夸耀有益吗？论反事实公平信用

IlanDoron-Arad Elchanan Mossel麻省理工学院
麻省理工学院

2026年11月24日

arXiv:2611.29083v1 [cs.GT] 21 Nov 2026

摘要

智力工作（如软件、法律文件和科学结果）通常建立在他人的工作之上，而且同一最终产品往往可以通过多种不同方式获得。在为这类工作分配功劳时，尚不清楚应依据工作的实际产出方式，还是也应考虑获得该成果的反事实方式。我们通过将 Shapley 值推广到能够考虑获得同一结果的反事实方式，来研究这一问题。我们证明，效率、零性和次模性（尤其是可加性）这些公平性公理，与激励创建者最大化社会福利彼此不相容。事实上，公平的功劳分配甚至可能只激励产生相对于最优值任意小的福利。我们还证明，若不要求次模性，则可以实现激励相容性；而四条经典公平性公理唯一确定了 Shapley 值，并排除了对反事实结果的依赖。

智力工作很少是从零开始创作的，同一个最终成果通常也可以通过不止一种方式获得。软件工程师可以通过基于现有的开源库构建功能来交付一个特性，也可以在内部重新实现该功能。律师可以从同事的判例出发起草合同，也可以从基本原理开始。科学家可以依赖某一部分早期工作来确立一个成果，也可以依赖另一部分早期工作。对于后一种情况，一个经典例子是：1735年，欧拉 [4] 利用牛顿恒等式 $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6$ 推导出了著名的巴塞尔问题恒等式，但后来有人又给出了一个基于勾股定理的简单几何证明 [5] ，证明了同样的结果。

这引出了一个问题：在向那些为一项智力工作作出贡献的人分配荣誉时，我们是否只应奖励那些实际被采用了其贡献的人（开源库、判例和牛顿的作者），还是也应奖励那些为同一成果提供替代路径的人（毕达哥拉斯）？ 此外，公平的荣誉分配是否能够激励创建者选择最大化社会福利的路径，也就是选择能够为社会带来最大价值的路径（例如最易维护的代码、最清晰的合同、最简单的证明）？ 换言之，荣誉是否应当追随实际贡献，还是也应考虑反事实方式

获得这一结果的可能性如何？答案并非不言自明。

Shapley 的开创性工作 [1] 表明，存在一种唯一的信用分配方式，能够满足四项基本公平性公理：效率、可加性、零性和对称性。在本文中，我们将 Shapley 的信用共享模型推广为能够核算反事实信用的模型。我们证明，信用分配的基本公平性公理与激励创建者最大化社会福利彼此不相容。事实上，公平信用可能唯一地激励出最糟糕的结果，其福利相对于最优结果可以任意小，并且无法迫使创建者选择使其边际贡献最小的选项。即使在考虑反事实替代方案时采用比 Shapley [1] 所使用的公理更弱的公理，这一结论仍然成立。下面我们将讨论相关的先前工作；随后定义模型，并陈述和证明我们的结果。

相关工作

信用分配规则将夏普利值 [1] 推广为随机顺序值 [6], 、受优先关系约束的值 [3], 以及非对称变体和因果变体 [7, 9]。科学信用分配也曾利用引文模式进行研究 [2]。反事实信用这一概念在强化学习中得到研究，其中替代行动被用于定义学习信号 [10, 11, 12,13, 14]。相比之下，我们要探讨的是：在实际结果的所有联盟价值保持不变的情况下，信用是否可以取决于反事实替代方案。这并不排除使用反事实替代方案来定义博弈本身。归因领域的其他不可能性结果关注特征归因解释模型行为的能力 [19]。其他相关工作研究了研究团队中的信用与激励 [16, 17], 、算法进展的信用 [15], 以及系统指标变化的因果归因 [18]。这些工作均未研究，获得同一结果的不同替代方式会如何影响对实际构造的信用分配。

模型与定义

固定参与者 $N = \{0, 1, \ldots, n\}$, $n \geq 1$，其中 0 是创建者 参与者；并设 X 是参与者 0 获得其结果的所有可能方式构成的有限非空集合。每个 $x \in X$ 都代表同一结果的一种不同构建方式（例如，某个软件功能的不同实现、某份合同的不同草稿，或某个定理的不同证明）。令 $\mathcal{V}$ 为所有归一化的单调价值 函数 $v\colon 2^N \times X \to \mathbb{R}_{\geq 0}$ 构成的集合，即满足以下条件的函数

$$\begin{aligned} &\bullet\ (\forall x \in X) && v(\varnothing, x) = 0, \\ &\bullet\ (\forall x \in X)\quad (\forall S \subseteq T \subseteq N) && v(S, x) \leq v(T, x). \end{aligned}$$

总价值 $v(N,\ x)$ 可能因构建方式 $x \in X$ 而异。我们将结果 $x \in X$ 的社会福利称为 $v(N,\ x)$。对于智力工作的署名，以上假设至关重要，因为我们将函数 v 视为参与者在 $S \subseteq N$ 中的贡献为获得 $x \in X$ 所作贡献的度量。事实上， $S = \emptyset$ 没有贡献，任何 S 的贡献值均为非负，且随着集合 S 的增大，贡献值只能增加。将价值函数相加记为

$$(\forall v, w \in \mathcal{V})\,(\forall S \subseteq N)\,(\forall x \in X) \qquad (v + w)(S, x) = v(S, x) + w(S, x).$$

不同的函数 $w \in \mathcal{V}$ 可被视为代表不同的结果，而 $v + w$代表它们的合并价值。我们用 $r \in X$ 表示参与者 0 的实际构造；其他元素 $X \setminus \{r\}$ 则是反事实构造。定义 i在 S 的基础上对 x所贡献的剩余价值为

$$(\forall i \in N)\,(\forall x \in X)\,(\forall S \subseteq N \setminus \{i\}) \qquad \Delta_i v(S, x) = v(S \cup \{i\}, x) - v(S, x).$$

公理。信用规则是 $\phi\colon\ \mathcal{V} \times X \to \mathbb{R}^N_{\geq 0}$。考虑以下信用规则必须满足的基本公平性公理。第一条公理称为零性，它断言，在任何反事实选项下都没有贡献的参与者不会获得功劳：

$$(\forall v \in \mathcal{V})(\forall r \in X)(\forall i \in N) \quad [(\forall x \in X)(\forall S \subseteq N \setminus \{i\})\,\Delta_i v(S, x) = 0] \implies \phi_i(v, r) = 0. \quad \text{(N)}$$

这将 Shapley 在 $|X| = 1$标准设定下的零性公理推广了。更基本的公平性公理包括效率和可加性，分别表示所有信用都分配给参与者，且信用在价值函数上具有可加性：

$$(\forall v \in \mathcal{V})(\forall r \in X) \quad \sum_{i \in N} \phi_i(v, r) = v(N, r), \tag{E}$$

$$(\forall v, w \in \mathcal{V})(\forall r \in X) \quad \phi(v + w, r) = \phi(v, r) + \phi(w, r). \tag{A}$$

前两条公理（即 (N) 和 (E)）在信用分配中非常自然：完全没有贡献的参与者不会获得信用，而且所有信用都会被分配。然而，可加性较难论证，因为价值函数之间可能存在依赖关系。因此，我们考虑一种较弱的可加性形式，称为次模性，它要求信用作为价值函数的函数满足边际收益递减性质。令 $v \vee w$ 和 $v \wedge w$ 分别表示逐点最大值和逐点最小值：

$$(\forall v, w \in \mathcal{V})\,(\forall S \subseteq N)\,(\forall x \in X) \qquad (v \vee w)(S,x) = \max\{v(S,x), w(S,x)\},$$
$$(\forall v, w \in \mathcal{V})\,(\forall S \subseteq N)\,(\forall x \in X) \qquad (v \wedge w)(S,x) = \min\{v(S,x), w(S,x)\}.$$

请注意，这两种运算在 $\mathcal{V}$下都封闭。信用向量之间的大小关系按分量比较。现在，来看次模性 公理：

$$(\forall v, w \in \mathcal{V})\,(\forall r \in X) \qquad \phi(v \vee w, r) + \phi(v \wedge w, r) \leq \phi(v, r) + \phi(w, r). \tag{SM}$$

这体现了信用归因中的边际收益递减：重叠的贡献合并考虑时，不应产生额外信用。可加性（A）是（SM）的一个特例：由于 $(v \vee w) + (v \wedge w) = v + w$，根据（A），（SM）中的等式成立。

激励相容信用函数。 以下假设历史和值函数 $v \in \mathcal{V}$ 保持不变，参与者 0 可以选择任意 $r \in X$，其效用为 $\phi_0(v, r)$。若参与者 $i \in N$ 对 $v \in \mathcal{V}$ 有贡献，则称其为有影响力 的参与者：

$$(\exists x \in X)\,(\exists S \subseteq N \setminus \{i\}) \qquad \Delta_i v(S,x) > 0.$$

若参与者 0 的激励与最大化社会福利的目标一致（只要她具有影响力），则信用规则 ϕ 是激励相容 的（IC）：

$$(\forall v \in \mathcal{V} \text{ such that player 0 is influential for } v) \quad \arg\max_{r \in X} \phi_0(v,r) \subseteq \arg\max_{r \in X} v(N,r). \tag{IC}$$

请注意，影响力条件是必要的，因为根据（N），没有影响力的参与者在每种构造中获得的信用都为零，即使这些构造具有不同的社会价值。

依赖反事实的信用。 定义价值函数对现实（非反事实）结果的限制：

$$(\forall v \in \mathcal{V})\,(\forall r \in X)\,(\forall S \subseteq N)\,(\forall x \in X) \qquad v_r(S,x) = v(S,x)\mathbf{1}_{\{x=r\}}.$$

注意，对于 $v \in \mathcal{V}$，函数 v_r 恰好构成 Shapley [1] 标准模型中仅有一个可能结果 $|X| = 1$ 时的归一化单调函数集合。信用规则 ϕ 若满足反事实相关，则

$$(\exists v \in \mathcal{V})\,(\exists r \in X) \qquad \phi(v,r) \neq \phi(v_r, r),$$

且 否则为反事实无关，也就是说，当

$$(\forall v \in \mathcal{V})\,(\forall r \in X) \qquad \phi(v,r) = \phi(v_r, r).$$

结果

我们的主要结果是，在满足 $|X| \geq 2$ 的反事实情境中，效率和零性这两个基本公平性公理与次模性相结合，会阻止信用分配遵循激励相容性这一性质。由于 (A) 特别属于 (SM)，因此也可以将该表述中的次模性替换为可加性。

定理 1。假设 $|X| \geq 2$。

1. 每个满足 *(N)* 的（IC）信用规则都是反事实相关的。

2. 每个满足 *(E)* 和（SM）的规则都是反事实无关的。

因此，(E)、(SM) 和 (N) 排除了 (IC)。更强地说，我们的证明表明，参与者的激励与公平性公理相结合，可能导致最优社会福利的比例任意小。形式化地，对于每个 $\varepsilon > 0$，都存在 $v \in \mathcal{V}$，使得参与者 0 具有影响力，并且存在 $b \in X$，使得每个满足 (E)、(SM) 和 (N) 的信用规则 ϕ 都满足$0 < v(N,\ b) < \varepsilon \max_{r \in X} v(N,\ r)$，并且

$$\arg\max_{r \in X} \phi_0(v, r) = \arg\min_{r \in X} v(N, r) = \{b\},$$
$$\arg\min_{r \in X} \phi_0(v, r) = \arg\max_{r \in X} v(N, r) = X \setminus \{b\}.$$

证明。对于 (1)，假设信用函数 ϕ 满足 (N) 且为反事实无关。固定$\varepsilon > 0$、 $b \in X$和 $0 < \delta < \min\{1, \varepsilon\}$。定义

$$(\forall S \subseteq N)\,(\forall x \in X) \qquad v(S, x) = \delta \mathbf{1}_{\{x = b,\, 0 \in S\}} + \mathbf{1}_{\{x \neq b,\, 1 \in S\}}.$$

对于每个 $x \in X$，函数 $v(\cdot,\ x)$ 都是归一化、非负且单调的，因此有 $v \in \mathcal{V}$。此外，$\Delta_0 v(\varnothing,\ b) = \delta > 0$，所以参与者 0 具有影响力。回顾 $v_r(S,\ x) = v(S,\ x)\mathbf{1}_{\{x=r\}}$。当 $r \neq b$ 时，参与者 0 在 v_r 中是无关参与者：

$$(\forall r \in X \setminus \{b\})\,(\forall x \in X)\,(\forall S \subseteq N \setminus \{0\}) \qquad \Delta_0 v_r(S, x) = 0.$$

因此，(N) 和反事实独立性给出

$$(\forall r \in X \setminus \{b\}) \qquad \phi_0(v, r) = \phi_0(v_r, r) = 0.$$

由于 $\phi_0(v,\ b) \geq 0$，我们有 b 使参与者 0 的贡献最大，尽管

$$(\forall r \in X \setminus \{b\}) \qquad v(N, b) = \delta < 1 = v(N, r).$$

因此，(IC) 不成立。若 (E) 也成立，则除 0 以外的每位参与者在 v_b 中都是无关参与者，即

$$(\forall i \in N \setminus \{0\})\,(\forall x \in X)\,(\forall S \subseteq N \setminus \{i\}) \qquad \Delta_i v_b(S, x) = 0,$$

所以 (N) 给出，对于每个 $i \in N \setminus \{0\}$，$\phi_i(v_b,\ b) = 0$。因此，

$$\phi_0(v, b) = \phi_0(v_b, b) = \sum_{i \in N} \phi_i(v_b, b) = v_b(N, b) = \delta.$$

因此，b 使参与者 0 的贡献唯一最大，并使福利唯一最小；所有其他结果都使她的贡献最小、福利最大。此外，

$$0 < v(N, b) = \delta < \varepsilon \max_{r \in X} v(N, r).$$

接下来，我们证明 (2)，这将在 (E)、(SM) 和 (N) 条件下证明一个更强的推论。为证明 (2)，假设 (E) 和 (SM) 成立。固定 $v \in \mathcal{V}$ 和 $r \in X$。 v_r 和 $v - v_r$ 均属于 $\mathcal{V}$：对每个$x \in X$，函数 $v_r(\cdot,\ x)$ 和 $(v - v_r)(\cdot,\ x)$，要么分别等于 $v(\cdot,\ x)$ 和零（当 $x = r$ 时），要么分别为零和 $v(\cdot,\ x)$。因此，两者均属于 $\mathcal{V}$。由于 $v_r(N,\ r) = v(N,\ r)$，我们有

$$(v - v_r)(N, r) = v(N, r) - v_r(N, r) = 0.$$

因此，根据信用规则 ϕ 的效率和非负性，我们有

$$\sum_{i\in N}\phi_i(v-v_r,r)=(v-v_r)(N,r)=0.$$

这意味着

$$(\forall i\in N)\qquad \phi_i(v-v_r,r)=0.$$

此外，$v_r(S,x)$ 和 $(v-v_r)(S,x)$ 至多有一个非零：

$$(\forall S\subseteq N)\,(\forall x\in X)\qquad v_r(S,x)=0\ \text{ or }\ (v-v_r)(S,x)=0.$$

因此，$v_r\vee(v-v_r)=v$ 和 $v_r\wedge(v-v_r)=0$。逐坐标应用 (SM)，得到

$$\begin{aligned}(\forall i\in N)\qquad \phi_i(v,r)+\phi_i(0,r)&\le\phi_i(v_r,r)+\phi_i(v-v_r,r)\\&=\phi_i(v_r,r).\end{aligned}$$

由 (E) 和非负性可知，$\sum_{i\in N}\phi_i(0,r)=0$，因此对每个 $i\in N$，都有 $\phi_i(0,r)=0$。因此，

$$(\forall i\in N)\qquad \phi_i(v,r)\le\phi_i(v_r,r).$$

另一方面，效率以及 $v_r(N,r)=v(N,r)$ 可得

$$\sum_{i\in N}\big(\phi_i(v_r,r)-\phi_i(v,r)\big)=v_r(N,r)-v(N,r)=0.$$

每一项都非负，因此每一项都为零。因此

$$(\forall i\in N)\qquad \phi_i(v,r)=\phi_i(v_r,r).$$

由于 $v\in\mathcal{V}$ 和 $r\in X$ 是任意的，这就证明了 (2)。 □

这一不可能性也适用于作为特殊情形的时间顺序信用 [3], ，因为它满足 (E)、(A) 和 (N)，因此也满足 (SM)。

最小化创建者的信用。假设存在一种更简单的情形，其中总价值$v(N,\ x)$与构造 x无关。在这种情况下，社会希望将授予当前创建者的信用降至最低，而创建者则希望将其最大化。定理 1(2)表明，这些目标可能完全对立。

推论1。设 $|X|\ge 2$。存在 $v\in\mathcal{V}$ 和 $b\in X$，使得 $v(N,x)=1$ 对于每个 $x\in X$都成立，并且每个满足 (E)、(SM) 和 (N) 的规则 ϕ 都满足

- $\arg\max_{r\in X}\phi_0(v,\ r)=\{b\}$ 和 $\max_{r\in X}\phi_0(v,\ r)=1$。
- $\arg\min_{r\in X}\phi_0(v,\ r)=X\setminus\{b\}$ 和$\min_{r\in X}\phi_0(v,\ r)=0$。

证明。固定 $b\in X$ 并定义

$$(\forall S\subseteq N)\,(\forall x\in X)\qquad v(S,x)=\mathbf{1}_{\{x=b,\,0\in S\}}+\mathbf{1}_{\{x\ne b,\,1\in S\}}.$$

这属于 $v\ \mathcal{V}$，并且对于所有 $x\in X,\ v(N,x)=1$。根据定理 1(2)，每个满足 (E) 和 (SM) 的规则 ϕ，对于每个 $r\in X$，都有 $\phi_0(v,r)=\phi_0(v_r,r)$。在 v_b 中，每个参与者 $i\ne 0$ 都是无关参与者；因此，(N) 和 (E) 蕴含 $\phi_0(v,b)=\phi_0(v_b,b)=v_b(N,b)=1$。对于每个 $r\ne b$，参与者 0 在 v_r 中是无关参与者，因此 (N) 给出 $\phi_0(v,r)=\phi_0(v_r,r)=0$。所声称的两个等式均由此得出。 □

没有次模性。 如果没有 (SM)，公理 (E)、(N) 和 (IC) 可以同时成立（自然地，比 (SM) 更强的 (A) 也必须舍弃）。从较高层面看，这类似于常见的致谢实践，例如科学引文或软件依赖列表：相关的先前工作会得到认可，但不会为个人贡献分配数值权重。

观察1。 存在一种同时满足(E)、(N)和(IC)的信用规则。

证明。 令 $A(v)$ 表示关于 v 的有影响力的参与者集合：

$$(\forall v \in \mathcal{V}) \qquad A(v) = \big\{ i \in N : \ (\exists x \in X)\, (\exists S \subseteq N \setminus \{i\}) \ \ \Delta_i v(S,x) > 0 \big\}.$$

定义

$$(\forall v \in \mathcal{V})\, (\forall r \in X)\, (\forall i \in N)$$
$$\phi_i(v,r) = \begin{cases} \dfrac{v(N,r)}{|A(v)|}, & A(v) \neq \varnothing \text{ and } i \in A(v), \\ 0, & \text{otherwise.} \end{cases}$$

由于 $v(N,\ r) \geq 0$，这定义了一个非负贡献分配规则。固定 $v \in \mathcal{V}$ 和 $r \in X$。如果 $A(v) = \varnothing$，则

$$(\forall i \in N)\, (\forall x \in X)\, (\forall S \subseteq N \setminus \{i\}) \qquad \Delta_i v(S,x) = 0.$$

令 i_1，.....，i_{n+1} 为 N 的一个排序，并令 $S_k = \{i_1, \ldots, i_k\}$ 对每个 $k \in \{1, \ldots, n+1\}$ 都满足 $S_0 = \varnothing$。因此

$$v(N,r) = v(\varnothing, r) + \sum_{k=1}^{n+1} \Delta_{i_k} v(S_{k-1}, r) = 0.$$

因此，$\sum_{i \in N} \phi_i(v,\ r) = 0 = v(N,\ r)$。如果 $A(v) \neq \varnothing$，则

$$\sum_{i \in N} \phi_i(v,r) = \sum_{i \in A(v)} \frac{v(N,r)}{|A(v)|} = v(N,r).$$

因此，(E) 在两种情况下都成立。如果 (N) 的前提对参与者 $i \in N$ 成立，即，

$$(\forall x \in X)\, (\forall S \subseteq N \setminus \{i\}) \qquad \Delta_i v(S,x) = 0,$$

则 $i \notin A(v)$。因此，对每个 $r \in X$，$\phi_i(v,\ r) = 0$，从而证明 (N)。最后，如果参与者 0 有影响力，则 $0 \in A(v)$ 和 $|A(v)| > 0$。因此

$$(\forall r \in X) \qquad \phi_0(v,r) = \frac{v(N,r)}{|A(v)|}.$$

由于 $|A(v)| > 0$ 与 r 无关，我们得到

$$\arg\max_{r \in X} \phi_0(v,r) = \arg\max_{r \in X} v(N,r).$$

因此，(IC) 成立。 □

若不满足 (SM)，唯一性通常无法成立：例如，在有影响力的参与者之间对 $v(N,r)$ 等额分配，以及在这些参与者之间按 $1 + \sum_{x \in X} v(\{i\},x)$ 的比例分配，都满足 (E)、(N) 和 (IC)。因此，需要额外公理来刻画唯一的信用规则。

反事实世界中 Shapley 的唯一性。上述观察并不能证明唯一性。现在我们将说明，加入对称性这一额外公理后，即使存在反事实替代方案，Shapley 的经典唯一性结果 [1] 仍然成立。根据定理 1，这种信用分配当然是反事实无关的。考虑额外的对称性公理。对于任意 $v \in \mathcal{V}$、$r \in X$，以及互不相同的 $i, j \in N$，对称性要求

$$\Big[(\forall x \in X)(\forall S \subseteq N \setminus \{i,j\})\, \Delta_i v(S,x) = \Delta_j v(S,x)\Big] \implies \phi_i(v,r) = \phi_j(v,r). \tag{S}$$

这四条公理 (E)、(A)、(N) 和 (S) 被视为信用分配的公平性公理 [1]。将实现博弈的 Shapley 值定义为

$$(\forall v \in \mathcal{V})\,(\forall r \in X)\,(\forall i \in N) \qquad \mathrm{Sh}_i(v(\cdot,r)) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \cdot \Delta_i v(S,r).$$

我们证明，即使存在反事实替代方案，实现博弈的 Shapley 值仍是唯一的公平信用分配。请注意，根据定理 1，在存在反事实结果时，这种分配并不满足 (IC) $|X| \geq 2$。

推论2. 满足 (E)、(A)、(N) 和 (S) 的唯一信用规则 ϕ为

$$(\forall v \in \mathcal{V})\,(\forall r \in X)\,(\forall i \in N) \qquad \phi_i(v,r) = \mathrm{Sh}_i(v(\cdot,r)).$$

证明。 根据定理 1(2)（若 $|X| = 1$，则显然成立），只需考虑实现博弈。Shapley 的唯一性结果 [1]适用于非负单调博弈；参见 [8, 第 2]节。 □

经费支持。 IDA 获 NSF DMS-2031883 资助，并获范内瓦·布什教职奖学金 ONR-N00014-20-1-2826（PI Mossel）支持。EM 部分获 NSF DMS-2031883、范内瓦·布什教职奖学金 ONR-N00014-20-1-2826、MURI N000142412742 以及西蒙斯研究员奖资助。

AI训练。 作者不授权在本文上进行AI训练。

参考文献

[1] L. S. Shapley。 n人博弈的一个值。载于 H. W. Kuhn 和 A. W. Tucker（编），《博弈论贡献》第二卷，第307–318页。普林斯顿大学出版社，1953年。[2] H.-W. Shen 和 A.-L. Barab´asi。科学中的集体贡献分配。《美国国家科学院院刊》，111(34):12325–12330，2014年。[3] U. Faigle 和 W. Kern。在优先约束下合作博弈的Shapley值。《国际博弈论杂志》，21:249–266，1992年。[4]L. Euler。《倒数级数之和》。《圣彼得堡科学院学报》，7:123–134，1740年。发表于1735年。[5] J. W¨astlund。利用欧几里得几何求平方倒数之和。查尔姆斯理工大学，2010年。https://www.math.chalmers.se/~wastlund/Cosmic.pdf。[6] R. J. Weber。博弈的概率值。载于 A. E. Roth（编），《*Shapley*值：献给*Lloyd S. Shapley*的论文集》，第101–120页。剑桥大学出版社，1988年。[7] C. Frye、C. Rowat 和 I. Feige。非对称Shapley值：将因果知识纳入模型无关可解释性。载于《神经信息处理系统进展》，第33卷，2020年。[8] R. J. Aumann 和 L. S. Shapley。《非原子博弈的值》。普林斯顿大学出版社，1974年。[9] T. Heskes、E. Sijben、I. G. Bucur 和 T. Claassen。因果Shapley值：利用因果知识解释复杂模型的个体预测。载于《神经信息处理系统进展》，第33卷，第4778–4789页，2020年。[10] J. Foerster、G. Farquhar、T. Afouras、N. Nardelli 和 S. Whiteson。反事实多智能体策略梯度。载于《*AAAI*人工智能会议论文集》，第32卷，第2974–2982页，2018年。[11] T. Mesnard、T. Weber、F. Viola、S. Thakoor、A. Saade、A. Harutyunyan、W. Dabney、T. S. Stepleton、N. Heess、A. Guez、E. Moulines、M. Hutter、L. Buesing 和 R. Munos。无模型强化学习中的反事实贡献分配。载于《第*38*届国际机器学习会议论文集》，《style id='36'>机器学习研究论文集第139卷，第7654–7664页，2021年。[12] A. Meulemans、S. Schug、S. Kobayashi、N. Daw 和 G. Wayne。如果是我，我会得到那个奖励吗？通过反事实贡献分析进行长期贡献分配。载于《神经信息处理系统进展》，第36卷，2023年。[13] J. Li、K. Kuang、B. Wang、F. Liu、L. Chen、F. Wu 和 J. Xiao。用于多智能体强化学习的Shapley反事实贡献。载于《第*27*届*ACM SIGKDD*知识发现与数据挖掘会议论文集》，第934–942页，2021年。[14] M. Li、K.-Z. Lee 和 E. Bareinboim。反事实Shapley贡献分配。《强化学习杂志》，第7卷，2026年。[15] L. Kotthoff、A. Fr´echette、T. P. Michalak、T. Rahwan、H. H. Hoos 和 K. Leyton-Brown。量化算法改进随时间的变化。载于《第*27*届国际人工智能联合会议论文集》（*IJCAI*），第5165–5171页，2018年。

[16] S. Ozerturk 和 H. Yildirim。信用归因与协作工作。《经济理论杂志》，195:105264，2021年。[17] S. Ozerturk 和 H. Yildirim。谁获得荣誉？信用归因、溢出效应与团队低效率。《博弈与经济行为》，154:246–266，2025年。[18] A. Sharma、H. Li 和 J. Jiao。反事实Shapley值：对系统指标变化进行归因。载于*NeurIPS 2022*因果机器学习现实世界影响研讨会，2022年。[19] B. Bilodeau、N. Jaques、P. W. Koh 和 B. Kim。特征归因的不可能定理。《美国国家科学院院刊》，121(2):e2304406120，2024年。