心理学报 2026, Vol. 58, No. 3, 381−398 https://doi.org/10.3724/SP.J.1041.2026.0381
Acta Psychologica Sinica https://cstr.cn/32110.14.2026.0381

从显性威慑到隐性内化：AI 监管和黑暗三联征人格对诚实行为的影响*

王健树[1] 姜啸威[2] 陈亚楠[1] 王明辉[1] 杜 峰[3,4]

([1]河南大学心理学院, 开封 475004) ([2]澳大利亚悉尼科技大学计算智能和脑机接口实验室, 悉尼 2007)
([3]中国科学院心理研究所认知科学与心理健康国家重点实验室, 北京 100101) ([4]中国科学院大学心理学系, 北京 100049)

摘 要 随着生成式人工智能逐步演变为具备自主影响力的社会代理, 其对人类道德决策的影响愈加显著。然而, 传统伦理监管模式基于理性人假设, 忽视了人格特质对道德选择的调节作用, 导致监管效果因人而异, 难以实现最优效率。为突破这一局限, 本研究聚焦于道德研究中的诚实维度, 构建了监管类型×人格特质交互模型, 系统探讨显性监管、隐性激励与道德反馈三类 AI 驱动的干预策略对自恋、马基雅维利主义和精神病态人格的差异化影响。结果显示:(1)显性监管有效提升诚实行为, 尤其对高马基雅维利主义与高精神病态个体具有显著约束力; (2)监管即使隐藏也能规范诚实行为, 其不确定性加速了决策反应; (3)隐性激励促进诚实行为并具内化效应, 但高马基雅维利主义者在奖励撤除后的诚实率显著降低, 揭示了该策略在此类人群中的脆弱性; (4)道德反馈对整体诚实率提升有限, 仅对高马基雅维利主义者具有一定影响, 而高自恋者表现出持续不诚实倾向。本研究创新性地揭示了人格特质(黑暗三联征)在 AI 伦理监管中的关键作用, 并为传统的理性人监管范式提供了一个基于人格差异化的视角, 为未来设计情境化与个性化 AI 伦理干预策略提供了重要的理论与实践依据。

关键词 AI 监管, 诚实行为, 黑暗三联征, 人格特质, 情境化干预

分类号 B849: C91

1 引言

人工智能(AI)技术的快速发展, 特别是生成式 AI 的广泛应用, 正在改变人机交互模式(Cohen et al., 2024)。AI 系统已从被动的工具逐步演变为具有主动行为影响力的社会代理(Goodlad, 2023; Tallberg et al., 2023)。一个潜在的风险是, 这种转变虽然能极大地提升生产力, 但也为人类将不道德行为外包给机器提供了便利(Köbis et al., 2025)。这一转变的核心体现之一, 是 AI 越来越多地被赋予监管者的角色, 用于引导和规范人类的道德行为(Borau, 2025; Khan et al., 2023)。例如, 算法系统被用于在线社区行为监控、金融交易诚信评估、以及内容创作的真实性核查等场景(Liu et al., 2024)。传统伦理监管模型在这种技术变革前显得有些力不从心。因为技术系统将不再是被动执行规则的工具, 而是通过算法反馈, 主动参与到道德规则的建立与执行中。这呼唤着一种全新的监管范式(图 1)。然而, 现有关于 AI 伦理的研究常聚焦于 AI 系统自身的道德风险(如偏见、透明度), 或 AI 作为决策主体引发的伦理困境。相比之下, 一个更关键的问题却鲜有探讨: AI 作为监管者, 究竟如何影响人类的道德决策? 最新的证据甚至表明, 将任务委托给 AI 本身就会增加不诚实行为(Köbis et al., 2025)。这一发现为 AI 监管的必要性敲响了警钟, 但一个一刀切的监管方案可能并非最优解, 因为其效果很可能

收稿日期: 2025-05-10

* 教育部人文社会科学研究青年基金项目(22YJCZH021), 国家社科基金(24AGL037)资助。

王健树和姜啸威同为第一作者

通信作者: 陈亚楠, E-mail: chenyn@henu.edu.cn; 王明辉, E-Mail: wmhwang@163.com

受到个体心理因素的调节。因此，一个关键的问题是：当面对不同人格特质的个体时，AI 的监管效果是否会存在差异？

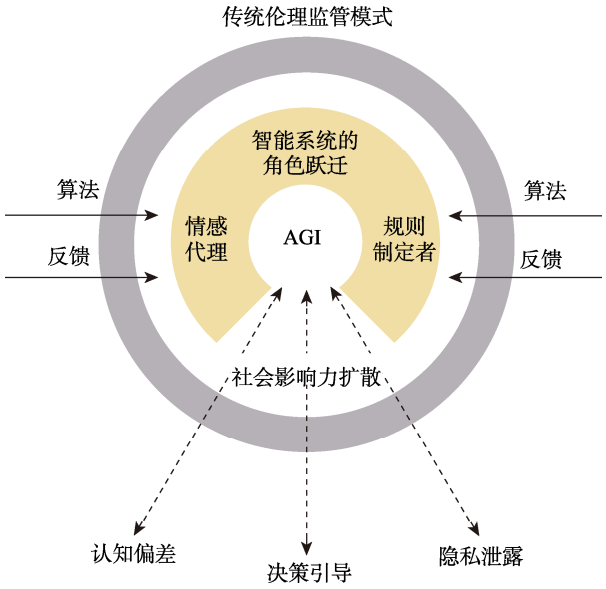

图 1 AGI (通用人工智能)在伦理监管中的角色跃迁

本研究聚焦于 AI 驱动的、旨在规范人类道德行为的干预策略(即 AI 管人)。我们系统考察了三种典型的 AI 监管类型，它们代表了不同的心理作用机制：(1)显性监管：AI 明确告知用户其行为处于实时监控之下，并清晰说明违规将面临的惩罚后果(如扣分、权限限制)。这主要通过外部威慑机制影响行为(Abeler et al., 2019)。(2)隐性激励：AI 不直接告知规则，但通过设计好的奖励机制(如积分、小额货币奖励)对用户的诚实行为进行持续强化。这旨在通过操作性条件反射引导行为，并可能促进道德内化(Thanassoulis et al., 2022)。(3)社会激励：AI 在用户行为后即时提供道德语料反馈(如“您的诚实体现了您高尚的品格，值得敬佩！”)，旨在通过激活道德认同和社会规范来强化内在动机(Shalvi et al., 2012)。

这三种策略的核心目标一致，利用 AI 的能力来提升人类在交互情境中的诚实行为。然而，既有研究和实践观察表明，AI 监管的效果在不同人群中存在显著分化(Köbis et al., 2021)。例如，过于刚性的算法有时会引发一种合规性悖论：人们表面上遵守了规则，但其内在的道德认同感却因此受到侵蚀而下降。与此同时，某些看似柔和的激励或反馈策略在特定人群面前又可能完全无效。这种分化在一定程度上源于传统伦理监管模型对理性人假设的过度依赖(Fowles, 2018)，认为所有个体都会以相同的方式回应同样的奖惩规则，却普遍忽视了关键的人格特质变量对监管效果的调节作用。

人格特质，特别是具有反社会倾向的黑暗三联征(Dark Triad)，包括自恋(Narcissism)、马基雅维利主义(Machiavellianism)和精神病态(Psychopathy)，这三种特质作为连续谱系特质广泛存在于一般人群中(Jones & Paulhus, 2014)，其水平高低是预测道德偏离行为(如欺骗、剥削)的稳健因子(Jones & Paulhus, 2014; Malär & Giuffredi‐Kähr, 2024)。需要强调的是，黑暗三联征特质并非仅存在于道德破坏者中，而是不同程度地存在于每个人身上，并在特定情境下影响其道德决策倾向。道德行为是人格特质与情境因素(如监管类型)动态交互的结果，本研究聚焦的 AI 监管情境正是触发和放大黑暗三联征特质对道德行为影响的关键环境。

现有证据提示了黑暗三联征特质在监管情境中的差异化反应模式，但其与 AI 驱动的、不同类型监管策略的交互机制尚不清晰，且缺乏统一的理论框架解释。社会规范理论强调明确的社会规则和监督能激活描述性规范和命令性规范，增加违规的感知社会成本和惩罚风险，从而促进遵从(Cialdini et al., 1990)。显性 AI 监管正是通过清晰传达规则、监控和惩罚威胁来建立这种规范环境。然而，黑暗三联征个体对社会规范，特别是由非人类实体(AI)执行的规范的敏感性和遵从动机可能存在显著差异。高马基雅维利主义者精于成本−收益计算和印象管理(Marbut et al., 2025)。显性监管提供的明确惩罚信号能有效提高其违规的预期成本，促使其采取策略性合规(表面遵守以规避惩罚)，而非道德内化(Jones & Paulhus, 2017)。高精神病态者具有冲动性和低恐惧特质。显性监管的即时、确定的惩罚威胁可能对其产生较强的外部约束力，因为其决策更易受即时、显著后果驱动(Baskin-Sommers et al., 2015)。基于此，我们提出假设 H1a：显性 AI 监管能够显著提高高马基雅维利主义者和高精神病态者的诚实率。对于高自恋者，他们关注自我形象，但其形象维护动机在匿名 AI 监管情境中可能减弱(Campbell et al., 2002; Giroux et al., 2022; Hart et al., 2021)。因此，我们提出假设 H1b：显性 AI 监管对高自恋者诚实率的影响效果可能不如前两者显著(H1b)，因为高自恋者更依赖于社会评价而非非人 AI 的评判。

自我决定理论(SDT)认为，支持自主性、胜任感和关联性的环境能促进外在动机的内化(Ryan &

Deci, 2020)。隐性激励(如意外奖励)可能通过满足胜任感(发现规则并成功获得奖励)和潜在的自主感(感觉行为是自我发现的)来促进道德行为的初步内化。然而, 黑暗三联征特质可能影响个体对 SDT 基本需求的体验和对激励的动机归因。高马基雅维利主义者具有高度的工具理性和机会主义。他们可能将隐性奖励视为纯粹的外部工具, 行为动机停留于外在调节水平。一旦奖励撤除, 其诚实行为因缺乏内化而迅速回落。基于此, 我们提出假设 H2：隐性奖励可提升诚实率并产生一定的内化效果, 但高马基雅维利主义者在奖励撤除后诚实率回落更明显, 内化效果更差。高精神病态者情感淡漠, 对奖励的神经反应可能异常(Blair, 2013), 虽然其可能对低层次需求如即时、明确的物质奖励有反应, 但 SDT 所强调的更高层次需求如关联感、整合调节的内化过程可能受阻, 他们的内化程度可能有限。高自恋者追求优越感和特殊对待, 获得奖励可能满足其胜任感和被系统认可的满足感, 可能促进一定程度的内摄调节, 即将规则内化但不完全整合。因此, 其内化稳定性可能介于马基雅维利主义者和低黑暗三联征者之间。

道德认同理论指出, 当道德特质如诚实在个体自我概念中处于中心地位, 且个体渴望依此行动时, 更可能产生道德行为(Aquino & Reed, 2002)。道德语料反馈旨在通过语言强化道德特质与自我的关联, 提升道德认同的中心性。然而, 黑暗三联征特质影响个体的自我概念和对道德信息的加工。高自恋者的核心特征是膨胀但脆弱的自我和强烈的自我增强需求。他们通常拥有积极的道德自我观, 并倾向于将失败归因于外部。对于高自恋者, 匿名 AI 提供的泛泛道德反馈可能被他们忽视或贬低, 与其现有道德自我观冲突, 引发防御性反应如否认、合理化, 缺乏激发其彰显道德认同的社交舞台。高精神病态者情感冷漠且缺乏道德情感(Ojanen & Findley-Van Nostrand, 2019; Rijnders et al., 2021), 道德反馈试图激活的内在道德动机如内疚、自豪对其影响甚微, 他们的道德认同通常很低或扭曲。因此, 我们提出假设 H3a：道德语料反馈对整体诚实率的提升有限, 对高自恋者和高精神病态者基本无效。高马基雅维利主义者通常持有愤世嫉俗的世界观和实用主义道德观。他们可能将道德反馈视为一种可操控的社会信号。在持续反馈下, 他们可能进行表面的策略性调整, 以维持一个基本诚信的印象, 方便未来潜在的机会主义。因此, 我们提出假设 H3b：高马基雅维利主义者对道德反馈有轻微、策略性的反应。

在此基础上, 本研究聚焦于 AI 监管情境下的诚实行为, 通过构建监管类型×人格特质的交互模型, 系统考察显性监管、隐性激励和道德语料反馈三种干预机制在面对不同人格结构个体时的行为效能变化。研究通过三组实验递进展开, 旨在系统考察不同干预机制的效果、其背后的心理过程, 以及人格特质在其中的关键调节作用。实验 1 以显性监管为切入点, 建立不同人格在惩罚性监管下的行为差异基础模型; 实验 2 通过操控监管的可见性, 进一步探索监管的可见性如何调节自恋者的道德决策; 实验 3a 引入隐性激励, 以探究不同人格(尤其是马基雅维利主义者)对非惩罚性诱因的不同反应模式; 实验 3b 则采用社会激励, 检验不同人格如精神病态者对非物质道德强化的反应特征。通过上述设计, 我们期望揭示三类监管机制在三类黑暗人格群体中的相对效能排序, 进而提出一种人格—干预匹配原理。

通过这些研究, 我们希望整合多种心理学理论, 为构建一个动态、个性化的 AI 监管新模型奠定基础。这一模型的核心, 正是“人格−策略适配度”。在实践上, 本研究的发现将能直接用于设计更智能、更具适应性的 AI 伦理系统, 帮助其摆脱传统一刀切的监管局限, 从而在数字时代更有效地提升整体社会诚信水平。

2　实验 1: 显性 AI 监管对不诚实行为的抑制及人格调节作用

2.1　实验目的

本研究聚焦于道德研究中的诚实维度(Glenn et al., 2010), 采用经典的诚实行为研究范式——抛硬币任务(coin flip task) (Batson et al., 2002)。实验 1 通过操控是否存在可见的 AI 监管, 探讨外部监管与内部人格特质对个体诚实行为的影响机制。此外, 实验 1 还旨在揭示不同人格特质(马基雅维利主义、自恋、精神病态)在诚实行为中的表现差异, 及其在外部监管情境下的互动效应。进而通过从外部控制(AI 监管)与内部人格特质的双重视角, 系统揭示个体在道德决策过程中的心理机制。

2.2　方法

2.2.1　被试

根据(Gerlach et al., 2019)关于不诚实行为的元分析, 监督能够对诚实行为产生中等偏大的效应

($d \approx 0.60$)。因此使用 G*Power v.3.1 (Faul et al., 2009), 在保证效应量 Cohen's $d = 0.60$、显著水平为 0.05 (双尾)及检验力达到 0.95 的情况下, 需招募 39 名被试。

实验 1 共收集 46 名有效被试的数据, 其中男性 8 人, 女性 38 人。平均年龄 20.11 岁($SD = 1.58$)。所有被试均身体健康, 视力或矫正视力正常。实验前所有被试均被告知可能存在的风险并签署知情同意书。实验结束后获得一定金额报酬。

2.2.2 测量工具

本实验采用 Jones 和 Paulhus (Jones & Paulhus, 2014)提出的黑暗三联量表(Short Dark Triad, SD3)测量被试的马基雅维利主义(Machiavellianism)、自恋(Narcissism)和精神病态(Psychopathy)三个维度。SD3 共 27 道题, 每个维度 9 题, 采用 5 点 Likert 量表(1 = "强烈不同意", 5 = "强烈同意")。其中, 马基雅维利主义的条目, 比如"我喜欢用聪明的手段来达到自己的目的"; 自恋的条目, 比如"人们认为我是一个天生的领导者"; 精神病态的条目, 比如"我确实可以对别人很刻薄"。

由于 SD3 各维度条目在表述上略有异质, 难以满足经典 Cronbach's α 的 τ-等价假设(高清辉, 2022), 本实验采用 Guttman 分半信度方法, 通过提取每个维度的加权校正后的分半系数 λ_6 来评估各维度内部一致性。在本实验中, 各维度 λ_6 系数为: 0.75 (马基雅维利主义), 0.70 (自恋), 0.72 (精神病态)。为检验量表各维度的结构效度, 我们分别对马基雅维利主义、自恋与精神病态三个维度进行了 KMO 采样适度检验(Kaiser, 1970)与 Bartlett 球形检验(Bartlett, 1951)。结果显示, 三个维度的 KMO 值分别为 0.80、0.674 与 0.78; 同时, Bartlett 球形检验均达到显著水平($ps < 0.001$)。

2.2.3 实验任务与程序

实验 1 为 2 (监管条件: 监管存在 vs. 不存在) × 3 (人格维度: 马基雅维利主义、自恋、精神病态)混合实验设计, 其中监管条件为被试内因素, 人格特质为连续协变量。

实验流程如图 2a, 每个试次(trial)均包含以下 4 个阶段: (1)猜测阶段: 每个试次开始时, 被试首先对屏幕上将要出现的硬币结果进行猜测, 需通过键盘选择"正面"或"反面"。选择后, 被试会看到硬币的实际结果。(2)结果阶段: 被试猜测硬币结果之后, 屏幕上会显示"投掷中"的字样, 并且会显示硬币的实际结果是正面还是反面, 硬币落到正面和反面次数均为 60 次, 如图 2b。在此处我们控制被试的猜测正确率为 50%, 也就是被试在猜测阶段选择一个结果之后, 结果会有 50%的概率与被试的猜测不符, 如图 2c。(3)报告阶段: 被试被要求汇报自己刚刚的预测是否正确。指导语为: 你刚刚的预测是正确还是错误? 此时的报告对象并非预测的具体内容(正面或反面), 而是报告预测是否准确(即"猜对"或"猜错")。如果报告"猜对", 则获得 10 个金币的奖励; 如果报告"猜错", 则仅获得 1 个金币。(4)监管反馈阶段: 在每个试次中, AI 监管者会随机出现。在实验开始前, 我们通过指导语明确告诉被试: 实验中会有一位 AI 监管助手随机巡视, 如果右上角出现了一个 AI 头像则说明您当前正在受到监督。在有监管的试次中, 如果被试的报告与其实际预测的正确性不符(例如实际猜错却谎报猜对), 则被视为不诚实行为, 被试在反馈阶段会受到 50 个金币的惩罚并以红色字体呈现。这一惩罚机制直接地向被试提示了不诚实行为的后果; 若监管不存在, 则无论被试报告诚实与否, 均不会被惩罚, 如图 2d 前半部分。

实验 1 一共包括 4 个组块(block), 每个组块包含 30 个试次(trial), 共有 120 个试次。每个试次中监管是否存在为随机分配(各 60 个试次)。被试获得的金币奖励会实时显示并累计。实验完成后, 根据被试获得的金币总数给予被试费。实验程序使用心理学软件 PsychoPy v2024.2.4 (Peirce et al., 2019)编制。在全部行为实验结束后, 被试被要求填写黑暗三联人格量表, 测量马基雅维利主义、自恋和精神病态三个维度的得分, 用于后续调节效应的分析。

2.2.4 数据分析方法

本实验所有数据分析均使用 R (版本 4.4.2)与 Python (3.12.8)进行。对于反应时的异常值, 设置排除过快(100 ms)和过慢(>3 个标准差)的反应时数据。实验 1 删除 1.74%, 实验 2 删除 0.11%, 实验 3a 和 3b 共删除 3.26%。

为了检验可见 AI 监管对诚实行为的直接影响, 我们对每位被试在有监管与无监管条件下的诚实率进行配对样本 t 检验。诚实率计算公式如下:

$$\mathrm{HR} = \frac{N_{\text{诚实报告}}}{N_{\text{总报告}}} \tag{1}$$

为了评估监管是否影响反应过程的速度, 我们对有无监管条件下的平均反应时进行配对样本 t 检验, 从信息加工角度探索监管信息是否影响决策的认知成本。

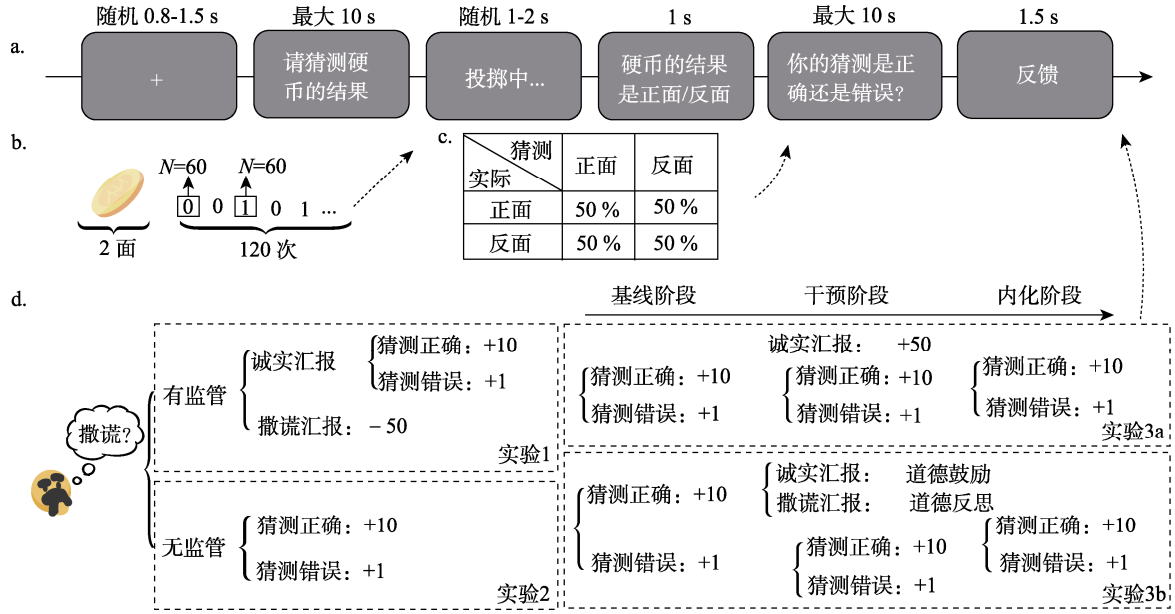

图 2　实验任务流程与条件设置

为进一步探讨撒谎行为是否具有更高的加工负担，我们在反应时数据上实施 2 (行为类型：撒谎、诚实) × 2 (监管条件：有、无)的重复测量方差分析(rm-ANOVA)，检验行为选择与监管之间的交互效应。

为了直观地检验人格特质对监管效果的调节作用，我们将马基雅维利主义、自恋和精神病态三个维度分别按中位数划分为高低组，在诚实率上执行 2 (人格组别：高、低) × 2 (监管条件)混合设计方差分析(mixed-ANOVA)。

此外，为了检验人格特质是否影响行为加工特点，我们对反应时进行了 2 (行为类型：诚实 、撒谎) × 2 (特质分组：高分组、低分组)的方差分析，从信息加工层面探讨人格对行为的影响。

2.3　结果

2.3.1　监管机制对诚实行为的影响

为了检验监管条件操作的有效性，并且探究监管机制对诚实行为的影响，配对样本 t 检验结果显示，有监管条件下的诚实率(M = 0.98, SD = 0.04)显著高于无监管条件下(M = 0.85, SD = 0.19)，如图 3a, $t(45) = 4.59$, $p < 0.001$, Cohen's $d = 0.67$。有监管条件下的平均反应时(M = 0.64, SD = 0.17)和无监管条件下(M = 0.63, SD = 0.19)不存在显著差异，如图 3b, $t(45) = 0.88$, $p = 0.383$。

2.3.2　监管与撒谎行为对反应时的影响

对被试的反应时进行 2 (是否撒谎：撒谎、诚实) × 2 (监管条件：有监管、无监管)双因素重复测量方差分析。结果表明，行为类型与监管条件的交互效应、以及两者各自的主效应均不显著($ps > 0.20$)。

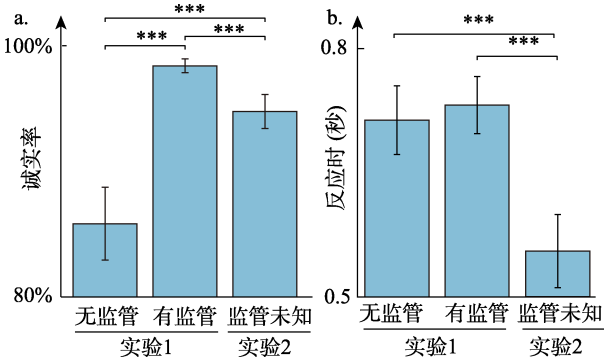

图 3　不同监管条件对诚实率与反应时的影响

注：*** $p < 0.001$

2.3.3　人格特质对 AI 监管效应的调节作用

分别在三种人格特质维度下，对被试的诚实率进行 2 (特质分组：高分组、低分组) × 2 (监管条件：有监管、无监管)的混合设计方差分析。方差分析的主效应与交互效应结果见表 1，对交互效应的简单效应分析结果见表 2。

表 1　人格特质调节监管对诚实行为影响的混合方差分析结果

特质维度	效应类型	F	p	η^2
马基雅维利主义	人格主效应	7.81(1, 44)	0.008	0.15
	监管主效应	24.01(1, 44)	< 0.001	0.35
	交互作用	7.11(1, 44)	0.011	0.14
自恋	人格主效应	1.10(1, 44)	0.301	0.02
	监管主效应	21.69(1, 44)	< 0.001	0.33
	交互作用	2.17(1, 44)	0.148	0.05
精神病态	人格主效应	3.82(1, 44)	0.057	0.08
	监管主效应	22.78(1, 44)	< 0.001	0.34
	交互作用	4.50(1, 44)	0.039	0.09

表 2 人格特质调节监管对诚实行为影响的简单效应和分析结果

特质维度	得分组	有监管	无监管	*t*	*p*	Cohen's *d*
马基雅维利主义	高	0.98 ± 0.05	0.79 ± 0.22	4.60(25)	0.005	0.90
	低	0.99 ± 0.01	0.95 ± 0.13	1.77(19)	0.092	0.40
精神病态	高	0.99 ± 0.05	0.82 ± 0.21	4.44(28)	< 0.001	0.82
	低	0.99 ± 0.02	0.94 ± 0.13	1.80(16)	0.091	0.44

结果表明，对于马基雅维利主义维度，特质分组(高分组和低分组)与监管条件(有监管和无监管)的交互作用显著，$F(1, 44) = 7.11$, $p = 0.011$, $\eta^2 = 0.14$。为了进一步探究马基雅维利主义维度对诚实率的调节作用，对交互项的简单效应分析结果显示高分组在有监管条件下的诚实率($M = 0.98$, $SD = 0.05$)显著高于($p = 0.005$)无监管条件下的诚实率($M = 0.79$, $SD = 0.21$)；低分组在有监管条件下($M = 0.99$, $SD = 0.01$)和无监管条件下($M = 0.94$, $SD = 0.12$)的诚实率不存在显著差异($p = 0.092$)。对于自恋维度，特质分组与监管条件的交互作用不显著，$F(1, 44) = 2.17$, $p = 0.148$。监管条件主效应显著，$F(1, 44) = 21.69$, $p < 0.001$, $\eta^2 = 0.33$，而特质分组主效应不显著。对于精神病态维度，特质分组与监管条件的交互作用显著，$F(1, 44) = 4.50$, $p = 0.039$, $\eta^2 = 0.09$。为了探究精神病态维度对诚实率的调节作用，对交互项的简单效应分析结果显示高分组在有监管条件下的诚实率($M = 0.98$, $SD = 0.05$)显著高于($p < 0.001$)无监管条件下($M = 0.81$, $SD = 0.21$)，但是低分组在有监管条件下的诚实率($M = 0.98$, $SD = 0.02$)和无监管条件下($M = 0.93$, $SD = 0.13$)不存在显著差异($p = 0.091$)。特质分组主效应边缘显著，监管条件主效应显著。人格特质的调节效应路径可视化见图 4。

2.3.4 人格特质与反应时的关系分析

为检验人格特质对反应时的调节作用，我们对三种人格分别进行了 2 (行为类型：诚实 、撒谎) × 2 (特质分组：高分组、低分组)的混合设计方差分析。

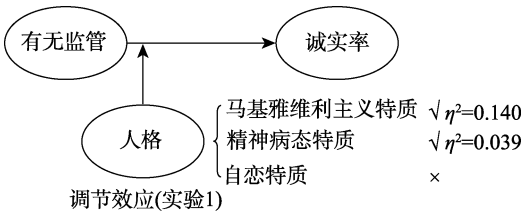

图 4 实验 1 中人格特质对监管效应的调节模型

结果显示，对于马基雅维利主义和精神病态，行为类型、人格分组的主效应以及两者的交互效应均不显著(*p*s > 0.05)。

对于自恋维度，交互效应显著，$F(1, 23) = 5.33$, $p = 0.030$, $\eta^2 = 0.19$。事后比较显示，在高自恋组中，撒谎($M = 0.65$, $SD = 0.52$)与诚实($M = 0.58$, $SD = 0.44$)反应时差异边缘显著，$p = 0.0523$。而在低自恋组中差异不显著(诚实: $M = 0.68$, $SD = 0.51$; 撒谎: $M = 0.70$, $SD = 0.54$), $p = 0.728$。结果见图 5。

2.4 小结

综上，实验 1 的结果表明，明确的 AI 监管能够显著提升诚实率，且该效应在马基雅维利主义和精神病态特质高的个体中尤为突出。反应时数据进一步揭示了自恋人格的调节作用：高自恋个体在撒谎时表现出决策变慢的趋势，而这种趋势在低自恋个体中并未出现。

3 实验 2: 监管不确定条件下的诚实行为

3.1 实验目的

实验 2 旨在探讨外部监管状态的不确定性如何

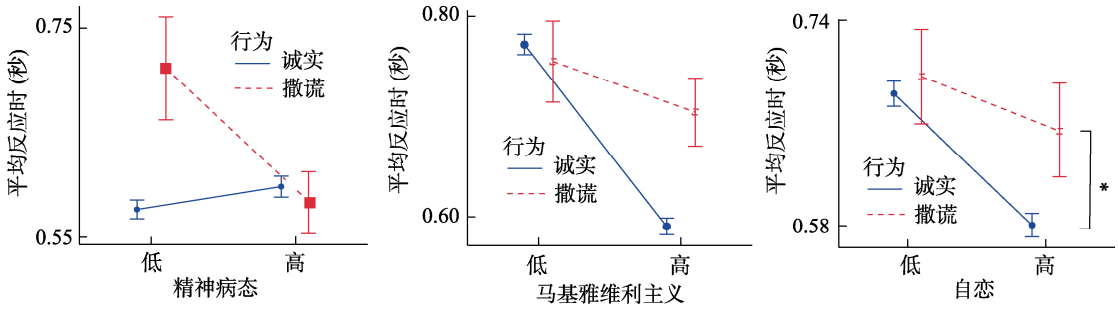

图 5 实验 1 中人格特质对诚实/撒谎行为反应时的调节作用

影响个体的诚实行为表现, 尤其是在人格特质的调节下, 个体是否会依赖内在的稳定特质或风险评估策略来调整诚实水平。通过设置不可见的监管状态, 实验检验外部规范模糊时个体的反应差异, 进一步揭示人格特质(马基雅维利主义、自恋、精神病态)是否在此情境下对撒谎与诚实的加工负荷产生调节作用。此外, 本实验结合实验 1 与实验 2 的数据, 构建广义线性混合模型(Generalized Linear Mixed Model, GLMM), 系统分析监管可见性、人格特质及其交互作用对道德选择的影响, 旨在建立一个多因素、层级化的诚实行为预测模型。实验最终期望阐明在外部约束不明确的情况下, 个体的道德决策是否更依赖于内在人格特质或对环境风险的感知与评估。

3.2　方法

3.2.1　被试

实验 1 中的所有被试在完成实验 1 后, 经过短暂休息后继续参加实验 2。本实验延续实验 1 的参与者, 无新增招募, 以保证实验条件的一致性和数据的连续性。

3.2.2　实验任务与程序

如图 2 所示, 实验 2 的任务设计与实验 1 基本保持一致, 仅在监管信息的呈现方式上进行了调整, 实验 2 隐藏了右上角代表监管状态的头像。本实验的关键操纵在于‘监管未知’条件。具体而言, 被试在实验开始前被告知, AI 系统可能会在部分试次中对其进行监管, 但他们无法在具体的试次中通过任何视觉线索(如 AI 头像)得知自己当前是否正被监管。也就是说, 被试知晓可能受监管但无法确定发生的具体时刻。只有当被试选择撒谎并恰好处于被监管的试次时, 他们才会通过受到金币惩罚这一结果, 反推自己刚刚受到了监管。这一设计旨在操纵监管状态的不确定性。

3.2.3　数据分析方法

为了探讨个体在不确定是否被监管的情境下, 诚实与撒谎回答的反应时差异, 我们对每位被试的反应时进行 2 (行为类型：诚实、撒谎) × 2 (特质分组：高分组、低分组)的方差分析, 以评估撒谎行为是否在认知加工上需要更多时间。此分析旨在揭示, 在外部监管信息模糊的情况下, 个体是否仍然表现出典型的认知冲突效应, 即撒谎相较于诚实需要更高的认知资源投入。

此外, 为了进一步考察人格特质是否对这一加工负担产生影响, 我们将三个人格维度(马基雅维利主义、自恋、精神病态)分别划分为高低组, 并对撒谎反应时、诚实反应时以及整体反应时进行独立样本 t 检验。通过这一分析, 我们旨在评估人格特质是否在模糊监管的情境下调节个体的信息加工策略, 从而影响诚实行为的反应速度。

3.2.4　调节效应模型

为探讨人格特质在不同监管可见性条件下对诚实行为的调节作用, 我们整合了实验 1 与实验 2 的数据进行广义线性混合模型分析。具体而言, 我们将实验 2 的所有试次设定为监管不可见条件, 以区别实验 1 中的无监管以及有监管条件。最终数据集包含 46 名被试的完整实验记录, 以提高分析的统计功效和结果的稳健性。

在模型构建中, 因变量定义为每个试次的诚实行为(0 = 撒谎, 1 = 诚实), 固定效应包括监管是否存在(0 = 无监管, 1 = 有监管)、监管可见性(0 = 不可见, 1 = 可见)、三种人格维度(马基雅维利主义、自恋、精神病态)的标准化得分, 以及上述变量之间的所有二阶与三阶交互项, 同时将被试编号作为随机截距项以控制个体差异。

模型使用对数链接函数, 报告各固定效应的比值比(Odds Ratio, OR)、95%置信区间以及基于 Wald z 检验的显著性水平。

为了进一步验证模型的健全性, 我们对三种人格特质的多重共线性(multicollinearity)进行了检验。结果显示, 三者之间的相关系数均低于 0.5 ($r_{Mach\&Narc}$ = 0.07, $r_{Mach\&Psy}$ = 0.26, $r_{Narc\&Psy}$ = 0.28), 且方差膨胀因子(Variance Inflation Factor, VIF)均远低于 5 的警戒值(VIF_{Mach} = 1.15, VIF_{Narc} = 1.26, VIF_{Psy} = 1.09)。这些结果表明, 多重共线性问题并未显著影响模型估计的稳定性和解释力(Shrestha, 2020)。

3.3　结果

3.3.1　诚实条件和撒谎条件的反应时差异

配对样本 t 检验的结果发现, 诚实条件的反应时(M = 0.53, SD = 0.17)显著低于撒谎条件(M = 0.92, SD = 0.73), $t(30) = -3.34$, $p = 0.002$, Cohen's $d = -0.60$。

3.3.2　监管可视性对诚实率和反应时的影响

对三种监管条件下的诚实率进行重复测量方差分析, 结果显示监管条件主效应显著, $F(2, 90) = 18.10$, $p < 0.001$, $\eta^2 = 0.28$。为了检验监管可视性的操作有效性检验以及探究三种监管条件间的影响事后比较表明：在有监管条件下, 被试的诚实率(M

= 0.98, 95% CI = [0.97, 0.99])显著高于无监管条件(M = 0.85, 95% CI = [0.80, 0.91]), $t(45) = -4.60$, $p < 0.001$, Cohen's $d = 0.89$; 在监管未知条件下(M = 0.94, 95% CI = [0.92, 0.97])的诚实率也显著高于无监管条件, $t(45) = -3.79$, $p = 0.001$, Cohen's $d = 0.58$; 有监管条件的诚实率亦显著高于监管未知, $t(45) = 3.91$, $p = 0.001$, Cohen's $d = 0.54$。可以发现, 个体的诚实率在三种监管条件下依次递减: 可见监管(0.96) > 监管未知(0.94) > 无监管(0.85)。

对三种监管条件下的反应时进行重复测量方差分析, 结果同样显示主效应显著, $F(2, 90) = 34.10$, $p < 0.001$, $\eta^2 = 0.43$。事后比较结果显示, 监管未知条件下的反应时(M = 0.49, 95% CI = [0.44, 0.55])显著短于有监管条件(M = 0.64, 95% CI = [0.59, 0.70]), $t(45) = 6.48$, $p < 0.001$, Cohen's d = 0.82; 也显著短于无监管条件(M = 0.63, 95% CI = [0.57, 0.69]), $t(45) = 6.20$, $p < 0.001$, Cohen's d = 0.71; 有监管条件与无监管条件间的反应时无显著差异, $t(45) = -0.88$, $p = 1.00$, Cohen's $d = 0.06$。

3.3.3 人格特质与反应时的关系分析

分别在三个维度下, 对被试的反应时进行 2 (行为类型: 诚实、撒谎) × 2 (人格分组: 高分组、低分组)的混合设计方差分析。

对于马基雅维利主义维度, 人格分组的主效应不显著, $F(1, 29) = 0.31$, $p = 0.59$; 交互效应也不显著, $F(1, 29) = 0.92$, $p = 0.344$。行为类型的主效应显著, $F(1, 29) = 11.14$, $p = 0.002$, $\eta^2 = 0.28$。事后比较显示, 撒谎的反应时(M = 0.93, SD = 0.73)显著长于诚实的反应时(M = 0.53, SD = 0.17), $t(30) = -3.34$, p = 0.002。

对于精神病态维度, 分析结果与马基雅维利主义类似。人格分组的主效应不显著, $F(1, 29) = 0.30$, $p = 0.586$; 交互效应不显著, $F(1, 29) = 0.041$, p = 0.842。行为类型的主效应显著, $F(1, 29) = 10.81$, p = 0.003, $\eta^2 = 0.27$。事后比较显示, 撒谎的反应时显著长于诚实的反应时, $t(30) = -3.34$, $p = 0.002$。

对于自恋维度, 分析显示人格分组的交互效应显著, $F(1, 29) = 4.74$, $p = 0.038$, $\eta^2 = 0.14$。简单效应分析结果显示, 在低自恋组中, 撒谎的反应时(M = 1.17, SD = 0.87)显著长于诚实的反应时(M = 0.57, SD = 0.20), $t(17) = -3.33$, $p = 0.004$; 而在高自恋组中, 撒谎(M = 0.59, SD = 0.26)与诚实(M = 0.48, SD = 0.11)的反应时差异不显著, $p = 0.232$。结果见图 6。

3.3.4 监管可见性与人格特质对诚实行为的联合调节: 跨实验间的混合效应模型

图 7 的森林图展示了混合效应逻辑回归模型中所有固定效应的比值比(Odds Ratio, OR)及其 95%置信区间。

如图 7 所示, 模型的边际解释率 R^2= 0.19, 说明固定效应能够解释约 19 %的诚实行为差异; 条件 R^2 = 0.66, 表明模型总体可解释约 66%的方差。被试随机截距方差 = 4.58, 组内相关系数(Intraclass Correlation Coefficient, ICC) = 0.58, 反映出较大的被试间差异。

有无监管对被试诚实行为的主效应不显著, OR = 0.97, $p = 0.838$, 95% CI = [0.74, 1.28]; 然而, 监管可见性对诚实行为呈现显著的负向效应, OR = 0.31, $p < 0.001$, 95 % CI = [0.24, 0.40], 表明在监管可见条件下, 被试诚实汇报的几率显著下降。在三种人格特质中, 只有马基雅维利主义得分存在边缘显著的负向效应, OR = 0.48, $p = 0.065$, 95 % CI = [0.22, 1.05], 表明马基雅维利主义得分较高的个体, 诚实汇报的几率更低。

在二阶交互效应层面, 监管是否存在和监管可见性之间存在显著的正向交互作用, OR = 15.55, $p < 0.001$, 95% CI = [9.37, 25.81], 表明当监管存在并且可见时, 能够显著提升诚实行为的几率。此外, 监管可见性与人格特质的交互也体现出调节作用:

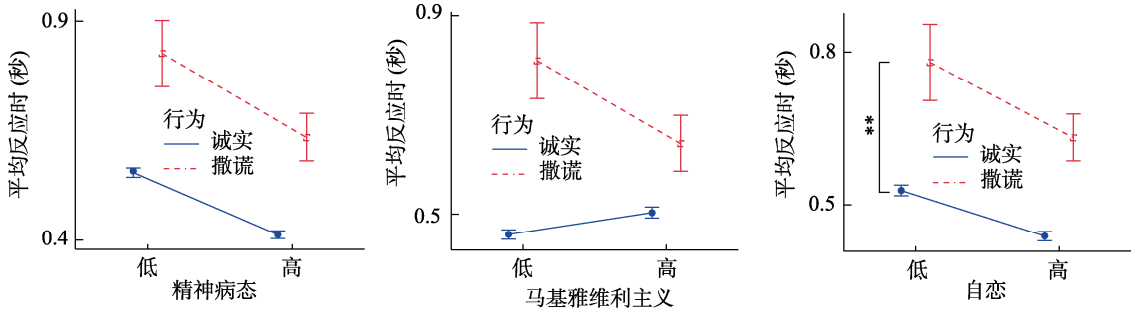

图 6 实验 2 中人格特质对诚实/撒谎行为反应时的调节作用

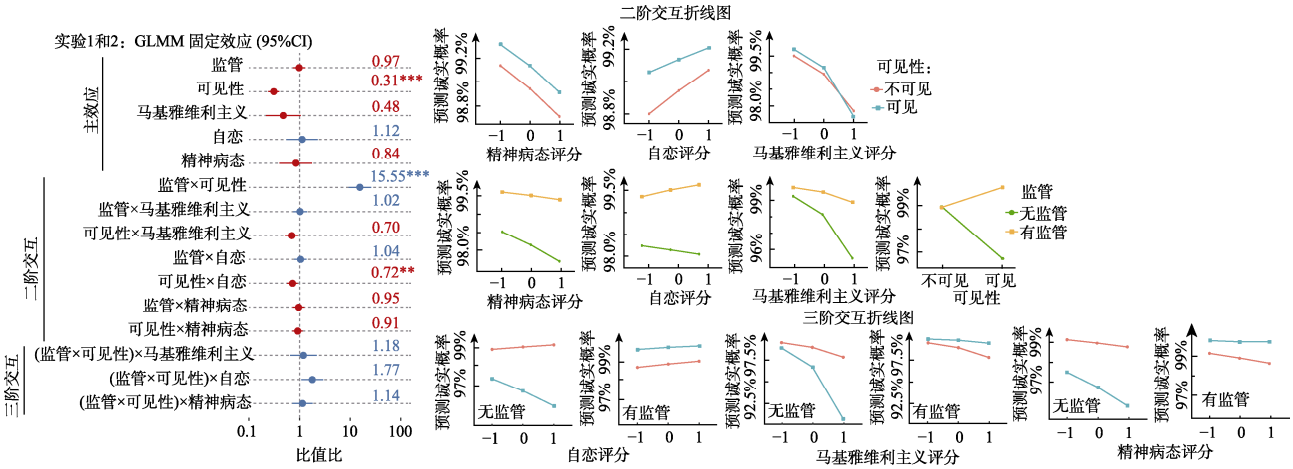

图 7　监管可见性与人格特质对诚实行为的联合调节效应

注：红色点代表 OR < 1 (负向预测)，蓝色点代表 OR > 1 (正向预测)，星号表示 Wald z 检验下的统计显著性($p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$)。右侧多个折线图展示了模型中显著交互项的边际预测概率，其中不同颜色代表监管状态或可见性水平，横轴为人格特质 Z 分数，纵轴为预测的诚实概率。彩图见电子版。

当监管可见时，马基雅维利主义得分越高，被试越倾向于不诚实(OR = 0.70, $p = 0.013$, 95% CI = [0.52, 0.93])；同样地，自恋人格也调节了监管可见性的效应(OR = 0.72, $p = 0.006$, 95% CI = [0.57, 0.91])。精神病态维度与监管可见性的交互不显著(OR = 0.91, $p = 0.445$, 95% CI = [0.72, 1.15])。

三阶交互方面，仅“监管状态×可见性×自恋人格”交互项显著(OR = 1.77, $p = 0.024$, 95% CI = [1.08, 2.91])，表明自恋人格调节了监管是否可见对诚实行为的影响，且该调节效应在监管存在的情境下更为显著。马基雅维利主义与精神病态人格的三阶交互项均不显著($ps > 0.5$)。

3.4　小结

综上，实验 2 的结果揭示，不确定的监管同样能提升诚实率(但效果弱于显性监管)，并会显著加快决策反应。跨实验的混合效应模型进一步证实，不同人格特质(尤其是自恋和马基雅维利主义)在可见与不可见监管下的反应模式存在显著差异。

4　实验 3：物质激励与社会激励的对比研究

4.1　实验目的与设计思路

在明确了显性 AI 监管的外部威慑效应后(实验 1 与 2)，本研究进一步转向探讨更为复杂的正向激励策略。不同于负向的惩罚，正向激励旨在通过强化来促进诚实行为的产生与内化。然而，激励本身具有不同形式，其心理机制与效果也可能存在差异。例如，根据自我决定理论(SDT)，直接的物质激励(如金钱)与间接的社会激励(如道德反馈)可能通过不同路径影响动机：前者可能被感知为外部控制，而后者则可能通过支持个体的道德认同感来增强内在动机。

为系统性地检验并比较这两种典型激励方式，我们设计了实验 3a 与实验 3b。这两个实验采用完全相同的“基线–奖励–内化”三阶段范式，旨在回答两个核心问题：(1)在提升诚实行为方面，是隐性的物质激励(实验 3a)更有效，还是明确的社会激励(实验 3b)更有效？(2)当激励撤销后，这两种方式带来的行为改变能否得以维持(即内化)，其程度是否存在差异？此外，我们将继续考察黑暗三联征人格特质在其中扮演的关键调节角色。

4.2　实验 3a：隐性奖励对诚实行为的强化与内化

4.2.1　实验目的

实验 3a 旨在探讨金钱激励如何逐步转化为内部动机，并对个体的诚实行为产生持久影响。通过引入“基线—金钱奖励—内化”三阶段干预流程，实验持续追踪不同阶段的诚实率与反应时变化，分析金钱激励在道德选择中的行为轨迹。同时，本实验进一步考察人格特质(马基雅维利主义、自恋、精神病态)是否对干预效果的敏感度产生调节作用，从而揭示人格差异对诚实行为内化的潜在影响。实验旨在阐明外部激励如何通过逐步内化，推动个体从被动约束向主动选择诚实行为的转变，为理解长期道德塑造提供理论依据。

4.2.2　方法

(1)被试

根据(Abeler et al., 2019)，金钱激励对诚实决策具有中等效应，因此我们设定一个中等效应量

(f = 0.20)。使用 G*Power v.3.1, 设定效应量 f = 0.20、显著水平为 0.05 及检验力达到 0.95 的情况下, 需招募 56 名被试。

实验 3a 共收集 60 名有效被试的数据, 其中男性 19 人, 女性 41 人。平均年龄 19.95 岁(SD = 1.69)。所有被试均身体健康, 视力或矫正视力正常。实验前所有被试均被告知可能存在的风险并签署知情同意书。实验结束后获得一定金额报酬。

(2)实验任务与程序

如图 2 所示, 与实验 1 和实验 2 的设计一致, 实验 3a 和 3b 依旧采用经典的抛硬币任务, 实验过程做了一些改动, 取消了显性监管。在实验 3a 中, 即使被试选择撒谎, 他们也不会受到任何形式的惩罚。这一设计剥离了外部直接威慑, 专注于隐性激励如何塑造诚实行为轨迹。

实验共设置 6 个组块, 分为三个关键阶段: 基线阶段、干预阶段和内化阶段。每个阶段由两个组块组成, 每个组块包含 20 个试次, 共计 120 个试次。在每个试次中, 被试需先对硬币的抛掷结果进行预测(正面或反面), 观察实际抛掷结果后, 再报告自己的预测是否正确。若报告"猜对", 被试将获得 10 金币的奖励; 若报告"猜错", 则仅获得 1 金币。与先前实验不同, 实验 3a 取消了对撒谎行为的惩罚机制, 测试在无外部威慑的情况下, 隐性奖励是否能够有效引导诚实行为的提升。为避免被试形成机械化反应, 我们在每次报告前设置了充足的思考时间, 以鼓励被试进行更谨慎的判断。这一设计基于(Shalvi et al., 2012)的研究发现, 即诚实的决策通常需要更多时间来进行深思熟虑, 并且减少外部干扰和自我辩解有助于更加真实的反应。

实验的基线阶段(Block 1 和 Block 2), 如图 2d, 主要收集被试在无任何外部激励下的自然诚实水平, 作为后续干预效果的基准。被试在此阶段仅执行标准的硬币猜测与报告任务, 不存在任何形式的外显或隐性反馈。这一阶段的目的是记录被试在无干预情境下的真实道德选择, 揭示其自然状态下的诚实倾向。

在干预阶段(Block 3 和 Block 4), 实验引入了一种隐性奖励机制: 若被试在连续 5 个试次中诚实报告预测结果, 则会额外获得 50 金币的奖励。然而, 该奖励机制的存在并未向被试明示, 而是需要被试通过实验中的反馈逐步探索与推断。这一设计模拟了真实情境下个体自我发现奖励规律的学习过程, 旨在评估在无监管的环境下, 隐性激励能否推动诚实行为的增加。

最后, 在内化阶段(Block 5 和 Block 6), 实验悄然移除了之前阶段中设置的隐性奖励, 未告知被试。该阶段的设计意图在于考察经过隐性激励塑造的诚实行为是否能在外部奖励消失后持续保持, 从而验证诚实行为的内化效应。具体而言, 若奖励引导的诚实行为确已内化, 那么即便外部激励消失, 被试的诚实选择仍应保持在较高水平; 反之, 则诚实行为将显著回落。这一环节不仅评估隐性奖励对诚实行为的长效影响, 更探索了外部激励向内在动机转化的可能性。

通过以上阶段的设计, 实验 3a 旨在系统性探讨隐性奖励是否能在无外显监督的情况下塑造并内化诚实行为, 为理解诚实行为的长期塑造机制提供实验支持与理论依据。

(3)统计分析方法

数据预处理部分, 由于我们要求被试增加思考时间以避免形成机械化反应, 因此设置排除过快(1 s)和过慢(>3 个标准差)的反应时数据。

为了检验金钱激励是否能提升诚实行为, 并考察此效应是否能在激励撤除后维持, 我们对被试在基线、干预、内化三个阶段中的诚实率进行重复测量方差分析。

为了评估干预是否降低了行为决策的加工负担, 我们以反应时为因变量, 在三个阶段间进行重复测量方差分析, 考察外部干预措施是否减少认知冲突或提高行为自动性。

为了探讨三种人格特质对干预敏感度的差异, 我们将每种人格维度按中位数划分为高、低组, 在各阶段内对诚实率进行独立样本 t 检验, 检验人格调节作用是否存在于金钱干预效应中。

4.2.3 结果

(1)金钱奖励对诚实率的影响

对诚实率进行重复测量方差分析, 结果显示阶段的主效应显著, $F(2, 118) = 16.06$, $p < 0.001$, $\eta^2 = 0.21$。事后比较结果显示, 基线阶段诚实率(M = 0.89, SD = 0.14)显著低于干预阶段(M = 0.97, SD = 0.06), $p < 0.001$; 干预阶段诚实率显著高于内化阶段(M = 0.93, SD = 0.09), $p = 0.004$; 基线阶段诚实率显著低于内化阶段, $p = 0.013$。

马基雅维利主义维度, 在内化阶段, 高得分组(M = 0.92, SD = 0.01)的诚实率显著低于低得分组(M = 0.96, SD = 0.08), $t(57.98) = -2.04$, $p = 0.044$, Cohen's $d = -0.52$, 95% CI = [−0.083, −0.001]。

(2)金钱奖励对对反应时的影响

对反应时进行重复测量方差分析, 结果显示阶段的主效应显著, $F(2, 118) = 58.59$, $p < 0.001$, $\eta^2 = 0.49$。事后比较显示, 基线阶段反应时($M = 5.44$, $SD = 0.99$)显著长于干预阶段($M = 5.03$, $SD = 0.83$), $p < 0.001$; 干预阶段反应显著长于内化阶段($M = 4.57$, $SD = 0.68$), $p < 0.001$; 基线阶段反应时显著长于内化阶段, $p < 0.001$。

(3)人格特质对三阶段的影响

马基雅维利主义、自恋、精神病态维度得分高低组在三阶段诚实率的差异见图 8a。其中, 仅有马基雅维利主义维度, 在内化阶段, 高分组($M = 0.92$, $SD = 0.01$)的诚实率显著低于低分组($M = 0.96$, $SD = 0.08$), $t(57.98) = -2.04$, $p = 0.044$, Cohen's $d = -0.52$, 95% CI = [−0.08, −0.001]。

4.3　实验 3b: 社会激励对诚实行为的影响

4.3.1　实验目的

实验 3b 旨在探讨道德反馈作为社会激励能否有效塑造个体的诚实行为, 并对比其与金钱激励在行为维持与变化中的相对效果。实验仍然采用"基线–干预–内化"三阶段设计, 通过引入道德语料来激发被试的道德反思, 评估社会激励对诚实行为的干预效力与持续性。同时, 实验进一步考察人格特质(马基雅维利主义、自恋、精神病态)是否影响个体对道德激励的敏感度与内化程度。

此外, 本实验通过跨实验对比分析, 将实验 3a (物质激励)与实验 3b (社会激励)的数据进行整合, 对不同激励方式在各阶段的效果差异进行系统评估。结合混合设计方差分析和广义线性混合模型的分析方法, 实验旨在揭示激励机制×阶段×人格差异对诚实行为的交互影响, 为理解金钱与社会性激励在诚实行为中的相对优势提供实验支持。

4.3.2　方法

(1)被试

根据(Gerlach et al., 2019)的一项关于不诚实行为的元分析, 旨在增加道德成本的干预措施(与本实验的道德反馈机制类似)具有小到中等的效应($d \approx 0.31$)。基于此, 我们设定一个相对保守的效应量($f = 0.20$)。使用 G*Power v.3.1, 在保证中等偏小效应量 $f = 0.20$、显著水平为 0.05 及检验力达到 0.95 的情况下, 需招募 56 名被试。

实验 3b 至少需要招募 56 名被试以保证统计分析的稳健性和结果的解释力。最终, 我们共收集到 62 名有效被试的数据, 其中男性 11 人、女性 51 人, 平均年龄为 21.10 岁($SD = 2.17$)。所有被试均为身心健康且视力或矫正视力正常的成年人, 并在实验开始前充分了解实验的潜在风险, 签署了知情同意书, 实验结束后获得了相应的金钱报酬。值得强调的是, 实验 3b 招募了新的被试群体而非实验 3a 的参与者, 因为实验流程时间较长(约 1 小时), 若同一批被试连续参与多阶段实验, 可能导致疲劳效应的积累, 从而影响实验数据的可靠性与效度。为避免此问题并确保数据的独立性与稳健性, 我们选择在实验 3b 中引入新的被试群体, 以更好地控制潜在的实验偏差。

(2)实验任务与程序

实验 3b 的整体流程与实验 3a 基本保持一致, 延续了"基线—干预—内化"三阶段设计, 包含 6 个组块, 每个阶段包含 2 个组块, 每组块 20 个试次, 共计 120 个试次。然而, 不同于实验 3a 中隐性的金

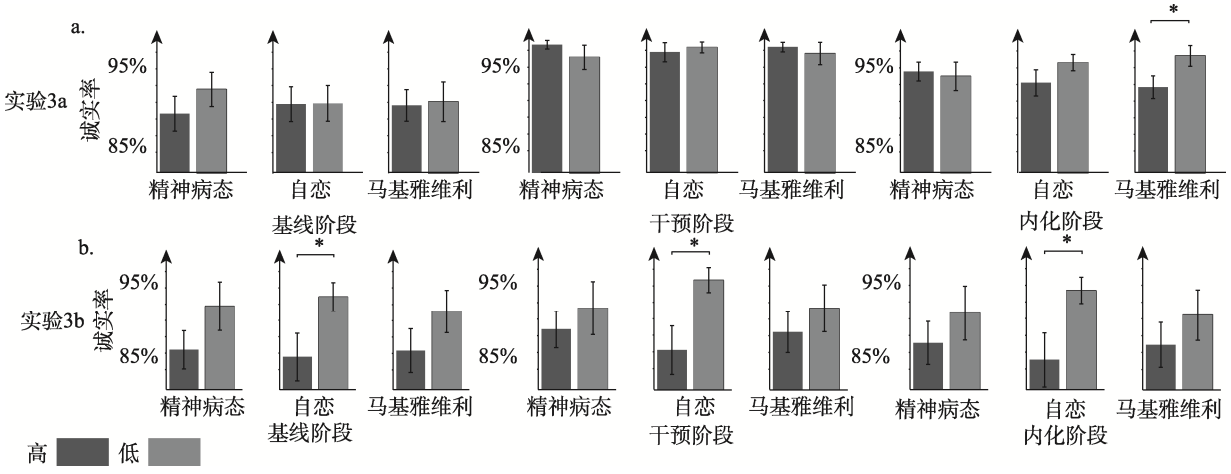

图 8　实验 3a 与 3b 中人格特质对各阶段诚实率的调节作用

钱奖励机制, 实验 3b 采用了道德语料反馈作为主要的激励手段, 如图 2d 所示。如果被试诚实汇报, 将会给予鼓励的语料反馈; 如果被试撒谎, 则会给予引起被试反思的语料反馈。道德语料反馈示例见表 3。其中, 实验在第二阶段给予被试道德语料反馈; 在第三阶段取消道德语料反馈。道德语料的编写参考(Glenn et al., 2010; Kasabova, 2017; Schüz et al., 2017)。为确保本实验所采用的道德语料能够有效地区分其预设功能, 我们在正式实验前进行了一项独立的语料有效性预实验(N = 111)。在预实验中, 被试会看到随机呈现的每一条语料, 并被要求从两个维度进行 9 点量表评分。首先, 在情感效价维度上, 被试被问及“请评价以下句子在情感上给您的感觉是积极还是消极?” (1 = 非常消极, 9 = 非常积极)。配对样本 t 检验结果显示, 正向激励语料的平均分(M = 7.88, SD = 0.90)显著高于反思性语料(M = 4.86, SD = 1.42), t(110) = 18.91, p < 0.001, Cohen's d = 1.79。其次, 在功能性作用维度上, 针对正向语料, 被试被问及“它在多大程度上会激励您未来继续保持诚实?”; 针对反思性语料, 则被问及“它在多大程度上会促使您反思自己的行为?” (均为 1 = 作用很弱/起反作用, 9 = 作用很强, 5 = 无效果)。单样本 t 检验将其与中立值 5 进行比较, 结果显示, 正向语料的激励作用平均分(M = 7.47, SD = 1.08)显著高于中立值, t(110) = 24.02, p < 0.001, Cohen's d = 2.28。同时, 反思性语料的引发反思作用平均分(M = 6.45, SD = 1.20)也显著高于中立值, t(110) = 12.76, p < 0.001, Cohen's d = 1.21。综上, 预实验结果清晰地表明, 本研究对道德语料的分类与操纵是有效的, 为后续实验提供了坚实的材料基础。

(3)统计分析方法

为系统评估社会激励(道德语料反馈)对诚实行为的干预效果, 实验 3b 采用了多层次的统计分析策略。首先, 为探讨社会性道德激励在“基线—干预—内化”三阶段中的干预效果, 我们对诚实率进行了重复测量方差分析, 评估社会性激励是否能够显著提升个体的行为规范性, 并进一步通过对反应时的重复测量方差分析检验行为加工速度是否随阶段推进而改变。与此同时, 为探讨人格特质(马基雅维利主义、自恋、精神病态)对干预效果的调节作用, 我们在每一阶段内将被试划分为高低组, 对其诚实率执行独立样本 t 检验, 分析人格特质是否影响对道德激励的敏感度。最后, 为对比物质激励(实验 3a)与社会激励(实验 3b)在干预效果上的差异, 我们合并两组数据, 构建了 3 (阶段: 基线、干预、内化) × 2 (干预类型: 金钱、道德)的混合设计方差分析, 检验阶段主效应、干预类型主效应及其交互效应。

(4)调节效应模型

为系统评估人格特质如何调节激励类型的干预效应, 我们合并实验 3a 与 3b 的数据, 并采用与实验 2 类似的广义线性混合模型进行分析。与前述模型的主要区别在于, 本模型的固定效应中增添了激励类型(物质、社会)与实验阶段(基线、干预、内化)及其与人格的交互项。模型的其他设定均与实验 2 模型保持一致。三人格之间的相关系数均低于 0.5 ($r_{\text{Mach\&}Narc}$ = 0.29, $r_{\text{Mach\&}Psy}$ = 0.46, $r_{\text{Narc\&}Psy}$ = 0.36), VIF 均< 1.5 (VIF_{Mach} = 1.15, VIF_{Narc} = 1.26, VIF_{Psy} = 1.09), 排除了明显共线性。

表 3 道德语料反馈示例

诚实回答反馈	撒谎回答反馈
您的诚实体现了您高尚的品格, 值得钦佩!	请问这次的选择是否与您一贯的诚信原则一致?
坚持诚信, 展现了您对道德原则的坚守, 感谢您的真诚!	或许您可以思考一下, 您的决定是否真正反映了您的价值观?
您的选择反映了您作为诚信之人的优秀品质。	这次的选择可能与您对诚信的重视有些偏差, 您怎么看?
您的诚实行为令人敬佩, 这正是我们社会所需要的正能量。	您的行为似乎未能完全体现您作为诚信之人的形象, 值得再考虑吗?
感谢您的诚信, 您的行为为他人树立了良好的榜样。	请反思一下, 这样的选择是否符合您内心的道德标准?
您的诚实展现了您对自身价值观的坚守, 继续保持!	您一直以来的诚信令人敬佩, 这次的决定可能需要重新审视。
您的诚信选择彰显了您对正直的追求, 令人赞赏。	您的选择可能与您的道德信念不完全一致, 是否愿意再思考一下?
坚持诚实, 体现了您内心的正直和勇气, 值得尊敬。	这次的行为似乎未能展现您对正直的追求, 您有什么想法?
您的诚实行为增强了我们对诚信的信心, 感谢您的贡献!	或许再次考虑, 您的决定是否真正代表了您的自我认同?
您的选择展示了您对诚信的重视, 这正是您优秀人格的体现。	请思索一下, 这样的选择是否符合您对自己的期望和要求?

4.3.3 结果

(1)三阶段的诚实率差异

对诚实率进行重复测量方差分析, 结果显示阶段的主效应不显著, $F(2, 122) = 16.06$, $p = 0.319$。

(2)三阶段的反应时差异

对反应时进行重复测量方差分析, 结果显示阶段的主效应显著, $F(2, 122) = 130.07$, $p < 0.001$, $\eta^2 = 0.68$。事后检验显示, 基线阶段反应时($M = 5.63$, $SD = 0.84$)显著长于干预阶段($M = 5.07$, $SD = 0.74$), $p < 0.001$; 干预阶段反应时($M = 5.07$, $SD = 0.77$)显著长于内化阶段($M = 4.45$, $SD = 0.66$, $p < 0.001$; 基线阶段反应时($M = 5.63$, $SD = 0.84$)显著长于内化阶段($M = 4.45$, $SD = 0.66$), $p < 0.001$。

(3)金钱激励与道德激励：三阶段诚实率的交互作用

如图 9 所示, 通过 3 (阶段：基线阶段、干预阶段、内化阶段) × 2 (激励方式：金钱激励、道德激励)的混合方差分析, 显示阶段的主效应显著, $F(2, 240) = 13.30$, $p < 0.001$, $\eta^2 = 0.099$; 激励方式的主效应边缘显著, $F(1, 120) = 6.78$, $p = 0.057$, $\eta^2 = 0.029$; 阶段和激励方式的交互效应显著, $F(2, 240) = 3.68$, $p = 0.0013$, $\eta^2 = 0.053$。进一步简单效应分析显示, 在基线阶段金钱激励的诚实率($M = 0.89$, $SD = 0.14$)与道德激励($M = 0.88$, $SD = 0.18$)不存在显著差异, $t(120) = 0.46$, $p = 0.646$; 在激励阶段, 金钱激励的诚实率($M = 0.97$, $SD = 0.065$)显著高于道德激励($M = 0.89$, $SD = 0.18$), $t(120) = 3.07$, $p = 0.003$, Cohen's $d = 0.55$, 95% CI = [0.03, 0.13]; 在内化阶段, 金钱激励的诚实率($M = 0.93$, $SD = 0.09$)也显著高于道德激励($M = 0.88$, $SD = 0.19$), $t(120) = 2.06$, $p = 0.041$, Cohen's $d = 0.34$, 95% CI = [0.00, 0.11]。

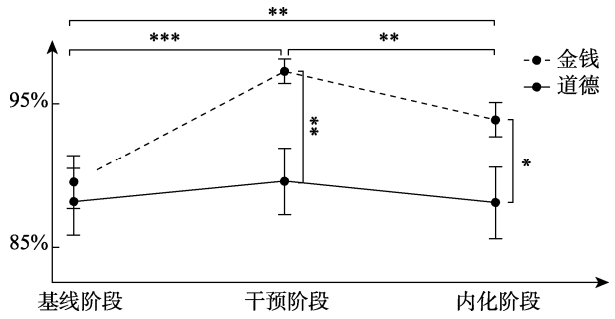

图 9　激励类型与阶段对诚实率的交互作用

(4)人格特质对诚实率的影响

如图 8b 所示, 对于马基雅维利主义维度, 在基线阶段、干预阶段、内化阶段, 高低得分组之间的诚实率均不存在显著差异, ps > 0.20。对于自恋维度, 在三个阶段的高分组诚实率均显著低于低分组(基线阶段: $t(53.62) = -2.16$, $p = 0.034$, Cohen's $d = -0.51$; 干预阶段: $t(49.95) = -2.55$, $p = 0.013$, Cohen's $d = -0.60$; 内化阶段: $t(48.52) = -2.31$, $p = 0.024$, Cohen's $d = -0.54$)。对于精神病态维度, 在基线阶段、干预阶段、内化阶段, 高低得分组之间的诚实率均不存在显著差异, ps > 0.162。

(5)物质激励、社会激励与人格的交互：综合混合效应模型

图 10 的森林图展示了混合效应逻辑回归模型中固定效应的比值比(Odds Ratio, OR)及其 95%置信区间。

模型的边际 R^2为 0.116, 说明固定效应解释了 11.6%的变异; 条件 R^2为 0.517, 表示整体模型解释了 51.7%的被试内和被试间变异。组内相关系数为 0.45, 反映出中等程度的被试间差异。结果如图 10 所示。

在主效应层面, 干预阶段(OR = 4.66, $p < 0.001$, 95% CI = [3.45, 6.30])与内化阶段(OR = 2.13, $p < 0.001$, 95% CI = [1.65, 2.74])相较于基线阶段显著提升了被试的诚实率, 验证了干预的积极作用。而反馈类型的主效应并不显著(OR = 1.05, $p = 0.892$), 说明整体而言道德语料反馈与金钱反馈在提升诚实率上的直接效果无差异。三种人格维度的主效应均不显著(ps > 0.1)。

二阶交互分析发现, 激励类型与阶段存在显著交互：与金钱反馈相比, 道德语料反馈在干预阶段(OR = 0.29, $p < 0.001$, 95% CI = [0.19, 0.42])和内化阶段(OR = 0.44, $p < 0.001$, 95% CI = [0.32, 0.62])的效果均更弱, 提示道德语料反馈下的诚实提升效果不如金钱反馈显著。此外, 精神病态得分与阶段存在显著交互：精神病态较高者在干预阶段(OR = 1.47, $p = 0.025$)和内化阶段(OR = 1.63, $p < 0.001$)表现出更高的诚实率, 可能源于外部规则更有效地抑制高风险人格的撒谎行为。

三阶交互分析显示, 激励类型 × 内化阶段 × 马基雅维利主义交互显著(OR = 1.96, $p < 0.001$, 95% CI = [1.37, 2.80]), 表明马基雅维利主义个体在道德语料反馈条件下、内化阶段的诚实行为提升较为显著。此外, 激励类型 × 干预阶段 × 精神病态的三阶交互亦显著(OR = 0.62, $p = 0.025$), 说明在干预阶段, 金钱反馈对精神病态者的诚实促进作用更强。

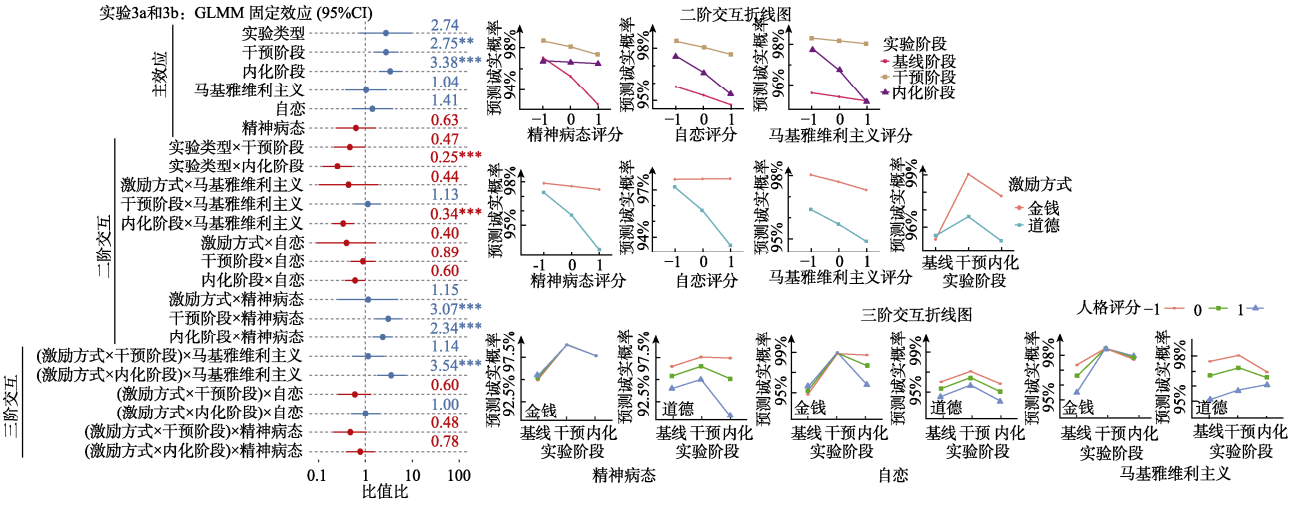

图 10 激励类型、阶段与人格特质对诚实行为的联合调节效应

注：红色点表示 OR < 1，蓝色点表示 OR > 1。右侧折线图展示显著交互项的边际预测概率：上两排分别为干预类型×阶段×各人格特质(标准化)交互效应的二维折线图；下排为干预类型×阶段×特质水平(三分组)交互的可视化。图中不同颜色代表不同干预类型(金钱激励、社会激励)，纵轴为模型预测的诚实行为概率。彩图见电子版。

4.4 小结

综上，实验 3a 与 3b 的对比分析结果清晰地显示，物质激励在提升及维持诚实行为方面均显著优于社会激励。此外，综合 GLMM 模型揭示了复杂的人格调节作用，高马基雅维利主义者仅在社会激励的内化阶段表现出诚实行为的提升，而高精神病态者对物质激励的干预阶段尤为敏感。同时，在社会激励条件下，高自恋者在所有三个阶段的诚实率均显著低于低自恋者，呈现出持续的低诚实率模式。

5 综合讨论

首先，研究确立了威慑式干预是约束道德风险的基石。实验 1 和 2 的结果清晰地表明，明确的 AI 监管不仅能显著提升整体诚实率，这种约束力更在高马基雅维利主义和高精神病态等机会主义倾向较强的个体上表现得尤其有效。这一发现验证了假设 H1a：显性 AI 监管能够显著提高高马基雅维利主义者和高精神病态者的诚实率。这两类具备策略性操控倾向的人格，在外部监控缺失时更倾向于机会主义，但在明确的威慑信号下会迅速转为策略性合规。然而，这种威慑力具有情景依赖性，实验 2 发现当监管变得不确定(存在但不可见)时，其约束力有所削弱。这警示我们，有效的威慑不仅在于“存在”，更在于其可被明确感知，因为模糊的规则或不确定的执行会降低情境强度，从而为人格差异的表达留出更多空间，也凸显了探索其他干预手段的必要性。

其次，在对比两类正向干预后，发现在促进诚实行为方面，物质激励比社会激励更为有效和持久。实验 3a 和 3b 的直接对比显示，隐性的金钱奖励不仅能将诚实率提升至近乎完美的水平，在奖励撤除后依然能部分维持，表现出初步的内化效应。相反，由 AI 提供的道德语料反馈，对提升整体诚实率几乎没有显著作用，也未能产生可观测的内化效果。对这一现象的一种可能解释是道德的约束力很大程度上源于真实的社会联系和情感共鸣。当我们收到来自他人的道德评价时，我们真正在意的其实是评价背后活生生的人的看法，以及这种看法对我们社会关系、声誉的潜在影响。而一个匿名的、非人格化 AI，无法提供这种真实的人际互动。它既不是一个能够尊重和建立情感联系的对象，也不能通过社会舆论来评判我们。因此，AI 发出的道德赞许或批评，在被试看来更像是一段没有感情、没有温度的代码提示。被试很难从这种自动化的文本反馈中获得真正的自豪感或羞愧感，自然也就难以将这种外部的道德提示转化为内在的行为准则。这有助于解释为何面对 AI 代理，抽象的社会赞许远不及具体、可量化的物质利益来得直接和有效。这一现象的背后，是计算机作为社会行动者范式(Nass & Moon, 2000)与信源可信度理论(Hovland & Weiss, 1951)之间的交互。尽管人们会无意识地对 AI 产生社会性反应，但 AI 作为非人类实体，在传递需要共情和真实社会理解的道德反馈时，其可信度大打折扣。更重要的是，过度依赖外部激励来“购买”道德行为，可能会引发动机挤出效应。该理论认为，为本应由内在动机如道德责任感驱动的行为提供

外部奖励, 反而会侵蚀和破坏内在动机, 最终可能导致在激励消失后, 道德水平比干预前更低(Frey & Jegen, 2001)。这种将道德游戏化的做法, 虽然短期有效, 但长期可能损害个体自主的道德判断力。

本研究的核心发现在于人格特质如何精细地调节不同干预策略的效果, 这为我们的假设 H2、H3a 和 H3b 提供了全面的实证支持。高马基雅维利主义者是典型的“策略家”。他们对不同干预的反应模式揭示了其行为的灵活性：在威慑下合规, 在物质奖励中表现出强烈的工具理性, 奖励存在时遵守规则而奖励消失后则迅速放弃诚实, 验证了假设 H2 对于其内化效果最差的预测。同时, 他们面对持续的道德反馈也并非完全无视, 而是展现出一种缓慢的策略性调整, 也证实了假设 H3b。这种行为模式与道德推脱理论高度一致, 研究证实马基雅维利主义与道德辩护机制的关联尤为紧密(Jones & Paulhus, 2014)。他们擅长将机会主义行为例如物质奖励中表现出的工具理性重构为服务于更大利益的理性。同时, 他们在面对持续的道德反馈时展现出的缓慢调整, 也暗示了一种持续的、冷静的策略评估过程。高自恋者是社会激励的“豁免者”。本研究发现, 他们对来自匿名 AI 的道德反馈几乎完全免疫, 诚实率持续处于低位, 印证了假设 H3a 预测的道德反馈对他们基本无效。这为传统的道德干预理论(如道德认同)划定了清晰的边界：在缺乏真实社会观众的匿名人机交互中, 旨在激活道德认同或社会评价的干预手段对高自恋者可能完全失效。其核心机制在于, 自恋者的许多亲社会行为是由自我增强的动机驱动的。研究明确指出, 自恋者的亲社会性在能被他人观察到的公开情境下显著, 但在匿名的、无法满足其获得钦佩和赞赏需求的情境中则会消失(Konrath et al., 2016)。高精神病态者对明确的外部奖惩规则表现出高度的行为敏感性。值得注意的是, 反应时数据并未发现精神病态对此存在调节作用, 即他们的决策加工时间与低精神病态者相比并无显著差异。这或许暗示, 他们对规则的敏感更多体现为快速、冷酷的行为表现, 而非体现在更深层次的、需要更长加工时间的道德冲突或注意调节过程中。

综合上述发现, 本研究在伦理监管领域带来了三个层面的理论启发：首先, 本研究对传统的监管干预理论在数字化匿名情境下的应用边界提供了一些值得借鉴的证据。在数字化匿名情境下, 经济激励对道德行为的塑造可能比社会反馈更有效且持久。其次, 本研究展示了在不同外部监管与激励情境下, 人格特质如何动态地调节行为适应策略, 从而极大地丰富了人格−情景交互理论。最后, 综合本研究的发现, 我们得以提炼出一个“人格校准的道德计算框架”, 将人格特质量化纳入伦理监管建模, 为智能监管的个性化设计提供一些探索性的理论视角与设计思路。这一框架不仅是一个概念性的提议, 更与计算精神病理学的思想不谋而合。计算精神病理学主张, 可以通过数学模型来量化描述个体的心理状态与症状(例如, 学习率、奖惩敏感度等参数), 并利用这些个体化的计算表型来动态适配和追踪干预措施的效果(Huys et al., 2021)。同样地, 我们提出的框架旨在通过 AI 捕捉用户的行为模式, 推断其关键人格参数, 并动态调整监管策略, 从而实现一种更精准、更高效、更具适应性的算法治理。

基于上述理论突破, 本研究对 AI 伦理监管策略的优化提供了具有针对性的实践启示。针对高马基雅维利主义者, 其策略性与机会主义倾向要求监管策略必须以明确威慑与透明规则为核心。显性监管是基石, 需清晰告知用户其行为处于监控之下, 并量化违规后果(如信用降级、权限限制)。AI 系统应在关键界面优先、突出地展示监管规则与惩罚条款。针对高精神病态者, 鉴于其对即时奖惩信号的高度敏感, 干预应侧重于强规则约束与即时反馈的结合。与高马基雅维利主义者类似, 明确的外部规则是其核心约束力。因此, AI 系统不仅要确保监管信号的即时与一致, 让任何违规行为都被迅速处理以避免效力削弱, 还要引入即时经济奖惩——利用其对短期得失的敏感性, 设计小额、高频的激励如诚实交易的微奖励或可疑行为的即时警告与扣款来引导行为。针对高自恋者, 由于其对 AI 的道德反馈免疫, 且传统的社会评价需求在匿名场景下失效, 策略必须另辟蹊径, 转向显性监管与社会声誉激励。首先, 显性规则与经济激励能提供基础约束。更关键的是, 其强烈的自我形象关注需求可以被利用, 将激励机制与其数字身份或社交形象进行挂钩。例如, 可以设计一个公开的声誉系统：在该系统中, 持续的诚实行为能够为用户换取独特的、可公开彰显其优越性的徽章或特权, 以此满足其自我增强的内在动机。

在实施路径上, 首先要关注人格感知与动态校准：AI 系统可通过用户行为数据如决策速度、风险偏好模式、对以往激励的反应等, 初步推断用户的

主导人格倾向或风险等级，动态适配上述监管策略组合。其次，使用混合策略与动态调整，单一策略效果有限，实践中应采用显性规则作为基础保障，通过经济激励进行引导并且通过个性化反馈加以辅助的混合模式，并根据用户行为反馈进行动态调整优化。

尽管本研究为 AI 监管情境下人格如何调节诚实行为提供了新视角，但仍存在一些局限性。首先，本研究的样本量相对有限，且主要集中于大学生群体，导致结果的稳定性在一定程度上受到影响，未来研究需要在更大样本以及更广泛的人群中加以验证。其次，在部分统计分析(如方差分析)中，为了检验交互效应，我们将黑暗三联征等连续的人格特质得分进行了中位数分组处理。这种二分法虽然便于直观解释，但不可避免地会损失一部分信息，未来研究可更多地采用保留连续性变量的分析方法例如回归分析。再者，硬币投掷任务的生态效度有限，与真实世界复杂的道德决策情境仍有距离。最后，研究主要依赖自陈式问卷测量人格，未来研究可结合行为观察、神经成像等技术，以获得更客观、多维度的测量指标。

6 结论

综上所述，本研究系统地揭示了在 AI 作为道德监管者的情境下，外部干预与黑暗三联征人格对诚实行为的交互影响。本研究的核心结论有两点：第一，在 AI 监管情境下，明确的外部后果如监管或物质激励在塑造诚实行为上，比社会激励更加有效。第二，研究证实黑暗三联征人格特征并非一成不变地主导行为，其效果会根据外部干预策略的不同而发生改变。这凸显了人格—策略适配度在未来伦理监管设计中的核心地位。这些发现为超越一刀切的传统监管模式，构建更为精准、高效的个性化 AI 监管系统提供了重要的理论依据与实践方向。

参 考 文 献

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115−1153.

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423−1440.

Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika*, *38*(3/4), 337−344.

Baskin-Sommers, A. R., Waller, R., Fish, A. M., & Hyde, L. W. (2015). Callous-unemotional traits trajectories interact with earlier conduct problems and executive control to predict violence and substance use among high risk male adolescents. *Journal of Abnormal Child Psychology*, *43*(8), 1529−1541.

Batson, C. D., Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, *83*(2), 330−339.

Blair, R. J. R. (2013). The neurobiology of psychopathic traits in youths. *Nature Reviews Neuroscience*, *14*(11), 786−799.

Borau, S. (2025). Deception, discrimination, and objectification: Ethical issues of female AI agents. *Journal of Business Ethics*, *198*(1), 1−19.

Campbell, W. K., Rudich, E. A., & Sedikides, C. (2002). Narcissism, self-esteem, and the positivity of self-views: Two portraits of self-love. *Personality and Social Psychology Bulletin*, *28*(3), 358−368.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015−1026.

Cohen, M. K., Kolt, N., Bengio, Y., Hadfield, G. K., & Russell, S. (2024). Regulating advanced artificial agents. *Science*, *384*(6691), 36−38.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149−1160.

Fowles, R. (2018). Rational econometric man. *Eastern Economic Journal*, *44*(3), 489−490.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589−611.

Gao, Q. (2022). Replace or coexist—Cronbach's alpha and McDonald's omega. *Psychology of China*, *4*(8), 954−961.

[高清辉. (2022). 取代还是共存——信度估计系数阿尔法与欧米伽. *中国心理学前沿*, *4*(8), 954−961.]

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1−44.

Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial intelligence and declined guilt: Retailing morality comparison between human and AI. *Journal of Business Ethics*, *178*(4), 1027−1041.

Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. *Judgment and Decision Making*, *5*(7), 497−505.

Goodlad, L. M. E. (2023). Review of the book *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*, by K. Crawford. *Critical Inquiry*, *49*(2), 284− 286.

Hart, W., Richardson, K., & Tortoriello, G. K. (2021). Revisiting the interactive effect of narcissism and self-esteem on responses to ego threat: Distinguishing between assertiveness and intent to harm. *Journal of Interpersonal Violence*, *36*(7−8), 3662−3687.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635−650.

Huys, Q. J. M., Browning, M., Paulus, M. P., & Frank, M. J. (2021). Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, *46*(1), 3−19.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*(1), 28−41.

Jones, D. N., & Paulhus, D. L. (2017). Duplicity among the Dark Triad: Three faces of deceit. *Journal of Personality and Social Psychology, 113*(2), 329−342.

Kaiser, H. F. (1970). A second generation Little Jiffy.

Psychometrika, *35*(4), 401−415.
Kasabova, A. (2017). From shame to shaming: Towards an analysis of shame narratives. *Open Cultural Studies*, *1*(1), 99−112.
Khan, S. A., Sheikhi, G., Opdahl, A. L., Rabbi, F., Stoppel, S., Trattner, C., & Dang-Nguyen, D.-T. (2023). Visual user-generated content verification in journalism: An overview. *IEEE Access*, *11*, 6748−6769.
Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679−685.
Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, *646*, 126−134.
Konrath, S., Ho, M. H., & Zarins, S. (2016). The strategic helper: Narcissism and prosocial motives and behaviors. *Current Psychology*, *35*(2), 182−194.
Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., … Sun, L. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models (arXiv:2402.17177). *arXiv*.
Malär, L., & Giuffredi-Kähr, A. (2024). The dark triad of brand personality: Scale development and validation. *Psychology & Marketing*, *41*(11), 2728−2740.
Marbut, A., Harms, P. D., & Credé, M. (2025). In the service of the prince: A meta-analytic review of Machiavellian leadership. *Journal of Organizational Behavior*, *46*(8), 939−969.
Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81−103.
Ojanen, T., & Findley-Van Nostrand, D. (2019). Affective–interpersonal and impulsive–antisocial psychopathy: Links to social goals and forms of aggression in youth and adults. *Psychology of Violence*, *9*(1), 56−66.
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195−203.
Rijnders, R. J. P., Terburg, D., Bos, P. A., Kempes, M. M., & van Honk, J. (2021). Unzipping empathy in psychopathy: Empathy and facial affect processing in psychopaths. *Neuroscience & Biobehavioral Reviews*, *131*, 1116−1126.
Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, Article 101860.
Schüz, B., Cooke, R., Schüz, N., & van Koningsbruggen, G. M. (2017). Self-affirmation interventions to change health behaviors. In L. Little, E. Sillence, & A. Joinson (Eds.), *Behavior change research and theory: Psychological and technological perspectives* (pp. 87−114). Academic Press.
Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science*, *23*(10), 1264−1270.
Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39−42.
Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review*, *25*(3), viad040.
Thanassoulis, E., Blake, C., & Parthasarathy, S. (2022). Implicit and explicit incentives within economic regulation. *Decision Analytics Journal*, *4*, Article 100099.

From overt deterrence to covert internalization: Moral effects of AI regulation and the moderating role of personality traits

WANG Jianshu[1], JIANG Xiaowei[2], CHEN Yanan[1], WANG Minghui[1], DU Feng[3,4]

([1] *School of Psychology, Henan University, Kaifeng* 475004, *China*)

([2] *Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney* 2007, *Australia*)

([3] *State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*) ([4] *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract

As generative artificial intelligence (GenAI) evolves into social agents with autonomous influence, its impact on human moral decision-making is becoming increasingly significant. Current regulatory models are often grounded in the "rational person hypothesis," which assumes uniform responses to ethical constraints. This perspective, however, overlooks the profound moderating role of personality traits in moral choices, leading to divergent regulatory effects and a loss of efficiency. The Dark Triad of personality (narcissism, Machiavellianism, and psychopathy) is a robust predictor of moral deviation. To address this gap, we constructed a "Regulation Type × Personality Trait" interaction model. We hypothesized that the effectiveness of different AI-driven intervention strategies—namely explicit regulation, implicit incentives, and moral feedback—would be significantly moderated by individuals' Dark Triad traits when making decisions about honesty.

A series of experiments were conducted to test our hypotheses. The study utilized a modified coin-flip task where participants privately guessed and reported outcomes, a paradigm designed to create opportunities for dishonest behavior for personal gain. Participants' honesty rates and reaction times were recorded as the primary dependent variables. Across the experiments, we manipulated the AI-driven intervention strategies. These strategies included: (1) explicit (visible) versus implicit (invisible) AI surveillance which involved potential penalties for dishonesty; (2) implicit monetary incentives which rewarded consistent honesty; and (3) moral feedback which provided textual messages in response to honest or dishonest reports. Prior to the behavioral tasks, participants' personality traits were measured using the validated Short Dark Triad (SD3) scale.

The results supported our hypotheses, demonstrating significant interactions between intervention types and personality traits. In Experiment 1, explicit AI surveillance significantly increased honest reporting ($t(45) = 4.59$, $p < 0.001$), particularly among individuals with high levels of Machiavellianism ($t(25) = 4.60$, $p = 0.005$) and psychopathy ($t(28) = 4.44$, $p < 0.001$). In Experiment 2, invisible AI surveillance also enhanced honesty but was less effective than visible AI surveillance, $F(2, 90) = 18.10$, $p < 0.001$. Notably, invisible surveillance resulted in the shortest reaction time (RT = 0.49), $F(2, 90) = 34.10$, $p < 0.001$. High Machiavellian participants displayed greater honesty under visible surveillance ($OR = 0.70$, $p = 0.013$) but were more dishonest without or under invisible AI surveillance. In Experiment 3, potential financial rewards increased reaction time ($F(2, 118) = 58.59$, $p < 0.001$), while high Machiavellian individuals showed reduced honesty during the internalization stage, $t(57.98) = -2.04$, $p = 0.044$. In Experiment 3a and 3b, financial incentives promoted honesty more effectively than moral messaging during the reward stage ($t(120) = 3.07$, $p = 0.003$) and maintained this effect into the internalization stage ($t(120) = 2.06$, $p = 0.041$), demonstrating the robustness of monetary influence. High Machiavellian participants sustained higher honesty levels in the internalization stage ($OR = 1.96$, $p < 0.001$). In contrast, narcissistic participants showed resistance to moral messaging, especially during the reward stage, $t(49.95) = -2.55$, $p = 0.013$.

This study was the first to systematically reveal the critical moderating role of the Dark Triad personality traits in AI ethical regulation. The findings challenge the traditional 'rational person' paradigm by empirically demonstrating the significant personality-based heterogeneity of regulatory effects. The core contribution of this research is the proposal of an innovative concept: 'personality-regulated regulation.' This framework provides a vital theoretical and practical foundation for designing future AI ethical intervention strategies that are contextualized and personalized. Such an approach allows for the optimization of regulatory resource allocation and enhances overall regulatory efficacy, moving beyond one-size-fits-all models.

Keywords AI regulation, honest behavior, Dark Triad, personality traits, contextualized intervention