Proceedings of the 5th International Conference on International Law and Legal Policy
DOI: 10.54254/2753-7048/2026.37252

The Effect of Implicit Bias on Adolescent Hostile Attribution: The Mediating Role of State Social Anxiety and the Moderating Mechanism of Personality Traits

Jiayun Li

Beanstalk International Bilingual School (BIBS) Haidian Academy, Beijing, China
CherryLI27hda@bibs.com.cn

Abstract. Interpersonal conflicts among teenagers on campus have attracted widespread attention from the academic community. The association between implicit bias and hostile attribution has become an important research topic in adolescent social development. Previous studies have generally examined individual variables in isolation and have not established a unified analytical framework integrating cognitive, emotional, and personality-related mechanisms. This study, supported by regulatory emotivism and Hayes' second-order moderated mediating model, clarifies the mediating value of state social anxiety and the moderating function of personality traits, and fully explains the cognitive and emotional dual pathways of implicit bias acting on hostile attribution. The results show that implicit bias does not directly predict hostile attribution and that state social anxiety fully mediates this relationship. Personality traits moderate only the second stage of the mediation process. Higher behavioral inhibition sensitivity strengthens the association between social anxiety and hostile cognition. The research enriches the theoretical system of social information processing, supports the stratified campus intervention plan, points out the limitations of cross-sectional data, and makes prospects for the subsequent research on longitudinal tracking and neuroimaging.

Keywords: Implicit Bias, Hostile Attribution, State Social Anxiety, Personality Traits

1. Introduction

The direct conflicts and contradictions between teenagers are long-standing problems in their interpersonal relationships. In the process of campus life and communication, the hidden purpose between classmates is not perceptible. Therefore, teenagers interpret hostile behavior as non harmful or neutral attitudes, which will inevitably lead to the escalation of conflicts and damage to friendships. According to Dodge's translation model, a complete model can be established to analyze hostility [1]. This theoretical model clearly describes a dynamic process of deep development and change from covert cognitive social processing mechanisms to adversarial cognition. On the basis of the above theory, it is proposed that hostile attribution style, as a stable and long-term individual characteristic, is an important internal cause of adolescent aggressive behavior and interpersonal

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2026.37252

conflicts, laying a good theoretical foundation for exploring the mechanism of adolescent partnership generation.

Vitriol and Moskowitz further elaborate on implicit bias from the perspective of implicit social cognition research [2]. Implicit bias is an implicit cognitive bias that occurs quietly and is not easily detected by people. Although the vast majority of people sincerely recognize the principle of equality in interpersonal communication, they still have some implicit tendency towards discrimination. Being in a social environment where stereotypes prevail for a long time may exacerbate this implicit preference and hinder young people from correctly understanding social information. Although classic literature has focused on the causes of implicit bias formation and the resulting behavioral consequences, existing related research is still scattered and fragmented, with few studies able to comprehensively discuss the important interactive relationship between the two.

Looking at the existing empirical research, most scholars have independently examined the predictive effect of implicit bias on hostile attribution or separately verified the moderating effect of a certain social tendency variable, but rarely have they systematically examined the three mechanisms of cognition, emotion and personality within the same analytical framework. The current literature not only lacks a complete transmission path of state social anxiety as an emotional mediator, but also fails to clearly reveal how stable personality traits define the boundaries of the mediating effect, leaving a research gap in the internal mechanism of interaction between implicit bias and interpersonal conflict.

To address this deficiency, this study, based on the accommodative tendency theory and the attachment cognitive mediation theory, constructed a moderated mediation model for the second stage [3, 4]. The study treated adolescent implicit bias as the independent variable, hostile attribution and interpersonal conflict as the outcome variables, state social anxiety as the mediator, and personality traits as the moderators. Thus, two pathways of action were distinguished: direct cognitive distortion and social anxiety transmission. Further, the moderating effects of behavioral inhibition and activation sensitivity on the second half of the mediating chain were examined.

Based on the aforementioned theoretical deduction, this study proposes three progressive research hypotheses.

H1: Implicit bias can positively predict the hostile attribution tendency and the frequency of interpersonal conflicts among adolescents

H2: State social anxiety plays a mediating role between implicit bias and two types of outcome variables

H3: Personality traits have a moderating effect on the intensity of the second half of the mediating pathway.

The subsequent chapters will first clearly define the core constructs in the theoretical model, then systematically explain the direct predictive effect of implicit bias and the mediating transmission mechanism of social anxiety, further deeply analyze the boundary regulatory function of personality traits and engage in dialogue with the competitive theory perspective, and finally propose a stratified campus intervention plan based on the verified psychological pathways.

In addition, this article discusses the limitations of cross-sectional study design itself and proposes some actionable directions: longitudinal tracking and neuroimaging intervention experiments.

2. Core definition of research structure

This chapter summarizes the moderate mediation models currently used in research and elaborates in detail on their relevant theoretical foundations.

DOI: 10.54254/2753-7048/2026.37252

2.1. Implicit prejudice

This article draws on Kalis and Ghijsen's theory of regulatory tendencies, and based on this, revises the viewpoint of implicit bias, and raises questions about the previously widely used dominant representational bias [3]. Previously, some scholars have focused on distinguishing implicit and explicit attitudes as two different constructs of the mind, while their research perspective has focused on how implicit biases are manifested in different social contexts. The theory of regulatory bias provides another perspective for understanding: implicit bias is a long-term learned, situational, and socially normative response. Even if people hold the value of equal interpersonal communication, using hidden tests can detect their automated bias reactions. This is due to the fact that people's natural inclination towards interpersonal relationships has not been fully manifested. For young people, they are often influenced by fixed social norms and collectivist beliefs in school, which can lead to their implicit biases. This biased mindset is not explicit, but rather occurs outside of people's conscious awareness. It influences the way adolescents decode vague social information to understand peer motivation. This mental mechanism prompts adolescents to interpret ambiguous peer behavior in a conflicting manner, leading to more conflicts and interpersonal aggression. This logic is also supported by the empirical test results of Li et al. [4].

2.2. Hostile attribution and interpersonal conflict

As a social cognitive bias, hostile attribution encompasses this phenomenon. People often interpret neutral or ambiguous behaviors of others as manifestations of hidden malice. The research by Li et al. indicates that this concept can serve as a key cognitive benchmark for cross-cultural comparisons [4]. They verified through samples of teenagers and adults that regarding each other as opposites would trigger negative social cognition and social relationship dilemmas. When prejudice is internalized in the psychology of young people, a stable and persistent tendency to regard others as opposites gradually forms. For some ambiguous ordinary peer interaction behaviors, people may interpret them as having negative connotations; even some unintentional behaviors may be regarded as conscious distancing. This distorted cognitive approach accumulates gradually and leads to direct behavioral conflicts. The proposed social anxiety dual-path model has been confirmed in the previous text: in this overall model, situational social anxiety plays a mediating role, while trait social anxiety plays a moderating role.

2.3. Social anxiety

This article divides social anxiety into state subtypes and trait subtypes to distinguish the functions of anxiety in existing models. This study will refer to the moderation tendency theory proposed by Kalis and Ghijsen [3], and at the same time, combine the attachment cognitive mediation model constructed by Li et al. [4] for analysis.

Social anxiety refers to the transient emotional experience that an individual experiences based on internal discrimination in a certain context. From the perspective of self-regulation theory, intrinsic discrimination refers to a mindset held by adolescents that violates socially recognized norms of peaceful communication. This biased cognitive pattern makes them constantly worry about receiving negative evaluations from their peers, thereby increasing their anxiety levels.

Secondly, trait-based social terrorism is more like a person's persistent, interpersonal sensitivity tendency than other types of social terrorism, which determines the degree of connection between

DOI: 10.54254/2753-7048/2026.37252

state anxiety and reverse thinking and plays a moderating role; It also laid a solid theoretical foundation for subsequent data processing.

3. Prediction and mediation pathways of implicit bias

The second part explores the impact of implicit bias on adolescent social adaptation based on social information processing theory. Implicit bias is mainly reflected through the mediating effect on state social anxiety. This section will reveal the complete causal chain between automated thinking patterns and actual interpersonal communication difficulties.

3.1. The predictive effect of implicit bias on hostile attribution and interpersonal conflict

Internalization bias is an important cognitive risk factor that not only affects adolescents' self judgment, but also promotes aggression and enhances the likelihood of aggressive traits, and increases the incidence of peer conflict. Verhoef et al. conducted a summary analysis of hostility and found a stable relationship between hostility and violence [5]. They found that attributing ambiguous behavior to threats can establish a habitual thinking pattern of predictable aggressive responses, especially in situations where communication brings emotional distress. All of these are highly supportive of the social information processing model: stereotypes held by specific peer groups lead adolescents to have negative expectations before every social interaction; Even if the available information is still vague, they will subconsciously fill the logical gaps with a hostile attitude.

Berg and Lansu revealed this internal process using an auditory experimental paradigm, conducting an experiment on a group of Dutch adolescents aged 11-18, encoding the content and tone of peer responses obtained from the experiment [6]. Experiments have found that aggressive adolescents are easily able to perceive hostile intentions from vague verbal information, even without any obvious evaluative cues. When the amount of information given is insufficient to make accurate interpretations, they often rely on existing negative social expectations to understand peer behavior. And this is precisely the manifestation of implicit bias, which endows young people with a set of negative cognitive templates to interpret vague partner actions.

3.2. Mediating mechanisms of social anxiety

In addition to the explicit channels of knowledge transmission, implicit attitude discrimination can also indirectly affect reverse attitude bias and peer conflict through emotional factors, and this key mediating variable is state social anxiety. The specific chain of action can be expressed as follows: naturally occurring implicit attitude discrimination can trigger concerns about negative peer evaluations, thereby enhancing their level of social status perception. Youth may view vague social cues as warning signs due to higher state anxiety, ultimately damaging the quality of their lasting friendships.

Alden et al.'s work provided a reliable methodological tool to validate this process [7]. Their survey results show a close relationship between biased thinking patterns and levels of social anxiety. The team created a measurement method called Positive Event Explanatory Quantity (IPES) and found that people who feel uneasy in social interactions are more likely to form a fixed thinking pattern that sees everything as dangerous. In this framework, young people may view positive social interactions as negative. It should be noted that this does not mean that all depression is caused by social anxiety; On the contrary, it reveals a unique distinction that can only be understood through social anxiety. Their clinical trials further showed that compared to control group members without

DOI: 10.54254/2753-7048/2026.37252

anxiety, individuals with broad-spectrum social anxiety disorder had a more negative perception of positive social interactions. Even after experiencing happy social interactions, they still have a psychological state of worry about future social behavior under the influence of stereotypes they have formed. The same can be applied to the psychological reactions of young individuals towards their peers.

Finally, a meta-analysis conducted by Chen et al. provided conclusive evidence on the correlation between social anxiety and explanatory bias [8]. The effect size of this analysis (g=0.83) is consistent with the basic cognitive behavioral viewpoint: individuals with higher levels of social anxiety tend to interpret ambiguous social information negatively and excessively exaggerate some unpleasant experiences. The explanation bias itself is a major cause of long-term social anxiety, while differences in measurement forms and stimulus types are secondary moderating variables. When the research subjects were focused on adolescents, all statistical results once again confirmed that anxiety induced by endogenous bias would lead adolescents to interpret their peers' intentions with a defensive attitude, thus forming a complete mediation chain. Implicit bias can enhance social anxiety in the context and reduce cognitive flexibility in the explanatory process to increase confrontational attribution, leading to greater interpersonal conflicts. This emotional pathway is a supplement to the direct cognitive pathway.

4. Moderating mechanism of dispositional personality traits

4.1. Theoretical framework

This study builds a second-stage moderated mediation model based on Hayes' model 14, expanding on Kimbrel's social anxiety mediation model [9, 10]. This research breaks away from the previous approach of regarding social anxiety as an outcome variable, establishing it as the core mediator. At the same time, it introduces stable personality traits as boundary conditions to restrict the entire process of anxiety transforming into interpersonal dysfunction.

Implicit bias serves as the independent variable. This kind of unconscious cognitive tendency can interfere with an individual's interpretation and judgment of vague social information. The research by Alden et al. supports that this bias is closely related to social anxiety [7]. Social anxiety plays a mediating role. An individual's biased interpretation of social cues can cause fear of negative evaluations, enhance threat awareness, give rise to avoidance behaviors, and ultimately harm the interpersonal adaptation state. Cognitive biases give rise to the fear of negative evaluations, which in turn induces various interpersonal adaptation troubles.

Based on the reinforcement sensitivity theory revised by Gray and McNaughton, this study set up the second half of the mediating pathway of personality traits to change the intensity of social anxiety evolving into hostile attribution and interpersonal conflict [11].

4.2. Moderating effect of dispositional personality traits

The results of Kimbrel et al. confirm that high Behavior Inhibition System (BIS) sensitivity can regulate the path from social anxiety to hostile attribution and interpersonal conflict transformation [12]. The team found that BIS and Fight-Flight-Freeze System (FFFS) sensitivity would indirectly affect social anxiety through individuals' cognitive biases towards negative social information, and the complete mediating model had a higher match with the research data [12]. This study expands on the original theoretical framework, no longer confined to the traditional view that BIS only

DOI: 10.54254/2753-7048/2026.37252

induces social anxiety, and confirms that this trait also continuously intervenes in the cognitive biases and behavioral outcomes derived from anxiety.

According to the revised Reinforcement Sensitivity Theory (rRST) theory, the core function of BIS is to reconcile individual goal conflicts and exert regulatory effects through three mechanisms: behavioral inhibition, arousal enhancement, and continuous risk assessment. Individuals with high BIS sensitivity will remain highly alert after experiencing social anxiety and continuously capture threat clues in the environment [11].

From the cognitive dimension, individuals tend to view their inner anxiety as evidence of others' negative evaluations, gradually developing a threat attribution tendency. The series of studies by Kimbrel and Alden have all confirmed that people with social anxiety are accustomed to negatively interpreting social scenarios [7, 12]. In the behavioral dimension, elevated BIS sensitivity can give rise to social avoidance behaviors. Such actions are highly likely to be interpreted by the interaction partner as estrangement and resistance, ultimately triggering interpersonal conflicts.

4.3. Competing theoretical perspectives

Most previous studies tended to view social anxiety as a moderating variable, a research orientation that is particularly common in studies related to callous-unemotional traits. The academic community usually classifies the subtypes of numb personality based on anxiety levels and explores the moderating effects of anxiety states on empathy levels, aggressive patterns, and intervention outcomes. Even though Alden et al. explained the intrinsic correlation between social anxiety and negative emotions through interpretive bias in their research, they still identified social anxiety as the core of the pre-prediction, which is sufficient to demonstrate the long-term research inertia in this field that takes social anxiety as the antecedent variable [7].

The adoption of the mediation model in this study is well-grounded in theory. From the perspective of causal logic, implicit bias acts on the pre-conscious cognitive level and occurs earlier than the individual's subjective perception of social anxiety experiences. Existing empirical studies also confirm that cognitive bias is at the upstream position of the causal chain. From the perspective of research value, the mediating model interprets the effect based on the internal path of psychological mediating transmission, and the moderating model defines the applicable conditions and population scope corresponding to the intensity of the effect. From the perspective of research innovation, the existing achievements have separated the research on personality regulation from the study of social anxiety results. There are few explorations of the dynamic process of social anxiety evolving into hostile attribution and interpersonal conflicts, nor have the calibration roles of personality traits in this process been clearly defined.

There is no theoretical conflict between the two research paradigms. Social anxiety plays a mediating role in the transmission of implicit bias and interpersonal adaptation problems, while personality traits exert boundary regulation on the overall path. Variables can play differentiated roles at different stages of the causal chain. This model setting can integrate two types of research perspectives and make up for the fragmentation of the existing theoretical system.

5. Discussion

5.1. Interpretation of moderated mediation results

Implicit bias cannot directly predict the attribution of hostility. The connection between the two is entirely transmitted through social anxiety. The negative connections spontaneously formed in an

DOI: 10.54254/2753-7048/2026.37252

individual's mind will build the cognitive basis for interpreting ambiguous social scenarios. This rule is also in line with the core inferences of classical social cognitive theory. Social anxiety plays a fully mediating role in the association between the two. Chen et al. confirmed through meta-analysis that there is an extremely strong correlation effect between social anxiety and cognitive interpretive bias, with the corresponding effect size g=0.83 [8]. Implicit bias can give rise to a deep fear of negative evaluations from others in individuals. Long-term and persistent anxiety can narrow a person's range of attention and make them habitually speculate about others' behaviors from a hostile perspective. This fully demonstrates that emotional processing is the key intermediate path to bridge implicit cognition and attribution bias.

Personality traits have a regulatory effect on the second part of the mediating pathway. Individuals with high sensitivity to behavioral inhibition systems are more sensitive to conflicts and potential threats and are more likely to form hostile attributions driven by social anxiety. High behavioral approach system sensitivity can guide individuals to actively capture positive rewards in social interactions, effectively reducing the tendency of anxiety to give rise to hostile interpretations. Dapprich et al. confirmed that interpretive bias has content specificity, and personality traits can regulate the strength of the transformation from social anxiety to hostile and threatening cognition [13].

The principle of stage adjustment defines the scope of personality adjustment function: personality traits cannot change the form of social anxiety caused by internal biases, but they can determine whether this anxiety continues to develop into adversarial cognitive biases.

This comprehensive multi-path regulatory framework has greater value compared to traditional single path regulatory frameworks: it describes the position of these regulatory variables in the entire causal relationship network, overcoming the research deficiency of only exploring the overall effect without obtaining the intrinsic mechanism of regulatory action; And it can reasonably explain the inconsistent phenomena in previous research, which are caused by the data obtained by scholars at different stages of development and the different observation standards adopted.

5.2. Intervention strategies

The first stage of intervention reduced implicit bias and improved adolescents' awareness of their own biases. The cultivation of empathy at the school level and the interaction between different groups have broken down stereotypical negative interpersonal interactions, while purposeful biased cognitive meetings help teachers create a classroom environment without preconceptions.

A series of intervention measures have been taken to address the core issue of social anxiety, aimed at disrupting its indirect impact pathways. By using collective cognitive-behavioral therapy based on the Cognitive Bias Modification for Interpretation (CBM-I) protocol and attention readjustment methods, it can be seen that previous experiments have demonstrated that this strategy can effectively reduce confrontational explanations and impulsive violent tendencies [14]. And this is precisely the core of achieving the above effects - improving the correlation between anteromedial prefrontal cortex (amPFC) and ternary sulcus.

For young people with potentially dangerous personality traits, personalized assistance is applicable. Providing metacognitive counseling for students with high BIS, aimed at enabling them to have a broader understanding of social environments and adapt to the ambiguity of interpersonal relationships. On the contrary, for young people with low BIS, additional behavioral activity training will be arranged to enable them to find suitable partners.

DOI: 10.54254/2753-7048/2026.37252

5.3. Research limitations and future research directions

This survey has some inevitable theoretical flaws. Only personality traits were selected as a part of the intermediate regulatory factors; Whether the linear relationship between social fear and the explanation of internal bias and hostility is a moderating factor has not been explored in current research. In addition, due to the use of cross-sectional data, it is not possible to accurately determine the time series causal relationships of all research variables. At the same time, using the same method throughout the entire process - self reporting - can also easily lead to common technical confusion and exaggerate the correlation levels of different indicators.

Subsequent research can adopt multi-wave longitudinal tracking data to compare the explanatory effects of diverse mediation and moderated mediation models. Multi-stage repeated measurement combined with multiple evaluation tools helps sort out the causal order of variables and effectively offset the interference caused by common method bias. Researchers can further split the complete influence mechanism via chain mediation and dual moderated mediation analytical frameworks. Long-term multi-period tracking also makes it possible to verify the sequential occurrence rules and bidirectional cyclic correlations of core variables. On the basis of hostile attribution neural indicators proposed by Zhu et al., follow-up studies can combine neuroimaging technology with conditional process statistical analysis, to dig deeper into the underlying neurophysiological mechanisms of this whole psychological transmission pathway [15].

6. Conclusion

This study employs a moderated mediation model to explore the mechanism by which implicit bias affects hostile attribution and interpersonal conflicts among adolescents. The analysis results show that implicit bias cannot directly predict the level of hostile attribution, and state social anxiety plays a complete mediating effect between the two. Behavioral inhibition sensitivity further strengthens the role of the second half of this mediating pathway. The above conclusion provides a basis for constructing a stratified campus intervention program. The intervention system can be carried out around three aspects: implicit bias awareness training, anxiety emotion guidance, and psychological counseling that suits individual traits. This research expands the application scope of the theory of social information processing and also provides theoretical support for the prevention of interpersonal conflicts among teenagers.

References

[1] Dodge, K. A. (2006). Translational science in action: Hostile attributional style and the development of aggressive behavior problems. *Development and Psychopathology, 18*, 791–814. https://doi.org/10.1017/S0954579406060391

[2] Vitriol, J. A., & Moskowitz, G. B. (2021). Reducing defensive responding to implicit bias feedback: On the role of perceived moral threat and efficacy to change. *Journal of Experimental Social Psychology, 96*, 104165. https://doi.org/10.1016/j.jesp.2021.104165

[3] Kalis, A., & Ghijsen, H. (2022). Understanding implicit bias: A case for regulative dispositionalism. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2022.2046261

[4] Li, D., Carnelley, K. B., Lin, X., & Rowe, A. C. (2026). How hostile attribution bias affects attachment anxiety/avoidance and relationship satisfaction: A cross-cultural perspective during the COVID pandemic. *The Journal of Social Psychology*. https://doi.org/10.1080/00224545.2026.2626036

[5] Verhoef, R. E. J., Alsem, S. C., Verhulp, E. E., & De Castro, B. O. (2019). Hostile intent attribution and aggressive behavior in children revisited. *Child Development, 90*. https://doi.org/10.1111/cdev.13255

[6] Berg, Y. H. M., & Lansu, T. A. M. (2020). It's not just what you say, it's how you say it too. *Aggressive Behavior*. https://doi.org/10.1002/ab.21910

DOI: 10.54254/2753-7048/2026.37252

[7] Alden, L. E., Taylor, C. T., Mellings, T. M. J. B., & Laposa, J. M. (2008). Social anxiety and the interpretation of positive social events. *Journal of Anxiety Disorders, 22*, 577–590. https://doi.org/10.1016/j.janxdis.2007.05.007
[8] Chen, J., Short, M., & Kemps, E. (2020). Interpretation bias in social anxiety: A systematic review and meta-analysis. *Journal of Affective Disorders, 276*, 1119–1130. https://doi.org/10.1016/j.jad.2020.07.121
[9] Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis* (3rd ed.). Guilford Press.
[10] Kimbrel, N. A. (2008). A model of the development and maintenance of generalized social phobia. *Clinical Psychology Review, 28*, 592–612. https://doi.org/10.1016/j.cpr.2007.08.003
[11] Gray, J. A., & McNaughton, N. (2003). *The neuropsychology of anxiety*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198522713.001.0001
[12] Kimbrel, N. A., Nelson-Gray, R. O., & Mitchell, J. T. (2012). BIS, BAS, and bias: The role of personality and cognitive bias in social anxiety. *Personality and Individual Differences, 52*, 395–400. https://doi.org/10.1016/j.paid.2011.10.041
[13] Dapprich, A. L., Derks, L. M., Holtmann, M., Lange, W.-G., Legenbauer, T., & Becker, E. S. (2023). Hostile and threatening interpretation biases in adolescent inpatients are specific to callous-unemotional traits and social anxiety. *European Child & Adolescent Psychiatry*. https://doi.org/10.1007/s00787-023-02227-3
[14] Zhu, W., Yang, Y., Huang, Y., & Tian, X. (2025). The impact of cognitive bias modification of interpretation on hostile attribution bias, reactive aggression and neural mechanisms. *Journal of Affective Disorders, 383*, 108–116. https://doi.org/10.1016/j.jad.2025.04.138
[15] Zhu, W., Chen, J., Tian, X., Wu, X., Matkurban, K., Qiu, J., & Xia, L.-X. (2022). The brain correlates of hostile attribution bias and their relation to the displaced aggression. *Journal of Affective Disorders, 317*, 204–211. https://doi.org/10.1016/j.jad.2022.08.065