《心理科学进展》审稿意见与作者回应

题目：第三方惩罚行为的认知神经机制
作者：郑好 陈荣荣 买晓琴

第一轮

审稿人 1 意见：

意见 1：正文第 1 页：“TPP 行为被认为是动力系统、情感系统、认知系统和执行系统交互作用的结果”：该划分方法中的各个类别存在交叉，因为动力（动机）并不独立于情感系统，而执行（尤其是执行控制）亦不独立于认知系统。其次，该说法基本上适用描述于人类的任何行为，因此其信息价值有限。

回应：感谢审稿专家的审阅。根据审稿专家的意见，我们对模型(图 3)及其原文对应表述做出如下改进。

首先，我们认为情绪和奖赏在第三方惩罚行为中起不同作用。情绪模型指出负性情绪(愤怒、厌恶等)是 TPP 的直接作用机制，负性情绪的产生是 TPP 的动力来源之一(Fehr & G ächter, 2002; Xiao & Houser, 2005)，而互惠模型认为对未来回报的期待(如声誉建立)以及惩罚后的权力体验感、满足感可以作为奖赏信号对个体进行内部强化(Izuma et al., 2008; Strobel et al., 2011)，促使个体做出下一次惩罚行为。我们认为情绪和奖赏都是 TPP 行为的动力来源，因此将两者进行整合。在该模型中，我们将情绪系统和奖赏系统共同作为 TPP 的动机系统。

其次，我们将认知系统划分为社会认知和执行控制两个子系统。Ginther 等人(2016)利用 fMRI 对意图评估、伤害程度评估和惩罚三个环节对应的大脑活动分开研究，发现三者具有不同的神经机制；Bellucci 等人(2017)也指出 TPP 是建立在责任评估和惩罚选择两个基本的认知功能之上的；在惩罚选择阶段，第三方还需要认知控制以进行情绪、公平以及自利之间的权衡(殷西乐 等, 2019)。因此，我们认为社会认知系统和执行控制系统共同组成了 TPP 的认知系统，前者负责责任评估(包括意图和伤害程度的评估)，后者负责认知控制以及执行惩罚。

审稿专家的问题让我们意识到我们在文章中表述不准确，我们在相应的部分做了如下补充和修改。原文修改为：“在该模型中，情绪系统和奖赏系统是 TPP 的动力来源，认知系统主要负责责任评估以及惩罚的选择。”根据审稿人的意见，我们在第 4 小节对模型及其解释进行了修改，修改后的模型见图 3，详见正文第 22-25 页。

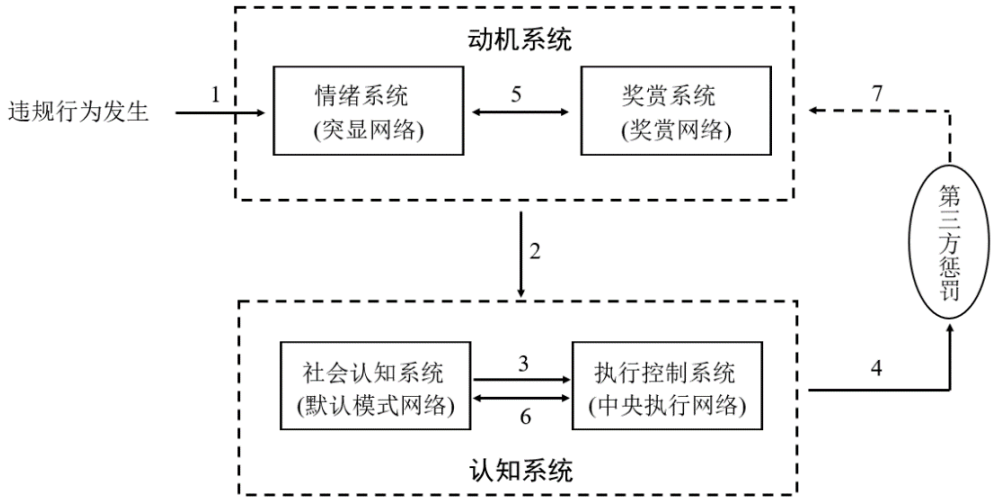

图 3 第三方惩罚行为的认知神经网络模型

注：单向箭头表示作用方向，双向箭头表示相互关系。

参考文献：

殷西乐, 李建标, 陈思宇, 刘晓丽, 郝洁. (2019). 第三方惩罚的神经机制：来自经颅直流电刺激的证据. *心理学报, 51*(5), 571–583.

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Monte, O. D., Knutson, K., Grafman, J., & Krueger, F. (2017). Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence. *Social Neuroscience*, *12*(2), 124–134.

Fehr, E., & G ächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience*, *36*(36), 9420–9434.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284–294.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: Neural and genetic bases of altruistic punishment. *NeuroImage*, *54*(1), 671–680.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7398–7401.

意见 2：第 1 页：“以 Fehr 为代表的一批学者指出有必要对违规行为进行约束，故在前人使用的范式基础上加入第三方因素”：不熟悉经济范式的读者可能无法理解这句话的逻辑。

回应：根据审稿专家的建议，我们在正文中对第三方惩罚范式的提出做出了更加详细的说明。具体内容如下：

“上个世纪以来，人们针对经济决策与分配行为开展了大量研究。行为经济学家最早在最后通牒博弈任务(Ultimatum Game, UG)中发现了利他惩罚(Thaler, 1988)。在该任务中存在提议者和响应者两方，他们需要就一定数量的资金如何分配达成一致。首先由提议者提出分

配方案，若响应者接受提议者提出的分配方案，则两人按照这一方式进行分配，反之，两人都不能获得任何资金。在这种情况下，拒绝可以看作是响应者对提议者的有代价的第二方惩罚。然而，在现实世界中，第二方往往是被动接受的角色。并且，如果仅存在第二方惩罚，能够维护的社会规范数量有限，因此，引入第三方惩罚能够扩大惩罚违规者的比例，更好地维护社会规范。以 Fehr 为代表的学者在实验室条件下证明了第三方惩罚的存在。Fehr 和 Fischbacher(2004a)在独裁者博弈任务(Dictator Game, DG)中引入第三方，DG 与 UG 的区别在于响应者不能拒绝独裁者(提议者)提出的分配方法，只能被动地接受，作为第三方观察者的被试在看到独裁者的分配方案后，可以通过付出一定代价(减少自己的钱数)来惩罚独裁者。随后，该范式成为 TPP 的重要研究方法之一，为在实验室条件下研究社会规范行为提供了一种新的思路。”

参考文献：

Fehr, E., & Fischbacher, U. (2004a). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Thaler, R. H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*, *2*(4), 195–206.

意见 3：“间接互惠无法解释个体在单试次或者匿名实验中的利他惩罚行为”，那么在这类实验设计中，TPP 的实际发生概率有多高，是否仍然能观察到稳定的利他惩罚现象？逻辑上，如果利他惩罚行为在单次或者匿名条件下消失了，那么间接互惠理论也就没有疏漏了。作者应该从前人文献中引用具体数据作为证明。

回应：感谢审稿专家的建议。根据间接互惠模型，个体执行有代价的第三方惩罚是为了获得声誉，从而在未来获取更多合作机会，因此间接互惠模型强调个体第三方惩罚的目的是获取间接利益。在单试次或者匿名情况下，由于个体几乎不存在未来的合作机会，或者无法辨识身份，因此并不存在声誉动机。以往研究揭示即使在这种情况下，个体依旧会执行第三方惩罚，并且发生概率高于 1/2。例如，Piazza 和 Bering(2008)发现在第三方完全匿名的条件下，71.4%的参与者选择牺牲 1/3 的资金去惩罚违规者；Feng 等人(2022)也发现在单次匿名实验中个体的平均惩罚率大于 50%。国内研究者杨莎莎和陈思静(2022)同样发现在单次匿名博弈中不惩罚或低惩罚的情况并不常见(3.15%)。此外，最近的研究发现即使是未涉足社会的仅八个月大的婴儿就已经能够使用目光来惩罚违规者(Kanakogi et al., 2022)。以上研究发现都无法用间接互惠来解释。

根据审稿专家的建议，我们在 2.1 小节补充了前人研究数据(见正文第 15 页)。

参考文献：

杨莎莎, 陈思静. (2022). 第三方惩罚中的规范错觉：基于公正世界信念的解释. *心理学报, 54(*3), 281–299.

Feng, C., Yang, Q., Azem, L., Atanasova, K. M., Gu, R., Luo, W., Hoffman, M., Lis, S., & Krueger, F. (2022). An fMRI investigation of the intention-outcome interactions in second- and third-party punishment. *Brain Imaging and Behavior*, *16*(2), 715–727.

Kanakogi, Y., Miyazaki, M., Takahashi, H., Yamamoto, H., Kobayashi, T., & Hiraki, K. (2022). Third-party punishment by preverbal infants. *Nature Human Behaviour*, *6*(9), 1234–1242.

Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, *6*(3), 487-501.

意见 4：第 3 小节“参与第三方惩罚的脑网络”与前文联系不紧密，不能对第 2 小节介绍的各种模型起到支持或者反对的作用，因此现版本前后部分的思路割裂颇严重。仅在 4.1 小节中，作者有提到网络神经科学与双系统模型之间的联系。

回应：感谢审稿专家的建议。我们对本文的整体逻辑梳理如下：

我们首先对 TPP 行为相关理论模型进行总结，包括互惠模型、情绪模型和强化学习视角下的双系统模型。其中，情绪模型和互惠模型为 TPP 的动机系统组成提供理论支持，双系统模型将情绪和认知因素结合起来，指出 TPP 的产生不仅来源于情绪的驱动，还依赖于社会认知和认知控制，为认知神经网络模型中的认知系统提供理论依据。

然后，我们对前人研究中神经层面的证据进行梳理，重点关注脑网络的证据。我们认为突显网络、默认模式网络、中央执行网络和奖赏网络在 TPP 中起到重要作用。其中，突显网络在整合负性情绪上的重要作用支持了情绪模型；默认模式网络在责任评估功能上体现了 TPP 是情感(伤害程度)和认知(意图推断)整合后的结果；中央执行网络中 dlPFC 的认知控制功能为强化学习视角下的双系统模型中的控制系统的存在提供了证据支持；奖赏网络是强化学习观点在 TPP 中神经层面上的体现。

最后，我们通过对以往研究结果的整合，提出 TPP 的认知神经网络模型：模型的整体构建源于双系统模型的观点，反馈通路的存在是强化学习在 TPP 中的体现，动机系统下的情绪系统和奖赏系统提高了情绪模型和互惠模型的解释力。

综上，情绪模型、互惠模型和双系统理论对模型的构建提供理论上的支持，脑网络的证据进一步为认知神经网络模型提供实证支持。

十分感谢审稿人提出的意见，我们意识到本文表述存在逻辑建构不清楚的问题。根据审稿人建设性的意见，我们在正文中对每个部分都做了相应的补充，详见正文第 19、21、23-24 页。

……

审稿人 2 意见：

本文综述了近十年的相关研究，对第三方惩罚行为的理论模型进行梳理，总结不同功能脑网络在第三方惩罚下的作用，并据此提出了第三方惩罚的认知神经网络模型，文章总体上条理清晰、综述全面，总结和展望部分全面地讲述了第三方惩罚这一问题未来可能的研究方向，但对于本文的文献梳理方法和提出的模型的可靠性仍有一些疑问：

意见 1：文章描述“本文综述了近十年来 TPP 相关的研究”但文中没有对文献选择的方法进行系统描述，特别是多篇文章的元分析，为了增强本文的可靠性建议使用 PRISMA 流程对文献进行筛选。

回应：感谢审稿专家宝贵的建议！我们在前言部分(正文第 14 页)对文献搜索过程进行了详细描述，补充了文献纳入与排除标准及 PRISMA 流程图，并将其放入正文附录图 S1。正文补充内容如下：

“因此，本文对近十年来与 TPP 相关的研究进行梳理。首先进行文献检索。英文文献检索使用 Web of Science、PubMed、ScienceDirect 数据库，TPP 的关键词为“third-party punishment”或“altruistic punishment”或“social punishment”，认知神经机制的关键词为“cognitive”或“neural bias”或“neural correlates”或“neuroimaging”或“fMRI”。中文文献检索使用知网、万方、维普数据库，TPP 的关键词为“第三方惩罚”或“利他惩罚”或“社会惩罚”，认知神经机制的关键词为“认知”或“神经机制”或“神经基础”或“脑成像”。同时，在阅读文后参考文献时利用滚雪球的方法检索文献进行查漏补缺。截止 2023 年 4 月，共检索到文献 1149 篇，文献检索的时间范围为 2013 年 4 月到 2023 年 4 月。经初筛、审查等阶段后，最终纳入文献数量为 60，其中涉及神经机制的文献 39 篇。”

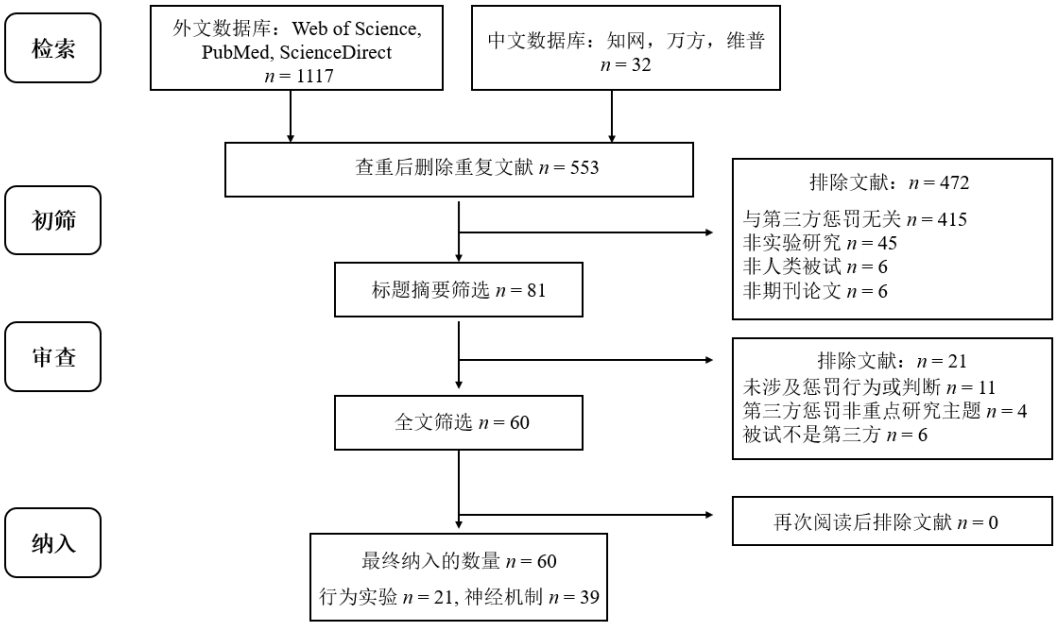

图 S1 PRISMA 流程图

注：*n* 代表文献数量。

意见 2：对于第四部分 “第三方惩罚的认知神经网络模型”，首先，作者没有列表说明这个模型的提出包括了哪些文献，及其详细信息；其次，既然作者想要基于前人文献提出自己的模型，应该对前人实验任务进行区分，比如，前人研究中个体的认知过程、情绪唤醒过程是否一致，以及被试年龄、被试数量；第三，作为总结并提出自己模型的一个部分比较缺少统计支撑，比如不同任务和认知状态所激活的脑区，其实可以用一些元分析来支持。

回应：感谢审稿专家的建议！根据此意见，对 TPP 相关文献经 PRISMA 流程筛选后，我们将与神经机制相关的 39 篇文章进一步分析，其中有 32 篇文献可以支持我们所提出的认知神经网络模型。为了让读者知悉被试信息、实验任务以及结果等，我们将文献详细信息表格纳入附录中(表 S1)，具体包括所使用的技术、被试信息、任务中个体的情绪唤醒和认知过程、相关 ERP 成分以及相关脑区的激活情况。

我们十分认可审稿专家所提到的元分析方法，我们认为该方法非常有意义。元分析在定性分析的基础上引入了定量的分析方法，经该方法得出的结论更具有普遍性和客观性。TPP 的产生分为不同的阶段，每个阶段所涉及的脑区和网络激活模式也各不相同，利用元分析能够处理分析大量的单个研究数据，做到有效的综合。早在 2020 年，Bellucci 等人就对 TPP 中持续激活的脑区和神经网络做了元分析，有助于更好地理解 TPP 行为背后的神经心理机制，但依旧对大脑区域和网络之间的相互作用缺乏整体性理解。基于以往研究不足，我们提

出 TPP 的认知神经网络模型。未来的研究可以更一步利用元分析的方法为我们提出的神经网络模型提供统计支持。十分感谢审稿专家的提议，我们将元分析的运用纳入 5.5 节未来展望部分，详见正文第 27 页。

参考文献：

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience and Biobehavioral Reviews, 113*, 426–439.

意见 3：“突显网络(salient network)、默认模式网络(default mode network)和中央执行网络(central executive network)(Bellucci et al., 2020; Krueger & Hoffman, 2016; Lo Gerfo et al., 2019)，分别参与“情绪产生”、“责任评估”和“惩罚选择”三个阶段。”首先，前文中没有提到这三个阶段，是引用还是作者提出的？其次，这三个网络和这三个阶段是一一对应吗？

回应：“情绪产生”、“责任评估”和“惩罚选择”三个阶段是基于前人文献概括得出(Bellucci et al., 2020; Krueger & Hoffman, 2016; Lo Gerfo et al., 2019)，且与突显网络、默认模式网络和中央执行网络一一对应。根据审稿专家的意见，我们在文章中相应位置修改了表述，具体内容如下：

“以往研究表明 TPP 包含情绪产生、意图和伤害程度评估以及选择惩罚阶段。结合前人研究中相关脑网络的功能与激活模式(Bellucci et al., 2020; Krueger & Hoffman, 2016; Lo Gerfo et al., 2019)，本文认为 TPP 行为的产生分为‘情绪产生’、‘责任评估’和‘惩罚选择’三个阶段，与之相对应的脑网络为突显网络(salient network)、默认模式网络(default mode network)和中央执行网络(central executive network)。此外，奖赏网络协作 TPP 加工过程，主要起价值表征、预期奖赏的作用。相关脑网络所包含的脑区及其位置见图 1。”

参考文献：

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience and Biobehavioral Reviews, 113*, 426–439.

Lo Gerfo, E., Gallucci, A., Morese, R., Vergallito, A., Ottone, S., Ponzano, F., Locatelli, G., Bosco, F., & Romero Lauro, L. J. (2019). The role of ventromedial prefrontal cortex and temporo-parietal junction in third-party punishment behavior. *NeuroImage*, *200*, 501–510.

Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences*, *39*(8), 499–501.

意见 4：“公平是一种默认的社会规范(Civai, 2013)。当违规行为发生，突显网络负责检测认知冲突以及感知个体愤怒、不公平厌恶等负性情绪”，突显网络监测的情绪只提到社会规范一个原因不够严谨。

回应：感谢审稿人的建议！根据审稿专家的意见，我们对以往研究进行总结，发现突显网络所检测的负性情绪主要有两个来源：观察到违规行为产生的愤怒、不公平厌恶(Pedersen et al., 2018)和预期自己未能惩罚违规者带来的内疚感(Nelissen & Zeelenberg, 2009)，这两种道德情绪分别以他人和自我为中心，其产生过程都与突显网络相关(Feng et al., 2016; Mclatchie et al., 2016)。根据审稿专家的建议，我们在 3.1 小节(正文第 18 页)补充了相关内容，具体内容如下：

“公平是一种默认的社会规范(Civai, 2013)，当违规行为发生时，个体会产生愤怒、不公平厌恶等负性情绪，这种愤怒和厌恶属于以他人为中心的道德情绪(Pedersen et al., 2018)。此外，当个体预期自己应当惩罚违规者以维护正义却没有这样做时，会产生以自我为中心的内疚感(Nelissen & Zeelenberg, 2009)，这种内疚感在一定程度上促进愤怒情绪的产生(Rothschild & Keefer, 2018)。因此，突显网络负责检测冲突并产生愤怒、厌恶、内疚等负性情绪(Bellucci et al., 2020; Buckholtz & Marois, 2012; Feng et al., 2016; Mclatchie et al., 2016)，主要脑区包括背侧前扣带皮层(dorsal anterior cingulate cortex, dACC)、前脑岛皮层(anterior insula cortex, AIC)和杏仁核。”

参考文献：

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience & Biobehavioral Reviews*, *113*, 426–439.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661.

Civai, C. (2013). Rejecting unfairness: Emotion-driven reaction or cognitive heuristic? *Frontiers in Human Neuroscience*, *7*, Article 126.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*(2), 663–677.

Mclatchie, N., Giner-Sorolla, R., & Derbyshire, S. W. G. (2016). ‘Imagined guilt’ vs ‘recollected guilt’: Implications for fMRI. *Social Cognitive and Affective Neuroscience*, *11*(5), 703–711.

Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*(7), 543–553.

Pedersen, E. J., McAuliffe, W. H. B., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General, 147*(4), 514–544.

Rothschild, Z. K., & Keefer, L. A. (2018). Righteous or self-righteous anger? Justice sensitivity moderates defensive outrage at a third-party harm-doer. *European Journal of Social Psychology*, *48*(4), 507–522.

意见 5：“由上我们推测突显网络在 TPP 中负责对违规行为进行监测以及产生情绪唤醒，主要参与到社会认知加工的早期阶段。”本段所引文献不能得出“主要在早期参与加工”这个结论。

回应：感谢审稿专家的意见。我们在此处的表述不够严谨，因此我们在 3.1 突显网络部分的最后(正文第 19 页)修改了表述以增加行文的准确性，具体修改如下：

“综上，突显网络在 TPP 中负责对违规行为进行检测，参与情绪加工并指导后续决策，为情绪模型提供了证据支持。然而，有研究者指出，以 AIC 为核心的突显网络启动并调节了大脑其他区域参与的认知-情感-动机过程(Menon & Uddin, 2010)。因此我们推测，突显网络在 TPP 过程中起到重要作用。”

参考文献：

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure and Function*, *214*(5–6), 655–667.

意见 6：图 2 TPP 中枢没有定义，并且需要对图二做进一步解释，图中有些实线有箭头，有些没有，比较迷惑。

回应：感谢审稿专家建议。根据审稿专家的意见，我们在正文部分补充了 TPP 中枢的定义，对图 2 进行修改并做了进一步解释。具体内容如下：

“前人研究发现，在与 TPP 相关的大脑区域之间存在一种独特的连接方式，dmPFC 是 TPP 激活模式的中枢(Bellucci et al., 2017; Feng et al., 2016)。中枢(hub)是指在格兰杰因果分析(Granger causality analysis)中与其他节点有最大数量因果联系的大脑区域，是信息交流的中心节点(Yang et al., 2023)，在这里体现为 dmPFC 与其他脑区之间有更多数量的功能连接。结合不同脑区的功能，我们对默认模式网络作用方式做出如下推测(图 2)：颞极(temporal pole, TP)负责理解违规行为，并向 dmPFC 提供伤害信息。dmPFC 在接收到伤害信息之后对伤害意图进行评估，并向其他区域传递信息，包括后扣带皮层(posterior cingulate cortex, PCC)、vmPFC 和 TPJ。其中，PCC 负责整合与违规行为相关的背景信息，vmPFC 负责编码伤害程

度，TPJ 负责推断意图，最后由 mPFC 整合伤害与意图两部分信息，形成“惩罚信号”(谢东杰, 苏彦捷, 2019)。”

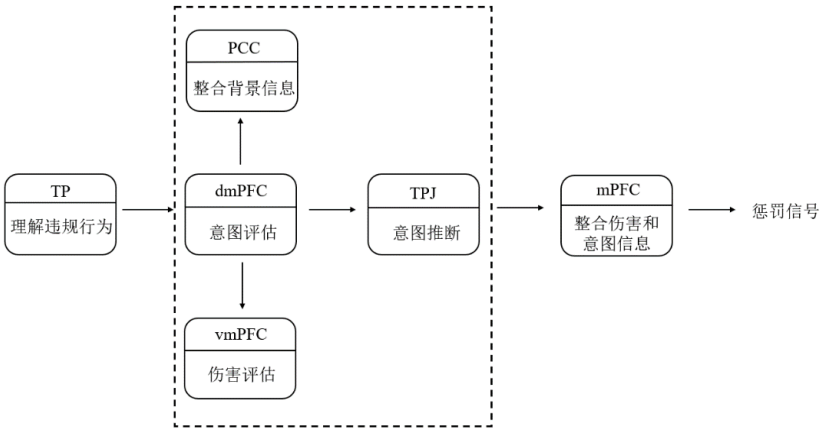

图 2 第三方惩罚行为在默认模式网络的作用路径

注：颞极(temporal pole, TP)，背内侧前额叶皮层(dorsomedial prefrontal cortex, dmPFC)，后扣带皮层(posterior cingulate cortex, PCC)，腹内侧前额叶皮层(ventromedial prefrontal cortex, vmPFC)，颞顶联合区(the temporoparietal junction, TPJ)，内侧前额叶皮层(medial prefrontal cortex, mPFC)，箭头代表信息传递方向。

参考文献：

谢东杰, 苏彦捷. (2019). 第三方惩罚的演化与认知机制. *心理科学*, *42*(1), 216–222.

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Monte, O. D., Knutson, K., Grafman, J., & Krueger, F. (2017). Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence. *Social Neuroscience*, *12*(2), 124–134.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*(2), 663–677.

Yang, C., Xiao, K., Ao, Y., Cui, Q., Jing, X., & Wang, Y. (2023). The thalamus is the causal hub of intervention in patients with major depressive disorder: Evidence from the Granger causality analysis. *NeuroImage: Clinical*, *37*, Article 103295.

意见 7：文章中图 2 和图 3 最好统一，比如图二用矩形代表脑区，图三用矩形代表认知过程。

回应：感谢审稿专家的建议！我们对图 2 和图 3 进行了修改，用直角矩形代表认知系统，用圆角矩形代表脑区。

第二轮

审稿人 1 意见：

作者对文章进行了大刀阔斧的修改，我认为之前的问题已经基本上解决了。

回应：感谢审稿专家的认可和肯定！

审稿人 2 意见：修改稿很好地回答了上一轮提出的问题。有一点建议：

图 3 提出了重要的理论模型，需要加以详细阐述。图中的数字 1-7 代表什么含义？是有时间的先后关系，还是单纯的数字区分？为什么有的箭头是实线，有的箭头是虚线？为什么有的是双向箭头，如箭头 5，而有的又是由两个单向箭头同时呈现（如 3 和 6）？为什么情绪系统是凸显网络？社会认知系统是默认模式网络？

回应：感谢审稿专家的肯定和宝贵建议！

首先，我们回答第一个问题，图 3 中数字 1-7 分别代表什么含义？

箭头 1 是违规行为发生后刺激诱发认知过程的起点，在模型中代表刺激输入。**箭头 2** 代表对动机系统(包括情绪系统和奖赏系统)信息的整合与传递。**箭头 3** 代表社会认知系统将整合后的“惩罚信号”传递给执行控制系统。具体而言，社会认知系统对伤害程度进行情感编码和对违规意图进行认知编码后形成“惩罚信号”，并将该信号进一步输送给执行控制系统。**箭头 4** 代表惩罚决策输出。**箭头 5** 表示在动机系统内部，情绪系统和奖赏系统的相互关系：情绪系统将负性情绪的产生等情绪信息传递给奖赏系统，使得个体更倾向于去寻求奖赏性刺激来缓解因负性情绪带来的不适感；同时，奖赏系统将以往的奖赏体验(如权力感和满足感)传递给情绪系统，以此引发积极的情绪体验，这种相互促进的关系为个体做出惩罚决策提供了动力。**箭头 6** 表示社会认知系统和执行控制系统之间的“拮抗”和“互补”关系。**箭头 7** 代表有奖赏参与的反馈通路：当第三方做出惩罚决定后，负性情绪的消解以及权力的体验感和满足感作为相应的奖励信号正反馈于奖赏系统，为下一次惩罚提供了动力(Strobel et al., 2011)。

根据审稿专家的建议，我们对理论模型进行了详细的注解，如图 3 所示。

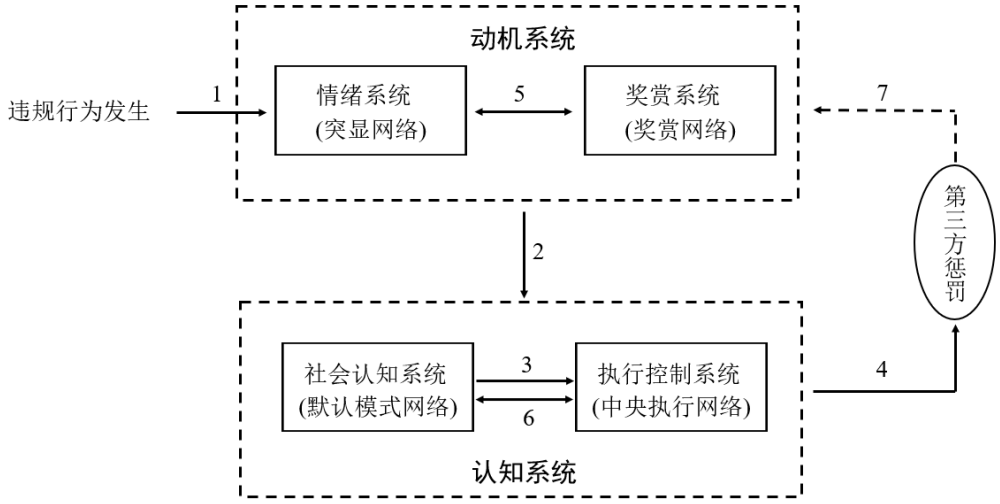

图 3 第三方惩罚行为的认知神经网络模型

注：虚线矩形框代表认知加工系统，实线矩形框代表各系统下的子系统，椭圆框代表行为决策。单向实线箭头(1、2、3、4)代表一次完整的惩罚过程，数字越大，阶段越后；箭头 1 代表刺激输入，箭头 2 代表对动机系统信息的整合与传递，箭头 3 代表社会认知系统将整合后的“惩罚信号”进一步输送给执行控制系统，箭头 4 代表惩罚决策输出。双向实线箭头(5、6)代表系统或网络之间的相互关系：箭头 5 代表情绪系统和奖赏系统之间的相互关系，箭头 6 代表默认模式网络和中央执行网络之间的“拮抗”和“互补”关系。虚线箭头(7)代表执行惩罚后的反馈过程，参与下一次第三方惩罚行为。

第二，是否有时间的先后关系？

模型中的数字具有一定的时间先后顺序。例如，**箭头 1、2、3、4** 代表一次完整的惩罚过程，以个体观察到违规行为开始，做出惩罚决定结束，数字越大，阶段越后。Treadway 等人(2014)的研究发现，对故意违规者的意图评估可能会抑制杏仁核的活动，减少因情绪冲动产生的惩罚。在 Krueger 和 Hoffman(2016)提出的 TPP 神经心理框架中，个体在面对违规行为时首先产生情绪反应，接着进行伤害和意图评估并将其整合到责任评估中，最后将“惩罚信号”转变为最终的惩罚决定。Qu 等人(2014)研究也指出，人们在 TPP 反应早期更可能受到情绪的驱动。由以上证据我们认为，情绪可能主要影响惩罚决策的早期阶段，是最终决策的重要驱动因素，但最终是否做出惩罚决定以及惩罚的强度还受到认知因素的影响，因此箭头 1-4 具有时间先后关系。但是对于**箭头 5 和 6** 代表的系统或网络之间的相互关系，与其它过程没有时间上的先后顺序。在整个 TPP 过程中，情绪系统和奖赏系统之间是相互配合、共同为 TPP 产生动力的，且前人发现任务态和静息态下默认模式网络(TPJ)和中央执行网络(dlPFC)之间均存在“拮抗”关系(Buckholtz et al., 2008; Zinchenko et al., 2021)，在功能上的互补性也并非特异于某个加工过程，因此箭头 5-6 没有时间先后关系。**箭头 7** 是惩罚决策执行

后的反馈过程，相关反馈信息作用于奖赏系统为下一次惩罚提供动力(Strobel et al., 2011)，在整个模型中处于最后阶段。

第三，为什么有的箭头是实线，有的箭头是虚线？

实线代表本次惩罚的作用过程或相互关系；虚线代表反馈通路，参与下一次第三方惩罚行为。

第四，为什么有的是双向箭头，如箭头 5，而有的又是由两个单向箭头同时呈现（如 3 和 6）？

单向箭头(1，2，3，4，7)表示单向信息传递过程以及传递方向，双向箭头表示系统间相互作用关系，如双向箭头 5 代表情绪系统和奖赏系统的相互关系，双向箭头 6 代表默认模式网络和中央执行网络之间的“拮抗”和“互补”关系。

第五，为什么情绪系统是凸显网络？

情绪系统涉及情绪的产生、处理和调节过程(Etkin et al., 2015; Pessoa, 2017)。以往研究发现，突显网络在 TPP 中主要负责检测冲突并参与负性情绪的加工(Bellucci et al., 2020; Feng et al., 2016; Mclatchie et al., 2016)，涉及脑区包括 dACC、AIC 和杏仁核。其中，AIC 与产生不公平厌恶反应有关(Hu et al., 2016; Krueger & Hoffman, 2016)，杏仁核负责根据受伤害程度产生情感唤醒信号并参与决定惩罚的严重程度(Stallen et al., 2018; Civai et al., 2019)，因此情绪系统对应的脑网络为突显网络。

最后，为什么社会认知系统是默认模式网络？

在 TPP 中，个体需要根据违规行为信息来评估伤害程度以及推断违规者的心理状态(Ginther et al., 2016; Treadway et al., 2014)。以往研究发现，默认模式网络主要负责对违规行为的伤害程度和伤害意图进行评估，主要包括 mPFC 和 TPJ(Lo Gerfo et al., 2019)。其中，TPJ 和 dmPFC 与各种社会认知功能相关，如他人意图推断(Feng et al., 2022)。若这两个脑区损伤或活动受到抑制，惩罚程度增加(Baumgartner et al., 2012, 2014; Moll et al., 2018)，这可能是个体对违规行为进行了合理的推测与解释。因此社会认知系统对应的脑网络为默认模式网络。

根据审稿专家的建议，我们在 4.1 小节对情绪系统和奖赏系统与脑网络的对应关系做了

进一步的说明 (正文见 28-29 页)；同样地，我们在 4.2 小节中对认知系统与脑网络之间的对应关系做进一步的说明(正文 29 页)，具体修改内容如下：

4.1 动机系统下的情绪系统和奖赏系统

情绪系统涉及情绪的产生、处理和调节过程(Etkin et al., 2015; Pessoa, 2017)。以往研究发现，突显网络在 TPP 中参与情绪加工并指导后续决策(Bellucci et al., 2020; Feng et al., 2016; Mclatchie et al., 2016)，奖赏网络的价值表征和预期奖赏的功能使得其与奖赏系统密切相关(Hu et al., 2015)。因此，情绪系统和奖赏系统对应的脑网络分别为突显网络和奖赏网络。值得注意的是，情绪系统和奖赏系统也并非完全独立(箭头 5)：情绪系统将负性情绪的产生等情绪信息传递给奖赏系统，使得个体更倾向于去寻求奖赏性刺激来缓解因负性情绪带来的不适感；同时，奖赏系统将以往的奖赏体验(如权力感和满足感)传递给情绪系统，以此引发积极的情绪体验，这种相互促进的关系为个体做出惩罚决策提供了动力。

4.2 认知系统内部关系及其对第三方惩罚行为的影响

认知系统包括社会认知系统和执行控制系统。在 TPP 中，个体需要根据违规行为信息来评估伤害程度以及推断违规者的心理状态(Ginther et al., 2016; Treadway et al., 2014)，而默认模式网络中与心智化相关的脑区如 TPJ 和 dmPFC 在意图推断方面起着重要作用(Feng et al., 2022)，且中央执行网络的核心脑区 dlPFC 的激活体现了认知控制在 TPP 中的重要性(殷西乐 等, 2019)。因此，社会认知系统和执行控制系统与默认模式网络和中央执行网络这两个脑网络相对应，分别参与“责任评估”和“惩罚选择”两个阶段。

参考文献：

陈瀛, 徐敏霞, 汪新建. (2020). 信任的认知神经网络模型. *心理科学进展*, *28*(5), 800–809.

殷西乐, 李建标, 陈思宇, 刘晓丽, 郝洁. (2019). 第三方惩罚的神经机制：来自经颅直流电刺激的证据. *心理学报*, *51*(5), 571–583.

Baumgartner, T., Bastian, S., Jrg, R., Gianotti, L. R. R., & Daria, K. (2014). Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Social Cognitive and Affective Neuroscience*, *9*(5), 653–660.

Baumgartner, T., Gtte, L., R Gügler, & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping*, *33*(6), 1452-1469.

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience and Biobehavioral Reviews*, *113*, 426–439.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*(5), 930–940.

Civai, C., Huijsmans, I., & Sanfey, A. G. (2019). Neurocognitive mechanisms of reactions to second- and third-party justice violations. *Scientific Reports*, *9*(1), Article 9271.

Etkin, A., Büchel, C., & Gross, J. J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience*, *16*(11), Article 11.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping*, *37*(2), 663–677.

Feng, C., Yang, Q., Azem, L., Atanasova, K. M., Gu, R., Luo, W., Hoffman, M., Lis, S., & Krueger, F. (2022). An fMRI investigation of the intention-outcome interactions in second- and third-party punishment. *Brain Imaging and Behavior*, *16*(2), 715–727.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience*, *36*(36), 9420–9434.

Hu, J., Blue, P. R., Yu, H., Gong, X., Xiang, Y., Jiang, C., & Zhou, X. (2016). Social status modulates the neural response to unfairness. *Social Cognitive and Affective Neuroscience*, *11*(1), 1–10.

Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, *9*, Article 24.

Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences*, *39*(8), 499–501.

Lo Gerfo, E., Gallucci, A., Morese, R., Vergallito, A., Ottone, S., Ponzano, F., Locatelli, G., Bosco, F., & Romero Lauro, L. J. (2019). The role of ventromedial prefrontal cortex and temporo-parietal junction in third-party punishment behavior. *NeuroImage*, *200*, 501–510.

Mclatchie, N., Giner-Sorolla, R., & Derbyshire, S. W. G. (2016). 'Imagined guilt' vs 'recollected guilt': Implications for fMRI. *Social Cognitive and Affective Neuroscience*, *11*(5), 703–711.

Moll, J., de Oliveira-Souza, R., Basilio, R., Bramati, I. E., Gordon, B., Rodr ́guez-Nieto, G., Zahn, R., Krueger, F., & Grafman, J. (2018). Altruistic decisions following penetrating traumatic brain injury. *Brain*, *141*(5), 1558–1569.

Pessoa, L. (2017). A network model of the emotional brain. *Trends in Cognitive Sciences*, *21*(5), 357–371.

Qu, L., Dou, W., You, C., & Qu, C. (2014). The processing course of conflicts in third-party punishment: An event-related potential study. *Psychology Journal*, *3*(3), 214–221.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K. W., & Sanfey, A. G. (2018). Neurobiological mechanisms of responding to injustice. *The Journal of Neuroscience*, *38*(12), 2944–2954.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: Neural and genetic bases of altruistic punishment. *NeuroImage*, *54*(1), 671–680.

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., Jones, O. D., & Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience*, *17*(9), 1270-1275.

Zinchenko, O., & Klucharev, V. (2017). Commentary: The emerging neuroscience of third-party punishment. *Frontiers in Human Neuroscience*, *11*, Article 512.

Zinchenko, O., Nikulin, V., & Klucharev, V. (2021). Wired to punish? Electroencephalographic study of the resting-state neuronal oscillations underlying third-party punishment. *Neuroscience, 471*, 1–10.

第三轮

审稿人 2 意见：已较好地回答本人提出的问题。

编委 1 意见：同意发表。

编委 2 意见：同意发表。

主编意见：根据编委和审稿专家的意见，建议发表。