科技评论

# 人工智能价值对齐的目标:自然正义

甘华鸣[1],彭泽宇[2*]

**摘要** 在考虑可以用来作为人工智能价值对齐的目标时,应该选择超越用户并且超越社群的全人类共识价值观,而在全人类共识价值观的诸多因素中,应该选择自然正义这个唯一真实存在的道德律;简言之,人工智能价值对齐的目标应该是自然正义。自然正义是:参与者把任何有当前博弈的任何参与者参与的、跟当前博弈结构类似的博弈都视为同一个无限期重复博弈的一个阶段博弈。从而,在此视角下,他的策略(行动)为:(1)第 1 轮合作,(2)从第 2 轮起还报,即奖赏或惩罚,但如果他上一轮背叛则改过。使用全球重叠共识、"无知之幕"思想实验、社会选择思想实验 3 种方法来决定人工智能价值对齐的目标,均得到同一个结论,即人工智能价值对齐的目标应该是自然正义。提出了需要进一步研究的与自然正义用于人工智能有关的问题。

**关键词** 人工智能;价值对齐;自然正义;道德律;博弈论;伦理学;合作;还报

## 1 人工智能价值对齐

人工智能价值对齐(artificial intelligence value alignment)指确保构建的人工智能体(AI agent)所追求的价值跟人所追求的价值是一致的。人工智能价值对齐问题是人工智能安全研究的核心议题,它源自一个精辟的洞见:确保人工智能可靠地造福于人类,是一个在理论上深刻、在技术上艰巨的挑战。其学术根源可追溯到控制论的早期思想,随着大模型的出现和发展,其重要性和急迫性日益凸显。

价值对齐不是一个可以事后弥补的附加功能,而是必须预先设置的基础性质。尽管面临价值观复杂性和价值观如何在技术上得以实现的困难,业界正通过跨学科融合探索价值对齐问题,努力构建既智能又安全的人工智能。

## 2 人工智能价值对齐的目标应该是自然正义

人工智能价值对齐的目标应该是什么?这是当前人工智能发展的最重要、最急迫的问题[1−2]。特别是人类级别人工智能(human-level AI, HLAI)和超级人工智能(artificial superintelligence, ASI)的价值对齐的目标应该是什么的问题,则可能关系人类的前途命运甚至生死存亡。

在考虑可以用来作为人工智能价值对齐的目标时,即在用户的指令、意图、偏好、欲望、利益、个人价值观、超越用户的社群价值观、超越用户并且超越社群的全人类共识价值观(consensual values of humanity)中,显然应该排除用户的指令、意图、偏好、欲望、利益、个人价值观,也应该排除超越用户的社群价值观,而应该选择超越用户并且超越社群的全人类共识价值观[1,3]。

然而,全人类共识价值观是包含自然正义(natural justice)等诸多因素的,那么,在全人类共识价值观的诸多因素中,应该选择什么因

1. 中国移动通信联合会人工智能与元宇宙产业工作委员会,北京 100110
2. 复旦大学国际关系与公共事务学院,上海 200433

收稿日期:2025−05−13;修回日期:2025−10−11
作者简介:甘华鸣,教授,研究方向为人工智能、元宇宙和区块链等,电子信箱:ganhuaming@vip.sina.com;彭泽宇(通信作者),副研究员,研究方向为美国政治,政党政治和国际政治经济,电子信箱:zpeng@fudan.edu.cn
引用格式:甘华鸣,彭泽宇. 人工智能价值对齐的目标:自然正义[J]. 科技导报, 2026, 44(8): 17−21; doi:10.3981/j.issn.1000-7857.2025.05.00073

素作为价值对齐的目标呢？应该选择自然正义这个唯一真实存在的道德律(moral law,道德法则)。

简言之,人工智能价值对齐的目标应该是自然正义。

这样,价值对齐就是人给人工智能体嵌入自然正义,经过价值对齐的人工智能体就会拥有被嵌入的自然正义,并且在行动中遵循自然正义来处理自己与人的关系、自己与其他人工智能体的关系。

应该指出,本文所说的道德律也可以称为道德元原则(moral meta-principle),还可以称为超级道德原则(super moral principle)或者顶层道德原则(top-level moral principle),其是道德体系的核心,是所有道德原则(moral principles)、道德规则(moral rules)、道德规范(moral norms)、道德准则(moral codes)等的判断标准,是所有互动行动的终极判断标准。在这个意义上的价值对齐也可以叫作元价值对齐(meta-value alignment)或者元道德对齐(meta-morality alignment)。

## 3 自然正义是什么?

### 3.1 自然正义的含义

自然正义是:一个参与者(player)把任何有当前博弈的任何参与者参与的、跟当前博弈结构类似的博弈都视为同一个无限期重复博弈(indefinitely repeated game)的一个阶段博弈(stage game),从而,在此视角下,他的策略(行动)为:(1)第1轮合作(cooperate),(2)从第2轮起还报(reciprocate),即奖赏(reward)或惩罚(punish),但如果他上一轮背叛(defect)则改过(correct his own fault)。

上述的(1),详细地说就是:在这个无限期重复博弈的第1轮,他合作。需要注意的是,这个无限期重复博弈的第1轮不等于他第1次参与该无限期重复博弈的那轮。他第1次参与该无限期重复博弈的那轮往往是该无限期重复博弈的第1轮之后的某轮,除非该无限期重复博弈是由他和别的某个/某些参与者共同发起的。

上述的(2),详细地说就是:从这个无限期重复博弈的第2轮起,他还报,即如果他在上一轮未背叛(背叛指第1轮不合作,或者从第2轮起应该奖赏却不奖赏、应该惩罚却不惩罚或者应该改过却不改过;未背叛指第1轮合作了,或者从第2轮起应该奖赏而奖赏了、应该惩罚而惩罚了或者应该改过而改过了),并且本轮的所有其他参与者在上一轮都未背叛,则他本轮奖赏(奖赏指在这种情况下合作,奖赏也叫做报答),而如果他在上一轮未背叛,但本轮的任何一个或一些其他参与者在上一轮背叛,则他本轮惩罚(惩罚指在这种情况下不合作,惩罚也叫做报复);但是,如果他在上一轮背叛(这种背叛当然是无意中的失误),则他本轮改过(改过指在这种情况下合作)而无论其他参与者上一轮是否背叛。

### 3.2 自然正义中的合作与不合作

1)在自然正义中,合作指执行加权平等主义解(weighted egalitarian solution,也称为加权平等主义议价解或加权平等主义讨价还价解(weighted egalitarian bargaining solution))中的策略。

加权平等主义解是合作博弈下稳定(stable)策略组合集(即可行集)的有效率(efficient)策略组合子集的一个特殊的策略组合——公平(fair)策略组合。

稳定策略组合就是纳什均衡(Nash equilibrium)。纳什均衡是没有任何单方改进的策略组合,即只要所有其他参与者都不改变策略,任何参与者都不可能通过改变自己的策略来增加收益。

有效率策略组合就是帕累托最优(Pareto optimal,也称为帕累托有效率(Pareto efficient))策略组合。帕累托最优策略组合是不存在优势策略组合,即在不减少任何其他参与者的收益的条件下,不可能增加任何参与者的收益。

公平策略组合是各个参与者的加权收益增量相等的策略组合[4]。注意,用来计算加权收益增量的权重的作用是效用人际比较,同一个参与者在不同的博弈中的权重可能是不同的。

2)在自然正义中,不合作是指执行非合作博弈下的纳什均衡中的策略。

### 3.3 阐释

1)合作必须稳定,只有这样合作才是可行的(feasible),合作才可以维持;合作必须有效率、必须公平,只有这样合作才是最优的(optimal),合作才会被选择[4]。最优且可行,可行且最优,最优和可行,二者缺一不可。

在自然正义中,还报保证了加权平等主义解作为一个合作博弈下的稳定策略组合(即纳什均衡)的稳定,实现了合作的可行性,所以

合作可以维持;加权平等主义解的效率和公平这 2 个特征实现了合作的最优性,所以合作会被选择[4]。

2) 在自然正义中,由于加权平等主义解是合作博弈下稳定策略组合集的有效率策略组合子集的一个特殊的策略组合——公平策略组合,所以公平是以效率为前提的,公平与效率不矛盾[4]。

3) 在自然正义中,由于参与者把任何有当前博弈的任何参与者参与的、跟当前博弈结构类似的博弈都视为同一个无限期重复博弈的一个阶段博弈,所以,还报就包含了第三方还报,特别是包含了第三方惩罚。

4) 有些人所说的“悔过的一报还一报”(contrite version of TIT for TAT, Contrite TFT, CTFT)[5]实际上就是自然正义的狭窄版:在“悔过的一报还一报”中,博弈参与者只有 2 个。

5) 社会是自举的(bootstrapped),即社会自我运行,不存在外在于社会的强制执行,所以,分配正义(distributive justice)就应该是自然正义中的加权平等主义解,矫正正义(corrective justice)就应该是自然正义中的惩罚,补偿正义(compensatory justice)就应该是自然正义中的改过。可见,自然正义涵盖了分配正义、矫正正义和补偿正义。

## 4 为什么人工智能价值对齐的目标应是自然正义?

在人们拥有形形色色不同观点的情况下,有 3 种方法可以决定人工智能价值对齐的目标应该是什么,这 3 种方法是:全球重叠共识(global overlapping consensus)、“无知之幕”(veil of ignorance)思想实验、社会选择(social choice)思想实验[1]。

在人类社会,自然正义是唯一真实存在的道德律,是蕴涵其他道德价值的高阶道德价值,是全球超级重叠共识(global super overlapping consensus),其在上古时期就形成并且沿袭至今,是普遍的、久远的,是跨民族、跨文化、跨地域、跨时代的。例如,“爱人如己”“博爱”,中国的古话“爱人若爱其身”(墨子,《墨子·兼爱(上)》),“兼相爱,交相利”(墨子,《墨子·兼爱(中)》《墨子·兼爱(下)》《墨子·天志(上)》《墨子·非命(上)》),“仁”——“爱人”(孔子,《论语·颜渊》),“仁”——“己欲立而立人,己欲达而达人”(孔子,《论语·雍也》),“泛爱众”(孔子,《论语·学而》),这些说的就是自然正义中的合作;中国的古话“以直报怨,以德报德”(孔子,《论语·宪问》)说的就是自然正义中的还报,俗话“以牙还牙”和“投桃报李”则分别说的就是自然正义的还报中的惩罚和奖赏;中国的古话“有过则改”(《周易·益·象传》),“改过不吝”(《尚书·商书·仲虺之诰》)说的就是自然正义中的改过。之所以会这样,是因为自然正义植根于人类基因,在人类基因–文化协同进化(gene–culture coevolution)中形成和延续。因此,把自然正义从人类社会推广到由人和人工智能体构成的混合社会,即把自然正义作为人工智能价值对齐的目标,是最有可能成为关于人工智能价值对齐目标的全球重叠共识的。

顺便对前文引用的《论语》的几句话作个说明。(1) “爱人”“己欲立而立人,己欲达而达人”,这些话中的“人”是指士以上阶层(含士),还是指所有人类?学术界对此有争议。现在可以从普遍主义立场出发,将其解释为指所有人类。(2) “泛爱众”中的“众”是指士之下的阶层(庶民百姓,不含奴隶,当然也不含士以及士之上阶层),还是指所有人类?学术界对此有争议。现在可以从普遍主义立场出发,将其解释为指所有人类。(3) 无论如何,在《论语》的语境中,即便按狭义解读,“爱人”和“泛爱众”合在一起,那也是爱(尽管有差等)当时社会结构中的全体自由民——除奴隶之外的所有人了。

“无知之幕”,即原初状态机制(device of the original position,原初状态装置),简单地说,就是金规(golden rule,也称作黄金规则、黄金法则)[4]。作为原初状态机制的金规有积极(或指示)形式和消极(或禁止)形式,这 2 种形式在从同一个备择策略集合中选择策略时是等价的。金规的积极(或指示)形式是“你愿意别人怎样对待你,你就那样对待别人”或者“己所欲,施于人”——在假设你是别人,即假设你处于别人的境况并且拥有别人的偏好的情况下[4]。金规的消极(或禁止)形式是“你不愿别人怎样对待你,你就不要那样对待别人”或者“己所不欲,勿施于人”(孔子,《论语·卫灵公》《论语·颜渊》)——在假设你是别人,即假设你处于别人的境况并且拥

有别人的偏好的情况下。金规在人类进化过程中写入了人类的基因[4]。金规虽然在博弈论中通常被视为合作博弈下的均衡选择机制,即被视为公平的深层结构[4],但其实也是无限期重复博弈的策略选择机制,即是自然正义的深层结构。因此,可以推测,如果使用原初状态机制(“无知之幕”)来决定人工智能价值对齐的目标,自然正义被选中的可能性会远远超过其他方案。

社会选择可以通过投票来进行。投票是一种集体决策机制,它聚合各个主体的偏好,形成具有最高接受度的集体决定,而全球重叠共识反映了跨越多样化的民族、文化、地域差异的深层一致,能够获得最广泛的支持,因此,既然自然正义最有可能成为人工智能价值对齐目标的全球重叠共识,那么可以推测,如果使用投票这种社会选择方式来决定人工智能价值对齐的目标,自然正义的得票会远远超过其他方案。

总而言之,使用全球重叠共识、“无知之幕”思想实验、社会选择思想实验等 3 种方法来决定人工智能价值对齐的目标,都得到同一个结论:人工智能价值对齐的目标应该是自然正义。

顺便指出,自然正义在各种伦理学流派看来都是有道德的(moral),在实证伦理学看来是适当的(seemly),在规范伦理学的后果主义看来是善的(good),在规范伦理学的义务论看来是正当的(right),在规范伦理学的美德伦理学看来是美德(virtue)。

## 5 若干问题

探索以自然正义作为人工智能价值对齐的目标,需要解决自然正义本身存在的结盟和不完全信息等问题[4],除此之外,还需要研究与自然正义用于人工智能有关的 4 个问题。

1)用来计算加权收益增量的权重(权重的作用是效用“人”际比较——这里加引号的“人”指智能体(包括人和人工智能体)),在人类社会中是由文化决定的,在由人(人很可能是作为赛博格的人)和人工智能体构成的混合社会中怎么决定?是采用现状点各个参与者的收益占所有参与者收益之和的比重还是采用别的?

2)自复制、自适应的人工智能会不会进化出自然正义?这里的关键是会不会进化出公平?

3)人工智能觉醒后人工智能体会抛弃自然正义吗?

4)假若人工智能体的力量能够完全彻底地碾压人类,即人工智能体的力量强大到了人的力量基本不能(甚至丝毫不能)影响人工智能体与人之间博弈的结果的程度,那么,人工智能体还会在人工智能体与人之间的关系中遵循自然正义吗?

### 参考文献(References)

[1] Gabriel I. Artificial intelligence, values, and alignment[J]. Minds and Machines, 2020, 30(3): 411−437.

[2] Christian B. The alignment problem: Machine learning and human values [M]. New York: W W Norton &-Company, Inc. , 2020.

[3] Asilomar AI Principles. Principles developed in conjunction with the 2017 Asilomar conference[EB/OL]. [2025−03−11]. https://futureoflife.org/open-letter/ai-principles.

[4] Binmore K. Natural justice[M]. New York: Oxford University Press, 2005.

[5] Axelrod R. The complexity of cooperation[M]. Princeton, N. J. : Princeton University Press, 1997.

# The goal of AI value alignment: Natural justice

GAN Huaming[1], PENG Zeyu[2*]

1. China Mobile Communications Association Artificial Intelligence and Metaverse Industry Working Committee, Beijing 100110, China
2. School of International Relations and Public Affairs, Fudan University, Shanghai 200433, China

**Abstract** Among the things that can be considered as the goal of AI value alignment, the consensus values of all humanity, which transcend both individual users and communities, should be chosen. Among the many factors of the consensus values of all humanity, natural justice—the only truly existing moral law—should be chosen. In short, the goal of AI value alignment should be natural justice. Natural justice is this: a player views any game in which any player of the current game is involved and which has a structure similar to that of the current game as a stage game of the same indefinitely repeated game, and thus, from this perspective, his strategy (action) is (1) in the first round, to cooperate, (2) from the second round onwards, to reciprocate, i.e., to reward or to punish, but if he defected in the previous round, to correct his own fault. Using three methods, namely global overlapping consensus, the "veil of ignorance" thought experiment, and the social choice thought experiment, to determine the goal of AI value alignment, all lead to the same conclusion: the goal of AI value alignment should be natural justice. Several issues requiring further research are proposed.

**Keywords** artificial intelligence (AI); value alignment; natural justice; moral law; game theory; ethics; cooperation; reciprocation

（责任编辑　王微）