**社工机构内吹哨伦理决策助手V1.2 技术说明文档**

1. **决策模型框架论述**
   1. 理论假设

模型以萨维奇的主观期望效用模型为依据（Savage，1972）。简单来说，在个体缺乏对客观概率认知的情况下，会基于自己的主观信念来评估不确定事件，并结合自己对结果的价值评估，做出最理性的选择。他认为每一个可能存在的“世界状态”都会被赋予一个主观概率，而每一个结果都会被主观赋予一个效用值，个体选择的行为是那个能带来最大主观期望效用的行为。然而以Ellsberg悖论为代表的行为理论却又反应出概率不确定性厌恶等因素导致了现实中存在着大量非理性行为（Ellsberg，1961），即使某个行为是符合主观期望的效用最大化的，但风险敏感却又将人们引向了非理性的决策。因此，本模型面向社工机构内吹哨伦理决策困境，尝试为主体给出尽可能符合主体心理预期的最高收益与最小风险的理性决策建议。

1.2 模型框架

本文采用了Paul和Kundu的基于分数阶熵的决策模型框架，建立社工机构内吹哨的伦理决策模型（Paul & Kundu，2026）。这一模型的初衷是用于在金融股市中寻求最佳的投资组合，兼具良好的投资回报与风险水平。而本文调整了这一模型，使得其可以在机构内吹哨难题中做出符合个体主观心理预期的最小风险水平与最大效用的决策。

首先，Shannon熵的一般形式如所示，其具体意义是某个随机事件的不确定性后果（Shannon，1948）。其中，是某个事件发生的概率，而某个事件发生后，它对主体带来的信息量就是，如式所示。例如，当某个事件发生的概率为1时，的取值就是0，即必然发生的事件即使发生也不会给主体带来更多信息价值。而当事件的发生概率无限接近于0时，取值极高，即小概率事件发生后，该事件对主体的信息量就会很大。但在现实中，结果不是固定的某一个，通常存在多种结果，因此对信息价值做加权平均后就得到式。显然，在这一体系下，常见事件出现得更频繁，所以在平均熵中权重更大；而小概率事件虽然单次信息量大，但出现少，所以平均贡献较小。因此，综合来说，熵可以看作主体主观上的所有可能结果的不确定性影响的加权平均，这也是本文对风险的定义。





在此基础上，考虑个体主观上的对风险承受能力，则可以得到Ubriaco熵函数（Ubriaco，2009），如式所示。与不同的是，考虑了每个不同个体对风险的不同承受能力，而对于具有不同风险承受能力的人来说，同一组行为带来的不确定性对具有不同风险承受能力的人的主观价值是不同的。在中，越大则说明一个人的风险承受能力越高，此时的会和不考虑风险承受能力影响的越来越相似。反之，则函数的增长速率减缓，熵对低概率事件的敏感度降低，更贴近风险规避者所期望的需求。



在社工机构内吹哨的伦理决策中，我们假设主体可以选择四种行为，直接举报、匿名举报、延迟行动（继续保持观察后采取行动）、保持沉默。本文分别将它们表示为、、、，那么可以选择的行动集合就是。

我们将问题被解决的概率定义为成功概率S，而主体主观上意识到的遭受无法承受的代价的概率，即有偿概率则被定义为D。那么与之相反的失败概率则是F，无偿概率就是ND，那么行动造成的四种后果就如式所示，这些后果的集合就是总状态空间。



假设成/败概率与有/无偿概率独立，对于每一种行动，有对应成功概率参数，与有偿概率参数，那么四种给定行动的状态概率分布就是，式就是采取第i种行为时，第j种情况发生的概率。



引入式中的分数阶熵后，行动的风险就可以表示为：



在考虑行为带来的效用方面，Waytz等提出了“公平-忠诚”权衡视角（fairness–loyalty tradeoff），是最受认可的组织内吹哨行为解释模型之一（Waytz et al.，2013）。这一模型认为，公平与正义的价值观强调人人平等。而忠诚则决定了优惠待遇，要求人在决策中偏向自己所处的群体。而公平和忠诚规范的冲突结果则决定了主体是否会做出组织内吹哨行为。公平的规范通常要求人们举报不法与违规行为，而忠诚规范则认为向第三方或者他人举报机构内的不良行为是一种背叛。因此，公平与忠诚之间的权衡直接影响了主体是否会做出吹哨的决策。

本文基于上述理念定义三个变量，公正价值、忠诚代价。另外，毫无疑问的是举报带来的收益会影响人的行为，因此定义感知收益。首先，本文定义公正价值，即一个人在多大程度上相信自己可以为公正的世界做贡献，是能帮助他人或者自己获得应得的东西的。这一定义受到了公正世界信念理论的启发（Lerner，1980），其测量以苏志强修订的中文版公正世界信念量表为基础（苏志强 et al.，2012），我们在此基础上进行了修改。在道德基础问卷的启发下，定义变量忠诚代价，意为个体在面对不当行为时，因揭发或者举报等行为而感受到的群体关系紧张、人际关系被破坏的程度。感知收益是值主体实施第i种特定行为后，主观认识到的可能从组织、社会或自身心理层面获得的各种积极结果。这一定义及测量工具以Near和Miceli的研究（Near & Miceli，1985）为基础，经过改编而来。由这三个指标组成的综合效用权衡指数表示为式。显然，当与增加时，随之增加。而随着的增加，随之下降。则意味着收益越大，而忠诚代价越小。



Nawrocki和Harding在研究证券投资问题时，曾分析了风险中性效用函数、风险厌恶效用函数与的作用（Nawrocki & Harding，1986）。Paul和Kundu的模型也同样应用了这三种效用函数（Paul & Kundu，2026）。据此，本文定义效用函数表示主体在采取的行动后的主观预期效用。通过如式表示的方式，即可得到。0.5这个阈值划分了三种不同的风险偏好类型，即风险中性型（）、风险敏感型（）、温和风险敏感型（）。



考虑到每个行为会带来后果，记作。定义在不同的下的主体预期效用为，则与的转化就成为了下一个问题。D.Kahneman和A.Tversky的前景理论曾提出了确定效应，即多数人在面临获利的时候都是风险规避的。而这一理论同时又提出反射效应，认为多数人在面临损失时又是风险喜好的（Kahneman & Tversky，1979）。受到这一理论启发，我们认为小概率事件发生对效用影响更大。再结合Ubriaco熵函数考虑每种情况发生概率的影响。显然，与是收益的情况，而与是损失的情况。而和代表的是问题解决维度，和代表是风险的维度。定义为效用收益-损失修正指数，其含义为决策者对收益和损失的敏感程度，在发生时效用增加一个单位的，反之发生时则减少。定义为风险效用修正指数，代表对损失的敏感程度，时增加一个单位的，反之发生时则减少。在这里采用Kahneman和Tversky给出的前景价值函数中的值赋予，就取值为0.88，而是2.25。那么四种状态后果的修正值就如式所示。那么就由所示的式子计算获得。





那么，本文中与行动相关的风险的预期效用分数熵（EU-FE）度量定义如式所示。是选择行为带来后果的情况下的期望效用，就是每一种后果发生概率与对应效用相乘后加总。



其中，是权衡参数，代表了风险考量在多大程度上能够左右一个人的决策。而当前行动效用越接近效用最大值，则效用越高。总体而言，决策分数代表风险与效用的差值，越小的决策分数意味着小的风险和更高的效用。加入方差项以后即可进一步获得EU-FEV，见式。最终选择一个最小者为最终决策，可以发现越小就意味着风险越低而效用越大。



1. **变量与测量**

本文通过网络平台招募的方式，通过百度贴吧、小红书、QQ、微信等渠道发放调查问卷收集了106份问卷，剔除注意力题回答异常等情况后，剩余101个有效样本。表1展示了本文中涉及的所有输入参数，以及其在问卷中对应的模块。

表1 模型输入参数及问卷中对应项

|  |  |  |
| --- | --- | --- |
| 参数 | 含义 | 问卷问题 |
|  | 个体主观上的对风险承受能力 | 你能在多大水平上承担选择可能带来的风险？ |
|  | 风险考量在多大程度上能够左右一个人的决策 | 你在做决策时，对风险的考虑在多大程度上影响你的决定 |
|  | 成功概率参数 | 假设你做出了以下选择，你认为问题被解决的概率有多大，请填写0-100之间的数值。 |
|  | 有偿概率参数 | 假设你做出了以下选择，你认为你有多大概率会承受难以承受的损失，请填写0-100之间的数值 |
|  | 公平价值 | 相关人的行为损害了社会公正 |
| 有些人没有得到他们应得的 |
| 不公正战胜了公正 |
| 遭受了不公正的人没有得到补偿 |
| 相关人员做决定时没有力求公正 |
|  | 忠诚代价 | 举报同事或朋友是一种对关系的背叛 |
| 即使举报是正确的，也会削弱我对团队的忠诚感 |
| 举报行为会被视为对团队利益的损害 |
| 考虑到我和朋友、同事的关系，我会犹豫是否继续举报 |
| 即使有不当行为，我仍会优先维护与团队的关系 |
| 举报会让我感觉自己不再完全属于这个群体 |
|  | 执行每一种行动后的感知收益 | 这会提高我在组织中的声望 |
| 这会给我带来职业晋升机会 |
| 这会使我获得满意的经济等物质奖励 |
| 这会让我觉得我在做正确的事情 |
| 这会使我获得心理上的满足或者释怀 |

的测量初始设置了6个问题，除了表1中的5个问题外，还有一题“案例中的不公正是偶然的，而不是必然”，但该题在单个因子上的载荷仅有0.32，而其它问题的载荷均大于0.65,因此删除该题。而在删除该题后，模块问卷的Cronbach系数从0.8272上升到了0.8719。用于测量的问卷模块的Cronbach系数为0.8599，所有问题的单因子载荷均达到0.60。

对于问卷部分的信效度检验，分四种行动的情况进行，在四种情况下，Cronbach系数分别是0.8004、0.8023、0.8462、0.9440。此外，不论是采取何种行动，部分问卷都可以被分成两个因子，前三个问题都属于第一个因子，且单因子载荷都大于0.70。而后两个问题被划分到了第二个因子，单因子载荷均大于0.55。可以发现问题“这会提高我在组织中的声望”“这会给我带来职业晋升机会”“这会使我获得满意的经济等物质奖励”可以被解释为对物质收益的感知，而“这会让我觉得我在做正确的事情”“这会使我获得心理上的满足或者释怀”可以解释为对心理收益的感知。而四种情况下的物质收益与心理收益感知问卷的Cronbach系数如表2所示。

表2 四种行动对应的物质收益与心理收益感知问卷的Cronbach系数

|  |  |  |
| --- | --- | --- |
| 行动 | 各维度的Cronbach系数 | |
| 物质收益感知 | 心理收益感知 |
| 直接举报 | 0.8895 | 0.7843 |
| 匿名举报 | 0.9235 | 0.8659 |
| 继续观察 | 0.9163 | 0.7118 |
| 保持沉默 | 0.9514 | 0.8515 |

可以看出，本研究设计的所有量表的内部一致性都大于0.70，达到“良好”标准。因子载荷均达到0.5以上，具备良好的结构效度。因此，本研究问卷的信度与效度均已达标。

1. **模型效果验证**

在模型做出决策MD的同时，通过问卷问题“你最后会选择怎么办？”收集调查对象最终做出的决策FC，那么模型决策与现实决策对应的效用期望就分别是和。那么两者之间的差异是。同理，定义模型决策和现实决策分别对应的风险（即）为和。那么两者的风险差值就是。

在对比模型建议决策与现实决策后发现，有47.52%的样本的值大于0，即模型建议决策可以提升47.52%的效用。与的比例分别是36.63%与14.85%（因保留两位小数，因此三者相加不严格等于100%）。在风险差上，7.92%的样本的，即他们在听从模型决策后，风险会降低。而与的样本比例是36.63%与54.46%，说明在听从模型决策后，有36.63%的人的决策风险是不变的，而有54.46%的人的风险反而上升。图1展示了和的总体分布情况。可以发现大多集中在0附近，说明大多数人效用变化不大，少数人显著提升或下降。则主要集中在0附近，绝大部分样本的值落在0-0.5之间，风险差值分布更集中。这初步表明本模型做出的决策趋向于通过适量增加风险来换取效用提升。

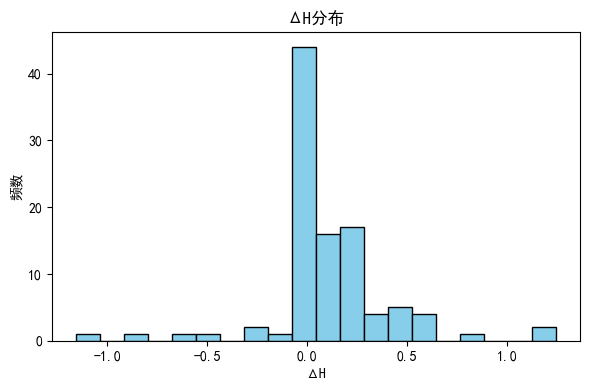
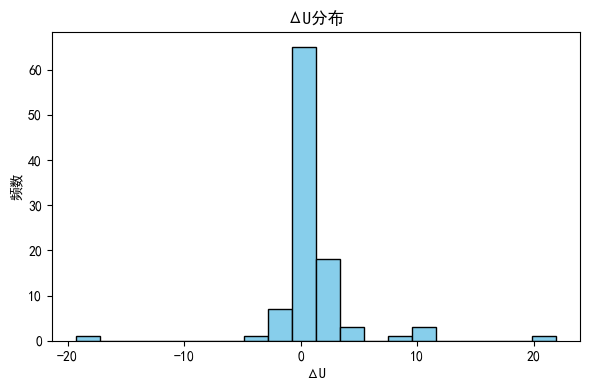


图1 与的总体分布状况

我们对人为与模型决策相对应的效用、风险等元素进行相关性分析，结果如图2所示。人为决策效用与模型决策效用之间的相关性极高（ρ=0.84，p < 0.05），说明模型在效用层面的判断与人类的选择高度一致。而人为决策风险与模型决策风险之间在一定程度上具有一致性（ρ=0.23，p < 0.05），但相关系数远不如人为决策效用与模型决策效用，说明本模型虽然可以在一定程度上解释人在现实情况下对风险的直觉判断，但仍需进一步改进以提升与现实的一致性，这可能是造成模型决策与现实决策的不一致性的重要原因。效用期望差值与人为决策效用负相关（ρ= -0.51，p < 0.05），这说明当人为效用数值大时，差值往往小，而当人为效用小时，差值则倾向于增大，表明模型更有助于那些原本效用水平低的人，而对原本效用高的人则可能带来效用下降。同理，风险差值与人为决策风险负相关（ρ= -0.71，p < 0.05），说明原本风险水平高的人群，在使用模型时获得的风险降低幅度更显著。效用差值与模型决策风险显著正相关（ρ= 0.52，p < 0.05），而风险差值与模型决策风险的相关性不显著。这说明在效用较低的情况下，模型的决策带来风险差值的上升的可能性越大。总体而言，本模型对效用与风险的考量决定了模型实现的“风险与效用的再分配机制”，简单来说模型在效用的评判上更接近人类，但在风险上并不符合多数人的直觉。对于效用较低的人群来说，模型可能更倾向做出提升效用，但是增加决策风险的决定。

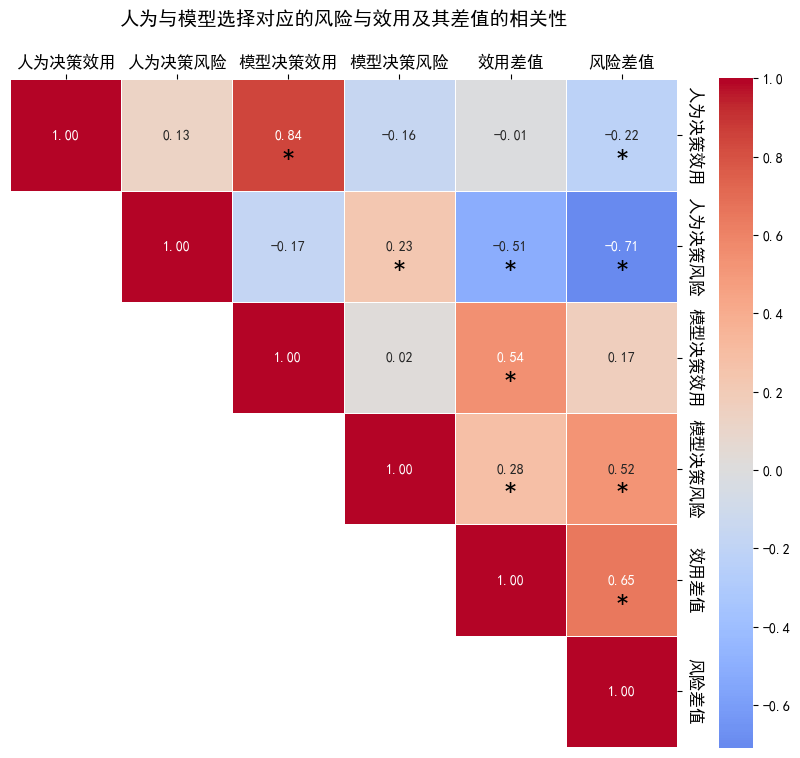


图2 人为与模型决策在风险与效用上的相关性结构

表3展示了每个样本的现实选择与模型推荐决策的效用与风险的差异分析结果。结果表明，模型选择和人为决策在造成的效用和风险方面都具有显著差异。模型推荐的决策在总体上提升了效用值，但也显著提高了风险。这进一步验证了模型选择牺牲风险来换取更高的效用的特性。

表3 人为决策与模型决策在效用与风险上的差异检验结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | 均值 | 标准差 | 中位数 | 平均差值 | T检验 | | Wilcoxon检验 | |
| T值 | P值 | W值 | P值 |
|  | 5.86 | 5.62 | 4.52 | 0.866 | 2.37 | 0.02 | 429 | 0.01 |
|  | 6.73 | 6.66 | 6.19 |
|  | 1.00 | 0.27 | 1.07 | 0.117 | 3.79 | 0.00 | 316 | 0.00 |
|  | 1.12 | 0.22 | 1.17 |

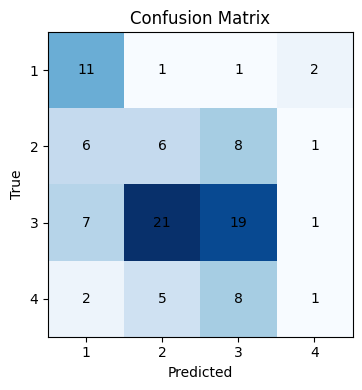


图3 模型决策与现实决策的混淆矩阵

图3中的混淆矩阵与表4的分类报告表系统展示了模型决策与现实决策的差异性与一致性情况。总体而言，有36.63%的人做出的决策和模型决策一致，表明模型可以解释36.63%的现实决策模式。在对具体类别的一致性与差异性方面，模型对行动1与行动3的决策与现实一致性较高，分别是42%和53%。而在行动2和4的决策建议方面和现实差异性较高。

表4 分类报告表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 行动类别 | 样本量 | Precision | Recall | F1-Score |
| 1 | 15 | 0.42 | 0.73 | 0.54 |
| 2 | 21 | 0.18 | 0.29 | 0.22 |
| 3 | 48 | 0.53 | 0.40 | 0.45 |
| 4 | 16 | 0.20 | 0.06 | 0.10 |
| 加权平均 |  | 0.39 | 0.37 | 0.36 |

注：另有1个受试者因数据填写异常问题，导致模型验证部分没有记录其数据。

整体而言，模型的引入不仅仅是对个体决策的简单替代，而是在风险与效用之间实现了一种风险与收益的再分配机制。模型通过在多数情境下提升效用水平，将原本分散、随机的决策过程重新塑造为一种更具备解释力的决策模型，个体承担少量的效用风险，换取更大概率上的群体效用增益。另一方面，效用变化的异质性也揭示出模型对不同人群的差异化调节作用。效用本来较低、风险较高的个体成为最大受益者，而效用较高群体则在一定程度上“补贴”了这种改善。这种结果不仅体现了模型在风险-效用权衡下的理性优化，更反映出一种近似于预期效用改进的逻辑，即在整体风险上升不大的前提下，通过局部让渡实现了对弱势群体的显著改善。因而，从更深层次看，模型效果不仅仅是数值意义上的风险下降，而是构建了一种在伦理决策中兼顾效率与公平的新均衡。

1. **未来展望**

根据模型验证结果初步可以断定，本模型能解释现实的决策逻辑的原因在于模型对于效用的估计和现实情况更为一致，造成差异的原因在于模型对风险的测量与常识一致性不足。未来的探索将会更进一步发掘风险的评估指标，使其尽量与现实一致，让本模型蕴含的理论思想更符合现实情况，解释与优化现实决策逻辑。在程序系统开发上，未来将引入大模型的评估，形成最终程序后小范围发放使用，通过使用者评价进一步评估模型效果。

**参考文献**

苏志强, 张大均 & 王鑫强 2012, “公正世界信念量表的修订及在大学生应用的信效度研究.” 中华行为医学与脑科学杂志 21(06).

Ellsberg, Daniel 1961, “Risk, Ambiguity, and the Savage Axioms.” *The Quarterly Journal of Economics* 75(4).

Kahneman, Daniel & Amos Tversky 1979, “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47(2).

Lerner, Melvin J. 1980, *The Belief in a Just World*. Boston, MA: Springer US.

Nawrocki, David N. & William H. Harding 1986, “State-Value Weighted Entropy as a Measure of Investment Risk.” *Applied Economics* 18(4).

Near, Janet P. & Marcia P. Miceli 1985, “Organizational Dissidence: The Case of Whistle-Blowing.” *Journal of Business Ethics* 4(1).

Paul, Poulami & Chanchal Kundu 2026, “Fractional Order Entropy-Based Decision-Making Models under Risk.” *Communications in Nonlinear Science and Numerical Simulation* 152.

Savage, L. J. 1972, *The Foundations of Statistics*. Dover Publications.

Shannon, C. E. 1948, “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27(3).

Ubriaco, Marcelo R. 2009, “Entropies Based on Fractional Calculus.” *Physics Letters A* 373(30).

Waytz, Adam, James Dungan & Liane Young 2013, “The Whistleblower’s Dilemma and the Fairness–Loyalty Tradeoff.” *Journal of Experimental Social Psychology* 49(6).