NEW RESEARCH AGGRESSIVE BEHAVIOR

Check for updates

Using Machine Learning to Determine a Functional Classifier of Retaliation and Its Association With Aggression

Robert James Richard Blair, PhD, Johannah Bashford-Largo, MEd, MA, PLMHP, Ahria J. Dominguez, BA, Melissa Hatch, BS, Matthew Dobbertin, DO, Karina S. Blair, PhD, Sahil Bajaj, PhD

Objective: Methods to determine integrity of integrated neural systems engaged in functional processes have proven elusive. This study sought to determine the extent to which a machine learning retaliation classifier (retaliation vs unfair offer) developed from a sample of typically developing (TD) adolescents could be applied to an independent sample of clinically concerning youth and the classifier-determined functional integrity for retaliation was associated with antisocial behavior and proactive and reactive aggression.

Method: Blood oxygen level–dependent response data were collected from 82 TD and 120 clinically concerning adolescents while they performed a retaliation task. The support vector machine algorithm was applied to the TD sample and tested on the clinically concerning sample (adolescents with externalizing and internalizing diagnoses).

Results: The support vector machine algorithm was able to distinguish the offer from the retaliation phase after training in the TD sample (accuracy = 92.48%, sensitivity = 89.47%, and specificity = 93.18%) that was comparably successful in distinguishing function in the test sample. Increasing retaliation distance from the hyperplane was associated with decreasing conduct problems and proactive aggression.

Conclusion: The current study provides preliminary data of the importance of a retaliation endophenotype whose functional integrity is associated with reported levels of conduct problems and proactive aggression.

Plain language summary: This study used a machine learning retaliation classifier developed from a sample of typically developing adolescents and applied it to data from an independent clinical sample. Goal directed aggression in the clinically concerning youth related to a failure to recruit the neural systems implicated in retaliation. The current study suggests a marker of retaliation response for use as a treatment target.

Diversity & Inclusion Statement: We worked to ensure sex and gender balance in the recruitment of human participants. We worked to ensure race, ethnic, and/or other types of diversity in the recruitment of human participants. We worked to ensure that the study questionnaires were prepared in an inclusive way. We worked to ensure sex balance in the selection of non-human subjects. We worked to ensure diversity in experimental samples through the selection of the cell lines. We worked to ensure diversity in experimental samples through the selection of the genomic datasets. Diverse cell lines and/or genomic datasets were not available. One or more of the authors of this paper self-identifies as a member of one or more historically underrepresented racial and/or ethnic groups in science. We actively worked to promote sex and gender balance in our author group. We actively worked to promote inclusion of historically underrepresented racial and/or ethnic groups in science in our author group. While citing references scientifically relevant for this work, we also actively worked to promote sex and gender balance in our reference list. While citing references scientifically relevant for this work, we also actively worked to promote inclusion of historically underrepresented racial and/or ethnic groups in science in our reference list. The author list of this paper includes contributors from the location and/or community where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Key words: proactive aggression; reactive aggression; retaliation

JAACAP Open 2025;3(1):137-146.

Aggression can be seen in all mammals and results in harm to others.[1] It incurs significant costs for victims, perpetrators, and society, and it peaks during adolescence.[2] Although aggressive behavior emerges early in development and can be relatively commonly displayed, some individuals show aggression that is atypically frequent and severe.[3] Notably, aggression is a leading cause of all child and adolescent referrals to mental health clinicians.[4]

An important distinction has long been made between reactive and instrumental aggression.[5] Reactive aggression describes a retaliatory form of aggression elicited in response to a perceived threat, frustration, or provocation[6] that is characterized by high levels of autonomic arousal and negative emotions such as anger and fear.[7,8] Instrumental aggression is goal-directed behavior (ie, the person uses aggression to achieve the their goals) and is not typically associated with heightened negative emotion.[6,7] One complication to this dichotomy is that while retaliatory, reactive aggression can be relatively planned[6]; the aggressor chooses their weapon and when they are going to attack. Indeed, common measures of reactive aggression such as the Point Subtraction Aggression Paradigm,[9] the Taylor aggression paradigm,[10] or our own retaliation paradigm[11] all assess this instrumental form of reactive aggression; the participant is provoked by the paradigm and frequently indicates anger/irritation at their treatment, but they choose via button press whether they will retaliate and/or the degree of retaliation.

Building on the work of animal researchers regarding the role of acute threat circuitry mediating automatic retaliatory aggression,[12] Blair[6,13] proposed a neurocognitive account of this form of reactive aggression implicating 3 sets of systems (see below). Others hold similar positions (eg, [14–16]), although some consider one of these sets of systems, the acute threat circuitry, as less important in human aggression.[17] Neural systems implicated include ventromedial prefrontal cortex (vmPFC)/rostromedial PFC (rmPFC), dorsomedial PFC (dmPFC), inferior frontal cortex (iFC), anterior insular cortex (aIC), and amygdala.[6,11,13,14,18,19] Suggestions are that vmPFC/rmPFC allow the representation of expected value[20] and are involved in the instrumental components of response selection; dmPFC, aIC/iFC, and caudate are implicated in response control generally and thus response selection[21]; and systems acutely responding to threat (amygdala, hypothalamus, periaqueductal gray [PAG]) mediate the reactive aggression response (for details, see [6,13]).

Recent empirical work from Blair's group used this theoretical model to examine the association of atypical responsiveness within the 3 components described above and observed and recorded aggressive episodes in residential care.[22] The study used a functional magnetic resonance imaging (fMRI) retaliation task paradigm previously used with adolescent participants.[11] Task-relevant neural systems were identified from the data of an independent sample of typically developing (TD) adolescents. These neural systems were then interrogated in a group of participants in residential care during performance of the retaliation paradigm.[22] Results indicated that aggressive adolescents showed reduced downregulation of activity within regions involved in representing the expected value of choice options (left vmPFC) and reduced recruitment of regions implicated in response control (right iFC and bilateral aIC) as a function of retaliation level. It was argued that this indicated that individuals with a greater propensity for aggression show reduced representation of the negative consequences of retaliation and associated reduced recruitment of regions potentially involved in overruling these negative consequences to engage in retaliation.

While the analysis pathway of the Mathur *et al.* study[22] was theoretically motivated, it was still at risk of a difficulty with standard univariate analysis approaches that has been termed blobology, ie, a lack of clarity with respect to determining the extent to which a regional activation in one study is replicated in another (eg, [23]). Indeed, it is notable that across studies purporting to examine the same neurocognitive architecture there are differences in which precise regions are regarded to be significantly implicated.[11,14,18,19] It is likely that in many of these studies other portions of the network would have been implicated, but not at a sufficient level to satisfy current multiple comparison thresholds. This concern with blobology comes at a time when there is greater interest in potential endophenotypes (cf. [24]) and, more generally, the possibilities of the Research Domain Criteria,[25,26] ie, the development of neurobiology-based biomarkers relevant to psychiatric disorders. Recently, there have been suggestions that multivariate analyses might provide greater hope with respect to endophenotype specification and the development of psychiatry-relevant neurocognitive biomarkers.[27,28] Specifically, it has been suggested that machine learning could be used to determine neurocognitive function classifiers—effectively markers of the functional integrity of specific functional processes (for additional details, see Supplemental 1.1, available online) (cf. [29,30]). The integrity of specific functions in individual patients can then be examined with such classifiers.

The goals of the current study were to determine the extent to which a retaliation classifier (retaliation vs unfair offer) with reasonably high accuracy can be developed from a sample of healthy TD adolescents and then applied to an independent sample of clinically concerning youth and the classifier-determined functional integrity for retaliation is associated with aggression levels as indexed by the conduct problems subscale of the Strengths and Difficulties Questionnaire Conduct Problems subscale (SDQ-CP)[31] and the Reactive Proactive Aggression Questionnaire (RPAQ).[32] In particular, we were interested in determining the association of functional integrity (here defined as distance from hyperplane [DFH]) with levels of reactive and proactive aggression, respectively. Our previous study examined observed aggression in a partially overlapping sample.[22] However, the aggressive episodes were not coded by observers with respect to whether they were reactive or proactive. The current study addresses this gap. We predicted that a classifier could be developed with reasonably high accuracy and that classifier-determined functional integrity (DFH) would be negatively associated with aggression levels. Based on the existing literature, predictions with respect to reactive vs proactive aggression were unclear. The retaliation task was developed to index a more reactive form of aggression.[11] As such, it is predicted that exaggerated integrity might be associated with increased reactive aggression. However, as noted above, the retaliation paradigm involves considerable goal direction with respect to response selection (cf. [6]). Reduced retaliation-decision$_{DFH}$ might thus be associated with greater levels of proactive aggression (see [6]).

METHOD

Participants

The current study involves a reanalysis of previously published data[22] as well as new data. Thus, of the 82 TD participants on whose data the support vector machine (SVM) learning was conducted (TD training sample) (mean [SD] age = 13.97 [2.48] years, 39 TD females), the data of 49 were used in the earlier study.[22] Similarly, of the 120 clinically concerning participants on whom the classifier was tested (mean [SD] age = 15.61 [2.09] years, 47 TD females), 43 (35.83%) had been previously involved in the earlier study.[22] Participants were recruited from a residential care facility and the surrounding community. Participants recruited from the residential facility had been referred for severe behavioral and mental health problems. Participants from the community were recruited through flyers or social media. Clinical characterization of all participants was done through psychiatric interviews by licensed and board-certified child and adolescent psychiatrists with the participants and their parents to adhere closely to common clinical practice. Institutional review board approval was acquired before data collection began. For details on exclusion criteria and consent procedure, see Supplement 1.2, available online.

Study Design and Measures

fMRI Paradigm. Participants were presented with a version of the retaliation task, which has been previously described[22] and which, in turn, is a version of the Ultimatum Game (Figure S1, available online). At the onset of each trial, the partner had \$20 and the participant had \$3. During the offer phase (3,000 ms), the partner offered to divide their \$20 with the participant either fairly (eg, \$10 to participant and \$10 to partner) or unfairly (eg, \$6, \$4, \$2 to participant and \$14, \$16, \$18 to partner). During the decision phase (4,000 ms), the participant used a button press to indicate whether they wished (button 1) to accept the partner's offer and give up their \$3 in exchange (eg, receive \$10 from the fair offer) or (buttons 2-4) to spend some of their \$3 to punish to the partner. Each decision corresponded to a different button. Participants had the option to spend \$1, \$2, or \$3 and cost the partner \$7, \$14, or \$21, respectively. Participants were informed that they would be playing with various partners and that their task was simply to decide how to respond to their partner's behavior. They were not directed to maximize profit, and they did not receive money from engaging in this task. For details on fMRI parameters, fMRI data processing, and individual-level analysis, see Supplement 1.4, available online.

Questionnaires. Both the parent-report SDQ-CP[31] and the RPAQ[32] were used to assess aggression. The Childhood Trauma Questionnaire (CTQ)[33] was used to assess severity of exposure to abuse and neglect (for further details, see Supplement 1.3, available online). IQ was assessed via the Wechsler Abbreviated Scale of Intelligence II (WASI-II) (perceptual reasoning and verbal comprehension subscales).[34]

Data Analysis

Behavioral data, correlation, and regression analyses were conducted using IBM SPSS version 25 (IBM Corp., Armonk, New York).

Demographic and Clinical Data. Demographic (age, IQ, and sex ratios) and clinical (self-report assessments) scores are presented for the 82 TD training sample participants on whom the SVM learning was conducted and the 120 clinically concerning participants.

Feature Creation. The whole brain was parcellated into 400 regional parcellations using Schaefer's atlas[35] (for additional details, see Supplement 1.5, available online). Subjectwise and hemispheric-wise blood oxygen level–dependent (BOLD) responses during the offer and retaliation phase were determined for each of the 400 regions. BOLD responses were modulated by offer level (offer phase) and retaliation level (retaliation phase). Subjectwise and hemispheric-wise BOLD response during the offer and retaliation phase were also determined for 10 bilateral subcortical regions (caudate, putamen, nucleus accumbens, hippocampus, and amygdala) as well as a unilateral 5-mm sphere focused on the PAG coordinates (x, y, z = 10, −21, 4) taken from previous work using the retaliation task.[11] The whole-brain parcellation into 400 bilateral cortical and 11 subcortical regions was performed using the mri_surf2surf, mris_anatomical_stats, and aparcstats2table pipelines following the FreeSurfer recon-all pipeline.[36]

Feature Selection and Machine Learning Analysis. Feature selection and SVM learning was conducted on the data from the TD training sample. For details, see Supplement 1.5, available online (cf. [37]). The results of this training were then applied to the clinically concerning participants to determine its performance in an independent dataset.

Correlational Analyses. Correlation analyses were conducted to examine associations between the aggression, demographic, maltreatment, diagnostic category, and DFH variables. These were conducted for the clinical test sample; comparable analyses for the whole sample are included in Supplement 1.6, available online.

Multiple Regression Analyses Examining Prediction for Aggression. To determine the association between DFH (decision and offer trials) and aggression history (as indexed by the SDQ-CP subscale and reactive and proactive RPAQ subscale scores), we employed 3 stepwise multiple regression analyses with each of the 3 aggression history measures as the outcome variables. Retaliation-decision$_{\mathrm{DFH}}$ and offer$_{\mathrm{DFH}}$ were entered as predictor variables as well as other variables associated with aggression: IQ, sex, age, neglect, and abuse (measured by CTQ).

Multiple Regression Analysis Predicting DFH for Retaliation Decisions by Diagnostic Category. A fourth exploratory multiple regression analysis sought to examine the extent to which psychiatric diagnosis (conduct disorder [CD], oppositional defiant disorder [ODD], attention-deficit/hyperactivity disorder, posttraumatic stress disorder, generalized anxiety disorder, and major depressive disorder), as predictor variables, could predict DFH for retaliation decision trials.

TABLE 1 Participant Characteristics

	Typically developing training sample (n = 82)		Clinical sample (n = 120)			
	Mean	**(SD)**	**Mean**	**(SD)**	***F***	***p***
Dimensional variables						
Age, y	13.97	(2.48)	15.61	(2.09)	25.95	.000
IQ	107.51	(13.28)	101.67	(11.79)	1.80	.001
Abuse[a]	16.69	(3.84)	23.60	(1.51)	32.07	.000
Neglect[b]	11.37	(2.38)	17.57	(7.84)	47.57	.000
Categorical variables						
	n	(%)	n	(%)	χ^2	*p*
Male sex	39	(48)	73	(61)	3.50	.063
CD	0	(0)	61	(51)	83.94	.000
ODD	0	(0)	76	(63)	14.23	.000
ADHD	0	(0)	74	(62)	13.61	.000
MDD	0	(0)	17	(35)	13.40	.000
GAD	0	(0)	42	(48)	43.72	.000
PTSD	0	(0)	21	(38)	17.22	.000

Note: ADHD = attention-deficit/hyperactivity disorder; CD = conduct disorder; GAD = generalized anxiety disorder; MDD = major depressive disorder; ODD = oppositional defiant disorder; PTSD = Posttraumatic stress disorder.
[a]*Abuse determined by Childhood Trauma Questionnaire abuse score (sexual + physical + emotional abuse scores).*
[b]*Neglect determined by Childhood Trauma Questionnaire neglect score (physical + emotional neglect scores).*

RESULTS

Demographic and Clinical Data

Demographic and clinical data for the sample are provided in Table 1. As can be seen, there were significant differences between the TD training set and the clinical sample in demographic variables (age and IQ) and experience of maltreatment.

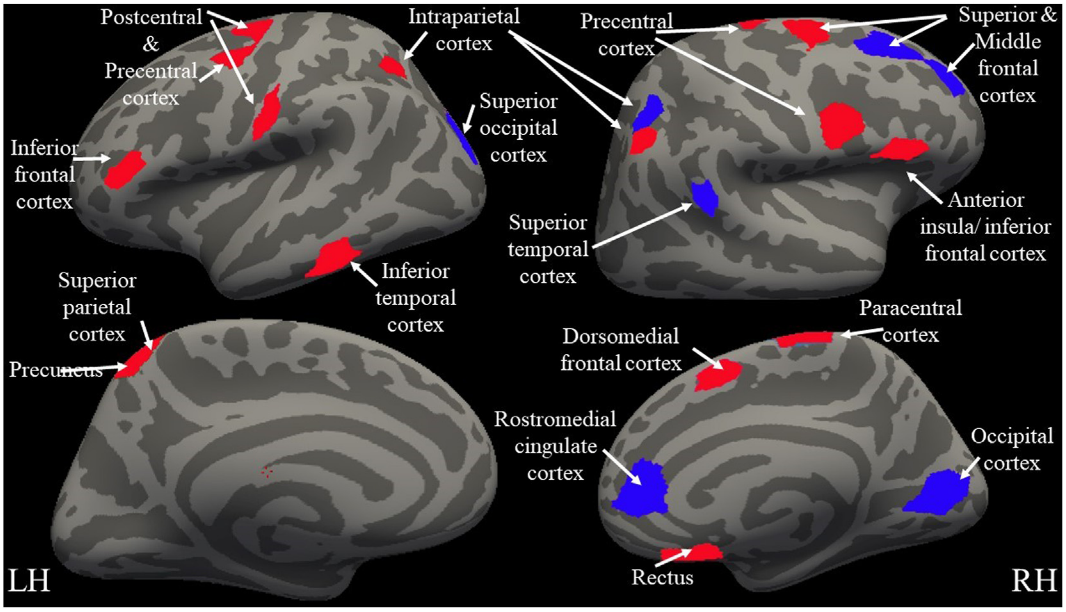
Feature Identification

Of 27 features identified (Figure 1; Table S1, available online), 19 features (including dmPFC and caudate) showed significantly greater responses as a function of retaliation rather than offer level, while 8 (including rmPFC and nucleus accumbens) showed significantly greater responses as a function of offer rather than retaliation level (Figure 1).

Correlational Analyses

Table 2 shows the results of the correlation analyses (clinical sample only). These analyses revealed significant positive correlations between the aggression variables (SDQ-CP subscale and scores on both the reactive and the proactive RPQ subscales), externalizing diagnoses, IQ and age (though not sex), maltreatment (abuse and neglect exposure), as well as retaliation DFH (retaliation-decision$_{\mathrm{DFH}}$) (this was not significant for reactive aggression as indexed by the RPQ). Retaliation-decision$_{\mathrm{DFH}}$ was not significantly associated with demographic variables but was significantly associated with ODD diagnostic status (but not attention-deficit/hyperactivity disorder and only at trend levels for CD). Table S2, available online, shows the results of the same analysis but also including the TD training sample. The majority of the findings seen with the clinical sample were also seen in the data from the entire sample.

FIGURE 1 Features Selected From the Typically Developing Training Sample



Note: The 27 features, identified via least absolute shrinkage and selection operator (see Supplement 1.5, available online) differentiating the blood oxygen level–dependent response as a function of retaliation level vs offer level (nucleus accumbens, caudate, hippocampus, and periaqueductal gray are not depicted). Red regions correspond to regions showing greater modulation as a function of retaliation level relative to offer level. Blue regions correspond to regions showing greater modulation as a function of offer level relative to retaliation level. LH = left hemisphere; RH = right hemisphere.

Generalized Model Performance and Feature Identification

With respect to the BOLD response data for responses during retaliation vs offer, our SVM learning analysis identified a hyperplane during training with accuracy of 92.48%, sensitivity of 89.47%, and specificity of 93.18%. Notably, sensitivity (correct recognitions of retaliation data) and specificity (correct recognitions of offer level) were high for the independent clinical participant sample (accuracy = 83.33%, sensitivity = 79.17%, and specificity = 87.50%).

Multiple Regression Analyses Examining Prediction for Aggression

The stepwise regression to predict SDQ-CP subscale score was significant ($F_{2,111} = 7.08$, $p = .001$; $R^2 = 0.113$). Participants' predicted SDQ-CP subscale scores were equal to 12.076 − (0.354 retaliation-decision$_{\text{DFH}}$ + 0.066 IQ); for additional details see Supplement 1.6.1, available online.

The stepwise regression to predict the proactive RPQ subscale score was significant ($F_{1,82} = 4.31$, $p = .041$; $R^2 = 0.05$). Participants' predicted proactive RPQ subscale scores were equal to 7.206 − 0.464 retaliation-decision$_{\text{DFH}}$; for additional details see Supplement 1.6.2, available online.

The stepwise regression to predict the reactive RPQ subscale score was significant ($F_{1,82} = 4.30$, $p = .041$; $R^2 = 0.05$). Participants predicted proactive RPQ subscale scores were equal to 8.309 + 0.900 offer$_{\text{DFH}}$; for additional details see Supplement 1.6.3, available online.

Multiple Regression Analysis Predicting DFH for Retaliation Decisions as a Function of Diagnostic Category

The stepwise regression to predict retaliation-decision$_{\text{DFH}}$ was significant ($F_{1,82} = 4.304$, $p = .041$; $R^2 = 0.050$). Participants' predicted retaliation-decision$_{\text{DFH}}$ level was equal to 8.309 + 0.900 ODD (with ODD diagnosed scored as a 1, not scored as a 0); for additional details, see Supplement 1.6.4, available online.

The results of all 4 multiple regressions performed on the entire sample, not only the participants with externalizing/internalizing diagnoses, are reported in Supplement 1.7, available online. This second series of multiple regressions largely mirrored the results of the main analyses (see detailed description in Supplement 1.7, available online).

DISCUSSION

The current study aimed to determine the extent to which a retaliation classifier could be developed and classifier-determined functional integrity for retaliation (retaliation-decision$_{\text{DFH}}$) would be associated with reactive and/or proactive aggression. There were 2 main findings: First, a

TABLE 2 Results of Correlation Analyses (Clinical Sample Only)

	Aggression scales			Diagnostic categories						Demographic variables			Maltreatment		Distance from hyperplane	
	SDQ-CP	**RPAQ Reactive**	**RPAQ Proactive**	**CD**	**ODD**	**ADHD**	**MDD**	**GAD**	**PTSD**	**Age**	**Sex**	**IQ**	**Abuse[a]**	**Neglect[b]**	**Offer[c]**	**Retaliation decision[d]**
SDQ-CP	1	0.776***	0.770***	0.729***	0.736***	0.529***	0.099	0.309***	0.278***	0.316***	0.117	−0.306***	0.345***	0.420***	0.087	−0.216*
RPAQ Reactive	0.776***	1	0.834***	0.556***	0.653***	0.471***	0.048	0.240***	0.292***	0.245***	0.094	−0.238***	0.268***	0.321***	0.223*	−0.147
RPAQ Proactive	0.770***	0.834***	1	0.509***	0.604***	0.443***	0.057	0.285***	0.325***	0.227***	0.064	−0.192*	0.333***	0.332***	0.086	−0.223*
CD	0.729***	0.556***	0.509***	1	0.647***	0.462***	0.111	0.221***	0.271***	0.343***	0.091	−0.233***	0.339***	0.400***	0.080	−0.136
ODD	0.736***	0.653***	0.604***	0.647***	1	0.555***	0.096	0.257***	0.238***	0.336***	0.141*	−0.188***	0.345***	0.437***	0.158	−0.266***
ADHD	0.529***	0.471***	0.443***	0.462***	0.555***	1	0.103	0.319***	0.078	0.115	0.185***	−0.072	0.313***	0.285***	0.092	−0.094
MDD	0.099	0.048	0.057	0.111	0.096	0.103	1	0.328***	0.423***	0.202***	−0.195***	−0.014	0.429***	0.413***	0.124	0.114
GAD	0.309***	0.240***	0.285***	0.221***	0.257***	0.319***	0.328***	1	0.505***	0.232***	−0.105	−0.08	0.466***	0.292***	−0.039	0.083
PTSD	0.278***	0.292***	0.325***	0.271***	0.238***	0.078	0.423***	0.505***	1	0.208***	−0.119	−0.084	0.583***	0.446***	0.052	0.041
Age	0.316***	0.245***	0.227***	0.343***	0.336***	0.115	0.202***	0.232***	0.208***	1	0.078	−0.276***	0.228***	0.182***	0.090	−0.020
Sex	0.117	0.094	0.064	0.091	0.141*	0.185***	−0.195***	−0.105	−0.119	0.078	1	−0.048	−0.086	0.073	0.008	0.009
IQ	−0.306***	−0.238***	−0.192*	−0.233***	−0.188***	−0.072	−0.014	−0.08	−0.084	−0.276***	−0.048	1	−0.220***	−0.200***	−0.115	0.060
Abuse[a]	0.345***	0.268***	0.333***	0.339***	0.345***	0.313***	0.429***	0.466***	0.583***	0.228***	−0.086	−0.220***	1	0.655***	0.039	−0.029
Neglect[b]	0.420***	0.321***	0.332***	0.400***	0.437***	0.285***	0.413***	0.292***	0.446***	0.182***	0.073	−0.200***	0.655***	1	0.054	−0.044

Note: *ADHD = attention-deficit/hyperactivity disorder; CD = conduct disorder; GAD = generalized anxiety disorder; MDD = major depressive disorder; ODD = oppositional defiant disorder; PTSD = Posttraumatic stress disorder; RPAQ Proactive = Reactive Proactive Aggression Questionnaire Proactive subscale; RPAQ Reactive = Reactive Proactive Aggression Questionnaire Reactive subscale; SDQ=CP = Strengths and Difficulties Questionnaire Conduct Problems subscale.*
[a]*Abuse determined by Childhood Trauma Questionnaire abuse score (sexual + physical + emotional abuse scores).*
[b]*Neglect determined by Childhood Trauma Questionnaire neglect score (physical + emotional neglect scores).*
[c]*Offer distance from hyperplane.*
[d]*Retaliation decision distance from hyperplane.*
$^{*}p < .05$; $^{***}p < .001$.

high-accuracy hyperplane distinguishing the BOLD response to retaliation decision level vs unfair offer level in a sample of TD adolescents could be transferred to an independent sample of clinically concerning adolescents to accurately distinguish these 2 functional responses. Second, there were significant negative associations between retaliation-decision$_{DFH}$ and general conduct problems (SDQ-CP subscale score) and level of proactive, rather than reactive, aggression.

Our first aim was to determine the extent to which we could use SVM learning to identify a hyperplane differentiating the neural responses to retaliation decision level vs unfair offer level in a TD sample that would successfully identify these responses in an independent sample of clinically concerning adolescents. This was clearly achieved. While retaliation-decision$_{DFH}$ was associated with individual differences in aggression in this clinically concerning sample, the hyperplane was highly accurate at distinguishing the functional responses in the group as a whole.

Related to our first aim was the extent to which the features making up the hyperplane would mirror those previously identified in the literature. Three sets of systems have been implicated in previous work (eg, [6,11,13,14,18,19]). The first set includes vmPFC, which allows the representation of expected value.[20] An rmPFC region consistent with this was one of the features identified. Consistent with previous work, activity within this region decreased as a function of retaliation level in the current study (it showed greater activity during the offer phase). Nucleus accumbens, a region frequently implicated in reward value coding (eg, [38]), was identified as a feature and showed decreased responding as a function of retaliation level. However, previous work has not typically implicated nucleus accumbens in retaliation[6,11,13,14,18,19] except for one study that reported increased nucleus accumbens activity during retaliation following provocation.[39] The second set includes dmPFC, iFC, and aIC. These regions are consistently implicated in response control/selection.[21] An additional region implicated in response control, caudate, has received relatively little attention in the previous retaliation literature (though see [17,40]). However, caudate was identified as a feature in the current study (showing greater responses during retaliation). The third set includes the acute threat circuitry (amygdala, hypothalamus, and PAG) implicated in animal aggression (eg, [12]). Some have argued that these regions are less implicated in human reactive aggression (eg, [17]), and certainly hypothalamus has not been well documented in human fMRI work (though see [41]). Moreover, amygdala findings may be more common in examinations of retaliation in patient samples[11,18] than such studies in healthy participants.[17,40] However, PAG has been relatively consistently documented particularly in work with this paradigm[11,22,40] and was a feature identified in the current study. In short, the features making up the hyperplane largely showed consistency with the previous literature.

Our second goal was to determine the extent to which the DFH (ie, functional integrity) for retaliation decision level might be associated with conduct problems and proactive aggression and/or reactive aggression. One view has been that increased acute threat responsiveness accompanied by dysfunctional representation of reinforcement value and/or response control would be associated with an increased risk for reactive aggression.[6,13–16] Several studies have shown that individuals at selective risk for reactive aggression show this heightened responsiveness to emotional stimuli and particularly social provocation.[11,18,42,43] However, as noted, retaliation can be goal directed—the weapon (or button press) is selected.[6] As such, the systems implicated in response selection (either via representing value or via response control) may play a disproportional role. The idea is that individuals showing reduced representation of the negative consequences of retaliation will show associated reduced recruitment of regions potentially involved in overruling these negative consequences to engage in retaliation (see [22]). Notably, previous work has suggested that individuals showing reduced representation of the negative consequences of aggression, particularly the distress of the victim, are at particular risk of proactive aggression.[6,44–46] In short, the current findings are notably compatible with those of Mathur *et al.*[22] (there was also partial participant overlap; see "Method"). Importantly, they extend the findings of Mathur *et al.*[22] to indicate that reduced expression of the neural response during retaliation increased the risk for conduct problems and particularly proactive aggression.

While this study has several strengths, including the use of a well-established fMRI task assessing retaliation and a novel approach to analysis, several limitations should be considered. First, there were no indications that a heightened retaliation response—greater DFH—was associated with reactive aggression as might have been expected.[11,18,42,43] A heightened response to unfair offers was associated with reactive aggression. If this is a robust finding, it would be theoretically interesting. However, we note that while the retaliation-decision$_{DFH}$–conduct problems/proactive aggression findings were maintained when the whole sample was considered (ie, the TD training data was added) (see Supplement 1.7, available online), the offer response reactive aggression association was not. In short, we are very cautious about interpreting this finding. Instead, we believe that the retaliation-decision$_{DFH}$–conduct problems/proactive aggression findings reflect the makeup of our participant sample, in which more than 50% met the criterion for a CD diagnosis. We believe that a clinical sample in which there was a greater proportion of patients with internalizing conditions, without CD, would provide greater numbers of participants with a more selective presentation of reactive aggression (cf. reactive aggression internalizing). We believe the application of the hyperplane to data from such a sample would reveal an association between increased retaliation-decision$_{DFH}$ and reactive aggression. Second, we did not implement a structured or semistructured diagnostic interview. As such, our diagnostic reliability can be challenged (cf. [47]). Of course, while the use of structured interviews such as the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS)[48] improve reliability scores, κ values often remain below 0.8 indicating nonoptimal reliability.[49] As such, there is always concern regarding the accuracy of associations between diagnosis and functional impairment. Indeed, that concern led to the design of this study. Importantly, here functional integrity (DFH) was related to conduct problems and forms of aggression using assessments of these behaviors of proven test-retest reliability (eg, [50,51]). The predictive associations between diagnostic status and retaliation-decision$_{DFH}$ were purely exploratory. Third, recent work has challenged the use of contrast-based data, rather than individual regressor based analyses, in individual differences research because of data reliability concerns. It is important to note here, though: Group-based analyses are relatively insensitive to the reliability concerns of contrast-based analyses (see [28,52]), and the determination of the hyperplane was derived from a group-based analysis. Also, all the individual difference analyses in this study were based on a single condition—distance of the retaliation response from the hyperplane (ie, not a contrast against another condition). Fourth, increasing the size of the training dataset has advantages in machine learning analyses (eg, [53]), and the training set here ($n = 82$) could be considered small. However, it is important to note that training set size is partly a function of data complexity,[53] and the critical question is applicability of the obtained hyperplane to independent datasets. In the current study, the hyperplane allowed the functional characterization of a heterogeneous group of patients. Fifth, it could be argued that the current task indexes altruistic punishment rather than retaliation (cf.[54]). Altruistic acts are engaged in for the well-being of others (eg, [55]), and altruistic punishment refers to punishment to the perpetrator by the victim, which incurs a cost to the victim,[56] to reinforce the social norm (cf.[57]). Notably, as documented in the recent review by Boccadoro *et al.*,[54] contrasts associated with altruistic punishment (eg, unfair vs fair offers and rejected vs accepted offers on the Ultimatum Game) induce activations within the second set of systems that Blair[6] implicates in reactive aggression, ie, dmPFC, iFC, and aIC—the regions consistently implicated in response control/selection.[21] This altruistic punishment vs retaliation concern cannot be discounted. It rests on the assumption that individuals punished in our task according to a motivation to reinforce social norms (ie, for the benefit of others) rather than as an angry retaliatory response to provocation (the unfairness of the offer). However, motivations are difficult to reliably determine. Importantly, though, "studies consistently demonstrate that unfair offers induce anger, and that rejecting such offers relates to aggression."[58(p84)] As such, the paradigm remains highly relevant to understanding aggression.

With respect to clinical implications of the current work, it is important to remember that currently interventions designed to decrease aggression are not tailored to the specific difficulties of the individual (eg, [59,60]). Thus, while we may be able to show biological-level differences between children showing difference types of aggression (reactive vs instrumental[6]), that information does not help the clinician working with the individual child. Yet, the treatment implications for a child with very high threat responsiveness and reactive aggression are rather different from the treatment implications for a child who retaliates due to indifference to their victims.[6] Critically, this article demonstrates that it is possible to develop a classifier that identifies the functional integrity of systems engaged during retaliation. For potential clinical application, the next step would be to determine percentile scores of this integrity (based on training set DFH standard deviations from the training set mean DFH) that can be applied to individual participant data as treatment targets and indices of individual treatment responsiveness. Of course, it is critical to determine that this signal shows high test-retest reliability.[28,61] However, should such reliability be demonstrable, functional classifiers such as that developed in this study could lead to an individualized approach to intervening with aggressive adolescents. Moreover, it will be critical to determine that such functional classifiers apply equally accurately across diverse populations or whether optimization will require population-specific classifier development.

In conclusion, we showed that SVM learning could accurately distinguish responses during retaliation from responses to unfair offers that could be validated in an independent patient sample. Patients showed a negative association between the functional integrity of their retaliation response (retaliation-decision$_{DFH}$) and conduct problems, particularly proactive aggression. As such, the distance from the hyperplane of neural response during retaliation could be an important biomarker at the individual level for treatment stratification and prediction of a patient's aggression propensity.

CRediT authorship contribution statement

Robert James Richard Blair: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Johannah Bashford-Largo:** Writing – review & editing, Project administration, Investigation. **Ahria J. Dominguez:** Writing – review & editing, Project administration, Data curation. **Melissa Hatch:** Writing – review & editing, Project administration, Data curation. **Matthew Dobbertin:** Writing – review & editing, Conceptualization. **Karina S. Blair:** Writing – review & editing, Project administration, Conceptualization. **Sahil Bajaj:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization.

This article is part of a special series devoted to addressing aggressive behavior as a focus of psychiatric attention and how its manifestations and treatment needs may vary across psychiatric disorders. The series is edited by Guest Editor Joseph Blader, PhD, Deputy Editor Robert Findling, MD, MBA, and Editor Manpreet K. Singh, MD, MS.

Accepted May 29, 2024.

Robert James Richard Blair is with the Copenhagen University Hospital, Copenhagen, Denmark, and the University of Copenhagen, Denmark. Johannah Bashford-Largo, Matthew Dobbertin, and Karina S. Blair are with Boys Town National Research Hospital, Boys Town, Nebraska. Johannah Bashford-Largo is also with the University of Nebraska—Lincoln, Lincoln, Nebraska. Ahria J. Dominguez and Melissa Hatch are with the University of Nebraska Medical Center, Omaha, Nebraska. Sahil Bajaj with the University of Texas MD Anderson Cancer Center, Houston, Texas.

This work was supported by Boys Town National Research Hospital. This research was in part supported by the National Institute of Mental Health under award number K22-MH109558 (to R.J.R.B.). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The research was performed with permission from Boys Town Institutional Review Board.

Consent has been provided for descriptions of specific patient information.

Data Sharing: Deidentified participant data are available by request: robert.james.blair@regionh.dk.

The authors wish to thank all participants of this study.

Disclosure: Robert James Richard Blair, Johannah Bashford-Largo, Ahria J. Dominguez, Melissa Hatch, Matthew Dobbertin, Karina S. Blair, and Sahil Bajaj have reported no biomedical financial interests or potential conflicts of interest.

Correspondence to Robert James Richard Blair, PhD, Child and Adolescent Mental Health Centre, Capital Region of Psychiatry, Gentofte Hospitalsvej 3A, 1. Sal, 2900 Hellerup, Capital Region of Denmark, Copenhagen, Denmark; e-mail: robert.james.blair@regionh.dk

https://doi.org/10.1016/j.jaacop.2024.04.007

REFERENCES

1. Austerman J. Violence and aggressive behavior. Pediatr Rev. 2017;38(2):69-80. https://doi.org/10.1542/pir.2016-0062
2. Farrington DP. Early predictors of adolescent aggression and adult violence. Violence Vict. 1989;4:79-100.
3. Keenan K, Shaw DS, Delliquadri E, Giovannelli J, Walsh B. Evidence for the continuity of early problem behaviors: application of a developmental model. J Abnorm Child Psychol. 1998;26:443-454.
4. Coghill D. Editorial: do clinical services need to take conduct disorder more seriously? J Child Psychol Psychiatry. 2013;54:921-923.
5. Dodge KA. Social cognition and children's aggressive behaviour. Child Dev. 1980;51: 162-170.
6. Blair RJ. The motivation of aggression: a cognitive neuroscience approach and neurochemical speculations. Motiv Sci. 2022;8(2):106-120. https://doi.org/10.1037/mot0000247
7. Scarpa A, Raine A. Psychophysiology of anger and violent behavior. Psychiatr Clin North Am. 1997;20(2):375-394. https://doi.org/10.1016/s0193-953x(05)70318-x
8. Puhalla AA, McCloskey MS. The relationship between physiological reactivity to provocation and emotion dysregulation with proactive and reactive aggression. Biol Psychol. 2020;155:107931. https://doi.org/10.1016/j.biopsycho.2020.107931
9. Cherek DR, Moeller FG, Schnapp W, Dougherty DM. Studies of violent and nonviolent male parolees: I. laboratory and psychometric measurements of aggression. Biol Psychiatry. 1997;41:514-522.
10. Taylor SP. Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. J Pers. 1967;35:297-310.
11. White SF, VanTieghem M, Brislin SJ, *et al*. Neural correlates of the propensity for retaliatory behavior in youths with disruptive behavior disorders. Am J Psychiatry. 2016; 173(3):282-290. https://doi.org/10.1176/appi.ajp.2015.15020250
12. Panksepp J, Knutson B, Burgdorf J. The role of brain emotional systems in addictions: a neuro-evolutionary perspective and new 'self-report' animal model. Addiction. 2002; 97(4):459-469.
13. Blair RJ. The roles of orbital frontal cortex in the modulation of antisocial behavior. Brain Cogn. 2004;55(1):198-208.
14. Repple J, Pawliczek CM, Voss B, *et al*. From provocation to aggression: the neural network. BMC Neurosci. 2017;18(1):73. https://doi.org/10.1186/s12868-017-0390-z
15. Bertsch K, Florange J, Herpertz SC. Understanding brain mechanisms of reactive aggression. Curr Psychiatry Rep. 2020;22(12):81. https://doi.org/10.1007/s11920-020-01208-6
16. Lischinsky JE, Lin D. Neural mechanisms of aggression across species. Nat Neurosci. 2020;23(11):1317-1328. https://doi.org/10.1038/s41593-020-00715-2
17. Dugré JR, Potvin S. Neural bases of frustration-aggression theory: A multi-domain meta-analysis of functional neuroimaging studies. J Affect Disord. 2023;331:64-76. https://doi.org/10.1016/j.jad.2023.03.005
18. da Cunha-Bang S, Fisher PM, Hjordt LV, *et al*. Violent offenders respond to provocations with high amygdala and striatal reactivity. Soc Cogn Affect Neurosci. 2017;12(5): 802-810. https://doi.org/10.1093/scan/nsx006
19. Skibsted AP, Cunha-Bang SD, Carré JM, *et al*. Aggression-related brain function assessed with the point subtraction aggression paradigm in fMRI. Aggress Behav. 2017;43(6): 601-610. https://doi.org/10.1002/ab.21718
20. O'Doherty JP, Lee SW, McNamee D. The structure of reinforcement-learning mechanisms in the human brain. Curr Opin Behav Sci. 2015;1:94-100.
21. Hannah R, Aron AR. Towards real-world generalizability of a circuit for action-stopping. Nat Rev Neurosci. 2021;22(9):538-552. https://doi.org/10.1038/s41583-021-00485-1
22. Mathur A, Bashford-Largo J, Elowsky J, *et al*. Association between aggression and differential functional activity of neural regions implicated in retaliation. J Am Acad Child Adolesc Psychiatry. 2023;62(7):805-815. https://doi.org/10.1016/j.jaac.2023.01.021
23. Hanson SJ. The failure of blobology: fMRI misinterpretation, maleficience and muddle. Front Hum Neurosci. 2022;16:870091. https://doi.org/10.3389/fnhum.2022.870091
24. Miskowiak KW, Seeberg I, Kjaerstad HL, *et al*. Affective cognition in bipolar disorder: a systematic review by the ISBD targeting cognition task force. Bipolar Disord. 2019; 21(8):686-719. https://doi.org/10.1111/bdi.12834
25. Insel TR. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. Am J Psychiatry. 2014;171(4):395-397. https://doi.org/10.1176/appi.ajp.2014.14020138

26. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 2013;11:126. https://doi.org/10.1186/1741-7015-11-126
27. Han X, Ashar YK, Kragel PA, *et al*. Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. Neuroimage. 2022;247:118844. https://doi.org/10.1016/j.neuroimage.2021.118844
28. Blair RJ, Mathur A, Haines N, Bajaj S. Future directions for cognitive neuroscience in psychiatry: recommendations for biomarker design based on recent test re-test reliability work. Curr Opin Behav Sci. 2022;44:101102. https://doi.org/10.1016/j.cobeha.2022.101102
29. Speer SPH, Keysers C, Barrios JC, *et al*. A multivariate brain signature for reward. Neuroimage. 2023;271:119990. https://doi.org/10.1016/j.neuroimage.2023.119990
30. Yu H, Koban L, Crockett MJ, Zhou X, Wager TD. Toward a brain-based bio-marker of guilt. Neurosci Insights. 2020;15:2633105520957638. https://doi.org/10.1177/2633105520957638
31. Goodman R. The Strengths and Difficulties Questionnaire: a research note. J Child Psychol Psychiatry. 1997;38(5):581-586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x
32. Raine A, Dodge K, Loeber R, *et al*. The Reactive-Proactive Aggression Questionnaire: differential correlates of reactive and proactive aggression in adolescent boys. Aggress Behav. 2006;32(2):159-171. https://doi.org/10.1002/ab.20115
33. Bernstein DP, Ahluvalia T, Pogge D, Handelsman L. Validity of the Childhood Trauma Questionnaire in an adolescent psychiatric population. J Am Acad Child Adolesc Psychiatry. 1997;36(3):340-348. https://doi.org/10.1097/00004583-199703000-00012
34. Wechsler D. Wechsler Abbreviated Scale of Intelligence. Second Edition (WASI-II). San Antonio, TX: NCS Pearson; 2011.
35. Schaefer A, Kong R, Gordon EM, *et al*. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb Cortex. 2018;28(9):3095-3114. https://doi.org/10.1093/cercor/bhx179
36. Fischl B, Salat DH, Busa E, *et al*. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002;33:341-355.
37. Bajaj S, Blair KS, Dobbertin M, *et al*. Machine learning based identification of structural brain alterations underlying suicide risk in adolescents. Discov Ment Health. 2023;3(1):6. https://doi.org/10.1007/s44192-023-00033-6
38. O'Doherty JP, Cockburn J, Pauli WM. Learning, reward, and decision making. Annu Rev Psychol. 2017;68:73-100. https://doi.org/10.1146/annurev-psych-010416-044216
39. Chester DS, DeWall CN. The pleasure of revenge: retaliatory aggression arises from a neural imbalance toward reward. Soc Cogn Affect Neurosci. 2016;11(7):1173-1182. https://doi.org/10.1093/scan/nsv082
40. White SF, Brislin SJ, Sinclair S, Blair JR. Punishing unfairness: rewarding or the organization of a reactively aggressive response? Hum Brain Mapp. 2014;35(5):2137-2147. https://doi.org/10.1002/hbm.22316
41. Fanning JR, Keedy S, Berman ME, Lee R, Coccaro EF. Neural correlates of aggressive behavior in real time: a review of fMRI studies of laboratory reactive aggression. Curr Behav Neurosci Rep. 2017;4(2):138-150. https://doi.org/10.1007/s40473-017-0115-8
42. Siep N, Tonnaer F, van de Ven V, Arntz A, Raine A, Cima M. Anger provocation increases limbic and decreases medial prefrontal cortex connectivity with the left amygdala in reactive aggressive violent offenders. Brain Imaging Behav. 2019;13(5):1311-1323. https://doi.org/10.1007/s11682-018-9945-6
43. Herpertz SC, Nagy K, Ueltzhoffer K, *et al*. Brain mechanisms underlying reactive aggression in borderline personality disorder-sex matters. Biol Psychiatry. 2017;82(4):257-266. https://doi.org/10.1016/j.biopsych.2017.02.1175
44. Blair RJR. The neurobiology of aggression. In: Charney DS, Nestler EJ, eds. Neurobiology of Mental Illness. Oxford: Oxford University Press; 2004:1076-1085.
45. De Brito SA, Forth AE, Baskin-Sommers AR, *et al*. Psychopathy. Nat Rev Dis Primers. 2021;7(1):49. https://doi.org/10.1038/s41572-021-00282-1
46. Viding E, Sebastian CL, Dadds MR, *et al*. Amygdala response to preattentive masked fear in children with conduct problems: the role of callous-unemotional traits. Am J Psychiatry. 2012;169(10):1109-1116. https://doi.org/10.1176/appi.ajp.2012.12020191
47. Regier DA, Narrow WE, Clarke DE, *et al*. DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry. 2013;170(1):59-70. https://doi.org/10.1176/appi.ajp.2012.12070999
48. Kaufman J, Birmaher B, Brent D, *et al*. Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry. 1997;36(7):980-988.
49. de la Peña FR, Villavicencio LR, Palacio JD, *et al*. Validity and reliability of the Kiddie Schedule for Affective Disorders and Schizophrenia Present and Lifetime Version DSM-5 (K-SADS-PL-5) Spanish version. BMC Psychiatry. 2018;18(1):193. https://doi.org/10.1186/s12888-018-1773-0
50. Baviskar S, Diop-Christensen A, Ebsen FC, From KJ, Mackrill T. Structural validity and reliability of the Danish self-report Strengths and Difficulties Questionnaire among male and female students in vocational education and training. J Child Adolesc Ment Health. 2020;32(2-3):87-98. https://doi.org/10.2989/17280583.2020.1830780
51. Fossati A, Raine A, Borroni S, *et al*. A cross-cultural study of the psychometric properties of the Reactive-Proactive Aggression Questionnaire among Italian nonclinical adolescents. Psychol Assess. 2009;21(1):131-135. https://doi.org/10.1037/a0014743
52. Chen G, Pine DS, Brotman MA, Smith AR, Cox RW, Haller SP. Trial and error: A hierarchical modeling approach to test-retest reliability. Neuroimage. 2021;245:118647. https://doi.org/10.1016/j.neuroimage.2021.118647
53. Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. Neuropsychopharmacology. 2021;46(1):176-190. https://doi.org/10.1038/s41386-020-0767-z
54. Boccadoro S, Wagels L, Puiu AA, *et al*. A meta-analysis on shared and distinct neural correlates of the decision-making underlying altruistic and retaliatory punishment. Hum Brain Mapp. 2021;42(17):5547-5562. https://doi.org/10.1002/hbm.25635
55. Pfattheicher S, Nielsen YA, Thielmann I. Prosocial behavior and altruism: a review of concepts and definitions. Curr Opin Psychol. 2022;44:124-129. https://doi.org/10.1016/j.copsyc.2021.08.021
56. Du E, Chang SW. Neural components of altruistic punishment. Front Neurosci. 2015;9:26. https://doi.org/10.3389/fnins.2015.00026
57. Fehr E, Gachter S. Altruistic punishment in humans. Nature. 2002;415(6868):137-140. https://doi.org/10.1038/415137a
58. Gilam G, Abend R, Shani H, Ben-Zion Z, Hendler T. The anger-infused Ultimatum Game: a reliable and valid paradigm to induce and assess anger. Emotion. 2019;19(1):84-96. https://doi.org/10.1037/emo0000435
59. O'Farrelly C, Watt H, Babalis D, *et al*. A brief home-based parenting intervention to reduce behavior problems in young children: a pragmatic randomized clinical trial. JAMA Pediatr. 2021;175(6):567-576. https://doi.org/10.1001/jamapediatrics.2020.6834
60. Munshi KR, Oken T, Guild DJ, *et al*. The use of antiepileptic drugs (AEDs) for the treatment of pediatric aggression and mood disorders. Pharmaceuticals (Basel). 2010;3(9):2986-3004. https://doi.org/10.3390/ph3092986
61. Haines N, Kvam PD, Irving LH, *et al*. Theoretically informed generative models can advance the social, behavioral, and brain sciences: lessons from the reliability paradox. Preprint. Posted online August 24, 2020. https://doi.org/10.31234/osf.io/xr7y3. PsyArXiv.