Cognitive and Neural Mechanisms of Third-Party Punishment

Authors: The user has provided a single Chinese word "郑好". This is a name, likely a person' s name. The task is to translate it according to the instructions for scientific academic translation.

Let me analyze the requirements: 1. Preserve ALL LaTeX commands and placeholders - not applicable here as there are none 2. Preserve ALL citation commands - not applicable 3. **PRESERVE ALL ···paragraph wrapper tags EXACTLY** - not provided, but I need to check if this is a complete paragraph 4. Maintain academic tone and formal scientific writing style 5. Use precise technical terminology - obey the glossary strictly 6. Preserve section structure and paragraph organization 7. Translate all content completely - do not omit any information

The input is just "郑好". This appears to be a name. In academic contexts, Chinese names are typically transliterated using Pinyin. "郑好" would be "Zheng Hao" in Pinyin.

However, I need to consider the context. The user hasn' t provided any tags or other structural elements. The instructions say "Return ONLY the translated text for the given input (no explanations, no quotes, no headings you invent)."

Since this is just a name, and there are no paragraph tags provided, I should just provide the transliteration. But I need to be careful about the format. The instructions emphasize preserving tags, but none are present in the input.

Let me re-read the instructions carefully: - "Return ONLY the translated text for the given input" - "Keep one output paragraph per input paragraph; do not merge or split paragraphs." - "**Maintain exact count and IDs of ···tags**"

The input is just "郑好" without any tags. So I should just output the translation of this name. In academic writing, Chinese names are typically written in Pinyin with the surname first: "Zheng Hao".

But wait, I should consider if this could be something else. "郑好" could be: 1. A person' s name (most likely) 2. A phrase meaning "exactly good" or "precisely good" - but this is unlikely in an academic context 3. Part of a larger sentence that was cut off

Given that the user just provided this single term and the instructions are about translating academic papers, it' s most likely a name that would appear in a, The user has provided a very short input: "陈荣荣". This appears to be a Chinese name. Let me analyze the requirements:

1. The user wants me to translate from Simplified Chinese to English with highest accuracy and academic tone
2. I must preserve ALL LaTeX commands and placeholders exactly
3. I must preserve ALL citation commands ([?], ??, (??), etc.) exactly
4. **PRESERVE ALL …paragraph wrapper tags EXACTLY**
5. Maintain academic tone and formal scientific writing style
6. Use precise technical terminology - obey the glossary strictly
7. Preserve section structure and paragraph organization
8. Translate all content completely - do not omit any information

However, the input "陈荣荣" is just a name. It doesn' t contain: - Any LaTeX commands - Any placeholders - Any citation commands - Any …tags - Any technical terminology from the glossary (机器学习, 深度学习) - Any section structure or paragraph organization

This is simply a Chinese name that needs to be transliterated into English. In academic contexts, Chinese names are typically transliterated using the Pinyin system, with the surname first followed by the given name.

"陈荣荣" would be: - Surname: 陈 (Chén) -> Chen - Given name: 荣荣 (Róngróng) -> Rongrong

In academic writing, Chinese names are usually written as: Chen Rongrong (with a space between surname and given name, and the given name often capitalized as one word or with a hyphen).

The most standard format would be: Chen Rongrong

Since there are no tags in the input, I should not add any. The output should be just the transliterated name.

But wait, let me re-read the instructions carefully. The user says "Translate from Simplified Chinese to English" . The input is just "陈荣荣" . There' s no context, no paragraphs, no LaTeX, nothing.

The CRITICAL REQUIREMENTS say: - **PRESERVE ALL …paragraph wrapper tags EXACTLY** - these are structural markers that identify paragraphs and MUST remain untouched with their IDs

But there are no such tags in the input. So I should just output the, The user has provided "买晓琴" which is a Chinese name. This is not a scientific paper or academic text with LaTeX commands, citations, or tags. It' s simply a name.

According to the instructions: - I am a professional scientific translator specializing in academic papers - I need to translate from Simplified Chinese to English with highest accuracy and academic tone - I must preserve LaTeX commands, placeholders, citation commands, and tags - The glossary only contains "机器学习 => machine learning" and "深度学习 => deep learning"

However, "买晓琴" is just a name. In academic contexts, Chinese names are typically transliterated (Pinyin) and formatted according to academic conventions. The surname is "买" (Mǎi) and the given name is "晓琴" (Xiǎoqín).

In academic writing, Chinese names are usually written in one of these formats: 1. Family name first: Mai Xiaoqin 2. Given name first: Xiaoqin Mai (Western order) 3. With hyphen: Xiao-qin Mai

The most common and academically appropriate format is to keep the family name first: Mai Xiaoqin. Sometimes the family name is written in all caps: MAI Xiaoqin, but that's less common in running text.

Since there are no tags, LaTeX commands, or any other structural elements, and it's just a name, I should simply provide the standard academic transliteration.

The user might be asking for the translation/romanization of this name for use in an academic paper. The most appropriate response is "Mai Xiaoqin".

Wait, let me re-read the instructions carefully. The user just wrote "买晓琴". This could be: 1. A test to see how I handle a simple input 2. The beginning of a larger text that got cut off 3. A request to translate just this name

Given the strict requirements about preserving tags and other elements, and the fact that this input has none of those, I should just provide the direct translation/transliteration.

The name "买晓琴" would be: - Surname: 买 (Mǎi) - Given name: 晓琴 (Xiǎoqín)

In academic English, the standard is: - Mai Xiaoqin (surname first, capitalized) - Or sometimes: Xiaoqin, Mai Xiaoqin

Date: 2023-09-13T00:00:00+00:00

Abstract

Third-party punishment (TPP) refers to punitive actions taken by individuals as third parties or observers against norm violators to uphold social norms. While numerous studies have provided insights into the neural mechanisms underlying TPP behavior, few have examined the holistic role played by different functional brain networks in this process. This paper reviews TPP-related research from the past decade, summarizes relevant theoretical models and brain networks, and proposes a cognitive neural network model of TPP that systematically explains and integrates the neural mechanisms underlying TPP

Full Text

Preamble

The Cognitive and Neural Mechanisms of Third-Party Punishment Behavior

ZHENG Hao, CHEN Rongrong, MAI Xiaoqin
(Department of Psychology, Renmin University of China, Beijing 100872, China)

Abstract

Third-party punishment (TPP) refers to punitive behaviors that individuals, acting as uninvolved third parties or observers, impose on norm violators to uphold social norms. While numerous studies have provided insights into the neural mechanisms underlying TPP behavior, few have examined the integrated role of different functional brain networks. This paper reviews TPP-related research from the past decade, summarizes relevant theoretical models and brain networks, and proposes a cognitive neural network model of TPP that systematically explains and integrates the neural mechanisms behind this behavior. In this model, the emotional and reward systems serve as the motivational sources of TPP, while the cognitive system is primarily responsible for responsibility assessment and punishment selection. The reward network, salient network, default mode network, and central executive network participate in distinct cognitive processing stages. This model establishes connections between psychological and cognitive-neural levels of TPP research, offering a more holistic and comprehensive explanation of the mechanisms underlying TPP behavior. Future research should employ meta-analysis or machine learning-based analytical methods to explore third-party intervention preferences and their underlying cognitive neural mechanisms across different contextual information and more complex social situations.

Keywords: third-party punishment, cognitive neural mechanisms, brain networks, fMRI

The user wants me to translate the "## 1 Introduction" section from Chinese to English. Let me carefully read the text and identify any issues.

Looking at the text, I see it' s already mostly in English, but there are some Chinese characters mixed in. The user specifically mentions that Chinese characters were detected: 经, 罚, 三, 他, 认, 惩, 会, 像, 制, 机, 神, 利, 成, 脑, 知 (and 5 more). These appear to be fragments that need to be removed or properly integrated.

Let me go through the text sentence by sentence to identify and fix these issues:

1. "The establishment and enforcement of social norms represent a distinguishing feature of human society (Köster et al., 2022), yet norm violations remain widespread in daily life." - This is fine, no Chinese characters.

2. "Increasing institutional constraints can enhance compliance with norms (Fehr & Schurtenberger, 2018)." - Fine.
3. "Social punishment is a common moral and institutional constraint behavior (Fehr & Fischbacher, 2004b) that can be categorized into second-party punishment and third-party punishment (TPP)." - Fine.
4. "Unlike second-party punishment, which follows a" tit-for-tat" logic, TPP occurs when individuals, as uninvolved third parties or observers, punish violators to maintain social norms (Fehr & Fischbacher, 2004a; Kanakogi et al., 2022)." - Fine.
5. "Generally, such punishment does not yield direct personal benefits and incurs certain costs, making it a form of altruistic punishment (Fehr & Fischbacher, 2003)." - Fine.
6. "TPP constrains and regulates human behavior, thereby promoting and maintaining social fairness and cooperation, which has attracted widespread attention from researchers (Fehr & Gächter, 2002; Kim et al., 2021; Martin et al., 2021; Yang et al., 2022)." - Fine.
7. "Since the last century, numerous studies have investigated economic decision-making and distribution behaviors." - Fine.
8. "Behavioral economists first discovered altruistic punishment in the Ultimatum Game (UG) (Thaler, 1988)." - Fine.
9. "In this two-player task, participants must agree on how to divide a sum of money." - Fine.
10. "The proposer offers a division scheme, which the responder can accept or reject." - Fine.
11. "If accepted, both receive money according to the proposal; if rejected, neither receives anything." - Fine.
12. "In this context, rejection can be viewed as a costly second-party punishment by the responder against the proposer." - Fine.
13. "However, in the real world, second parties often play a passive role." - Fine.
14. "Moreover, second-party punishment alone can only maintain a limited number of social norms." - Fine.
15. "Therefore, introducing third-party punishment can expand the scope of punishment for violators and better maintain social norms." - Fine.
16. "Scholars such as Fehr demonstrated the existence of TPP under laboratory conditions." - Fine.
17. "Fehr and Fischbacher (2004a) introduced a third party into the Dictator Game (DG)." - Fine.

18. “Unlike UG, the responder in DG cannot reject the dictator’ s (proposer’ s) allocation and must passively accept it.” - Fine.
19. “As a third-party observer, participants could punish the dictator by paying a cost (reducing their own money) after seeing the allocation scheme.” - Fine.
20. “This paradigm subsequently became one of the important research methods for TPP, providing a new approach to studying social norm behavior in laboratory settings.” - Fine.
21. “Researchers have manipulated variables such as the social status of dictators and recipients (Cui et al., 2019; Ouyang et al., 2021), the degree of allocation unfairness (Sun et al., 2015), the cost of punishment (Cheng et al., 2022), and the social distance and group relationships between third parties and the two players (Bernhard et al., 2006; Liu et al., 2018) to further explore whether TPP behavior is influenced by social context or background information.” - Fine.
22. “In addition to DG, commonly used paradigms in TPP research include public goods games (Zhou et al., 2017; Tang et al., 2022), trust games (Konishi & Ohtsubo, 2015), prisoner’ s dilemma (Lergetporer et al., 2014), justice games (Civai et al., 2019), and the Chicago Moral Sensitivity Task (Kim et al., 2021; Meidenbauer et al., 2018).” - Fine.
23. “Researchers typically focus on whether third parties intervene in norm violations and what factors influence such intervention.” - Fine.
24. “As research has progressed, investigators have begun to infer the cognitive neural mechanisms underlying TPP behavior from electrophysiological and functional imaging evidence.” - Fine.
25. “However, due to methodological limitations, most evidence has been confined to single components or isolated brain region activations.” - Fine.
26. “Moreover, no comprehensive understanding has yet been formed regarding the connection between the cognitive processes and brain functional networks underlying TPP behavior.” - Fine.
27. “Therefore, this paper reviews TPP-related research from the past decade.” - Fine.
28. “We first conducted a literature search using Web of Science, PubMed, and ScienceDirect for English literature, with TPP keywords including” third-party punishment,” “altruistic punishment,” or “social punishment,” and cognitive neural mechanism keywords including “cognitive,” “neural bias,” “neural correlates,” “neuroimaging,” or “fMRI.” - Fine.
29. “For Chinese literature, we searched CNKI, Wanfang, and VIP databases using TPP keywords”第三方惩罚”(third-party punishment), “利他惩罚”(altruistic punishment), or “社会惩罚” (social punishment), and cognitive neural

mechanism keywords "认知" (cognitive), "神经机制" (neural mechanism), "神经基础" (neural basis), or "脑成像" (brain imaging)." - This contains Chinese characters but they are part of the search terms and their translations, which should be preserved as is.

30. "We also used a snowballing method by examining references in retrieved papers to identify additional studies." - Fine.
31. "As of April 2023, we retrieved 1,149 papers published between April 2013 and April 2023." - Fine.
32. "After initial screening and review, 60 papers were included, of which 39 involved neural mechanisms." - Fine.
33. "Literature inclusion and exclusion criteria and the PRISMA flowchart are provided in Appendix Figure S1." - Fine.
34. "Based on previous research, this paper first introduces theoretical models related to TPP behavior, interpreting TPP from a theoretical perspective." - Fine.
35. "Second, we summarize evidence regarding brain networks involved in TPP, with particular attention to coordination between brain regions." - Fine.
36. "Finally, we construct a cognitive neural network model of TPP based on existing theories and research, providing a new dynamic perspective for understanding TPP behavior." - Fine.

Wait, the user mentioned that Chinese characters were detected, but when I read through the text, I don't see any obvious Chinese characters mixed in the English sentences. Let me look more carefully at the text provided.

Actually, looking at the text again, I notice that the user said "These appear to be: mixed technical content and proper nouns". This suggests that the Chinese characters might be embedded in the text in a way that's not immediately visible to me as I'm reading it. The user wants me to ensure NO Chinese characters remain.

Let me re-examine the text character by character to see if there are any hidden Chinese characters. The user listed specific characters: 经, 罚, 三, 他, 认, 惩, 会, 像, 制, 机, 神, 利, 成, 脑, 知 (and 5 more).

Looking at the text again, I see that in sentence 29, there are Chinese terms with their English translations in parentheses. This is appropriate and should be kept. But maybe there are other places where Chinese characters are mixed in.

Let me check the references and author names: - Köster et al., 2022 - Fehr & Schurtenberger, 2018 - Fehr & Fischbacher, 2004b - Fehr & Fischbacher, 2004a - Kanakogi et al., 2022 - Fehr & Fischbacher, 2003 - Fehr & Gächter, 2002 - Kim et al., 2021 - Martin et al., 2021 - Yang et al., 2022 - Thaler, 1988 - Cui et al., 2019 - Ouyang et al., 2021 - Sun et al., 2015 - Cheng et al., 2022 - Bernhard et al., 2006 - Liu et al., 2018 - Zhou et al., 2017 - Tang et al., 2022 - Konishi & Ohtsubo, 2015 - Lergetporer et al., 2014 - Civai et al., 2019 - Meidenbauer et al., 2018

These all look like proper Romanized names. The Chinese names like "Cui" , "Ouyang" , "Sun" , "Cheng" , "Liu" , "Zhou" , "Tang" , "Yang" are correctly Romanized.

Wait, I think I understand now. The user is saying that the text CONTAINS Chinese characters that need to be removed. But when I look at the text provided, it appears to be clean English with only the intentional Chinese terms in sentence 29 (which are properly formatted with translations).

Let me check the text again very carefully. The user said "These appear to be: mixed technical content and proper nouns" . This might mean that the Chinese characters are appearing as artifacts, perhaps from OCR or some encoding issue.

Looking at the text character by character in my mind: - "The establishment and enforcement of social norms represent a distinguishing feature of human society (Köster et al., 2022), yet norm violations remain widespread in daily life." - No Chinese characters here.

Actually, I think the issue might be that the user is giving me a text that has Chinese character artifacts embedded in it, but when I read it in the prompt, they might not be visible due to formatting. The user wants me to produce a clean translation without any Chinese characters.

Given that the text appears to be already in English (except for the intentional Chinese terms in sentence 29), my task is to: 1. Ensure the text is fluent and readable academic English 2. Remove any hidden Chinese character artifacts 3. Preserve all the citations, math (though there is none here), figure/table markers (none here) 4. Keep the structure as "## 1 Introduction"

Since I don' t see any obvious Chinese characters in the text provided (except the intentional ones in sentence 29), I should just polish the English to make it more fluent and academic, while preserving all the content.

Let me review the text for any awkward phrasing or improvements:

Original: "The establishment and enforcement of social norms represent a distinguishing feature of human society (Köster et al., 2022), yet norm violations remain widespread in daily life."

This is good.

Original: "Increasing institutional constraints can enhance compliance with norms (Fehr & Schurtenberger, 2018)."

Good.

Original: "Social punishment is a common moral and institutional constraint behavior (Fehr & Fischbacher, 2004b) that can be categorized into second-party punishment and third-party punishment (TPP)."

Good.

Original: "Unlike second-party punishment, which follows a" tit-for-tat" logic, TPP occurs when individuals, as uninvolved third parties or observers, punish violators to maintain social norms (Fehr & Fischbacher, 2004a; Kanakogi et al., 2022)."

Good.

Original: "Generally, such punishment does not yield direct personal benefits and incurs certain costs, making it a form of altruistic punishment (Fehr & Fischbacher, 2003)."

Good.

Original: "TPP constrains and regulates human behavior, thereby promoting and maintaining social fairness and cooperation, which has attracted widespread attention from researchers (Fehr & Gächter, 2002; Kim et al., 2021; Martin et al., 2021; Yang et al., 2022)."

Good.

Original: "Since the last century, numerous studies have investigated economic decision-making and distribution behaviors."

Good.

Original: "Behavioral economists first discovered altruistic punishment in the Ultimatum Game (UG) (Thaler, 1988)."

Good.

Original: "In this two-player task, participants must agree on how to divide a sum of money."

Good.

Original: "The proposer offers a division scheme, which the responder can accept or reject."

Good.

Original: "If accepted, both receive money according to the proposal; if rejected, neither receives anything."

Good.

Original: "In this context, rejection can be viewed as a costly second-party punishment by the responder against the proposer."

Good.

Original: "However, in the real world, second parties often play a passive role."

Good.

Original: "Moreover, second-party punishment alone can only maintain a limited number of social norms."

Good.

Original: "Therefore, introducing third-party punishment can expand the scope of punishment for violators and better maintain social norms."

Good.

Original: "Scholars such as Fehr demonstrated the existence of TPP under laboratory conditions."

Good.

Original: "Fehr and Fischbacher (2004a) introduced a third party into the Dictator Game (DG)."

Good.

Original: "Unlike UG, the responder in DG cannot reject the dictator' s (proposer' s) allocation and must passively accept it."

Good.

Original: "As a third-party observer, participants could punish the dictator by paying a cost (reducing their own money) after seeing the allocation scheme."

Good.

Original: "This paradigm subsequently became one of the important research methods for TPP, providing a new approach to studying social norm behavior in laboratory settings."

Good.

Original: "Researchers have manipulated variables such as the social status of dictators and recipients (Cui et al., 2019; Ouyang et al., 2021), the degree of allocation unfairness (Sun et al., 2015), the cost of punishment (Cheng et al., 2022), and the social distance and group relationships between third parties and the two players (Bernhard et al., 2006; Liu et al., 2018) to further explore whether TPP behavior is influenced by social context or background information."

Good.

Original: "In addition to DG, commonly used paradigms in TPP research include public goods games (Zhou et al., 2017; Tang et al., 2022), trust games (Konishi & Ohtsubo, 2015), prisoner' s dilemma (## 2 Theoretical Models of Third-Party Punishment

TPP is a complex altruistic behavior, and revealing its underlying cognitive mechanisms can enhance our understanding of its emergence and development. Previous studies have found that emotion, cognitive control, and contextual factors influence TPP behavior. However, current theoretical understanding of TPP remains insufficient, with most interpretations based on theoretical models of social value decision-making. Therefore, we first summarize theories that can explain TPP behavior to form a theoretical understanding. These models include the reciprocity model reflecting individual cooperation and fairness preferences, the emotion model based on intuitive processing, and the dual-system model that integrates emotional and cognitive factors from a reinforcement learning perspective.

2.1 Reciprocity Model

"I help you, you help me." Reciprocity refers to returning others' behaviors in a similar manner and is crucial for promoting human cooperation (Guala, 2012). TPP behavior involves two reciprocity principles: weak reciprocity and strong reciprocity (Fehr & Fischbacher, 2003). In TPP, weak reciprocity generally refers to indirect reciprocity, where individuals use this costly altruistic behavior as a "trustworthy" signal to imply their own noble qualities of fairness and justice, thereby establishing reputation within the group to obtain more opportunities and cooperation in future interpersonal interactions, ultimately achieving indirect benefits (Jordan, Hoffman et al., 2016; Rai, 2022). Unlike weak reciprocity, which expects benefits, strong reciprocity posits that individuals are motivated to punish unfair behavior to maintain fairness (Ciaramidaro et al., 2018), even when punishment is costly and these costs may not be compensated. The existence of strong reciprocators ensures group benefits and enables extensive cooperation among people, consistent with evolutionary principles (Buckholtz & Marois, 2012).

The reciprocity model explains the mechanism of TPP from a macro perspective, suggesting that individuals punish based on considerations of cooperation and fairness. However, this model has limitations. First, indirect reciprocity emphasizes that the purpose of third-party punishment is to obtain indirect benefits. In single-trial or anonymous situations where individuals have almost no future cooperation opportunities or cannot be identified, reputation motives do not exist. Previous research has revealed that even under these conditions, individuals still engage in TPP, with probabilities exceeding 1/2. For example, Piazza and Bering (2008) found that under completely anonymous third-party conditions, 71.4% of participants chose to sacrifice one-third of their money to punish violators. Feng et al. (2022) also found that in single anonymous experiments, individuals'average punishment rate exceeded 50%. Domestic researchers Yang and Chen (2022) similarly found that non-punishment or low-punishment scenarios were uncommon (3.15%) in single anonymous games. Furthermore, recent studies have shown that even eight-month-old infants, who have not yet entered society, can use gaze to punish violators (Kanakogi et al., 2022). These findings cannot be explained by indirect reciprocity. Second, strong reciprocity does not account for retaliatory behaviors resulting from punishment, which are not conducive to group development and stability. Finally, strong reciprocity punishment is a voluntary behavior whose intensity is highly sensitive to punishment costs. If punishment costs are too high, third parties, as unaffected individuals, will rationally reduce their demand to implement punishment at their own expense. Thus, understanding TPP behavior solely from the perspective of reciprocity, cooperation, and fairness has certain limitations.

2.2 Emotion Model

The "affect-as-information" model posits that affect is a heuristic tool that simplifies judgment and can serve as an information source during decision-making to guide subsequent choices (Bright & Goodman-Delahunty, 2006; Zhao et al., 2022). Negative emotion generation is one motivational source of TPP (Fehr & Fischbacher, 2004a; Fehr & Gächter, 2002; Xiao & Houser, 2005). When individuals observe norm violations, they experience a series of negative emotions, including anger toward violators (Jordan, McAuliffe, & Rand, 2016) and disgust toward selfish intentions (McAuliffe et al., 2015) and unfairness (Raihani & McAuliffe, 2012; Sun et al., 2015). At this point, individuals' initial intention may be to alleviate their negative emotions through punishing violators, while maintaining social norms is merely an "additional benefit" (de Quervain et al., 2004).

However, some researchers have found that negative emotions do not necessarily lead to punishment behavior. For example, Qu et al. (2014) used a third-party punishment DG task and found that when negative emotions conflicted with economic losses caused by punishment costs, individuals sometimes chose "not to punish." Additionally, even when individuals' emotional activation was the same, punishment severity still varied. Research has found that even when individuals perceived the same degree of allocation unfairness—that is, had the same level of inequity aversion—the final punishment severity still changed according to different punishment costs (Cheng et al., 2022). These studies suggest that emotions may influence certain stages of response judgment, with the final punishment decision being a comprehensive consideration of various factors, including emotion, fairness, and self-interest trade-offs.

2.3 Dual-System Model from a Reinforcement Learning Perspective

Reinforcement learning (RL) refers to individuals learning through receiving environmental feedback information and continuously adjusting behavioral strategies, representing a method of reward learning (Köster et al., 2022; Morris et al., 2017). Dual-system theory posits that social decision-making is influenced by two systems: automatic processing and controlled processing (Chung et al., 2023). The former belongs to a bottom-up heuristic processing process, where negative emotion generation prompts punishment to become an automatic dominant response (Mussel et al., 2018), corresponding to the model-free (MF) strategy in RL. The latter belongs to top-down processing, where individuals choose optimal behavioral approaches under rational guidance and experience (Zhou et al., 2014), corresponding to the model-based (MB) strategy in RL.

Previous research indicates that individuals' final decisions represent a dynamic balance between the two systems (Zhang et al., 2018). MF makes decisions in an intuitive and impulsive manner (with minimal cost) while being influenced and controlled by MB, reflecting individuals' exploration of adaptive behaviors through interactions between different systems (Gershman et al., 2014; Lee et al., 2014).

In TPP, when a norm violation occurs, individuals' automatic affective evaluation activates punishment behavior, reflecting the emotional drive for punishment (Qu et al., 2014). When intuitive responses conflict with economic interests, MB exerts cognitive control over MF, and the final punishment decision results from individuals' "deliberation" (Yin et al., 2019). After punishment behavior occurs, the resulting sense of satisfaction and power experience (Delgado et al., 2003; Strobel et al., 2011; Yamagishi et al., 2017) helps individuals establish higher self-esteem, which, combined with the resolution of negative emotions and anticipation of future rewards, serves as a reward signal for internal reinforcement, prompting individuals to engage in the next punishment behavior.

This model presents traditional dual-system theory from a reinforcement learning perspective, offering a novel angle to explore the emergence and development of TPP behavior. Specifically, this model not only combines emotional and cognitive factors but also suggests that TPP should be a dynamic process involving feedback and reinforcement, where individuals learn from each feedback instance and eventually form stable behavioral patterns. However, this model does not address individuals' speculation about the causes of norm violations. Some research suggests that intention assessment of deliberate violators may inhibit amygdala activity and reduce punishment driven by emotional impulse (Treadway et al., 2014). This suggests that future applications of this model to explain TPP behavior mechanisms should also consider the important factor of evaluating violators' mental states.

The above theoretical models explain the mechanisms of TPP behavior from different perspectives. However, existing theoretical models have certain limitations in explaining TPP behavior. Increasingly, researchers view TPP as a dynamic process involving feedback. "Dynamic" refers to two characteristics: multi-systemic nature and temporality (Kozlowski & Ilgen, 2006). As a complex social behavior, TPP likely results from interactions among multiple systems including emotion, cognition, and reward, satisfying the multi-systemic characteristic. Additionally, the existence of reward feedback allows individuals to adjust TPP behavior during each reinforcement episode. ERP studies have also found different brain electrical components at various stages, such as the medial frontal negativity (MFN) peaking around 200–350 ms after stimulus and the late positive component (LPC) peaking around 300–900 ms, with amplitudes correlating with subsequent punishment severity (Cui et al., 2019; Qu et al., 2014), satisfying the temporality characteristic.

As early as 2011, Strobel et al. proposed that a cognitive-affective-motivational network serves as the driving force behind strong reciprocity behavior. However, due to the lack of emotion ratings in their experiments, they could not directly determine which emotions influenced TPP behavior. Subsequently, Buckholtz and Marois (2012) noted that successful TPP execution requires two independent cognitive mechanisms: responsibility assessment and punishment selection, which partially integrated cognition and behavioral decision-making but did not incorporate emotional factors. Currently, many studies aim to demonstrate that affect and cognition shape human prosocial behavior, yet the specific decision-making processes and mechanisms of interaction between emotion and cognition remain unclear (Rahal & Fiedler, 2022), with relevant research evidence scattered and lacking holistic understanding. Neural-level evidence can help provide more detailed understanding of the complex cognitive and affective processes involved in decision-making. Therefore, we review functional neuroimaging and electrophysiological evidence related to TPP, focusing particularly on intra-network and inter-network connections to establish a neural foundation for understanding TPP behavior.

3 Brain Networks Involved in Third-Party Punishment

Previous research indicates that TPP involves emotion generation, intent and harm degree assessment, and punishment selection stages. Combining the functions and activation patterns of relevant brain networks from previous studies (Bellucci et al., 2020; Krueger & Hoffman, 2016; Lo Gerfo et al., 2019), this paper proposes that TPP behavior generation involves three stages: "emotion generation," "responsibility assessment," and "punishment selection," with corresponding brain networks being the salient network, default mode network, and central executive network. Additionally, the reward network collaborates in TPP processing, primarily serving value representation and reward expectation functions. The brain regions included in these networks and their locations are shown in Figure 1 [Figure 1: see original paper].

Figure 1. Brain networks related to third-party punishment and their anatomical locations. Note: dorsal anterior cingulate cortex ($dACC$), anterior insula cortex (AIC), amygdala, posterior cingulate cortex (PCC), dorsomedial prefrontal cortex ($dmPFC$), temporoparietal junction (TPJ), dorsomedial prefrontal cortex ($dlPFC$), ventrolateral prefrontal cortex ($vlPFC$), posterior parietal cortex (PPC), intraparietal sulcus (IPS), ventral striatum (VS), ventromedial prefrontal cortex ($vmPFC$), ventral tegmental area (VTA).

3.1 Salient Network

Fairness is a default social norm (Civai, 2013). When norm violations occur, individuals experience anger, inequity aversion, and other negative emotions. This anger and disgust are other-centered moral emotions (Pedersen et al., 2018). Additionally, when individuals expect they should punish violators to maintain justice but fail to do so, they experience self-centered guilt (Nelissen & Zeelenberg, 2009), which promotes anger generation to some extent (Rothschild &

Keefer, 2018). Therefore, the salient network is responsible for detecting conflict and generating negative emotions such as anger, disgust, and guilt (Bellucci et al., 2020; Buckholtz & Marois, 2012; Feng et al., 2016; Mclatchie et al., 2016), with key brain regions including the dorsal anterior cingulate cortex ($dACC$), anterior insula cortex (AIC), and amygdala.

Previous research indicates that $dACC$ plays an important role in monitoring cognitive conflict (Wang et al., 2017), while AIC is related to negative emotion representation (Shenhav et al., 2016; Singer et al., 2009). Craig (2009) integrated these two regions into an insula functional model, suggesting they jointly participate in emotional processing. When individuals observe unfair behavior, $dACC$ and AIC monitor norm violations or threats (Feng et al., 2016), with AIC marking violation signals and generating inequity aversion responses (Civai et al., 2012; Hu, Blue et al., 2016). Research has found that AIC activation correlates positively with the degree of allocation unfairness (Zhong et al., 2016). Related ERP studies have also found increased MFN amplitude sourced near the anterior cingulate cortex (ACC) during this process, a component sensitive to violations of social expectations and norms (Mothes et al., 2016; Sun et al., 2015; van der Helden et al., 2010; Wu et al., 2011). These findings suggest that $dACC$ and AIC may correspond to early evaluation processing of outcomes. Additionally, as one of the important brain regions for social and emotional processing, the amygdala in TPP is responsible for generating affective arousal signals based on the victim' s degree of harm (Buckholtz & Marois, 2012; Krueger & Hoffman, 2016) and participates in determining punishment severity (Stallen et al., 2018). In short, AIC and amygdala reflect two different aspects of response decision-making to unfair behavior: the former reflects individuals' social preferences and punishment willingness, while the latter reflects individuals' emotional experiences and influences punishment severity (Civai et al., 2019). In summary, the salient network in TPP is responsible for detecting norm violations, participating in emotional processing, and guiding subsequent decisions, providing evidence support for the emotion model. However, some researchers have noted that the AIC-centered salient network initiates and modulates cognitive-affective-motivational processes involving other brain regions (Menon & Uddin, 2010). We therefore speculate that the salient network plays an important role in the TPP process.

3.2 Default Mode Network

Before making punishment decisions, individuals need to assess the harm degree and intent of norm violations and integrate this information into responsibility assessment to form a "punishment signal." The network involved in this process is called the default mode network, primarily including the medial prefrontal cortex ($mPFC$) and temporoparietal junction (TPJ). Among them, the ventromedial prefrontal cortex ($vmPFC$) is responsible for evaluating the degree of harm (Bellucci et al., 2017), and enhanced functional connectivity with the amygdala suggests these two regions may jointly encode the emotional aspects of harm degree (Treadway et al., 2014). The dorsomedial prefrontal cortex ($dmPFC$) and TPJ are related to inferring others' mental states (Jamali et al., 2021; Morese et al., 2016; Xie et al., 2020; Yang, Shao et al., 2019). $dmPFC$ and TPJ in TPP are responsible for assessing the violator' s intent, and their activation degree and functional connectivity strength correlate negatively with punishment severity (Baumgartner et al., 2012, 2014; Moll et al., 2018; Zinchenko et al., 2019). This may reflect individuals' reasonable speculation and explanation of norm violations. When harm occurs unintentionally, the TPJ-$mPFC$ loop inhibits amygdala activity, reducing punishment severity (Treadway et al., 2014). This demonstrates the important role of evaluating violators' mental states in TPP behavior implementation.

Previous research has found a unique connectivity pattern among brain regions related to TPP, with $dmPFC$ serving as a hub in TPP activation patterns (Bellucci et al., 2017; Feng et al., 2016). A hub refers to a brain region with the maximum number of causal connections to other nodes in Granger causality analysis, representing a central node for information exchange (Yang et al., 2023). Here, it reflects that $dmPFC$ has more functional connections with other brain regions. Combining the functions of different brain regions, we speculate on the operation mode of the default mode network as follows (Figure 2 [Figure 2: see original paper]): The temporal pole (TP) is responsible for understanding norm violations and provides harm information to $dmPFC$. After receiving harm information, $dmPFC$ assesses harm intent and transmits information to other regions, including the posterior cingulate cortex (PCC), $vmPFC$, and TPJ. Among them, PCC integrates background information related to norm violations, $vmPFC$ encodes harm degree, TPJ infers intent, and finally $mPFC$ integrates harm and intent information to form a "punishment signal" (Xie & Su, 2019).

Figure 2. The functional pathway of third-party punishment behavior in the default mode network. Note: temporal pole (TP), dorsomedial prefrontal cortex ($dmPFC$), posterior cingulate cortex (PCC), ventromedial prefrontal cortex ($vmPFC$), temporoparietal junction (TPJ), medial prefrontal cortex ($mPFC$). Arrows represent information transmission direction.

Notably, both $dmPFC$ and TPJ play roles in intent inference and are key brain regions for cognitive mentalizing (Feng et al., 2022). Additionally, $vmPFC$ and amygdala are related to affective mentalizing (Anne et al., 2012), leading some to propose that these four brain regions jointly constitute a mentalizing network (Feng et al., 2016; Glass et al., 2016). The mentalizing network can infer others' mental states, reduce monitoring of norm violations, and decrease the value calculation level of social norms. In fact, Bellucci et al. (2017) consider the default mode network and mentalizing network to essentially belong to the same network, while other researchers believe they should have distinct roles in TPP, jointly influencing TPP behavior (Lo Gerfo et al., 2019). Although there are inconsistent views on their relationship, it is undeniable that the two networks share similarities in both structure and function. The default mode network focuses more on overall responsibility assessment of norm violations, while the mentalizing network emphasizes inferring others' mental states. In other words, the default mode network may achieve responsibility assessment through the mentalizing network, with the two complementing and interacting with each other.

3.3 Central Executive Network

Buckholtz et al. (2015) proposed that responsibility assessment and punishment selection belong to two independent cognitive mechanisms, with the latter corresponding to the central executive network. The central executive network in TPP is responsible for transforming the "punishment signal" from the default mode network into "punishment behavior" (Lo Gerfo et al., 2019; Zinchenko & Klucharev, 2017). Relevant brain regions include the dorsomedial prefrontal cortex ($dlPFC$), posterior parietal cortex (PPC), and intraparietal sulcus (IPS).

When the central executive network receives the "punishment signal" from the default mode network, IPS is responsible for linking the scenario with punishment behavior, PPC constructs a punishment type hierarchy, and finally $dlPFC$ selects specific punishment from this hierarchy, representing the final output (Buckholtz et al., 2008; Glass et al., 2016; Krueger & Hoffman, 2016). Numerous studies have indicated that $dlPFC$ plays an important role in cognitive control—that is, inhibiting negative emotions and self-interest tendencies (Knoch et al., 2006, 2008; Sanfey et al., 2003; Luo et al., 2013), a finding confirmed in TPP as well (Feng et al., 2016; Yin et al., 2019). Related ERP studies have found that late component LPC amplitude correlates with cognitive effort (Cui et al., 2018; Johnson et al., 2008), indicating that final punishment decisions result from a trade-off between negative emotions and self-interest mechanisms. Additionally, $dlPFC$ is related to the integration and selection of goal-directed behaviors (Luo et al., 2013), and inhibiting $dlPFC$ activity using repetitive transcranial magnetic stimulation (rTMS) interferes with final punishment decisions (Buckholtz et al., 2015). From this, we speculate that $dlPFC$, as a core brain region of the central executive network, reflects the important role of cognitive control in the later stages of response judgment and provides neural-level evidence for the existence of a control system in the dual-system model.

Furthermore, another important brain region in the central executive network is the ventrolateral prefrontal cortex ($vlPFC$). Glass et al. (2016) noted that although $vlPFC$ is important for successful TPP implementation, it is not intrinsically linked to it. However, a meta-analysis by Bellucci et al. (2020) found that TPP consistently activates $vlPFC$. Unfortunately, no studies have directly examined the relationship between $vlPFC$ and TPP behavior. Previous research has confirmed that $vlPFC$ is a key brain region in prosocial behavior regulation (Yang, Zheng et al., 2019). We therefore speculate that $vlPFC$ may play an important role in inhibiting punishment impulses aimed at harm and promoting more equitable sanctions. Future research should further investigate the function of $vlPFC$ in TPP by targeting this region.

3.4 Reward Network

TPP behavior generation involves reward processing, particularly internal rewards. According to RL theory, the satisfaction and power experience individuals obtain by punishing violators serve as internal rewards that prompt the next punishment behavior, representing the proximal mechanism for the evolution of TPP (Zhang et al., 2013). Brüne et al. (2021) found that Huntington' s disease patients could understand unfair behavior itself but might exhibit reduced TPP behavior due to impaired reward experience ability, demonstrating the important role of reward processing in TPP. At the neural level, many studies have found that TPP activates brain regions related to reward processing, mainly including the ventral striatum (VS) and $vmPFC$ (Hu et al., 2015), both connected to the ventral tegmental area (VTA), forming the mesolimbic and mesocortical pathways, respectively (Ikemoto, 2010).

In the mesolimbic pathway, VS plays an important role in expecting future rewards (O' Doherty et al., 2004), and its activation degree and functional connectivity with $vmPFC$ increase during TPP (Hu et al., 2015). In the mesocortical pathway, $vmPFC$ is a key brain region for affective and cognitive processing (Naqvi et al., 2006; Liu et al., 2022). The previously mentioned $vmPFC$ assessment of harm degree reflects its function of integrating emotion into TPP (Asp et al., 2019). In cognitive processing, $vmPFC$ is responsible for subjective value assessment in decision-making (Ruff & Fehr, 2014), and its activation positively reinforces social rewards (Zhong et al., 2016). This evidence suggests that the reward network in TPP is primarily responsible for value representation and reward expectation.

Furthermore, VTA is rich in the neurotransmitter dopamine, which plays an important role in neural transmission in both the mesolimbic and mesocortical pathways (Wise & Rompre, 1989). Specifically, dopamine is responsible for encoding reward prediction errors (Schultz, 2007, 2013). When the brain monitors outcomes worse than expected, this signal is encoded by ACC and the dopamine system, which may explain why MFN appears under expectancy violation conditions in EEG studies (Enge et al., 2017; Wu & Luo, 2011), accounting for the phenomenon of increased MFN amplitude when individuals observe unfair allocations (Ouyang et al., 2021). Numerous studies have found that brain regions such as $dlPFC$ and $dACC$ in TPP are modulated by input signals from midbrain dopaminergic neurons (Lockwood et al., 2016). Genetic variation in dopamine functional molecules can explain individual differences in neural activation during TPP. The COMT Met allele is related to the ability to experience rewards, and carriers of this gene show stronger VS activation during TPP (Strobel et al., 2011). This demonstrates that dopamine levels have important influences on TPP decision-making, and this process is closely related to reward processing.

4 Cognitive Neural Network Model of Third-Party Punishment

Previous research has extensively investigated the neural mechanisms of TPP, but most have focused only on independent activations of various brain regions when TPP behavior occurs. In recent years, scholars have conducted meta-analyses on consistently activated neural networks in TPP (Bellucci et al., 2020), but understanding of interactions between brain regions and networks remains lacking. Moreover, the relationship between the reward network, as the region most closely related to motivation in TPP, and other brain networks is still unclear. As a complex social decision-making behavior, TPP activates extensive and complex brain regions. Therefore, it is necessary to integrate psychological and neural-level evidence to form a more systematic understanding. To explain the mechanisms of TPP behavior from a more comprehensive perspective, this paper integrates previous research findings, summarizes individuals' emotional arousal, cognitive processes, and brain activation patterns during TPP (detailed literature information see Appendix Table S1), and proposes a cognitive neural network model of TPP (Figure 3). In this model, the emotional and reward systems jointly serve as the motivational system for TPP, providing power for TPP behavior, with corresponding brain networks being the salient network and reward network. The social cognition system and executive control system, as two subsystems of the cognitive system, participate in the "responsibility assessment" and "punishment selection" stages, with corresponding brain networks being the default mode network and central executive network. Various components in the model cooperate and interact with each other, with the executive control system finally making decisions about whether to punish and punishment intensity. The resulting feedback information further acts on the internal loop, and individuals learn from each feedback to form optimal behavioral decisions. Understanding this model includes four aspects: (1) the emotional and reward systems under the motivational system; (2) internal relationships within the cognitive system and their influence on TPP behavior; (3) automatic and controlled processing pathways; and (4) the feedback pathway involving the reward network.

Figure 3 [Figure 3: see original paper]. Cognitive neural network model of third-party punishment behavior. Note: Dashed rectangles represent cognitive processing systems, solid rectangles represent subsystems under each system, and ellipses represent behavioral decisions. Unidirectional solid arrows (1, 2, 3, 4) represent a complete punishment process, with larger numbers indicating later stages: Arrow 1 represents stimulus input, Arrow 2 represents integration and transmission of motivational system information, Arrow 3 represents the social cognition system transmitting the integrated "punishment signal" to the executive control system, and Arrow 4 represents punishment decision output. Bidirectional solid arrows (5, 6) represent relationships between systems or networks: Arrow 5 represents the relationship between the emotional and reward systems, and Arrow 6 represents the "antagonistic" and "complementary" relationship between the default mode network and central executive network. Dashed arrow (7) represents the feedback process after punishment implementation, which participates in the next TPP behavior.

4.1 Emotional and Reward Systems Under the Motivational System

The emotion model posits that negative emotion generation is one source of motivation for TPP (Fehr & Gächter, 2002; Xiao & Houser, 2005), while the reciprocity model suggests that anticipation of future rewards (such as reputation establishment) and the power experience and satisfaction after punishment can serve as reward signals for internal reinforcement, prompting the next punishment behavior. Therefore, both emotional and reward systems participate in the motivational generation process of TPP and jointly constitute the motivational system of TPP. The emotional system involves emotion generation, processing, and regulation (Etkin et al., 2015; Pessoa, 2017). Previous research has found that the salient network participates in emotional processing and guides subsequent decisions in TPP (Bellucci et al., 2020; Feng et al., 2016; Mclatchie et al., 2016), while the reward network' s functions of value representation and reward expectation make it closely related to the reward system (Hu et al., 2015). Therefore, the emotional and reward systems correspond to the salient network and reward network, respectively. Notably, the emotional and reward systems are not completely independent (Arrow 5). The emotional system transmits emotional information such as negative emotion generation to the reward system, making individuals more inclined to seek reward stimuli to alleviate discomfort caused by negative emotions. Simultaneously, the reward system transmits previous reward experiences (such as power and satisfaction) to the emotional system, thereby triggering positive emotional experiences. This mutually reinforcing relationship provides motivation for individuals to make punishment decisions.

4.2 Internal Relationships Within the Cognitive System and Their Influence on Third-Party Punishment Behavior

The cognitive system includes the social cognition system and executive control system. In TPP, individuals need to assess harm degree and infer violators' mental states based on violation information (Ginther et al., 2016; Treadway et al., 2014), while brain regions in the default mode network related to mentalizing, such as TPJ and $dmPFC$, play important roles in intent inference (Feng et al., 2022), and activation of $dlPFC$, the core brain region of the central executive network, reflects the importance of cognitive control in TPP (Yin et al., 2019). Therefore, the social cognition system and executive control system correspond to the default mode network and central executive network, respectively, participating in the "responsibility assessment" and "punishment selection" stages. Some researchers have noted that unlike second-party punishment, which is more emotion-driven, TPP emphasizes more on intent inference of violators, whose importance may exceed that of emotion (Ginther et al., 2016; Treadway et al., 2014), reflecting the important role of the social cognition system in the TPP process. Additionally, whether individuals finally punish and the intensity of punishment are influenced by the executive control system. Therefore, the roles of the social cognition system and executive control system in TPP behavior are progressive rather than independent. At the neural level, Buckholtz et al. (2008) first found an "antagonistic" relationship between TPJ and $dlPFC$ activity: when TPJ activity increased, $dlPFC$ showed deactivation, but when individuals decided to punish, its activity increased, reflecting the biphasic neural activity of $dlPFC$. Additionally, Zinchenko et al. (2021) found that overall resting-state connectivity between right $dlPFC$ and right TPJ negatively correlated with TPP levels. Both task-state and resting-state studies indicate an "antagonistic" relationship between the default mode network (TPJ) and central executive network ($dlPFC$) (Arrow 6). More specifically, this biphasic neural activity of $dlPFC$ may underlie the central executive network' s inhibitory control over the default mode network (Zinchenko & Klucharev, 2017). Meanwhile, the two networks also exhibit complementarity (Chen et al., 2020): the default mode network represents "intuitive" social cognition, reflecting responsibility assessment dominated by cognition and emotion, while the central executive network typically represents "rational" logical thinking, with its role in cognitive control reflecting rational thinking and weighing in the brain. The combination of "intuition" and "rationality" demonstrates the functional complementarity between the default mode network and central executive network.

4.3 Automatic and Controlled Processing Pathways

Combining the dual-system model mentioned above, this model includes two pathways: automatic processing and controlled processing (hereinafter referred to as Pathway 1 and Pathway 2). Both pathways start with the emotional system and output final TPP behavior through the executive control system. The difference is that after observing norm violations (Arrow 1), in Pathway 1, the emotional system directly transmits integrated negative emotions to the cognitive system (Arrow 2), and the executive control system executes the punishment decision. At this point, individuals may punish to resolve negative emotions (Arrow 4). In Pathway 2, factors of rational control are added: information from the emotional system first undergoes harm degree and intent assessment by the social cognition system, then brain regions in the executive control system weigh emotions and self-interest (Arrow 3), and finally execute the punishment decision (Arrow 4), corresponding to the three stages of "emotion generation," "responsibility assessment," and "punishment selection" in TPP.

4.4 Feedback Pathway Involving the Reward Network

During TPP decision-making, individuals need to constantly collect and evaluate environmental information to adjust behavioral strategies, which depends on feedback mechanisms. Specifically, after third parties make punishment decisions (Arrow 4), the resolution of negative emotions and the experience of power and satisfaction serve as corresponding reward signals that positively feedback to the reward system (Arrow 7), providing motivation for the next punishment. According to RL theory, individuals learn from reward information provided by feedback, comprehensively consider various factors to estimate maximum reward, and eventually form stable punishment patterns. From this, we can see that the existence of feedback mechanisms enables the smooth progression of TPP behavior emergence and development. When feedback mechanisms are damaged, individuals may exhibit abnormal behavior, demonstrating that feedback mechanisms play a crucial role in the cognitive neural network model of TPP.

5 Summary and Future Directions

This paper reviews TPP-related research from the past decade, summarizing the cognitive neural mechanisms of TPP behavior from theoretical and neural network perspectives, and proposes a cognitive neural network model of TPP based on previous studies. This model presents the emergence and development process and neural correlates of TPP behavior from a more macroscopic perspective. In this model, the emotional and reward systems are the motivational sources of TPP, the cognitive system is primarily responsible for responsibility assessment and punishment selection, and the reward network, salient network, default mode network, and central executive network participate in different cognitive processing stages. Meanwhile, the existence of TPP behavior feedback mechanisms allows individuals to optimize behavioral strategies through each feedback learning experience, eventually forming stable behavioral patterns. However, current TPP research still has some limitations and unresolved issues.

5.1 Effects of Neurotransmitters and Hormones on Third-Party Punishment Behavior

Cognitive neural systems are influenced by neurotransmitters and hormones (Hu, Scheele et al., 2016). Investigating the relationships among TPP behavior, neurotransmitters, and hormones can help us understand TPP behavior from a more microscopic perspective. This paper only briefly analyzed the genetic mechanisms behind individual punishment behavior from the perspective of dopamine levels and their genetic variation, providing insights into revealing the neurobiochemical basis of TPP behavior. In addition to dopamine, some studies have found that individuals with low serotonin levels exhibit more altruistic punishment behavior (Emanuele et al., 2008), and 5-HTTLPR L/L gene carriers show greater TPP intensity (Enge et al., 2017). However, Enge et al.' s study did not exclude the interaction effects between dopamine and serotonin, and the mechanism of how serotonin influences TPP behavior remains unclear. At the hormonal level, peptide hormone oxytocin (OXT) has been proven to relate to human prosocial behavior (Marsh et al., 2021). However, few studies have examined the relationship between exogenous OXT and TPP behavior. Krueger et al. (2012) found that OXT only affected harm degree perception but not

TPP intensity. Overall, research on how neurotransmitters, hormones, and their genetic polymorphisms influence TPP behavior is still in its preliminary stages. Future research from this perspective will help reveal the neurophysiological mechanisms of TPP behavior from a more microscopic angle.

5.2 Individual Differences in Third-Party Punishment Behavior

Previous studies have mostly analyzed samples as a whole, ignoring differences between individuals. For example, empathy, as the ability to understand and feel others' emotions, is one source of motivation for prosocial behavior and has important influences on TPP behavior (Ferguson et al., 2019). Individuals show higher empathy levels when facing poorer violators, are more tolerant of them, and therefore impose lower punishment intensity (Ouyang et al., 2021). However, empathy also shows huge individual differences, which may further manifest in neurophysiological responses (van Baar et al., 2021; Wang et al., 2022). Moreover, because traditional ERP analysis performs group averaging on all participant data, the final results may not reflect individual differences, leading to misinterpretations (Chen et al., 2023). Therefore, future research should incorporate individual differences such as empathy, social value orientation, and mentalizing ability into the study of cognitive neural mechanisms of TPP behavior to further explore the relationship between neurophysiological representations of TPP behavior and personality trait variables.

5.3 Quantitative Explanation and Prediction of Third-Party Punishment Behavior

The cognitive neural network model of TPP proposed in this paper can predict TPP behavior to some extent but remains at a qualitative level. Whether actual TPP behavior follows this pattern requires further verification. In recent years, increasing researchers have applied machine learning methods to cognitive processing fields. Taking multivariate decoding analysis as an example, this method can quantify information represented by single neural signals and be used to train and test classifiers to distinguish brain activity signals corresponding to specific events (Contini et al., 2017). These neural signals can come from neuroimaging signals such as functional magnetic resonance imaging (fMRI) or neurophysiological signals such as electroencephalogram (EEG) and magnetoencephalogram (MEG). Considering that TPP is a complex social behavior composed of multiple sub-stages, this method can further explore complex neural activity patterns related to TPP behavior. Therefore, future research needs to further optimize and develop relevant models, introducing machine learning-based analytical methods such as multivariate decoding analysis to construct mathematical models. This would allow investigation of neural signal differences across various situations, such as different punishment costs and degrees of unfair allocation, to provide quantitative explanations and predictions of TPP behavior.

5.4 Similarities and Differences in Cognitive Neural Mechanisms Between Third-Party Punishment and Compensation Behavior

From a reputation establishment perspective, when third-party individuals punish violators, it demonstrates their cooperative intentions. However, when individuals have other more positive ways to express their cooperative tendencies, reputation established through punishment may be harmful (Raihani & Bshary, 2015a). Specifically, when third-party individuals have options beyond "punishment" and "inaction," such as "helping the victim," the altruistic intensity of TPP weakens because punishment is then viewed as a competitive, "harmful to others without benefiting oneself" signal (Jordan, Hoffman et al., 2016), while helping receives more positive evaluation (Raihani & Bshary, 2015b; Li et al., 2018). In such complex social situations, introducing "third-party intervention" becomes particularly important, including both third-party punishment and third-party compensation (helping). Unlike punishment, which focuses on violators, third-party compensation focuses on victims—that is, third-party individuals sacrifice their own interests to compensate victims, truly benefiting them. Existing research shows that individuals with higher empathy levels are more likely to make helping decisions, and the neural mechanisms behind helping and punishment show both similarities and differences (Hu et al., 2015; Xie et al., 2022; Su et al., 2019). Because these two intervention forms focus on different targets, future research can examine individual third-party intervention preferences and their underlying neural mechanisms across different social contexts.

5.5 Using Meta-Analysis to Provide Statistical Support for the Cognitive Neural Network Model

The cognitive neural network model proposed in this paper can provide a more holistic and comprehensive theoretical explanation for the emergence and development of TPP behavior, but it still lacks statistical support. Meta-analysis introduces quantitative analytical methods on the basis of qualitative analysis, making conclusions more generalizable and universal. On one hand, TPP generation includes different stages, each involving different brain regions and network activation patterns. Meta-analysis can process and analyze large amounts of individual study data for effective synthesis (Dong, 2004). On the other hand, previous studies exploring the neuropsychological mechanisms of TPP have used different experimental paradigms and dependent variable measurement methods, and these methodological variations may lead to inconsistent results. Meta-analysis can further explore the sources of heterogeneity in characteristic variables across studies and better address this issue. Therefore, in future model refinement, meta-analysis can be used to comprehensively analyze and re-evaluate existing research results, providing quantitative data support for the model to increase its reliability.

References

Chen, X., Li, H., & Ding, Y. (2023). Exploring neural representation patterns in event-related EEG/MEG signals: Methods based on classification decoding and representational similarity analysis. *Advances in Psychological Science*, 31(2), 173–195.

Chen, Y., Xu, M., & Wang, X. (2020). A cognitive neural network model of trust. *Advances in Psychological Science*, 28(5), 800–809.

Dong, Q. (2004). *Psychological and Educational Research Methods* (Revised Edition). Beijing: Beijing Normal University Press.

Liu, Y., Duan, Y., Liu, H., Liu, J., & Wang, H. (2022). Neural mechanisms underlying differences in third-party punishment decision-making between gain and loss contexts: A study based on rTMS. *Psychological Science*, 45(4), 942–952.

Luo, Y., Feng, C., Gu, R., Wu, T., & Luo, Y. (2013). Fairness norms and their neural mechanisms in social decision-making. *Advances in Psychological Science*, 21(02), 300–308.

Su, Y., Xie, D., & Wang, X. (2019). The role of cognitive control in third-party punishment. *Advances in Psychological Science*, 27(8), 1331–1343.

Tang, J., Huang, X., Wu, S., & Cui, F. (2022). With great wealth comes great responsibility: The influence of fund amount and source on third-party punishment in public goods dilemmas. *Psychological Science*, 45(3), 665–671.

Wu, Y., & Luo, Y. (2011). Outcome evaluation in altruistic punishment: An ERP study. *Acta Psychologica Sinica*, 43(6), 661–673.

Xie, D., & Su, Y. (2019). The evolution and cognitive mechanisms of third-party punishment. *Psychological Science*, 42(1), 216–222.

Yang, S., & Chen, S. (2022). Normative illusion in third-party punishment: An explanation based on belief in a just world. *Acta Psychologica Sinica*, 54(3), 281–299.

Yin, X., Li, J., Chen, S., Liu, X., & Hao, J. (2019). Neural mechanisms of third-party punishment: Evidence from transcranial direct current stimulation. *Acta Psychologica Sinica*, 51(5), 571–583.

Zhang, H., Ma, H., Xu, F., Liu, Y., & Shi, Y. (2018). Fairness preferences in the ultimatum game: A perspective based on dual-system theory. *Advances in Psychological Science*, 26(2), 319–330.

Zhang, Y., Lin, Z., & Zhu, L. (2013). Human altruistic punishment: A cognitive neuroscience perspective. *Progress in Biochemistry and Biophysics*, 40(9), 796–803.

Anne, L., Frank, K., Olga, D. M., Matteo, P., Pulaski, S. J., Jeffrey, S., &

Jordan, G. (2012). Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. *Social Cognitive and Affective Neuroscience*, 7(8), 871-880.

Asp, E. W., Gullickson, J. T., Warner, K. A., Koscik, T. R., Denburg, N. L., & Tranel, D. (2019). Soft on crime: Patients with ventromedial prefrontal cortex damage allocate reduced third-party punishment to violent criminals. *Cortex*, 119, 33-45.

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), 655-663.

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.

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.

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group affiliation and altruistic norm enforcement. *American Economic Review*, 96(2), 217-221.

Bright, D. A., & Goodman-Delahunty, J. (2006). Gruesome evidence and emotion: Anger, blame, and jury decision-making. *Law and Human Behavior*, 30(2), 183-202.

Brüne, M., von Hein, S. M., Claassen, C., Hoffmann, R., & Saft, C. (2021). Altered third-party punishment in Huntington' s disease: A study using neuroeconomic games. *Brain and Behavior*, 11(1), Article e01908.

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.

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., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From blame to punishment: Disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron*, 87(6), 1369-1380.

Cheng, X., Zheng, L., Liu, Z., Ling, X., Wang, X., Ouyang, H., Chen, X., Huang, D., & Guo, X. (2022). Punishment cost affects third-parties' behavioral and neural responses to unfairness. *International Journal of Psychophysiology*, 177, 27-33.

Chung, J. C. Y., Bhatoa, R. S., Kirkpatrick, R., & Woodcock, K. A. (2023). The role of emotion regulation and choice repetition bias in the ultimatum game. *Emotion*, 23(4), 925-936.

Ciaramidaro, A., Toppi, J., Casper, C., Freitag, C. M., Siniatchkin, M., & Astolfi, L. (2018). Multiple-brain connectivity during third party punishment: An EEG hyperscanning study. *Scientific Reports*, 8(1), Article 6822.

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

Civai, C., Crescentini, C., Rustichini, A., & Rumiati, R. I. (2012). Equality versus self-interest in the brain: Differential roles of anterior insula and medial prefrontal cortex. *Neuroimage*, 62(1), 102-112.

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.

Contini, E . W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*, 105, 165-176.

Craig, A. D. B. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10(1), 59-70.

Cui, F., Wang, C., Cao, Q., & Jiao, C. (2019). Social hierarchies in third-party punishment: A behavioral and ERP study. *Biological Psychology*, 146, Article 107722.

Cui, F., Wu, S., Wu, H., Wang, C., Jiao, C., & Luo, Y. (2018). Altruistic and self-serving goals modulate behavioral and neural responses in deception. *Social Cognitive and Affective Neuroscience*, 13(1), 63-71.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, 305(5688), 1254-1258.

Delgado, M. R., Locke, H. M., Stenger, V. A., & Fiez, J. A. (2003). Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. *Cognitive, Affective, & Behavioral Neuroscience*, 3(1), 27-38.

Emanuele, E., Brondino, N., Bertona, M., Re, S., & Geroldi, D. (2008). Relationship between platelet serotonin content and rejections of unfair offers in the ultimatum game. *Neuroscience Letters*, 437(2), 158-161.

Enge, S., Mothes, H., Fleischhauer, M., Reif, A., & Strobel, A. (2017). Genetic variation of dopamine and serotonin function modulates the feedback-related negativity during altruistic punishment. *Scientific Reports*, 7(1), Article 2996.

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

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785–791.

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

Fehr, E., & Fischbacher, U. (2004b). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4), 185–190.

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

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7), 458–468.

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.

Ferguson, E., Quigley, E., Powell, G., Stewart, L., Harrison, F., & Tallentire, H. (2019). To help or punish in the face of unfairness: Men and women prefer mutually-beneficial strategies over punishment in a sexual selection context. *Royal Society Open Science*, 6(9), Article 181441.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, 143(1), 182–194.

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.

Glass, L., Moody, L., Grafman, J., & Krueger, F. (2016). Neural signatures of third-party punishment: Evidence from penetrating traumatic brain injury. *Social Cognitive and Affective Neuroscience*, 11(2), 253–262.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35(1), 1–15.

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., Scheele, D., Becker, B., Voos, G., David, B., Hurlemann, R., & Weber, B. (2016). The effect of oxytocin on third-party altruistic decisions in unfair situations: An fMRI study. *Scientific Reports*, 6, Article 20236.

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.

Ikemoto, S. (2010). Brain reward circuitry beyond the mesolimbic dopamine system: A neurobiological theory. *Neuroscience and Biobehavioral Reviews*, 35(2), 129–150.

Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others' beliefs in humans. *Nature*, 591(7851), 610–614.

Johnson, R., Henkell, H., Simon, E., & Zhu, J. (2008). The self in conflict: The role of executive processes during truthful and deceptive responses about attitudes. *NeuroImage*, 39(1), 469–482.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, 530(7591), 473–476.

Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 19(4), 741–763.

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.

Kim, M., Decety, J., Wu, L., Baek, S., & Sankey, D. (2021). Neural computations in children' s third-party interventions are modulated by their parents' moral values. *NPJ Science of Learning*, 6(1), Article 38.

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., & Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. *Cerebral Cortex*, 18(9), 1987-1990.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), 829–832.

Konishi, N., & Ohtsubo, Y. (2015). Does dishonesty really invite third-party punishment? Results of a more stringent test. *Biology Letters*, 11(5), Article 20150172.

Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G. K., & Leibo, J. Z. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3), Article e2106028118.

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7(3), 77-124.

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

Krueger, F., Parasuraman, R., Coughlin, L., Twieg, P., de Visser, E., Mccabe, K., O' Hara, M., & Lee, M. (2012). Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. *Social Cognitive and Affective Neuroscience*, 8(5), 494-498.

Lee, S. W., Shimojo, S., & O' Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3), 687-699.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., & Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19), 6916-6921.

Li, J., Li, S., Wang, P., Liu, X., Zhu, C., Niu, X., Wang, G., & Yin, X. (2018). Fourth-party evaluation of third-party pro-social help and punishment: An ERP study. *Frontiers in Psychology*, 9, Article 932.

Liu, Y., Bian, X., Hu, Y., Chen, Y.-T., Li, X., & Di Fabrizio, B. (2018). Intergroup bias influences third-party punishment and compensation: In-group relationships attenuate altruistic punishment. *Social Behavior and Personality: An International Journal*, 46(8), 1397-1408.

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.

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America*, 113(35), 9763-9768.

Marsh, N., Marsh, A. A., Lee, M. R., & Hurlemann, R. (2021). Oxytocin and the neurobiology of prosocial behavior. *The Neuroscientist*, 27(6), 604-619.

Martin, J. W., Martin, S., & McAuliffe, K. (2021). Third-party punishment promotes fairness in children. *Developmental psychology*, 57(6), 927-939.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, 134, 1-10.

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.

Meidenbauer, K. L., Cowell, J. M., & Decety, J. (2018). Children' s neural processing of moral scenarios provides insight into the formation and reduction of in-group biases. *Developmental Science*, 21(6), Article e12676.

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.

Moll, J., de Oliveira-Souza, R., Basilio, R., Bramati, I. E., Gordon, B., Rodríguez-Nieto, G., Zahn, R., Krueger, F., & Grafman, J. (2018). Altruistic decisions following penetrating traumatic brain injury. *Brain*, 141(5), 1558-1570.

Morris, A., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences*, 114(39), 10396-10401.

Morese, R., Rabellino, D., Sambataro, F., Perussia, F., Valentini, M. C., Bara,

Note: Figure translations are in progress. See original paper for figures.