《心理学报》审稿意见与作者回应

题目：算法歧视比人类歧视引起更少道德惩罚欲
作者：许丽颖，喻丰，彭凯平

第一轮

审稿人 1 意见：

该研究探讨了大数据和人工智能快速发展的新形势下，道德心理学研究面临的现实问题，具有新颖性和借鉴价值。研究过程及其表述方式较为规范，结论具有一定的有理性。

意见：有一个问题希望作者考虑：人们对某事物的“道德惩罚欲望”与约定俗成的“道德惩罚阈限”有着密切的联系——在高于或低于特定“道德惩罚阈限”的情况下，人们的“道德惩罚欲望”是会有显著差异的；因此，在脱离“道德惩罚阈限”的前提下，单纯地论述“道德惩罚欲望”及其差异，是否显得有些牵强？

回应：非常感谢审稿专家对本文的肯定以及提出的宝贵意见。针对“道德惩罚阈限”问题，我们在文章的局限与展望部分增加了如下内容：“此外，人们日常生活中约定俗成的“道德惩罚阈限”也可能对道德惩罚欲产生影响。即对于非人对象而言，人们对其道德惩罚欲可能本身就存在一定的阈限，哪怕其道德违规再严重也难以突破一定阈值，因此在之后的研究中也可以对相关阈限问题进行进一步的探讨。”

审稿人 2 意见：

作者通过多个精心设计的研究，探讨了人们对算法歧视和人类歧视的道德惩罚差异，选题相当新颖，结果可信，方法合理。具体意见如下：

意见 1：道德惩罚欲问卷共三个题目，有些题目（比如，你在多大程度上认为应该要求这个银行审理人（算法）恢复因其不道德行为所造成的损害）不太好理解：算法是程序，要求程序补偿损失感觉很诡异，程序怎么补偿损失，使用程序的是某个单位比如银行，即使发生法律纠纷受害人也会状告使用程序的公司而非状告程序本身。所以，即便这个题目是测对某个对象（银行审理人或程序）的道德惩罚欲，但实际上语义层面容易引发困惑，实际操作层面也不可行，换言之这里的道德惩罚欲在算法这个对象身上失去了作用，惩罚本身变得没有意义了。在惩罚本身对算法没有意义的情况下讨论惩罚的轻重，并认为对算法较轻的惩罚代表了被试对算法本身的道德惩罚欲较少，值得商榷。假如在一个密室逃脱游戏中，某参与者因为呼吸不畅而丧生，被试面对这样的场景要选择是否严厉惩罚“空气”，因为“空气”不足导致了受害人窒息而死，那么这样的问题设置或讨论也很奇怪。 诚然，决策似乎是算法做出的，但背后的参数设置却不是算法设定的，换言之算法本身在实现着某些人的理念，它充当了执行者的角色。即便是作恶，执行者也是帮凶，元凶巨恶是使用算法来牟利的自然人或法人。因此，讨论对算法的道德惩罚欲似乎不如讨论更实际的对涉及算法牟利或算法不当使用的“使用者”的惩罚更合理，也更有现实意义。作者若能在讨论中就这一问题发表高见，就更好了。但这个要求不一定合适，请作者斟酌考虑。

回应：审稿专家提到：“讨论对算法的道德惩罚欲似乎不如讨论更实际的对涉及算法牟利或

算法不当使用的‘使用者’的惩罚更合理，也更有现实意义”。这一问题在一定程度上与我们在文末讨论的道德归责主体相关，即人们可能会更倾向于让算法背后的人来承担责任并接受惩罚。为了进一步明晰这一点，我们增加了如下内容：

“而道德归责之后的惩罚也会随之受到影响，这一方面是由于责任与惩罚的相关性，即人们若倾向于将责任归因于人，那么对于人工智能的道德惩罚也会相应较少；另一方面，从道德惩罚本身的“惩戒”、“规范”等实际功用而言，对人工智能背后的“人”，尤其是涉及非法牟利或使用不当的“使用者”进行惩罚或许更合理，也更有现实意义。虽然如算法等人工智能现在还无法成为完全的道德主体，但是在其犯错时，我们依然还是可能惩罚它，如扫地机器人犯错可能招致人类脚踢的惩罚，而若算法犯错，人也可能惩罚承载其的智能设备，如摔手机等。当然，对算法探讨的核心也是为了人类福祉。因此，对于人类和人工智能之间道德责任乃至道德惩罚的分配都值得进一步研究。”

意见 2：另外，正文第一句话“歧视毋庸赘述，遍及生活、普遍存在，”建议修改为“歧视毋庸赘述，普遍存在，”，不然容易给人同义反复啰嗦累赘的不良印象。

回应：我们已将正文第一句话修改为“歧视毋庸赘述，普遍存在”。

第二轮

审稿人 3 意见：

本研究通过四项实验检验并证实了算法歧视比人类歧视引起更少道德惩罚欲，自由意志在其中起中介作用，以及拟人化倾向调节了歧视主体对道德惩罚欲的影响这三个主要假设。人工智能迅速发展的时代背景下研究这类问题具有很大的理论和应用价值，研究报告总体上写作逻辑清晰有条理，四个实验层层递进，为假设提供了汇聚性证据。但是本研究还存在一些局限和未能解决的问题，需要作者进一步考虑完善。以下是我的一些意见和疑问，供作者参考。

意见 1：建议在前言部分做假设推导过程中将算法歧视比人类歧视引起更少道德惩罚欲，自由意志在其中起中介作用，以及拟人化倾向在其中起调节作用分别标注为假设 1、假设 2 和假设 3，在研究概览部分或具体实验部分描述每个实验验证的具体假设，不要出现“验证主效应”这样的表述（比如第 9 页和第 10 页）。另外，在“实验 3 通过操纵被试的自由意志信念进一步验证了假设的理论机制”这句话中，假设的理论机制是什么意思？请考虑修改表述，以避免误解。

回应：非常感谢您的建议。

（1）我们在修改稿的前言假设推导部分明确提出了三个假设：“据此，本文提出假设 1：相较于人类歧视，人们对算法歧视的道德惩罚欲更小”、“据此，本文的假设 2 为：自由意志信念在歧视主体（人类 vs. 算法）对道德惩罚欲的影响中起中介作用”、“据此，本文提出假设 3：拟人化在歧视主体（人类 vs. 算法）对道德惩罚欲的影响中起调节作用”，并且对每一个假设作了具体阐述。

（2）在研究概览部分或具体实验部分，我们去掉了三处“验证主效应”的表述，以每个实验验证的具体假设代之：“实验 1 验证人们对算法歧视的道德惩罚欲是否小于对人类歧视的道德惩罚欲”、“实验 1 的目的是初步探讨与人类歧视相比，算法歧视是否会引发人们更少的道德惩罚欲”、“为进一步检验实验 1 和实验 2 结果的稳健性，实验 3 将实验情境材料设置为民族歧视情境”。

（3）我们去掉了两处“假设的理论机制”，将其修改为具体的机制表述：“实验 3 通过操纵

被试的总体自由意志信念，进一步检验自由意志信念是否是导致人们产生不同道德惩罚欲的机制”、“实验 3 通过对被试自由意志信念的操纵，进一步验证了自由意志信念是造成道德惩罚欲差异的机制”。

意见 2：Hofmann 等人(2018)等人采用经验取样的方法让被试回顾自己经历的道德或者不道德行为，并测量了对不道德行为的惩罚欲，因此测量题目中包含“不道德行为”这类字眼是可以理解的。但是本研究在测量对人类歧视和算法歧视的道德惩罚欲时使用了“道德惩罚”和“不道德行为”这类字眼似乎非常不妥当，研究者将“不道德行为”强加于人类/算法歧视行为，势必会引导被试并影响被试的作答。为何在研究过程中不采用相对中性的表达方式（比如“你认为…应该为这种行为受到多大程度的惩罚？你在多大程度上应该要求…恢复其行为所造成的损害”）？

回应：非常感谢您的意见，我们十分赞同。

（1）心理学研究讲求实验中的心理真实，因此说法以被试实际理解为主。为了改进并验证这一说法是否会有影响，我们在新加入的两个实验（实验 4 和实验 6）中对道德惩罚欲的测量条目进行了相应修改，采用了相对中性的表达方式，并在文中加入了相应说明。实验 4：“需要提到的是，为了改进前面 3 个实验的道德惩罚欲测量条目中‘道德惩罚’、‘不道德行为’等表述可能对被试作答的影响，本实验对测量条目进行了相应修改，将第一道题和第三道题分别修改为：‘你认为该人力资源经理（算法）应该为这种行为受到多大程度的惩罚？’、‘你在多大程度上认为应该要求该人力资源经理（算法）恢复因其行为所造成的损害？’”，实验 6：“在阅读完以上材料并进行操纵和注意检查后，三组被试分别报告了对人类歧视或算法歧视的道德惩罚欲，测量条目同实验 4（Hofmann et al., 2018），评价对象为赵广/奇智/R2000，$\alpha = 0.88$。” 希望能够弥补研究的不足。

（2）除了在新增实验中的修正，我们也在总讨论的“8.3 局限与展望”部分加入了如下讨论：“首先，在实验设计的细节方面存在一定不足。第一，实验 1、2、3、5 的道德惩罚欲测量直接采用了 Hofmann 等人（2018）的条目，其中“道德惩罚”和“不道德行为”的表述置于算法歧视行为可能会对被试的回答造成一定的影响。当然我们在实验 4 和实验 6 中变换为了更加中性的说法，并未影响主要结果。”

意见 3：考虑到不同被试对算法的看法和知识的差异可能影响他们对算法歧视的道德惩罚欲，作者测量了被试对算法的熟悉程度、了解程度和喜爱程度，并在分析结果时将这三个变量作为协变量控制。这一做法似乎并不妥当，也无法排除这三个变量对结果可能的影响。对算法的熟悉、了解和喜爱程度可能会影响被试对算法歧视的道德惩罚欲，但与对人类歧视的道德惩罚欲并无直接关系。如果想要检验它们是否会影响对算法歧视的道德惩罚欲，应该单独分析这些变量与对算法歧视的道德惩罚欲之间的关系。我认为在算法歧视组的实验材料中先向被试解释“算法”的含义并举例说明，以确保被试在完成实验前理解算法的含义可能是个更为规范的做法。希望作者考虑在讨论部分说明这一局限性。

回应：非常感谢您的意见和建议。

（1）协方差检验原本也是为了排除干扰之后检验自变量对因变量的影响。但是我们认为您提出的方法确实也能够解决这一问题。因此根据建议，为了检验被试对算法的熟悉程度、了解程度和喜爱程度是否会影响对算法歧视的道德惩罚欲，我们在删除协方差分析后加入了对这些变量与对算法歧视的道德惩罚欲之间关系的单独分析，补充内容如下：“为了排除被试对算法的熟悉程度、了解程度和喜爱程度可能对结果的影响，我们将算法组的道德惩罚欲评分与这些变量进行相关分析，结果表明相关均不显著，$r_{熟悉} = -0.13$, $r_{了解} = -0.10$, $r_{喜爱} = -0.15$, $ps > 0.05$。”

（2）我们非常赞同您关于在算法歧视组实验材料中加入对算法含义的解释及举例说明的建议，为了解决这一问题，我们新加了实验，在实验 4 和实验 6 中，我们在算法组的实验材料中加入了相应内容，并在文中作了如下说明：实验 4“为了更好地排除被试对算法的理解程度可能对实验结果造成的影响，两个算法组在民族歧视情境描述之前均加入了对“算法”含义的解释及举例说明（改编自维基百科和 Merriam-Webster，详见附录），以确保被试在完成实验前理解算法的含义”；实验 6“与实验 4 相同，为了更好地排除被试对算法的理解程度可能对实验结果造成的影响，两个算法组在歧视情境描述之前均加入了对“算法”含义的解释及举例说明，以确保被试在完成实验前理解算法的含义。”

（3）除了在新增实验中加入算法含义解释和举例说明，我们也在总讨论的“8.3 局限与展望”部分加入了相应内容：“第二，被试对算法的熟悉和了解程度可能对研究结果造成影响，但在一些实验（1、2、3、5）中我们并未解释算法的含义并举例说明，这在实验 4 和实验 6 中进行了弥补，但未来的研究仍需重点关注这一问题。”。

意见 4：实验 1 的信用卡申请评估情境中人类和算法的表述为“银行审理人（算法）”，而实验 2-4 的招聘情境中人类的表述改为了“人力资源经理李原/张沛/赵广”，而算法的表述依然是“算法”。“人力资源经理李原/张沛/赵广”属于具体明确的对象，而算法属于宽泛的概念，无具体明确的对象，这两种表述在具体/抽象维度的不同也会影响结果。例如，有研究发现，相比于一个身份不明的违规者，人们更倾向于惩罚一个确定的违规者(e.g., Small & Loewenstein (2005)。本研究似乎并不能排除这一可能的解释。如果将“人力资源经理李原/张沛/赵广”换成“人力资源经理”，结果会怎样？或者将算法表达得更为具体，结果又会怎样？实际上，Bigman (2020)的研究中也并没有简单地使用“算法”这一抽象的表述，而是用了“Dr. Smith (CompNet, an Artificial-Intelligence-based computer program)”这一更为具体明确的表述。

回应：非常感谢您的意见，我们很赞同您的想法。

（1）实际上在原有的实验 1 中，我们对两者的表述均为抽象维度，即“银行审理人（算法）”，另外 3 个实验的确存在您所说的表述的具体/抽象维度不同的问题。为了弥补这一局限性，我们在新加入的实验中重点解决了这一问题，在修改稿的实验 4 中，表述为“人力资源经理（算法）”，均为较抽象的表述，并且加入了如下说明：“实验 4 中对人类的表述有所不同，由于人名（如“张沛”）属于具体明确的对象，而“算法”则属于宽泛的概念，无具体明确的对象，因此为了统一人类和算法情境表述的具体/抽象程度，实验 4 中人类组的歧视主体仅表述为“人力资源经理”；在修改稿的实验 6 中，表述为“人力资源经理赵广（算法‘奇智’/算法‘R2000’）”，均为较具体的表述。希望能够尽量排除表述差异可能造成的影响。

（2）除了在新增的实验中统一表述的具体抽象程度之外，我们也在总讨论的“8.3 局限与展望”部分加入了如下内容：“第三，实验 2~4 的歧视情境描述存在歧视主体表述维度差异的问题，即“人力资源经理李原/张沛/赵广” 属于具体明确的对象，而算法属于宽泛的概念，无具体明确的对象。有研究发现，相比于一个身份不明的违规者，人们更倾向于惩罚一个确定的违规者（e.g., Small & Loewenstein, 2005），所以实验中歧视主体表述维度的差异可能对研究结果造成一定的影响。为此我们在实验 1 和实验 4 的情境材料中均采用较为抽象的表述，在实验 6 的情境材料中则均采用较为具体的表述，均得到了类似结果。

意见 5：实验 2 的结果部分：据我所知，SPSS 插件 PROCESS 的输出结果中呈现的是非标准化系数 b，为何图 1 的模型图中报告的都是标准化回归系数β ？是否存在结果报告错误？在进行中介效应检验时歧视主体的两个条件是如何编码的？是 0 =人类，1= 算法吗？这需要在文中说明。

回应：非常感谢您的提问。

（1）中介模型图中的标准化回归系数并非结果报告错误，是我们使用传统逐步回归方法再次进行中介效应检验的结果报告，由于表述不清对您造成的不便我们深感抱歉，并已在修改稿中加入了如下说明：“为了进一步验证中介效应的稳健性，我们又使用传统逐步回归方法进行了中介效应分析（温忠麟 等, 2004），结果见图 1”。

（2）歧视主体的两个条件编码为 0 =人类，1 = 算法，再次抱歉未说明，我们在修改稿中已经加入。

意见 6：实验 3 试图通过操纵高低自由意志信念来进一步验证自由意志这一心理机制。但是，采用 2（人类歧视，算法歧视）× 2（高自由意志信念，低自由意志信念）被试间设计得出的结果无法证明是低自由意志信念降低了对人类歧视的道德惩罚欲还是高自由意志信念增加了对人类歧视的道德惩罚欲？主要原因在于没有控制组。理想的操纵方式应该是将被试随机分配到三个条件：人类歧视，算法歧视，人类歧视+人类不存在自由意志的操纵（或算法歧视+算法存在自由意志的操纵），如果人类歧视组与算法歧视组存在道德惩罚欲的差异，而算法歧视组和“人类歧视+人类不存在自由意志的操纵”组（或者人类歧视组和“算法歧视+算法存在自由意志的操纵”组）无显著差异，则证明了自由意志确实是解释歧视主体影响道德惩罚欲的心理机制。

回应：谢谢您的意见和建议。为了解决这一问题，我们根据您的建议又重新增加了相关实验，将被试随机分配至三组：人类组、相信算法无自由意志组、相信算法有自由意志组。具体研究过程及结果详见修改稿中的实验 4。

意见 7：实验 4：（1）拟人化倾向高的个体是否更倾向于对算法进行自由意志的归因？这两个变量之间的相关程度如何？（2）在考查歧视行为主体与拟人化倾向的交互作用时，歧视行为主体是如何编码的？关于歧视行为主体的回归分析结果请具体描述，而不要用“主效应显著”；（3）图 3 的结果从另一方面来看，为何拟人化倾向更高的个体对人类歧视组的道德惩罚欲更低呢？这一差异是显著的吗？

回应：谢谢您的提问。

（1）我们在实验 5（原实验 4）的结果报告中增加了如下内容：“并且在算法组，被试的拟人化倾向与对算法的自由意志信念显著正相关，$r = 0.20$, $p = 0.044$”，即拟人化倾向高的个体是否更倾向于对算法进行自由意志的归因。

（2）在考查歧视行为主体与拟人化倾向的交互作用时，歧视行为主体的编码为：“人类组 = –1, 算法组 = 1”，已在修改稿中加入说明。关于歧视行为主体的回归分析结果，我们删除了“主效应显著”的表述，修改为：“人类组的道德惩罚欲显著高于算法组”、“拟人化倾向的高低对道德惩罚欲无显著影响”。

（3）为了回答您的问题，我们计算了拟人化倾向得分的均值和标准差，并将拟人化倾向得分大于 M+1SD 的被试作为高拟人化倾向组，将拟人化倾向得分小于 M–1SD 的被试作为低拟人化倾向组。以歧视行为主体（人类组 = –1, 算法组 = 1）和拟人化倾向（低拟人化倾向组 = 1, 高拟人化倾向组 = 2）作为自变量，以道德惩罚欲作为因变量进行方差分析。数据结果表明，人类组的道德惩罚欲（$M = 5.33$, $SD = 1.05$）显著高于算法组（$M = 4.60$, $SD = 1.53$），$F(1, 63) = 7.33$，$p = 0.009$，$\eta_p^2 = 0.10$），高低拟人化倾向组的道德惩罚欲差异不显著，$F(1, 63) = 0.64$，$p = 0.426$，$\eta_p^2 = 0.01$），歧视行为主体和拟人化倾向的交互作用显著，$F(1, 63) = 5.03$，$p = 0.029$，$\eta_p^2 = 0.07$。

简单效应分析发现，**在人类组，高低拟人化倾向组的道德惩罚欲评分无显著差异，$F(1, 63) = 1.45$，$p = 0.233$，$\eta_p^2 = 0.023$**；在算法组，高拟人化倾向组的道德惩罚欲评分（$M = 4.91$, $SD =$

1.49）高于低拟人化倾向组（M = 3.92, SD = 1.48），差异呈边缘显著，$F(1, 63) = 3.60$，$p = 0.062$，$\eta_p^2 = 0.05$。

由于以上结果与实验 5（原实验 4）目的关系不甚密切，为了控制文章字数，我们未将此部分加入正文中。如若需要，请您再批评指正，谢谢！

意见 8：在报告人类歧视组和算法歧视组的道德惩罚欲的差异时，建议报告 Cohen’s d 这一效应量以及该效应量的 95%CI，而不是报告算法歧视组和人类歧视组的道德惩罚欲的平均数的 95%CI。另外，为何不用独立样本 t 检验（本质上和单因素两个条件的 F 检验没有差异）并报告 Cohen’s d 这一效应量呢？

回应：谢谢您的建议！两水平的 one-way ANOVA 与 *t*-test 在结果上是等效的，而 Cohen’s *d* 和 partial eta square 也都能够作为同样标定效应量的指标。但我们接受您的建议，我们已经在修改稿中改为独立样本 *t* 检验并报告 Cohen’s *d* 效应量。

实验 1：“独立样本 *t* 检验结果显示，人类组的道德惩罚欲评分（M = 5.29, SD = 0.99）高于算法组（M = 4.97, SD = 1.34），差异呈边缘显著，$t(170) = 1.82$，$p = 0.073$，Cohen’s $d = 0.27$。”

实验 2：“独立样本 *t* 检验结果显示，人类组的道德惩罚欲评分（M = 5.11, SD = 1.14）显著高于算法组（M = 4.60, SD = 1.54），$t(170) = 2.44$，$p = 0.016$，Cohen’s $d = 0.38$。”

实验 3：“独立样本 *t* 检验结果显示，低自由意志信念组的自由意志信念（M = 5.85, SD = 1.90）显著低于高自由意志信念组（M = 6.54, SD = 1.49），$t(203) = -2.88$，$p = 0.004$，Cohen’s $d = -0.40$。说明自由意志信念操纵有效。”

实验 4：“独立样本 *t* 检验结果显示，相信算法有自由意志组对算法的自由意志信念（M = 3.70, SD = 1.73）显著高于相信算法无自由意志组（M = 2.65, SD = 1.34），$t(137) = 4.00$，$p < 0.001$，Cohen’s $d = 0.68$。说明我们对被试关于算法的自由意志信念操纵有效。”

实验 5：“独立样本 *t* 检验结果显示，人类组的道德惩罚欲评分（M = 5.29, SD = 0.97）显著高于算法组（M = 4.61, SD = 1.32），$t(197) = 4.17$，$p < 0.001$，Cohen’s $d = 0.59$。”

实验 6：“独立样本 *t* 检验结果显示，拟人化算法组的拟人化评分（M = 5.43, SD = 0.88）显著高于非拟人化算法组（M = 4.83, SD = 1.25），$t(136) = 3.31$，$p = 0.001$，Cohen’s $d = 0.56$。说明我们对算法拟人化的操纵有效。”

意见 9：文中有一些表述不当的地方需要修改。举例来说：（1）文中第 6 页“将用户的性别设置女性”应改为“将用户的性别设置为女性”；（2）文中第 7 页“desire to moral punishment”应改为“desire for moral punishment”，“歧视发出者”改为“歧视者”是否更好？（3）“在机器做出已经道德决策的情况下”这里表述不恰当。

回应：非常感谢您对本文的细致审查，我们做出了如下修改：

（1）“将用户的性别设置女性”已改为“将用户的性别设置为女性”；

（2）“desire to moral punishment”已改为“desire for moral punishment”；

（3）歧视发出者”已改为“歧视者”；

（4）“在机器做出已经道德决策的情况下”已改为“在机器已经做出道德决策的情况下”。

……

审稿人 4 意见：

本文探究了人面对算法歧视和人类歧视时，对歧视来源进行道德惩罚的欲望是否有差异，发现人们对算法的道德惩罚欲更低，并检验了其机制，即人相信算法相比人类而言缺乏

自由意志。研究选题新颖，具有理论和实践意义，几个研究之间有良好的逻辑关系，实验设计比较严谨，结论可靠，行文表达流畅。但还有一些问题需要作者考虑。

意见 1：为何不具有自由意志的算法不会受惩罚？

作者认为具有自我意志的个体，在作出不道德行为时更可能被人惩罚。但这背后可能的原因不止一种：（1）动机的视角：具有自我意志的个体作出不道德行为可能说明其具有不道德的动机（该个体是“坏”的），故而应受惩罚；（2）责任的视角：具有自我意志的个体能够自主选择，也应当自主承担责任，故而其作出不道德行为后应受惩罚；（3）惩罚效果的视角：具有自我意志的个体能够理解惩罚并反思，对其不道德行为进行惩罚更可能促使其产生积极变化。

而这些不同的机制似乎都可以解释本研究发现的对算法的宽容：算法没有人一样的不良动机（Bigman et al., 2020）、算法本身应承担的责任较小（多数责任可归于编制算法者）、算法受惩罚并不能促使其进步。这些不同机制涉及到的变量可能也不同。本文没有对此进行充分的阐述和区分，难以明确解释算法到底因哪些原因而更受宽容。

建议作者从理论上进行阐述。

回应：非常感谢您的建议，我们赞同您所说的可能存在其它解释机制。为了对这些可能的解释机制做出回应，进一步凸显本文研究的主题，我们在前言部分增加了如下内容：“当然，需要说明的是，本文所提出的自由意志信念并非人们对不同歧视主体（人类 vs. 算法）产生不同道德惩罚欲的唯一解释机制。例如，算法没有人一样的不良动机（Bigman et al., 2020）、算法本身应承担的责任较小（多数责任可归于编制算法者）、算法受惩罚并不能促使其进步等，这些都能够在一定程度上作为解释机制。但本文之所以重点关注自由意志，主要是由于上述可能的解释机制均与自由意志密切相关。第一，具有自由意志的个体作出不道德行为可能说明其具有不道德的动机（该个体是“坏”的），即判断个体具有自由意志可能是我们推测其动机的必要条件（e.g., Laming, 2004）。第二，具有自由意志的个体能够自主选择，也应当自主承担责任，即自由意志是个体承担责任的必要条件（e.g., Sinnott-Armstrong, 2014）。第三，具有自由意志的个体可能能够理解惩罚并反思，对其不道德行为进行惩罚更可能促使其产生积极变化，因此或许个体在一定程度上具有自由意志亦是惩罚能够对个体产生积极促进作用的必要条件。综上，我们认为自由意志与动机、责任和惩罚效果等因素密切相关，并且自由意志信念的解释机制在某种意义上可能更为基础，能够在一定程度上涵盖以上所述的其它解释机制。因此，本文对不同歧视主体（人类 vs. 算法）如何影响道德惩罚欲的机制探讨将着重检验自由意志信念的作用”。

参考文献：

Laming, D. (2004). *Understanding human motivation: What makes people tick?* Malden, MA: Blackwell.

Sinnott-Armstrong, W. (2014). *Moral psychology: Free will and moral responsibility.* MIT Press.

意见 2：存在与自由意志无关的竞争假设

（1）人的行为更容易被解释，而算法复杂和不透明。被试因为算法的内在逻辑难以甄别，也就更难判定歧视行为是不道德的，甚至可能认为其具有合理性，进而对其表现出宽容。

（2）人无法真正惩罚算法，即惩罚算法是不切实际的，但人可以惩罚人类。换言之，所谓惩罚算法是惩罚算法的载体，正如作者所说“扫地机器人犯错可能招致人类脚踢的惩罚，而若算法犯错，人也可能惩罚承载其的智能设备，如摔手机等”。但是，踢扫地机器人和摔手机并非惩罚算法本身。鉴于很难真正惩罚算法，因此当算法出现歧视后，人不怎么愿意惩罚它。

当然，两个调节研究（研究 3 和 4）可排除上述竞争假设。建议作者在文中合适处提及可能存在的竞争假设，并在两个调节研究之后讨论它们如何排除这些可能性。

回应：非常感谢您的建议。

我们已经在前言中加入了可能存在的竞争假设：“此外，可能还存在与自由意志无关的竞争假设。首先，人的行为更容易被解释，而算法复杂和不透明。被试因为算法的内在逻辑难以甄别，也就更难判定歧视行为是不道德的，甚至可能认为其具有合理性，进而对其表现出宽容。其次，人无法真正惩罚算法，即惩罚算法是不切实际的，但人可以惩罚人类。换言之，所谓惩罚算法是惩罚算法的载体，并非惩罚算法本身。鉴于很难真正惩罚算法，因此当算法出现歧视后，人不怎么愿意惩罚它。鉴于以上与自由意志无关的竞争假设存在，我们将会通过 4 个调节研究（研究 3~6）将之进行排除。”

并在讨论部分加入了相应内容讨论调节研究如何排除这些可能性：“此外需要提到的是，虽然可能存在与自由意志无关的竞争假设，如人的行为相对于算法更容易被解释、人无法真正惩罚算法等，但本文通过 2 个自由意志信念的调节研究（实验 3~4）和 2 个拟人化（与自由意志密切相关）的调节研究（实验 5~6）重复验证了本文所提出的自由意志信念机制，在一定程度上排除了上述竞争假设”。

意见 3：研究的创新性

Bigman 等人（2020）发现，人对作出歧视行为的算法（vs. 人类）的道德愤怒程度更低，并测量了被试在多大程度上认为“有歧视行为的算法应该被弃用”“采用该算法的公司应该道歉”。这些测题和本文中的道德惩罚具有一定的相似性。并且，道德愤怒是道德惩罚的前因，根据 Bigman 等人（2020）的发现也可以推测出本文的结果。作者应当详细说明本文和 Bigman 等人（2020）研究的区别、本文的独特创新之处。

回应：非常感谢您的建议。为了突出本文与 Bigman 等人（2020）研究的区别和创新之处，我们在讨论部分中增加了以下内容:“但需要强调的是，虽然同为对不同歧视主体（人 vs. 算法）的比较，但和 Bigman 等人（2020）的研究相比，本研究与之区别的独特创新之处主要在于以下几点：第一，因变量的差异，Bigman 等人（2020）主要探讨了道德情绪即道德义愤的不同，而本文的研究重点在于行为倾向即道德惩罚欲的不同。虽然道德义愤和道德惩罚具有一定程度的相关性，但不意味着两者完全等同，道德惩罚欲依然有其独特的研究价值。并且 Bigman 等人（2020）虽然在研究 5 中加入了与道德惩罚测量具有一定相似性的条目（如“有歧视行为的算法应该被弃用”“采用该算法的公司应该道歉”），但一方面这些条目并非完全针对算法，另一方面其研究结果也并未发现两组之间的显著差异。第二，机制的差异，Bigman 等人（2020）的研究着重探讨了动机机制，而本文则重复验证了自由意志信念的机制，并且从一定意义上来说自由意志机制比动机机制更为基础，因为个体具有自由意志是判断其动机的必要条件（Laming, 2004）；第三，调节的差异，Bigman 等人（2020）的研究并未进行调节变量的探索，他们在文章的局限和未来方向部分强调了拟人化倾向可能会扮演的调节作用，而本文则通过两个研究，从个体的拟人化倾向和算法本身的拟人化两个方面重复验证了拟人化的调节作用”。

意见 4：研究 3 通过操纵被试对人类自由意志的信念，检验自由意志信念是否是导致人们对歧视主体（人类 vs.算法）有不同惩罚欲的机制。作者是用调节的手段检验机制，而非用调节的手段检验“中介效应”。建议修改说法。

回应：非常感谢您的建议。我们已经对文中相关内容进行修改：实验 3 引入部分“为了增加实验结果的稳健性，实验 3 再次丰富了歧视类型，关注民族歧视问题，并且通过对被试总体自由意志信念的操纵，进一步探讨自由意志信念是否是造成道德惩罚欲差异的机制”；实验

3 讨论部分“实验 3 通过对被试自由意志信念的操纵，进一步验证了自由意志信念是造成道德惩罚欲差异的机制”。

意见 5：研究 3 中，作者以阅读文章的方式操纵被试的自由意志信念。该研究存在实验设计上的问题：自由意志信念的操纵短文均是关于“人类是否有自由意志”，而与算法无关，该操纵仅影响判断人类的被试。因此，研究 3 的结果仅能说明，若被试不相信人有自由意志，则对人进行道德惩罚的欲望降低；不能说明对人和对算法的道德惩罚欲望之差异是由“人们相信人类比算法有更多自由意志”所致。也许解决第 5 个问题的一条思路是，提升判断者对算法的自由意志信念，即采用单因素设计（人类/相信算法无自由意志/相信算法有自由意志），考察人类组与相信算法有自由意志之间的差异是否较人类组与相信算法无自由意志的差异有所减小。请作者考虑该设计是否合理，能否解决现有问题。另外，该设计具有应用价值：它提供了消除人与算法差异的一种实践做法。

回应：非常感谢您的建议。由于您在第 6 条意见中提供了解决第 5 个问题的思路，因此我们将这两条合并回答。为了解决实验 3 的问题，我们根据您的建议增加了相关实验，将被试随机分配至三组：人类组、相信算法无自由意志组、相信算法有自由意志组。具体研究过程及结果详见修改稿中的实验 4。

意见 6：作者在介绍研究 4 的目的和逻辑时，多次写道：“如果将算法拟人化，是否会增加人们对算法自由意志的归因，从而调节歧视行为主体对道德惩罚欲的影响呢？”读至此处，我以为作者会采用单因素设计（算法-非拟人化/算法-拟人化/人类）或 2（歧视来源：算法/人类）× 2（拟人化：是/否）的设计，从而拟人化算法。但其实，作者用“拟人化个体差异量表”区分被试在多大程度上倾向于把物品拟人化。如此做逻辑上没问题，但不能准确反映前文所述的“如果将算法拟人化……”建议作者修改表述。当然，如果能补充拟人化算法的实验最好，因为该实验也提供了消除人与算法差异的可行做法。

回应：非常感谢您的建议。

（1）我们修改了涉及实验 5（原实验 4）的相应表述：实验 4 讨论部分“那么，既然算法歧视比人类歧视引发较少道德惩罚欲的原因是人们认为算法比人类拥有较少的自由意志，那么拟人化倾向的个体差异是否会对这种效应起到调节作用呢？为了回答这一问题，在实验 5 中，我们将继续探索歧视行为主体影响道德惩罚欲的边界条件，考察拟人化倾向可能存在的调节效应”；实验 5 引言部分：“理论上人类相较于算法有更高的自由意志，那么拟人化倾向较高的人是否会更倾向于增加对于算法自由意志的归因，从而对歧视行为主体对道德惩罚欲的关系造成影响呢？实验 5 旨在回答这一问题”。

（2）由于人类无法再区分为是否拟人化，因此我们没有采用 2（歧视来源：算法/人类）× 2（拟人化：是/否）的实验设计，而是增加了一个单因素设计的实验，将被试随机分配至三组：人类组、拟人化算法组、非拟人化算法组。具体实验过程和结果详见实验 6。

意见 7：研究 4 中作者采用的拟人化问卷题目基本都测量的是“某种客体是否有自我意志”。然而拟人化包括了外观、动机、行为等方面，作者采用的问卷似乎更像是“自我意志问卷”，未必能完整反映被试将客体拟人化的倾向。

回应：非常感谢您的提问。您说的很对，将自由意志赋予非人对象的确是拟人化的一个重要方式，也是本文所采用的拟人化个体倾向差异量表的其中一个部分，但是除了这个部分之外，该量表还包括以下方面：有心灵（mind），如“普通的电脑在多大程度上有它自己的心灵？”、有意识（consciousness），如“普通的机器人在多大程度上有意识？”、有意图（intentions），如“奶牛在多大程度上是有意图的？”、能体验情绪（experience emotions），如“一台电视

机在多大程度上可以体验情绪？”。这些都在 Waytz 等人（2010）的文章中有详细说明。该量表发表在国际知名心理学期刊《Perspectives on Psychological Science》，并且至今已有 723 次引用（Google Scholar），是较为成熟和广受认可的拟人化个体倾向量表。

参考文献：

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232.

意见 8：研究 4 报告简单斜率分析结果时，高/低拟人化倾向分组是以 M ± SD 区分的吗？应当明确报告。
回应：非常感谢您的提醒。高/低拟人化倾向分组是以 M ± 1SD 区分的，我们已经在简单斜率分析的图中加入了相应标注。

意见 9：建议将“自由意志”统一表述为“自由意志信念”，因为作者关心的是“人们是否相信自由意志的存在”，而非“自由意志”本身。
回应：非常感谢您的建议，我们已在修改稿中将“自由意志”统一表述为“自由意志信念”。

意见 10：除 p 值外，其他统计数值请保留到小数点后两位。
回应：谢谢您的提醒，我们已在修改稿中将除 p 值以外的其它统计数值保留到小数点后两位。

意见 10：因变量的量程为 1-7，图表纵坐标不可包含 0，因为量表中无 0-1 这一段。
回应：谢谢您的提醒，我们已在修改稿中对图 2 进行了相应修改，并在新增的图中也避免了这一问题。

意见 12：讨论部分 6.1“这以结果”应为“这一结果”。
回应：非常感谢您的细致审查，我们已在修改稿中将“这以结果”改为“这一结果”。

第三轮

审稿人 3 意见：

作者新增加的两个实验以及对文章的细致修改基本解决了本人在上一轮审稿时所提出的问题和研究局限，论文质量得到了很大提升。不过这轮修改稿中还存在几处表达错误和小问题，建议作者考虑修改：

意见 1：“调节研究”这个词在文中多处出现，但是这个表达不是很清楚，建议作者修改表述；
回应：谢谢您的建议，“调节研究”一词为另一审稿人建议使用，为使表达更清楚，我们将文中的“调节研究”修改为了“调节效应研究”。

意见 2：实验 4 并没有额外操纵被试对算法无自由意志的信念，称为“算法组”比“相信算法无自由意志组”似乎更恰当；
回应：谢谢您的建议，已修改相应文字和图表。

意见 3：实验 6 为单因素三水平被试间设计，但在文中作者写道 “随机分配至人类组和算

法组”（7.1.1 部分）、“两组分别为人类组、拟人化算法组以及非拟人化算法组”（7.1.2 部分），这里似乎存在写作失误，建议作者修改表述；
回应：谢谢您的细致审阅，写作失误已修改。

意见 4：“5.2.2 歧视行为主体与关于算法的自由意志信念的交互作用”和“7.2.1 算法拟人化的调节作用”这两部分的结果并没有考查两个变量之间的交互效应，标题容易造成误解，建议作者修改。另外，既然实验 4 和实验 6 有明确的目的和假设，在比较不同组别的道德惩罚欲的差异时为何采用事后多重比较（post-hoc comparison）而不是计划对比（planned contrast）？
回应：谢谢您的建议，已将两个标题分别修改为“7.2.2 算法拟人化的影响”和“5.2.2 关于算法的自由意志信念的影响”。并已将实验 4 和实验 6 中的结果报告从事后多重比较改为计划对比。

意见 5：建议作者在文末增加“研究结论”部分。
回应：谢谢您的建议，已在文末增加“研究结论”部分。

审稿人 4 意见：

作者根据审稿意见对文章进行了修改，对多数审稿意见作出了有力答复，有效提升了文章的质量。不过还有一些问题需要作者考虑：

意见 1：增加实验 4 后，实验 3 的理论意义显得不大。第一，两个实验的目的类似，均为通过操纵自由意志信念检验效应的机制。第二，如之前的审稿意见所言，实验 3 设计存在不足，无法充分检验机制，即“人们对算法的道德惩罚欲高于对人类的道德惩罚欲是由于人们认为算法相比人类缺乏自由意志”。第三，从结果来看，实验 3 操纵“总体的自由意志信念”不影响被试对算法的惩罚欲；但实验 4 中操纵“对算法的自由意志信念”则能够影响被试对算法的惩罚欲。这些不一致的结果使人疑惑“总体的自由意志信念”是怎样的概念，究竟如何影响被试对算法的看法，两种操纵方式间的关键区别何在。如之前的审稿意见所提出的，实验 3 采用的阅读材料主要涉及人类的自由意志，似乎不像是“总体的自由意志信念”。作者可考虑是否需要保留实验 3，如希望保留，则需要对此作出更充分的解释。
回应：非常感谢您的建议，为了保持研究的完整性并更好地说明实验的意义及不足之处，我们将涉及到实验 3 的“总体自由意志信念”表述修改为“自由意志信念”，并在实验 3 的讨论部分增加了如下内容：“但实验 3 存在一些不足之处，首先，实验 3 未能充分检验本研究所提出的机制，即‘人们对算法的道德惩罚欲高于对人类的道德惩罚欲是由于人们认为算法相比人类缺乏自由意志’；其次，实验 3 对自由意志信念的操纵并未影响被试对算法的惩罚欲，因此这种操纵可能并未对人们关于算法的自由意志信念产生影响。因此，为了更直接地检验对算法的自由意志信念是否是造成人们对不同歧视主体（人类 vs. 算法）有不同道德惩罚欲的机制，实验 4 将直接对人们关于算法的自由意志信念进行操纵。”

意见 2：实验 5 对中介和调节变量进行了测量，可以考虑采用 Model 7 或 14 做前半段或后半段调节的中介模型。
回应：谢谢您的建议，但由于我们并未提出相应的假设，即在实验 5 中对中介效应进行前半段或后半段调节的假设。我们现在的假设是独立的调节效应与中介效应，鉴于我们的假设实际上已然满足研究逻辑，再加入前半段或者后半段调节对研究假设的推进意义不大，且现在

再加上类似模型有用后来数据结果去重新框架研究假设之感（当然我们按照审稿人要求做了类似模型检验，结果并非是完全和想象一致的，在这样的情况下再做就真的需要按照结果来编故事，所以我们放弃了），因此我们在结果报告中并未采用这两个模型进行分析。但是非常感谢您的建议！

意见 3：摘要应该将效应的中介机制和调节的结果描述得更为具体。例如“潜在机制是人们认为算法（与人类相比）更缺乏自由意志（实验 2-4），且人们的拟人化倾向越强，对算法的道德惩罚欲越强（实验 5-6）”。
回应：非常感谢您的建议，已根据您的建议对摘要如下修改：“结果发现：相对于人类歧视，人们对算法歧视的道德惩罚欲更少（实验 1~6），潜在机制是人们认为算法（与人类相比）更缺乏自由意志（实验 2~4），且个体拟人化倾向越强或者算法越拟人化，人们对算法的道德惩罚欲越强（实验 5~6）。”

意见 4：其他细节问题： （1）p. 23 “究竟是什么决定了我们会否想要对行为者进行道德惩罚……”“会否”可改为“是否”。
回应：谢谢您的建议，已将“会否”改为“是否”。

意见 5：p. 37 “实验 5 旨在回答这一问题。并且在实验 5 中进一步丰富了歧视的类型，关注年龄歧视问题。”该句表达不通顺。
回应：谢谢您的建议，已将此句修改为：“实验 5 旨在回答这一问题。此外，在实验 5 中我们着眼于年龄歧视问题，从而也进一步丰富了研究的歧视类型。”

意见 6：p. 40 “实验 6 为单因素三水平被试间实验设计，两组分别为人类组、拟人化算法组以及非拟人化算法组”一句中，“两组”应为“三组”。
回应：谢谢您指出，已将“两组”修改为“三组”。

意见 7：参考文献中温忠麟等人(2004)那篇（包括中英文），期号多了个 0。
回应：谢谢您指出，已修改。

第四轮

编委意见：

作者们好，文章经过几轮修改和实验的增加，有了很大提高，基本达到可发表的程度。还有一些小的方面希望作者能继续修改：

意见 1：文章现在有些过长，希望能精简一下文字，尤其是对实验和实验结果的描述；
回应：谢谢编委老师的建议，我们已经对文章进行了精简。

意见 2：图的分辨率不是很高，看上去比较模糊，可以提高；另外，图中的 Error bars 没有注解是 SD 还是 SE，还有，图 4 中只给了均值，没有 Error bars；
回应：谢谢编委老师的建议，我们已经提升了图片的分辨率、在图中标注了 95%CI，并且在图 4 中加入了置信区间。

意见 3：英文摘要过于冗长，远超字数限制，须删减。

回应：谢谢编委老师的建议，我们注意到学报的投稿指南中提到“英文摘要需详细，1 页 4 段，约 500 个单词”，我们将英文摘要的字数从 563 删减到了 446。