

MORAL DISENGAGEMENT AS A MEDIATOR BETWEEN PROBLEMATIC SOCIAL MEDIA USE AND ANTISOCIAL BEHAVIOR AMONG YOUNG ADULTS

Noman Gulfam ^{a*}, Farooq Ahmad ^b, Samra Javed ^c, Nida Iqbal ^d, & Hina Rizwan ^e

^aMSc Forensic Psychology, University of Westminster, London, United Kingdom. Email: nomangulfam4@gmail.com

^bStudent, Bangor University, Wales, United Kingdom. Email: farooq7.lm@outlook.com. ORCID iD: 0009-0006-9618-2591

^cAutism Spectrum Disorder Educator, Fitzalan High School, Cardiff, Wales, United Kingdom, and Associate Professor, Alghazali University, Karachi, Pakistan. Email: samrajaved@gmail.com

^dMPhil Clinical Psychology, The University of Lahore, Lahore, Pakistan. Email: nidaiqbal628@gmail.com. ORCID ID: 0009-0008-5135-0496

^eClinical Psychologist, Trauma-Informed Therapist and Program Specialist, Aurattech and Therapybeans (Private Space), Pakistan. Email: hinarizwan9041@outlook.com. ORCID ID: 0009-0004-2230-4889

*Correspondence to: Noman Gulfam, MSc Forensic Psychology, University of Westminster, London, United Kingdom. Email: nomangulfam4@gmail.com

KEYWORDS:

Problematic Social Media Use, Moral Disengagement, Antisocial Behavior, Statistical Mediation, Young Adults, Cross-Sectional Design

CITE THIS ARTICLE:

Gulfam, N., Ahmad, F., Javed, S., Iqbal, N., & Rizwan, H. (2026). Moral disengagement as a mediator between problematic social media use and antisocial behavior among young adults. *Pakistan Journal of Positive Psychology*, 3(7), 52-64.

ABSTRACT

Moral disengagement, the cognitive maneuvers through which conduct that would ordinarily be inhibited by self-sanction is reconstrued as acceptable, has been proposed as a moral-cognitive correlate of both problematic social media use and antisocial conduct, but the three constructs have rarely been examined together in young adults, and evidence on this configuration among Pakistani young adults was not located. The present study therefore examined whether moral disengagement statistically mediates the association between problematic social media use and antisocial behavior. A cross-sectional correlational design with simple mediation analysis was used with Pakistani young adults enrolled in English-medium higher education institutions in Lahore, recruited through non-probability purposive sampling using a self-administered English questionnaire. Problematic social media use was measured with the Bergen Social Media Addiction Scale, moral disengagement with the short form of the Propensity to Morally Disengage Scale, and antisocial behavior with the Subtypes of Antisocial Behavior Questionnaire scored as a total; all measures showed acceptable internal consistency. Data were analyzed using descriptive statistics, Pearson correlations, and the PROCESS macro for simple mediation, with the indirect effect evaluated through a percentile bootstrap confidence interval. All three variables were positively and significantly correlated, and the indirect association through moral disengagement was supported, its confidence interval excluding zero, although it was modest and the direct association remained significant. The findings are consistent with a positive indirect association through moral disengagement, a pattern described as partial statistical mediation. Because the design is cross-sectional, they cannot establish temporal precedence or causality.

I. INTRODUCTION

Social media platforms now mediate a substantial share of young adults' communication, information seeking, and social comparison. Alongside ordinary use, a pattern of dysregulated engagement has been described in which use becomes preoccupying and difficult to control. The Bergen Social Media Addiction Scale operationalizes this pattern through six components derived from the addiction literature, one item each for salience, mood modification, tolerance, withdrawal, conflict, and relapse, with reference to the preceding twelve months

(Andreassen et al., 2016). This conceptualization treats problematic social media use as a dimensional pattern of dysregulated engagement rather than as a diagnostic category, and the present study adopts that reading throughout. Antisocial behavior refers to actions that harm others, violate social norms, or infringe on others' personal or property rights. The construct is heterogeneous. Burt and Donnellan (2009) argued that physical aggression, non-aggressive rule-breaking, and social aggression constitute meaningfully distinct though correlated components, each with different developmental trajectories, demographic correlates, and personological underpinnings, and developed the Subtypes of Antisocial Behavior Questionnaire to assess all three efficiently and in the public domain. They also noted that a measure suitable for college students was needed, given the elevated rates of adolescence-limited antisocial behavior in that population. University-enrolled young adults are therefore an appropriate population in which to study variation in such conduct, provided that the measurement is understood as self-reported behavioral frequency rather than as a clinical or diagnostic determination.

Bandura (1999) described moral agency as regulated by self-sanctions tied to personal standards, and specified a set of psychosocial maneuvers through which those self-sanctions can be selectively disengaged from injurious conduct. These include the cognitive restructuring of harmful conduct through moral justification, sanitizing language, and advantageous comparison; the disavowal of personal agency through the diffusion or displacement of responsibility; the disregarding or minimizing of injurious effects; and the attribution of blame to, and dehumanization of, those who are harmed. Moore et al. (2012) operationalized this propensity as a general individual difference in adults, developing an eight-item short form in which each item represents one mechanism and which functions as a unidimensional measure. Two features of that instrument are relevant here: it treats moral disengagement as a relatively stable propensity rather than a momentary state, and it was developed and evaluated primarily in adult organizational contexts. The association between moral disengagement and harmful conduct is the best established of the three relationships examined in this study. In a meta-analysis of 27 independent samples comprising 17,776 participants aged 8 to 18 years, Gini et al. (2014) reported a positive overall association between moral disengagement and aggressive behavior, $r = .28$, 95% CI [.23, .32], with larger effect sizes for adolescents than for children and effect sizes that did not vary by type of aggressive behavior, gender, or publication status. That synthesis concerns children and adolescents rather than young adults. Evidence in the present age range is thinner but directionally consistent: Jeong et al. (2024) found, in 434 college students, that moral disengagement was positively associated with cyberbullying perpetration, cyberbullying victimization, passive bystanding, and reinforcing behavior online, although that study assessed online bullying-related behavior rather than the broader antisocial domain.

The theoretical case for relating problematic social media use to moral disengagement rests on features of digitally mediated interaction rather than on quantity of use as such. Suler (2004) proposed that online settings loosen ordinary interpersonal restraint through dissociative anonymity, invisibility, asynchronicity, solipsistic introjection, dissociative imagination, and the minimization of authority, and characterized the result as a shift in the constellation of affect and cognition brought to interaction rather than the disclosure of an underlying true self. Read alongside Bandura's (1999) account, these conditions describe settings in which justificatory cognitions may be exercised with fewer immediate social consequences. Empirical work positioning these two constructs in sequence is limited and concentrated outside the young-adult range. Colella et al. (2025) examined 895 Italian and Spanish preadolescents ($M_{\text{age}} = 11.23$, $SD = 1.064$) and reported that problematic social media use and moral disengagement serially mediated the association between cybervictimization and cyberbullying perpetration, with a correlation of .373 between the two mediators. Those authors also stated explicitly that the reverse ordering, in which moral disengagement precedes problematic use, has been less explored and was not examined in their study. The ordering adopted in the present model therefore has precedent in the literature, but that precedent is structural rather than evidential: it is drawn from a preadolescent sample, a different mediator instrument, a cyberbullying outcome, and a cross-sectional design.

Evidence on the association between problematic social media use and antisocial conduct is mixed, and the inconsistency is most visible in the work closest to the present population. Shahid et al. (2024), studying 162 Pakistani young adults aged 18 to 25 with the same problematic-use instrument used here, reported a positive but non-significant correlation between social media addiction and aggression, together with a significant negative conditional effect of social media addiction on aggression and a significant interaction with social connectedness.

That study assessed aggression rather than antisocial behavior and estimated conditional rather than unconditional effects, so it is not a direct comparison; it does, however, indicate that the direction and magnitude of this association should not be treated as settled. Three gaps motivate the present model. First, the three constructs have most often been examined in pairs, or with moral disengagement positioned in relation to cyberbullying specifically, rather than in relation to a general antisocial-behavior composite in young adults. Second, evidence examining this specific three-variable configuration among Pakistani young adults was not located, so the study is exploratory with respect to that population rather than a replication within it. Third, no verified Urdu or Pakistani validation was identified for any of the three instruments used, which constrains what a single study in this setting can establish about measurement.

The study therefore examines whether moral disengagement statistically mediates the association between problematic social media use and antisocial behavior. The design is cross-sectional, and this places a firm limit on interpretation. Maxwell and Cole (2007) demonstrated that cross-sectional approaches to mediation typically generate substantially biased estimates of longitudinal parameters even under the ideal condition of complete mediation, and Maxwell et al. (2011) extended this to the more common case of partial mediation, showing that a cross-sectional analysis can imply a substantial indirect effect even when the true longitudinal indirect effect is zero. Accordingly, the analysis tests whether the data are consistent with an indirect association; it does not test a mechanism, and it does not establish temporal precedence. The study addresses four questions: whether problematic social media use is associated with moral disengagement; whether moral disengagement is associated with antisocial behavior; whether problematic social media use is associated with antisocial behavior; and whether moral disengagement statistically mediates the association between problematic social media use and antisocial behavior. The corresponding hypotheses are:

- H1. Problematic social media use is positively associated with moral disengagement.
- H2. Moral disengagement is positively associated with antisocial behavior.
- H3. Problematic social media use is positively associated with antisocial behavior.
- H4. Moral disengagement statistically mediates the association between problematic social media use and antisocial behavior.

II. METHOD

The study employs a quantitative, cross-sectional, correlational design analyzed through simple mediation. All three variables are specified as concurrent measurements obtained on a single occasion. The design is appropriate for estimating the pattern of association among the three constructs and for decomposing a total association into direct and indirect components. It is not suited to establishing temporal order, and no such claim is made. The design is not longitudinal, prospective, or experimental.

Participants

The target population is Pakistani young adults aged 18 to 25 enrolled in English-medium higher education institutions in Lahore, Pakistan. Sampling is non-probability purposive, with the questionnaire battery self-administered online. The analyzed dataset comprises 250 cases. Of these, 144 (57.6%) are recorded as female and 106 (42.4%) as male, with ages ranging from 18 to 25 years ($M = 21.17$, $SD = 1.85$). Year of enrollment is distributed across first year ($n = 61$, 24.4%), second year ($n = 58$, 23.2%), third year ($n = 62$, 24.8%), fourth year ($n = 51$, 20.4%), and postgraduate level ($n = 18$, 7.2%). Reported daily social media use is concentrated in the 3–4-hour band ($n = 102$, 40.8%), followed by 1–2 hours ($n = 61$, 24.4%), 5–6 hours ($n = 47$, 18.8%), more than 6 hours ($n = 30$, 12.0%), and less than 1 hour ($n = 10$, 4.0%). The most frequently recorded primary platform is Instagram ($n = 86$, 34.4%), followed by WhatsApp ($n = 65$, 26.0%), TikTok ($n = 36$, 14.4%), YouTube ($n = 29$, 11.6%), Facebook ($n = 23$, 9.2%), and X or other platforms ($n = 11$, 4.4%). These characteristics are reported descriptively only and do not enter the mediation model as covariates. Inclusion requires that a respondent is aged 18 to 25 years, is currently enrolled in an English-medium higher education institution, has been an active user of at least one social media platform during the reference period, and provides the required consent response before proceeding. Exclusion applies where a respondent falls outside the stated age range, is not enrolled in the specified educational setting,

declines the required consent response, or leaves more than 10% of items missing on any study scale. No further eligibility conditions are imposed.

Sample size was determined with reference to power for the indirect effect rather than for the individual regression paths, because power for a product of coefficients is not adequately established by entering a mediation model into conventional regression-based power software. A Monte Carlo power analysis was conducted for a simple mediation model with a standardized a path of .25 and a standardized b path of .20, implying an indirect effect of .050, with $\alpha = .05$, two-tailed inference, and 3,000 replications per candidate sample size. Estimated power for the indirect effect was .740 at $N = 200$, .771 at $N = 210$, .798 at $N = 220$, and .834 at $N = 230$. A target of $N = 250$ was adopted as a conservative value above the estimated minimum required for power of .80 against the smallest association considered worth detecting, with the additional margin reflecting the expectation that measurement error in a brief mediator measure attenuates the observed indirect effect. The analytical N was fixed at 250 before the analyses reported below and was not revised thereafter.

Measures

The entire battery is administered in English. This is justified by the sampling frame, whose medium of instruction is English. No translation, back-translation, or cultural adaptation was undertaken, and no Urdu or Pakistani validation is claimed for any instrument. For none of the three measures was a validated Urdu or Pakistani version identified; each is therefore used in its original English form. [I] The Bergen Social Media Addiction Scale (BSMAS; Andreassen et al., 2016) comprises six items, each corresponding to one component of addictive use (salience, mood modification, tolerance, withdrawal, conflict, relapse), rated from 1 (*very rarely*) to 5 (*very often*) with reference to the past twelve months. The total is the unweighted sum of the six items, with a possible range of 6 to 30 and higher scores indicating greater problematic use. The instrument is brief, does not require a proprietary manual, and has been used widely in university samples. [II] The Propensity to Morally Disengage Scale, eight-item form (PMD-8; Moore et al., 2012), comprises eight items, one for each of the eight mechanisms described by Bandura (1999), rated from 1 (*strongly disagree*) to 7 (*strongly agree*). The total is the unweighted sum, with a possible range of 8 to 56 and higher scores indicating a greater propensity to morally disengage. The measure functions as a unidimensional index of a general propensity in adults. It was developed and evaluated primarily in adult organizational contexts rather than in university student populations, a constraint considered in the Discussion. [III] The Subtypes of Antisocial Behavior Questionnaire (STAB; Burt & Donnellan, 2009) comprises 32 items covering physical aggression (10 items), social aggression (11 items), and rule-breaking (11 items), each rated from 1 (*never*) to 5 (*nearly all the time*). The instrument was developed expressly to be freely available to researchers in the public domain. In the present study the outcome is the STAB total score, the unweighted sum of all 32 items, with a possible range of 32 to 160. The three components are reported descriptively and for internal consistency only and are not analyzed as separate outcomes. The decision to use the total score departs from the instrument's original three-component scoring and is treated as a limitation rather than as a psychometric improvement.

Procedure

The specified procedure is a single-occasion, self-administered online questionnaire. Respondents first view an information page describing the purpose of the study, the voluntary nature of participation, the approximate completion time, the absence of any requirement to answer every item, and the right to withdraw at any point before submission without giving a reason. Participation proceeds only after the required consent response is provided. The battery is presented in a fixed order: a brief demographic section, followed by the BSMAS, the PMD-8, and the STAB. No identifying information is collected, responses are anonymous, and the data are treated as confidential and reported only in aggregate. The procedure is framed in accordance with the ethical principles of the Declaration of Helsinki and the ethical standards of the American Psychological Association concerning voluntary participation, informed consent, anonymity, and the right of withdrawal. No incentive is offered. All scale scores are unweighted item sums, computed from item-level responses. The BSMAS total is the sum of its six items; the PMD-8 total is the sum of its eight items; the three STAB component scores are the sums of their 10, 11, and 11 items respectively, and the STAB total is the sum of the three component scores, which is arithmetically identical to the sum of all 32 items. No items are reverse-scored, as all items on all three instruments are positively keyed. No item or scale scores are standardized, transformed, or weighted, and no covariates are computed. Data

preparation involved verification that every item response fell within the legitimate response range of its instrument and that every total fell within the possible score range of its scale.

Statistical Analysis

Analyzes followed a fixed sequence: screening of item ranges, score reconstruction, and distributional inspection; descriptive statistics including means, standard deviations, observed minima and maxima, skewness, and kurtosis; Cronbach's alpha computed from item-level responses for each scale and for the three STAB components; Pearson product-moment correlations among the study variables; and simple mediation analysis using PROCESS Model 4 (Hayes, 2022). In the mediation model, X is problematic social media use, M is moral disengagement, and Y is the STAB total score. No covariates are included in either equation. The model yields the *a* path from X to M, the *b* path from M to Y conditional on X, the direct effect *c'* from X to Y conditional on M, the total effect *c* from X to Y, and the indirect effect as the product *ab*. Inference on the indirect effect rests on a percentile bootstrap confidence interval based on 5,000 bootstrap samples, with the indirect association treated as statistically supported when the 95% interval excludes zero. The indirect effect is not evaluated by requiring each constituent path to reach significance individually, and no normal-theory test of the product term is reported. Univariate outliers were screened using standardized scores with a threshold of $|z| = 3.29$, and multivariate distance was assessed using Mahalanobis D^2 against the $\chi^2(3)$ critical value at $p < .001$. No additional analyzes, subgroup comparisons, moderation tests, or alternative model specifications were conducted.

III. RESULTS

Data Screening and Analytical Properties of the Dataset

The analyzed dataset comprises 250 cases and 46 item-level variables (BSMAS = 6 items, PMD-8 = 8 items, STAB = 32 items). All item cells are complete, so no missing-data estimation or case exclusion was required and no imputation was applied. Scale totals were computed as unweighted item sums, and every observed total fall within the legitimate score range of its instrument: BSMAS 6-28 (possible range 6-30), PMD-8 8-40 (possible range 8-56), and STAB total 34-100 (possible range 32-160). The STAB total equals the exact sum of its physical-aggression, social-aggression, and rule-breaking components for every case. Univariate extreme values were examined using standardized scores; no case exceeded $|z| = 3.29$ on problematic social media use, moral disengagement, or antisocial behavior. Multivariate distance across the three variables was assessed with Mahalanobis D^2 ; no case exceeded the critical $\chi^2(3) = 16.27$ at $p < .001$. Skewness and kurtosis values for the three study variables fall within ± 1 and ± 2 respectively, indicating that the observed distributions showed no extreme univariate departures on these two indices, which does not in itself establish that all assumptions of ordinary least squares estimation were satisfied.

Table I: Sample Characteristics (N = 250)

Characteristic	<i>n</i>	%
Gender		
Female	144	57.6
Male	106	42.4
Year of enrollment		
First year	61	24.4
Second year	58	23.2
Third year	62	24.8
Fourth year	51	20.4
Postgraduate	18	7.2
Daily social media use		
Less than 1 hour	10	4.0
1-2 hours	61	24.4
3-4 hours	102	40.8
5-6 hours	47	18.8
More than 6 hours	30	12.0

Primary platform		
Instagram	86	34.4
WhatsApp	65	26.0
TikTok	36	14.4
YouTube	29	11.6
Facebook	23	9.2
X / other	11	4.4

Note. Age range 18-25 years, $M = 21.17$, $SD = 1.85$. Percentages are based on $N = 250$ and sum to 100.0 within rounding.

Table 1 indicates that the 250 cases fall within the specified eligibility band of 18 to 25 years ($M = 21.17$, $SD = 1.85$), with female cases in the majority at 144 (57.6%) against 106 male cases (42.4%). Year of enrollment is distributed relatively evenly across the four undergraduate years, which together account for 232 cases (92.8%), while postgraduate enrollment is comparatively infrequent at 18 cases (7.2%). Reported daily social media use clusters in the middle of the distribution, with the 3–4-hour band the most frequently endorsed at 102 cases (40.8%) and the two extreme bands the least frequent, at 10 cases (4.0%) for less than 1 hour and 30 cases (12.0%) for more than 6 hours; 179 cases (71.6%) report three or more hours of daily use. Platform use is concentrated in two services, with Instagram ($n = 86$, 34.4%) and WhatsApp ($n = 65$, 26.0%) together accounting for 151 cases (60.4%), and X or other platforms recorded least frequently ($n = 11$, 4.4%). These characteristics are reported descriptively and were not entered into the mediation model as covariates.

Table II: Descriptive Statistics and Internal Consistency (N = 250)

Variable	M	SD	α	Range
Problematic social media use (BSMAS)	14.84	4.35	.806	6-30
Moral disengagement (PMD-8)	23.36	6.08	.739	8-56
Antisocial behavior, total (STAB)	58.50	12.72	.909	32-160
Physical aggression	20.42	5.53	.824	10-50
Social aggression	23.00	6.35	.864	11-55
Rule-breaking	15.08	3.17	.804	11-55

Note. α = Cronbach's alpha computed from item-level responses. The three STAB components are reported for descriptive and reliability purposes only; the inferential model uses the STAB total score.

As Table 2 shows, mean scores on all three focal measures fall below their respective scale midpoints, with problematic social media use averaging 14.84 ($SD = 4.35$) against a possible range of 6 to 30, moral disengagement averaging 23.36 ($SD = 6.08$) against a possible range of 8 to 56, and the antisocial behavior total averaging 58.50 ($SD = 12.72$) against a possible range of 32 to 160. Observed ranges are narrower than the possible ranges on every measure, most markedly for moral disengagement, whose observed maximum of 40 leaves the upper portion of the scale unrepresented. Among the STAB components, social aggression records the highest mean relative to its possible range ($M = 23.00$, $SD = 6.35$) and rule-breaking the lowest ($M = 15.08$, $SD = 3.17$), with the rule-breaking distribution the most positively skewed of the six variables (skewness = 0.75) and its observed minimum resting on the scale floor of 11. Skewness and kurtosis values are modest throughout. Internal consistency, computed from item-level responses rather than adopted from prior literature, is acceptable for the BSMAS ($\alpha = .806$), for the three STAB components ($\alpha = .804$ to $.864$), and for the STAB total ($\alpha = .909$), while the PMD-8 yields the lowest coefficient in the set ($\alpha = .739$), which meets the conventional .70 threshold without being strong; its consequences for the precision of the mediator-related estimates are considered in the Discussion. The coefficient obtained for the STAB total is reported as a property of the aggregate score only. Because the STAB was developed to differentiate physical aggression, social aggression, and rule-breaking as distinguishable components and the present analysis aggregates them into a single outcome, this coefficient is not treated as evidence that aggregation is preferable to the instrument's original three-component structure.

Table III: Pearson Correlations Among Study Variables and STAB Components (N = 250)

Variables	1	2	3	4	5	6
1. Problematic social media use	—					
2. Moral disengagement	.258***	—				
3. Antisocial behavior (total)	.296***	.328***	—			
4. Physical aggression	.230***	.267***	—	—		
5. Social aggression	.251***	.279***	—	.563***	—	
6. Rule-breaking	.281***	.291***	—	.549***	.542***	—

*** $p < .001$

Table 3 shows that all three associations among the focal variables are positive and statistically significant, and that they fall in the small-to-moderate range. The strongest of the three is between moral disengagement and antisocial behavior ($r = .328, p < .001$), followed by problematic social media use and antisocial behavior ($r = .296, p < .001$) and problematic social media use and moral disengagement ($r = .258, p < .001$). The pattern across the STAB components is consistent in direction and similar in magnitude: problematic social media use correlates most strongly with rule-breaking ($r = .281$) and least strongly with physical aggression ($r = .230$), and moral disengagement likewise correlates most strongly with rule-breaking ($r = .291$) and least strongly with physical aggression ($r = .267$), so neither predictor is associated with one form of antisocial conduct to the exclusion of the others. The three components intercorrelate between $r = .542$ and $r = .563$, indicating substantial but incomplete overlap. These coefficients are unconditional bivariate associations and are reported independently of the mediation model; the indirect effect was evaluated directly within the specified model rather than inferred from this correlation pattern.

Table IV: Simple Mediation Model Coefficients (PROCESS Model 4; N = 250, 5,000 Bootstrap Samples)

Path	<i>B</i>	<i>SE</i>	<i>t</i>	<i>df</i>	<i>p</i>	β
<i>a</i> : PSMU → MD	0.360	0.086	4.20	248	< .001	.258
<i>b</i> : MD → ASB, PSMU controlled	0.564	0.127	4.46	247	< .001	.270
<i>c'</i> : PSMU → ASB, MD controlled	0.661	0.177	3.74	247	< .001	.226
<i>c</i> : PSMU → ASB, total effect	0.864	0.177	4.88	248	< .001	.296
Effect	Estimate	Boot <i>SE</i>	Boot 95% LL	Boot 95% UL		
Indirect (<i>ab</i>)	0.203	0.066	0.088	0.340		

Note. PSMU = problematic social media use; MD = moral disengagement; ASB = antisocial behavior (STAB total). No covariates were included. Model summaries: mediator equation $R^2 = .066, F(1, 248) = 17.65, p < .001$; outcome equation $R^2 = .155, F(2, 247) = 22.73, p < .001$; total-effect equation $R^2 = .088, F(1, 248) = 23.78, p < .001$. The confidence interval is a percentile bootstrap interval based on 5,000 resamples.

Table 4 reports the four regression paths of the mediation model together with the bootstrapped indirect effect. In the mediator equation, problematic social media use is positively associated with moral disengagement ($a = 0.360, SE = 0.086, t(248) = 4.20, p < .001, \beta = .258$), a one-point increase on the BSMAS corresponding to an increase of approximately 0.36 points on the PMD-8, and the predictor accounts for 6.6% of the variance in the mediator, $R^2 = .066, F(1, 248) = 17.65, p < .001$. In the outcome equation, both predictors are positively associated with antisocial behavior: moral disengagement is associated with the STAB total conditional on problematic social media use ($b = 0.564, SE = 0.127, t(247) = 4.46, p < .001, \beta = .270$), and the direct association of problematic social media use with antisocial behavior remains statistically significant with the mediator in the model ($c' = 0.661, SE = 0.177, t(247) = 3.74, p < .001, \beta = .226$), the two predictors together accounting for 15.5% of the variance in the outcome, $R^2 = .155, F(2, 247) = 22.73, p < .001$. The total effect is positive and statistically significant ($c = 0.864, SE = 0.177, t(248) = 4.88, p < .001, \beta = .296$), $R^2 = .088, F(1, 248) = 23.78, p < .001$, and its decomposition is exact, the direct and indirect components summing to the total ($0.661 + 0.203 = 0.864$). The standardized coefficients show that the two paths constituting the indirect route are of comparable magnitude ($\beta = .258$ and $\beta = .270$) and that the direct path is slightly smaller ($\beta = .226$) than the total association it partly replaces ($\beta = .296$), the reduction from c to c' corresponding to the portion of the association that runs through the mediator.

The indirect association of problematic social media use with antisocial behavior through moral disengagement is estimated at $ab = 0.203$ (bootstrap $SE = 0.066$), with a 95% percentile bootstrap confidence interval of [0.088, 0.340]. Because the interval does not contain zero, the indirect association is statistically supported. In completely standardized terms the indirect effect is .070, which is modest in magnitude. As a descriptive decomposition of the observed total association, the estimated indirect component represents approximately 23.5% of that association, leaving the larger part of the observed association outside the mediated pathway; this proportion describes the arithmetic partitioning of the estimates and carries no causal interpretation. Because the indirect effect is statistically supported while the direct effect also remains statistically significant, the observed configuration of estimates is described as partial statistical mediation. This label characterizes the pattern of coefficients obtained in the model. It is not evidence of a causal or temporal mediating mechanism: all three variables were specified as concurrent measurements within a cross-sectional design, which cannot establish temporal precedence among them. The results are consistent with a positive indirect association operating through moral disengagement, and consistency of this kind does not adjudicate between the specified ordering and alternative orderings of the same three variables.

Statistical Conclusions by Hypothesis

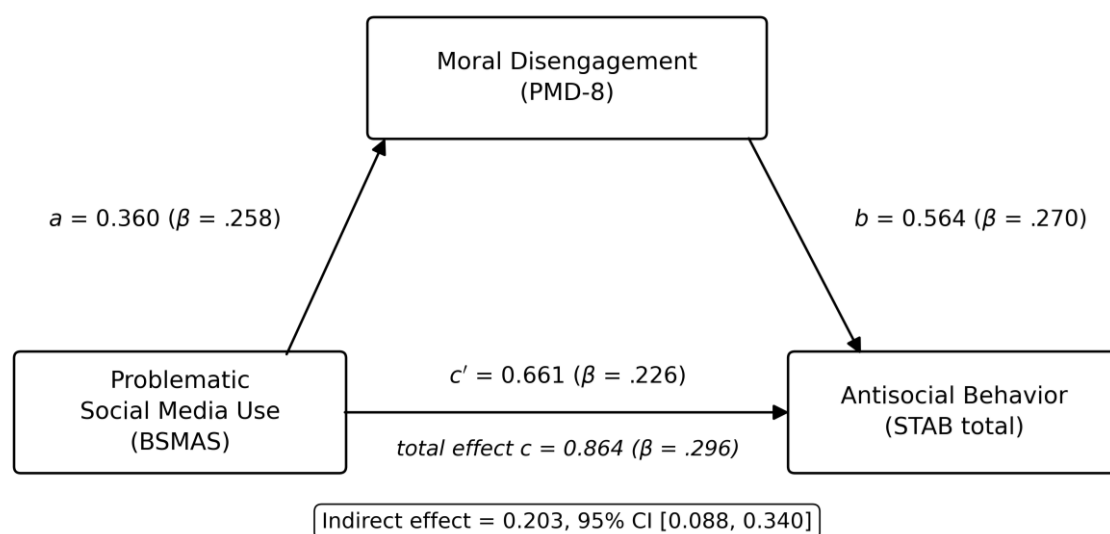
H1 (problematic social media use is positively associated with moral disengagement) was supported. The zero-order association was $r = .258$, $p < .001$. The corresponding unstandardized regression path in the mediation model, the a path, was 0.360 ($SE = 0.086$), $p < .001$, $\beta = .258$. These are two distinct estimates: the former is an unconditional bivariate correlation, the latter the regression coefficient estimated within the mediator equation of Model 4.

H2 (moral disengagement is positively associated with antisocial behavior) was supported. The zero-order association was $r = .328$, $p < .001$. The b path, which estimates the association between moral disengagement and antisocial behavior conditional on problematic social media use, was 0.564 ($SE = 0.127$), $p < .001$, $\beta = .270$. The conditional estimate is not interchangeable with the unconditional correlation.

H3 (problematic social media use is positively associated with antisocial behavior) was supported. The zero-order association was $r = .296$, $p < .001$, and the total effect in the mediation model was $c = 0.864$ ($SE = 0.177$), $p < .001$, $\beta = .296$; the standardized total effect and the zero-order correlation coincide here because the total-effect equation contains a single predictor.

H4 (moral disengagement statistically mediates the association between problematic social media use and antisocial behavior) was supported on the stated criterion, the 95% bootstrap confidence interval for ab [0.088, 0.340] excluding zero. The support obtained is for a statistical indirect association of modest magnitude within a cross-sectional design that cannot establish a causal or temporal mediating sequence.

Figure 1: Simple mediation model of the association between problematic social media use and antisocial behavior through moral disengagement ($N = 250$)



Unstandardized coefficients are reported with standardized coefficients in parentheses. All four path coefficients (a , b , c' , c) are statistically significant at $p < .001$. The indirect effect is based on 5,000 bootstrap samples; its inference rests on the bootstrap confidence interval rather than on a p -value.

IV. DISCUSSION

The present analysis examined whether moral disengagement statistically mediates the association between problematic social media use and antisocial behavior among young adults, using a cross-sectional correlational design and PROCESS Model 4. All four hypotheses received statistical support. Problematic social media use was positively associated with moral disengagement ($r = .258$; $a = 0.360$, $\beta = .258$, $p < .001$), moral disengagement was positively associated with antisocial behavior conditional on problematic social media use ($r = .328$; $b = 0.564$, $\beta = .270$, $p < .001$), and problematic social media use showed a positive total association with antisocial behavior ($r = .296$; $c = 0.864$, $\beta = .296$, $p < .001$). The indirect association through moral disengagement was statistically supported, $ab = 0.203$, 95% bootstrap CI [0.088, 0.340]. The direct association remained statistically significant with the mediator in the model ($c' = 0.661$, $\beta = .226$, $p < .001$), so the coefficient pattern corresponds to partial statistical mediation. Three features warrant emphasis before interpretation. First, all associations are small to moderate: problematic social media use accounted for 6.6% of the variance in moral disengagement, and the two predictors together accounted for 15.5% of the variance in antisocial behavior, leaving most variance in both unaccounted for. Second, the standardized indirect effect was .070, indicating a modest indirect association, while the significant direct effect indicates that the specified mediator did not account for the entire observed association. Third, the design is cross-sectional, so none of these estimates can establish temporal precedence or causality, and the statistical support obtained for the indirect association is support for a pattern of covariation rather than for a mechanism.

The positive association between problematic social media use and moral disengagement was statistically supported but modest. Theoretically, this pattern is compatible with accounts treating digital environments as contexts in which ordinary self-regulatory restraints operate less forcefully. Suler (2004) proposed that online environments generate a disinhibition effect through dissociative anonymity, invisibility, asynchronicity, solipsistic introjection, dissociative imagination, and the minimization of authority, characterizing this as a shift in the constellation of affect and cognition brought to interaction rather than the disclosure of an underlying true self. A plausible reading, offered as interpretation rather than as a demonstrated process, is that heavy and dysregulated engagement with such environments provides repeated occasions on which justificatory cognitions of the kind catalogued by Bandura (1999) may be exercised. The present analysis tested no such process; it estimated a single regression path between two concurrently specified variables. The reverse ordering is equally compatible with this data structure. Individuals with a higher propensity to morally disengage may gravitate toward, or persist longer in, the kinds of engagement the BSMAS indexes, in which case moral disengagement would be better positioned as an antecedent than as a mediator. A reciprocal relationship is also possible. Notably, Colella et al. (2025), who specified the same ordering used here, state explicitly that the reverse relationship has been less explored in the literature and was not tested in their study either. Cross-sectional estimates cannot distinguish among these orderings, and the support obtained for a path does not favour the specified direction over its alternatives.

The association between moral disengagement and antisocial behavior, conditional on problematic social media use, was statistically supported, and the corresponding zero-order association was the strongest in the matrix ($r = .328$). This aligns with the most established of the three relevant literatures. In a meta-analysis of 27 independent samples comprising 17,776 participants aged 8 to 18 years, Gini et al. (2014) reported a positive overall association between moral disengagement and aggressive behavior ($r = .28$, 95% CI [.23, .32]), with larger effect sizes for adolescents than children and effect sizes that did not vary by type of aggressive behavior, gender, or publication status. The magnitude observed here is closely comparable to that pooled estimate, with two qualifications: that meta-analysis concerned children and adolescents rather than young adults, and it concerned aggressive behavior rather than a composite encompassing physical aggression, social aggression, and rule-breaking. The convergence is therefore approximate. Evidence in young adults' points in the same direction. Jeong et al. (2024), in a sample of 434 college students, found moral disengagement positively associated with

cyberbullying perpetration, victimization, passive bystanding, and reinforcing behavior online. That study assessed online bullying-related behavior rather than the broader antisocial domain measured by the STAB, so it corroborates the direction of the association in this age group without providing a direct comparison for the present outcome.

The total association was positive and statistically supported ($r = .296$; $c = 0.864$). Here the comparison literature is mixed rather than convergent, and the divergence is sharpest in the evidence closest to the present population. Shahid et al. (2024), studying 162 Pakistani young adults aged 18 to 25 with the BSMAS, the Buss–Perry Aggression Questionnaire, and the Revised Social Connectedness Scale, reported a positive but non-significant correlation between social media addiction and aggression, alongside a significant negative conditional effect of social media addiction on aggression and a significant interaction with social connectedness. That pattern does not replicate the direction of the total association observed here. The studies differ in outcome construct, in analytical model, and in sample size, and a negative conditional effect estimated in a model containing an interaction term is not directly comparable to an unconditional total effect. The divergence is best treated as a genuine inconsistency in a small evidence base rather than explained away. Beyond this, evidence examining the specific three-variable pathway among Pakistani young adults was not located. The direct comparison literature for this model in this population is limited, and the present findings should not be read as either replicating or contradicting a local consensus, because no such consensus was identified. The bootstrap confidence interval for the indirect effect excluded zero, satisfying the stated criterion for statistical support. Interpreted strictly, the product of two regression coefficients estimated within the specified model is distinguishable from zero under resampling. This does not mean that moral disengagement transmits an effect, that it explains the association between the other two variables, or that any temporal sequence has been identified. The standardized indirect effect of .070 indicates a modest indirect association, and the significant direct path indicates that the specified mediator did not account for the entire observed association.

Methodological literature cautions that this configuration should not be interpreted as evidence of a causal mediating process. Maxwell and Cole (2007) demonstrated that cross-sectional approaches to mediation typically generate substantially biased estimates of longitudinal parameters even under the ideal condition of complete mediation. Maxwell et al. (2011) extended this to the more common partial-mediation case, showing that substantial bias can arise there as well, and that a cross-sectional analysis can imply a substantial indirect effect even when the true longitudinal indirect effect is zero, so that a variable appearing to be a strong mediator cross-sectionally may not be a mediator at all longitudinally. Since the present configuration is partial rather than complete mediation, that second result applies directly. The term partial statistical mediation is used here as a description of the coefficient pattern in Table 4 and Figure 1; it is not a claim about mechanism, and it does not adjudicate between the specified model and alternative arrangements of the same three variables. Within Bandura's (1999) social cognitive account, moral disengagement comprises cognitive maneuvers through which conduct that would ordinarily be inhibited by internal self-sanctions is reconstrued as acceptable, including moral justification, euphemistic labeling, advantageous comparison, displacement and diffusion of responsibility, distortion of consequences, dehumanization, and attribution of blame. The instrument used here was designed to capture this propensity as a general individual difference in adults, with its eight-item form representing each mechanism and showing a unidimensional internal structure (Moore et al., 2012). A theoretical reading consistent with both this framework and the observed pattern is that individual differences in justificatory cognition are relevant to whether transgressive conduct is enacted, and that digitally mediated contexts may be one setting among many in which such cognition is exercised.

Some precedent exists for positioning these two constructs in sequence. Colella et al. (2025) found that problematic social media use and moral disengagement serially mediated the association between cybervictimization and cyberbullying perpetration in 895 Italian and Spanish preadolescents ($M_{\text{age}} = 11.23$), reporting a PSMU–MD correlation of .373. That study locates the same constructs in the ordering specified here, which gives the present model precedent in the literature. It differs in four consequential respects: its participants were preadolescents rather than young adults, its outcome was cyberbullying perpetration rather than a general antisocial composite, its mediator was a 24-item Moral Disengagement Scale rather than the eight-item propensity measure used here, and it was itself cross-sectional. The precedent is therefore structural rather

than evidential. These readings are interpretations compatible with the observed pattern. The present analysis did not test any of them, and the statistically supported indirect effect should not be treated as retrospective confirmation of the theoretical account that motivated the model. Two features merit note. The sample size was determined by a power analysis targeted at the indirect effect rather than by convenience or a rule of thumb, which matters because power for a product of coefficients is not adequately established by entering a mediation model into conventional regression-based power software. Inference about the indirect effect rests on a percentile bootstrap confidence interval rather than on the joint significance of individual paths, consistent with contemporary practice (Hayes, 2022).

Design and temporal precedence. All three variables were specified as concurrent measurements, so the analysis cannot establish that problematic social media use precedes moral disengagement, or that moral disengagement precedes antisocial behavior. This constrains H4 specifically: a model in which moral disengagement is the antecedent and problematic social media use the mediator would fit this data structure equally well, and nothing in the present results favors one arrangement over the other.

Self-report and common-method variance. Because all three constructs were assessed through self-report during a single administration, common-method variance may have influenced the observed associations. The direction and magnitude of any such bias cannot be established from the present design.

Non-probability sampling and generalizability. Sampling was purposive and non-probability, drawn from English-medium university students aged 18 to 25 in a single city. The results cannot be generalized to Pakistani young adults as a population, to non-university young adults, or to young adults outside this age band. Findings should therefore be regarded as specific to this sampling frame.

Language of administration. The battery was administered in English, justified by the English-medium enrollment of the sampling frame. No translation or cultural adaptation was undertaken and none is claimed. English administration nonetheless raises the possibility that item comprehension varied with English proficiency, and that variation is unmeasured.

Absence of verified local validation evidence. For none of the three instruments was a validated Urdu or Pakistani version identified. All were used in their original English forms, and their psychometric properties in a Pakistani young-adult population remain unestablished. The internal consistency values obtained here do not constitute validation evidence for this population.

Mediator measurement precision. The PMD-8 yielded $\alpha = .739$. This meets conventional criteria; it does not establish adequate measurement, and it is not a basis for dismissing the mediator either. An eight-item measure with internal consistency at this level provides limited measurement precision, and unreliability in a mediator generally attenuates observed associations involving it. No correction for attenuation was applied and no estimate of its magnitude is offered. The consequence for interpretation is that the *b* path and the indirect effect are estimated with less precision than the reported standard errors alone convey. The measure was also developed and evaluated primarily in adult organizational samples rather than university student populations, so its performance in the present context is not established.

Aggregation of the outcome. The STAB was developed around three distinguishable components, physical aggression, social aggression, and rule-breaking, which show different demographic correlates and developmental patterns (Burt & Donnellan, 2009). The present analysis aggregates them into a single total score, as specified for this analysis. The internal consistency of the aggregate is not evidence that aggregation is preferable to the instrument's original three-component structure. The consequence is that the analysis cannot say whether the observed indirect association pertains to aggressive conduct, to rule-breaking, or to both.

Unmeasured confounding. No covariates were included. Other unmeasured characteristics may be associated with problematic social media use, moral disengagement, and antisocial behavior and could therefore contribute to the observed associations. If such a characteristic is substantially related to all three variables, both

the *a* and *b* paths, and hence the indirect effect, would be affected. The indirect effect cannot be interpreted as adjusted for shared antecedents.

Inference from the observed distributions. The descriptive and distributional properties reported in the Results should not be read as population estimates. In particular, the observed distribution of the rule-breaking component, which rests close to the scale floor, should not be taken to represent the distribution of such conduct among Pakistani young adults more generally.

The research implications follow from the design limitations. Because the central question concerns a putative sequence, it requires a design capable of ordering the variables in time; a multi-wave panel design with cross-lagged estimation would allow competing orderings to be compared rather than assumed. Replication in Pakistani young adults with instruments whose psychometric properties have been established in that population is a prerequisite for local inference, and the absence of verified Urdu or Pakistani validation for all three measures marks this as a measurement priority rather than a peripheral concern. Practical implications must be stated conservatively, and the magnitudes reported here support only weak ones. Associations of this size do not identify problematic social media use as a determinant of antisocial conduct, nor do they establish that intervening on moral disengagement would alter such conduct. What the pattern is consistent with is that justificatory moral cognition merits attention in psychoeducational work with young adults alongside patterns of digital engagement, rather than either being addressed in isolation. Any stronger claim, in particular a proposal to target moral disengagement in order to reduce antisocial behavior, would require experimental or longitudinal evidence that this analysis does not provide. Four directions follow. Longitudinal designs with at least two measurement occasions are needed to evaluate the temporal ordering this model assumes and to test competing orderings against one another. Multi-informant or behavioral indicators of antisocial conduct would reduce the shared-method variance affecting single-source estimates. Disaggregating the outcome, examining physical aggression, social aggression, and rule-breaking separately in a study powered for that purpose, would establish whether the association is general across forms of antisocial conduct or specific to one, which the present total-score analysis cannot address. Finally, since no validated Urdu version of any of these instruments was identified, systematic adaptation and validation work with Pakistani young adults would be a substantive contribution in its own right and a precondition for cumulative local research on this model.

V. CONCLUSION

Within a cross-sectional design, problematic social media use, moral disengagement, and antisocial behavior were positively and significantly interrelated among young adults, and the indirect association of problematic social media use with antisocial behavior through moral disengagement was statistically supported, $ab = 0.203$, 95% bootstrap CI [0.088, 0.340]. The direct association remained significant, so the coefficient pattern corresponds to partial statistical mediation. The standardized indirect effect of .070 is modest, and the design cannot establish temporal precedence or causality. The results are consistent with a positive indirect association through moral disengagement; they do not demonstrate that such a pathway operates, and they do not identify moral disengagement as the mechanism linking problematic social media use to antisocial behavior. Establishing whether the ordering specified here holds requires longitudinal evidence and locally validated measurement.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

The author received no funding from any organizations.

VI. REFERENCES

Andreassen, C. S., Billieux, J., Griffiths, M. D., Kuss, D. J., Demetrovics, Z., Mazzoni, E., & Pallesen, S. (2016). The relationship between addictive use of social media and video games and symptoms of psychiatric

- disorders: A large-scale cross-sectional study. *Psychology of Addictive Behaviors*, 30(2), 252–262. <https://doi.org/10.1037/adb0000160>
- Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3
- Burt, S. A., & Donnellan, M. B. (2009). Development and validation of the Subtypes of Antisocial Behavior Questionnaire. *Aggressive Behavior*, 35(5), 376–398. <https://doi.org/10.1002/ab.20314>
- Colella, G. M., Servidio, R. C., Palermi, A. L., Bartolo, M. G., García-Carrera, P., Ortega-Ruiz, R., & Romera, E. M. (2025). Cyberbullying perpetration and socio-behavioral correlates in Italian and Spanish preadolescents: A cross-national study and serial mediation analysis. *International Journal of Environmental Research and Public Health*, 22(3), 389. <https://doi.org/10.3390/ijerph22030389>
- Gini, G., Pozzoli, T., & Hymel, S. (2014). Moral disengagement among children and youth: A meta-analytic review of links to aggressive behavior. *Aggressive Behavior*, 40(1), 56–68. <https://doi.org/10.1002/ab.21502>
- Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.
- Jeong, R., Gilbertson, M., Riffle, L. N., & Demaray, M. K. (2024). Participant role behavior in cyberbullying: An examination of moral disengagement among college students. *International Journal of Bullying Prevention*, 6, 28–40. <https://doi.org/10.1007/s42380-022-00137-7>
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12(1), 23–44. <https://doi.org/10.1037/1082-989X.12.1.23>
- Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, 46(5), 816–841. <https://doi.org/10.1080/00273171.2011.606716>
- Moore, C., Detert, J. R., Klebe Treviño, L., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology*, 65(1), 1–48. <https://doi.org/10.1111/j.1744-6570.2011.01237.x>
- Shahid, M. S., Bashir, S., & Fatima, S. (2024). Social media addiction and aggression in Pakistani young adults: Social connectedness as a moderator. *Journal of Professional & Applied Psychology*, 5(3), 424–433. <https://doi.org/10.52053/jpap.v5i3.308>
- Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, 7(3), 321–326. <https://doi.org/10.1089/1094931041291295>