 2006, 26(1):107 - 129.

⑤Donald G. Saari and Vincent R. Merlin. "A geometric examination of kemeny's rule". *Social Choice & Welfare*, 2000, 17 (3):403-438.

⑥Beishui Liao and Marija Slavkovik and Leendert van der Torre. "Building Jiminy Cricket: An architecture for moral agreements among stakeholders", *Proceedings of the* 2019 *AAAI/ACM Conference on AI, Ethics, and Society*, 2019:147-153.

⑦Sanjay Modgil. "Reasoning about preferences in argumentation frameworks". *Artificial intelligence*, 2009, 173(9-10):901-934.

⑧Sanjay Modgil and Henry Prakken. "A general account of argumentation with preferences". *Artificial Intelligence*, 2013, 195:361-397.

⑨Leila Amgoud and Claudette Cayrol. "A reasoning model based on the production of acceptable arguments". *Annals of Mathematics and Artificial Intelligence*, 2002, 34(1-3):197-215.

⑩Souhila Kaci and Leendert van der Torre. "Preference-based argumentation: Arguments supporting multiple values". *International Journal of Approximate Reasoning*, 2008, 48(3):730-751.

⑪Trevor J. M Bench-capon. "Persuasion in practical argument using value-based argumentation frameworks". *J Logic & Computation*, 2003, 13(3):429-448.

⑫Jérôme Delobelle, Adrian Haret, Sébastien Konieczny, Jean-Guy Mailly, Julien Rossit, and Stefan Woltran. "Merging of abstract argumentation frameworks". In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2016.

⑬Chen, Weiwei, and Ulle Endriss. "Preservation of semantic properties in collective argumentation: The case of aggregating abstract argumentation frameworks." *Artificial Intelligence*, 2019,269:27-48.

⑭Martin Caminada and Gabriella Pigozzi. "On judgment aggregation in abstract argumentation". *Autonomous Agents and Multi-Agent Systems*, 2011, 22(1):64-102.

聚合和论辩推理相结合解决道德困境的研究,在该领域中尚属于创新的工作。

我们从上文的讨论中梳理出未来在偏好聚合和论辩推理相结合的研究方向上主要有三项工作。一是研究满足何种条件,基于某一特定的聚合函数,可以得到一个较为线性的社会偏好,从而基于现有的研究在论辩推理中获得唯一求解。二是通过定义一种新的论辩语义对个体框架的外延进行聚合,并使聚合后的结果与某一特定偏好聚合函数的论辩推理结果在一定条件下建立等价关系。第三则是更具有挑战性的工作:本文探讨的智能体面临的道德困境是利益相关方价值排序不一致场景下的规范冲突。但对于更为高级的智能体,如无人车、机器人管家等,它们在影响环境和社会时还必然受伦理规范的约束。因此,这样的智能体可能不仅面临着各利益相关方关于价值的偏好,还有更宏观层面的关于伦理规范的偏好。两者在智能体的行为规范上也可能存在冲突,例如无人车不能成为所有者报复他人的工具。如何找到伦理偏好与利益相关方价值偏好之间合理的结合方式,并最终形成“社会”的伦理偏好,以及如何采用形式论辩理论进行建模,并通过论辩推理得出符合伦理规范偏好和多数利益方价值偏好的求解,是未来的研究方向。

A Solution to Ethical Dilemmas Based on Preference Aggregation and Formal Argumentation

LIAO Beishui & LI Chonghui

(Department of Philosophy, Zhejiang University, Hangzhou, 310000, China)

Abstract: In the decision-making of an autonomous agent, for an ethical dilemma, different stakeholders have not only different values, but also different preferences over these values. In order to obtain an agreement on the individual preferences of different stakeholders, based on our previous work on argumentation-based ethical decision-making, this paper introduces a revised version of the moral council framework. In this framework, different levels of individual preferences are aggregated into a unified social ordering in terms of two methods from social choice theory, to obtain an ordering over the set of arguments, each of which is associated with a set of values. Then, by using argumentation-based reasoning, solutions of ethical dilemmas are defined.

Key words: formal argumentation; preference aggregation; social choice theory normative system; moral dilemma

(责任校对　钟丽)