




ARTICLE

<https://doi.org/10.1038/s41467-019-09962-9>

OPEN

# Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males

Jenifer Z. Siegel <sup>1,2</sup>, Suzanne Estrada<sup>2</sup>, Molly J. Crockett <sup>2,3</sup> & Arielle Baskin-Sommers <sup>2,3</sup>

Individuals exposed to community violence are more likely to engage in antisocial behavior, resulting in a dramatic increase in contact with justice and social service systems. Theoretical accounts suggest that disruptions in learning underlie the link between exposure to violence and maladaptive behaviors. However, empirical evidence specifying these processes is sparse. Here, in a sample of incarcerated males, we investigated how exposure to violence affects the ability to learn about the harmfulness of others and use this information to adaptively modulate trust behavior. Exposure to violence does not impact the ability to accurately develop beliefs about agents' harm preferences and predict their choices. However, exposure to violence disrupts the ability to form moral impressions that dissociate between agents with distinguishable harm preferences, and subsequently, the ability to adjust trust behavior towards different agents. These findings reveal a process that may explain the association between exposure to violence and maladaptive behavior.

<sup>1</sup>Department of Experimental Psychology, University of Oxford, Oxford OX1 2JD, UK. <sup>2</sup>Department of Psychology, Yale University, New Haven, CT 06520, USA. <sup>3</sup>These authors contributed equally: Molly J. Crockett, Arielle Baskin-Sommers. Correspondence and requests for materials should be addressed to M.J.C. (email: [molly.crockett@yale.edu](mailto:molly.crockett@yale.edu)) or to A.B.-S. (email: [arielle.baskin-sommers@yale.edu](mailto:arielle.baskin-sommers@yale.edu))

Exposure to community violence, whether it is witnessing someone get chased or hurt, hearing gunshots in the neighborhood, or being directly chased, assaulted or shot at, is a significant public health concern. In the United States, over three-quarters of youth have been exposed to some form of community violence in their lifetime<sup>1,2</sup>. In general, both cross-sectional and longitudinal research finds that exposure to violence places young people at risk for persistent academic underachievement<sup>3</sup>, physical health problems (e.g., difficulty sleeping, headaches, heart disease, immune disease<sup>4,5</sup>), mental health problems (e.g., depression, anxiety, post-traumatic stress, anti-social personality<sup>5-7</sup>), and interpersonal problems (e.g., problems with trust, lower levels of empathy<sup>8</sup>). Additionally, individuals exposed to violence are more likely to engage in antisocial behavior<sup>6,7</sup>, show earlier and more chronic aggressive behavior<sup>9</sup>, and hold beliefs that can normalize or romanticize aggression<sup>10</sup>. As a result, exposure to violence dramatically increases the likelihood of involvement in the justice and social service systems<sup>11</sup>.

Exposure to violence predisposes some individuals to diverse forms of negative life experiences and mental health problems, as well as comprises a prominent risk factor for a lifetime mired in aggression. Chronic exposure to violence, whether in a larger community or justice system context, shapes cognition in a way that is likely to distort perceptions of what is considered harmful behavior and how to react to harmful behavior. However, the precise social cognitive processes that may underlie these distortions in individuals exposed to violence is unclear. At the core of several theories about the relationship between exposure to violence and aggressive/antisocial behavior is the role of learning<sup>10,12-15</sup>. However, empirical evidence identifying and specifying the way in which learning is disrupted and can affect behavior in individuals exposed to violence remains elusive.

One aspect of learning that is especially relevant to adaptive social behavior is learning about whether other individuals might harm us. Harmfulness is a core dimension of moral character<sup>16,17</sup>. Research on social learning has shown that there are two, distinct, components of harmfulness learning. On the one hand, people use social cues to objectively update their beliefs about others' harmfulness by gradually accumulating information over time to predict future outcomes (i.e., in a Bayesian manner<sup>18</sup>). On the other hand, people form subjective impressions about moral character that emerge rapidly and effortlessly<sup>19,20</sup>. These beliefs and moral impressions are used to adaptively learn and decide whom to trust in social interactions<sup>16,21</sup>. For example, in a study by Siegel and colleagues<sup>18</sup>, participants entrusted more money to agents who were less willing to harm others for profit and ascribed better moral character (subjective impression) to those agents compared to agents who were more willing to harm for profit. Together, these components of learning about other's harmfulness serve as powerful informational tools; for the purpose of survival, humans are evolutionarily inclined to identify potential foes and avoid them through adaptive social decision-making<sup>22,23</sup>. However, life experiences, such as exposure to violence that for some individuals follows them continuously from the community to prison<sup>24</sup>, are likely to shape social learning and resulting social behaviors. Prior research linking exposure to violence to normalized views of aggression and aberrations in interpersonal functioning raises the possibility that exposure to violence may impact learning about the harmfulness of others, and by extension, behaviors that rely on trust. To date, however, there has been no research on exposure to violence and harm learning.

To examine the relationships among exposure to violence, harm learning, and trust behavior, we administer a harmfulness learning task<sup>18</sup> to a sample of incarcerated males. While a sample of currently incarcerated individuals is not the same as a sample

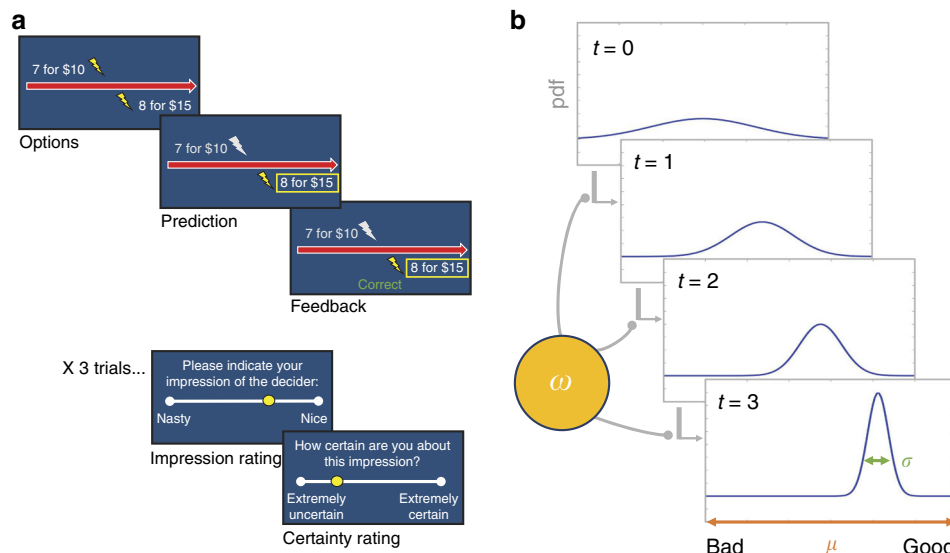
from the general population, this type of sample does serve as an informative sample in which to explore how differences in exposure to violence impact harm learning. It is well-documented that exposure to violence among the incarcerated covers the full continuum of potential experiences compared to the general population where scores are often restricted in range and narrowly centered around a few points within that range<sup>25-27</sup>. Moreover, by studying a sample of currently incarcerated individuals, we are better poised to investigate the variation in exposure to violence within a sample that is already demonstrating the theorized behavioral effects of such exposure.

In the task, participants predict and observe the choices of two agents who repeatedly decide whether to inflict painful electric shocks on another individual in exchange for money (Fig. 1a). The two agents substantially differ in their preferences towards harm (i.e., their exchange rate between money and pain). On each trial, participants predict the choice made by the agent and receive immediate feedback about their accuracy. Every three trials, participants rate their overall subjective impression of the agent's moral character (on a scale from nasty to nice) and their certainty of that impression. This task enables us to measure two distinct components of harm learning: the ability to develop accurate beliefs about the agents' objective exchange rates between money and pain (a quantity that is used to predict their choices), and the use of those estimates to form subjective, global impressions about other's moral character. After the harmfulness learning task, participants engage in a one-shot trust game<sup>28</sup> with each of the agents. All participants complete a battery that assesses exposure to violence using the Exposure to Violence Scale (ETV)<sup>29</sup>, as well as a clinical assessment measuring different aspects of antisociality to address potential confounds. We show that the ability to learn about harmfulness is not affected by exposure to violence. However, exposure to violence impairs the development of subjective impressions, and consequently, the ability to adapt trust behavior toward more harmful vs. less harmful agents.

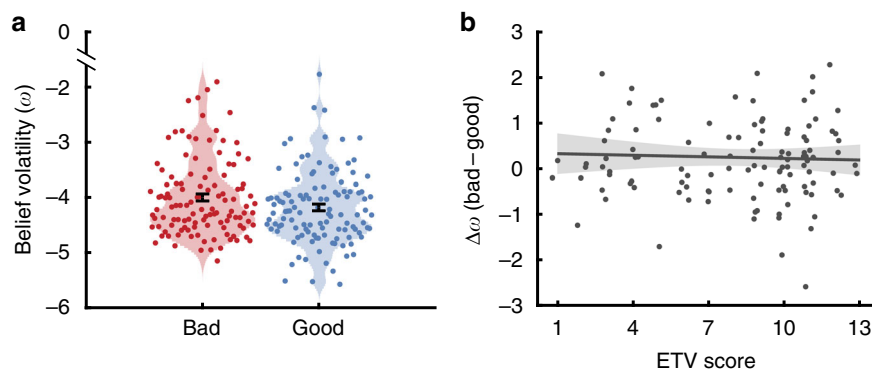
## Results

**Beliefs and predictions about harm preferences.** We first investigated participants' ability to develop accurate beliefs about the agents' objective harm preferences and predict their decisions. On average, participants predicted accurately 72% of the good agent's choices and 77% of the bad agent's choices. There was no relationship between ETV score and prediction accuracy for either agent (Spearman's  $\rho$ , good:  $\rho = -0.065$ ,  $p = 0.483$ ; bad:  $\rho = 0.043$ ,  $p = 0.639$ ). This suggests that participants with higher exposure to violence were equally motivated to learn the harm preferences of the agents, relative to those with lower exposure to violence.

Next, we examined how rapidly participants updated their beliefs about the agents' preferences in response to feedback. To this end, we fit a hierarchical Bayesian model for learning stable preferences under conditions of uncertainty to participants' predictions. The model defines how beliefs about an agent's harm preference evolve over time as a function of a participant-specific parameter  $\omega$ , capturing inter-individual differences in the rate of belief updating (see Methods and Fig. 1b)<sup>30</sup>. Formal model comparison indicated that the hierarchical Bayesian model outperformed two alternative Rescorla Wagner models in our sample of participants (see Methods and Supplementary Note 1). Replicating previous research<sup>18</sup>, beliefs about the bad agent's preferences were more rapidly updated in response to feedback, as indicated by a higher  $\omega$ , than beliefs about the good agent's preferences (signed rank test,  $Z = -2.328$ ,  $p = 0.020$ ; Fig. 2a). ETV score was not significantly related to  $\omega$  for either agent



**Fig. 1** Learning task and model. **a** Representation of the task schematic, created by the authors. Participants predicted sequences of choices for two agents (Decider A and Decider B). On each trial the agent chose between two options: more shocks inflicted on another person in exchange for more money, or fewer shocks for less money. After making their prediction, participants observed the agent’s actual choice along with feedback indicating whether they were correct or not in their prediction. Every third trial participants made a judgment about the agent’s moral character (ranging from nasty to nice) and indicated how certain they were about their judgment. **b** Model schematic for learning about a good agent, modified from Siegel et al.<sup>18</sup>. Beliefs about the agent’s harm preference are represented by probability distributions with a mean  $\mu$  and variance  $\sigma$ . Beliefs evolve over time as a function of Gaussian random walks whose step-size is governed by  $\omega$ , a participant-specific parameter that captures individual differences in the rate at which beliefs evolve over time,  $t$



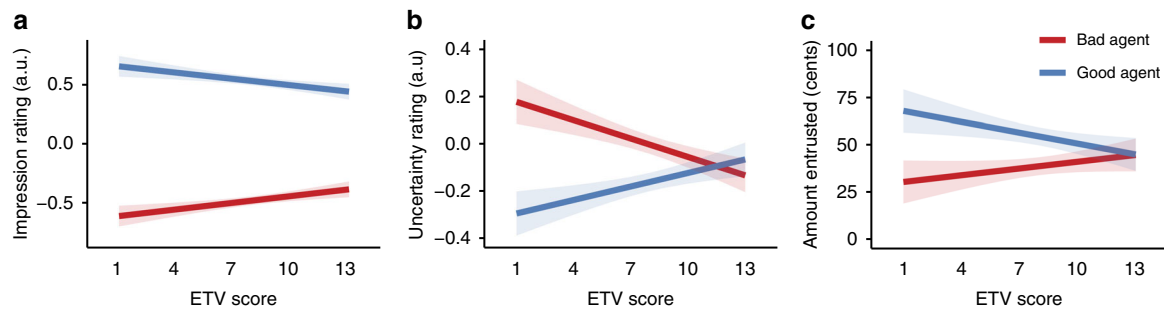
**Fig. 2** Objective harm learning does not covary with exposure to violence. **a** Beliefs about the bad agent’s harm preferences were more volatile than beliefs about the good agent’s harm preferences. **b** Between-agent asymmetries in belief updating ( $\Delta\omega = \text{bad agent belief volatility} - \text{good agent belief volatility}$ ) were not related to participant’s ETV score, suggesting that exposure to violence does not significantly impact the underlying processes of objective harm learning. Error bars represent standard error of the mean. Error bands represent 95% confidence intervals. Source data are provided as a Source Data file

(Spearman’s  $\rho$ , good:  $\rho = 0.025, p = 0.785$ ; bad:  $\rho = -0.014, p = 0.879$ ), nor was it related to the difference in  $\omega$  between good and bad agents ( $\Delta\omega: \rho = 0.008, p = 0.929$ ; Fig. 2b). Together, these results suggest that objective harm learning was largely intact in this sample and did not covary with exposure to violence.

**Subjective impressions of the agent’s moral character.** Despite the fact that ETV score did not impact learning the objective features of agents’ preferences, we observed a strong effect of ETV score on participants’ subjective global impressions of the agent’s moral character. A robust linear regression was used to predict subjective moral impression ratings as a function of agent (good vs. bad), ETV score, and their interaction. We report means and standard error of the mean (sem) as mean  $\pm$  sem. We included an additional regressor to control for trial number, and found no effects on impression ratings ( $\beta = 0.003 \pm 0.003, t = 0.955, p =$

0.340). Replicating previous research, in general, participants formed more favorable impressions of the good agent’s moral character than the bad agent’s moral character ( $\beta = -1.300 \pm 0.075, t = -17.284, p < 0.001$ ). Higher ETV scores predicted more negative impressions of the good and bad agents’ moral character ( $\beta = -0.018 \pm 0.006, t = -2.902, p = 0.004$ ). There was a significant interaction between ETV score and type of agent ( $\beta = 0.037 \pm 0.009, t = 4.222, p < 0.001$ ), indicating that for participants with higher ETV scores, there was less differentiation in their impressions of the good and bad agents’ moral character (Fig. 3a).

To further investigate the interaction between type of agent and exposure to violence on subjective impression ratings, we ran separate regressions on ratings for the good and bad agent. Specifically, we asked whether diminishing effects of agent with higher ETV scores were driven by the good agent, the bad agent, or both. These analyses revealed that the effects were not



**Fig. 3** Model estimates showing diminishing effects of agent with increasing exposure to violence. Participants with higher ETV scores showed less differentiation in their subjective impressions of good (blue line) vs. bad (red line) agent's moral character (**a**) and reported smaller discrepancies in the uncertainty of their impressions of good and bad agents (**b**). Higher ETV scores also resulted in smaller discrepancies in the amounts that participants entrusted with good vs. bad agents in a one-shot trust game (**c**). Y-axis in figures **a** and **b** denote standardized values (z-scored). Error bands represent 95% confidence intervals. Source data are provided as a Source Data file

specifically driven by either agent alone. Higher ETV scores predicted more favorable impressions of the bad agent ( $\beta = 0.019 \pm 0.007$ ,  $p = 0.004$ ) and less favorable impressions of the good agent ( $\beta = -0.019 \pm 0.006$ ,  $p = 0.001$ ; see Fig. 3a).

Of note, participants with higher ETV scores were no more likely to predict worse harm intentions of the agents in the task before they observed any of their choices (Spearman's  $\rho$ ,  $\rho = -0.082$ ,  $p = 0.361$ ) and were no less trusting of others in general (as indicated by scores on the General Trust scale<sup>31</sup>;  $\rho = -0.040$ ,  $p = 0.665$ ). These findings suggest that exposure to violence affects how participants in this sample form subjective impressions about other's moral character through observing their choices, rather than affecting prior beliefs about others.

**Certainty of subjective impressions.** Past work in non-incarcerated samples indicates that adults hold more certain positive impressions of others and more uncertain negative impressions, which is hypothesized to serve the adaptive social function of enabling people to more easily update negative impressions that turn out to be inaccurate<sup>18,32</sup>. Consistent with previous research, uncertainty decreased over time as participants were exposed to more information about the agents' harm preferences (Robust regression,  $\beta = -0.018 \pm 0.003$ ,  $t = -5.969$ ,  $p < 0.001$ ). Furthermore, participants expressed greater uncertainty in their impressions of the bad agent, relative to the good agent ( $\beta = 0.513 \pm 0.078$ ,  $t = 6.605$ ,  $p < 0.001$ ).

To investigate whether exposure to violence affected participants' uncertainty in their impressions of the agents' moral character, we performed a robust linear regression to investigate the effects of agent, ETV score, and their interaction on participants' ratings of uncertainty about their impressions of the agents. Participants with higher ETV scores were more uncertain in their impressions overall ( $\beta = 0.019 \pm 0.006$ ,  $t = 2.982$ ,  $p = 0.003$ ), and this interacted with the effect of agent ( $\beta = -0.045 \pm 0.009$ ,  $t = -4.973$ ,  $p < 0.001$ ; Fig. 3b). Consistent with our findings on subjective impressions, the interaction indicated that participants with higher ETV scores expressed smaller differences in their uncertainty ratings between good and bad agents, such that they became more uncertain that the good agent was good, and less uncertain that the bad agent was bad. Notably, smaller differences in impression ratings between good and bad agents predicted less discrepant uncertainty ratings (Spearman's  $\rho$ ,  $\rho = 0.336$ ,  $p < 0.001$ ). Finally, mirroring the subjective impression results, these effects were not specifically driven by either agent alone. Higher ETV scores predicted less uncertain impressions of the bad agent ( $\beta = -0.025 \pm 0.007$ ,  $t = -3.795$ ,  $p < 0.001$ ) and more uncertain impressions of the good agent ( $\beta = 0.019 \pm 0.006$ ,  $t = 3.126$ ,  $p = 0.002$ ; see Fig. 3b). Results from

our robust linear regressions did not change after controlling for age and education (see Supplementary Tables 4, 5).

**Trust behavior.** Although exposure to violence impaired participants' ability to form distinct subjective impressions of agents with different harm preferences, it is unclear whether this has consequences for social behavior. To address this question, we asked participants to engage in a one-shot trust game with each of the agents, after predicting all the agents' choices in the harm learning task (see Methods). Previous work has shown that non-incarcerated adults adjust their behavior in the trust game according to the harm preferences of the agent with whom they are interacting (i.e., people entrust significantly less money with those who treat others poorly than those who treat others well<sup>18,33</sup>).

To investigate whether adaptive trust behavior was diminished in participants with higher ETV scores, we entered the amount participants entrusted in a repeated measures general linear model with agent (good vs. bad) as the within-subject factor and ETV score as a continuous covariate. Consistent with previous research, participants entrusted more points with the good agent than the bad agent ( $F(1,119) = 6.202$ ,  $p = 0.014$ ,  $\eta^2 = 0.056$ ). The effect of agent was significantly moderated by ETV score ( $F(13,119) = 2.142$ ,  $p = 0.017$ ,  $\eta^2 = 0.210$ ; Fig. 3c). The interaction indicated that higher ETV scores predicted smaller discrepancies in the amount participants entrusted with the good vs. the bad agent. Specifically, those with higher ETV scores entrusted significantly less with the good agent (Spearman's  $\rho$ ,  $\rho = -0.220$ ,  $p = 0.016$ ), and consequently ended up earning fewer points overall ( $\rho = -0.325$ ,  $p < 0.001$ ). ETV scores did not significantly affect the amount that participants entrusted with the bad agent ( $\rho = 0.119$ ,  $p = 0.198$ ). Thus, exposure to violence was associated with maladaptive trusting behavior, specifically when interacting with those who are less willing to harm others, and this had a negative impact on their overall earnings.

Given the relationship between exposure to violence and differential trust behavior ( $\Delta\text{trust}$ , calculated as amount entrusted with good agent – amount entrusted with bad agent), it is possible that participants' final subjective impressions of the agents' moral character ( $\Delta\text{judgment}$ , calculated as final impression of the good agent – final impression of the bad agent) account for (i.e., mediate) that relationship. We found a significant indirect effect of  $\Delta\text{judgment}$  on  $\Delta\text{trust}$ , effect =  $-0.812$ , CI  $[-1.705, -0.054]$  (see Supplementary Note 2 for full mediation results and additional analysis), suggesting that impressions about other's moral character account for differences in social behavior among participants with higher levels of exposure to violence.



Despite these results, some might question the validity of the trust game in the current sample given their incarceration status. Therefore, we examined whether the extent to which participants adjusted their trust behavior according to the agents' harm preferences predicted social behavior in prison. Less discrepant behavior towards good and bad agents was associated with more behavioral violations in prison (Spearman's  $\rho$ ,  $\rho = -0.208$ ,  $p = 0.023$ ), and specifically with aggressive violations against persons ( $\rho = -0.217$ ,  $p = 0.020$ ). This suggests the ability to adjust trust behavior based on impressions of other's moral character, as measured by our task, captures variance in real-world social behavior.

However, it's possible that this relationship between less discrepant behavior towards good and bad agents and more behavioral violations in prison is largely explained by the relationship with ETV scores. Indeed, higher ETV scores predict more behavioral violations in prison (Spearman's  $\rho$ ,  $\rho = 0.450$ ,  $p < 0.001$ ). However, we predicted that this relationship would be mediated by the extent to which participants differentiated in their subjective impressions and trust behavior between the good and bad agent. Consequently, we applied a serial multiple mediation analysis using the PROCESS macros for SPSS<sup>34</sup> (model 6) that allowed us to determine the causal link between mediators with a specified direction of causal flow. We investigated whether the relationship between exposure to violence and prison violations was mediated by trust behavior ( $\Delta$ trust) as a function of impression sensitivity ( $\Delta$ judgment). ETV score was only a marginally significant predictor of prison violations after impression sensitivity and trust behavior were accounted for (direct effect =  $0.622 \pm 0.340$ ,  $p = 0.070$ ). The indirect effects were tested using a bootstrap estimation approach with 5000 samples. These results indicated the indirect serial coefficient was significant (indirect effect =  $0.099 \pm 0.071$ , 95% CI = [0.002, 0.274]; see Supplementary Note 3 for full mediation results and Supplementary Fig. 1), suggesting that disruptions in the ability to form distinguishable impressions resulting from higher ETV scores, translates into maladaptive trust behavior, which in turn leads to a greater number of direct violations in prison.

**Specificity of exposure to violence effects.** Previous work has shown that exposure to violence is associated with antisocial behavior and psychopathic traits<sup>10,35–37</sup>. Here we also found that ETV score was associated with increased antisociality, as indicated by higher scores on the Hare Psychopathy Checklist-Revised<sup>38</sup> (PCL-R) and increased symptoms of Antisocial Personality Disorder<sup>39</sup> (APD) (Spearman's  $\rho$ , PCL-R total score:  $\rho = 0.394$ ,  $p < 0.001$ ; APD symptoms:  $\rho = 0.261$ ,  $p = 0.004$ ). This leaves open the question of whether the effects of exposure to violence on subjective impressions observed here are a primary consequence of exposure to violence, or an indirect consequence of possessing characteristics that predispose exposure to violence, such as Psychopathy or Antisocial Personality Disorder. To assess whether ETV score had a direct, as opposed to indirect, effect on subjective impressions and uncertainty ratings, and social behavior, we entered each covariate (PCL-R total score and total number of APD symptoms) separately into our regressions with ETV score. Across all measures, we found that interactions between agent and ETV score remain significant even when we include the interaction between agent and each covariate (see Supplementary Tables 4–6 for all analyses including covariates). An alternative possibility is that the observed effects of exposure to violence on impressions reflect a general impact of traumatic experiences, rather than being specific to community violence. To investigate, we entered scores from the Childhood Trauma Questionnaire<sup>40</sup> (CTQ) into our regression with ETV score.

Again, we found that the interactions between agent and ETV score remain significant even after controlling for CTQ (see Supplementary Tables 4–6 for full analysis). Together, this suggests that being exposed to violence had a direct effect on subjective impressions of moral character and social behavior and that findings could not be entirely explained by antisocial psychopathology or childhood trauma.

## Discussion

The ability to infer other's intentions and predict their behavior is crucial for successful social interactions. In particular, learning whether others are likely to harm us is important for consequential social decisions like deciding whom to trust. However, there are environmental experiences that may impact how we learn about harm and use this information to make adaptive social decisions. Exposure to violence is one environmental experience that is associated with aberrations in beliefs about harm<sup>10,15</sup>. As a result, exposure to violence is related to behaviors that reflect a lack of trust and prosociality (e.g., aggression, crime), increasing contact with systems of social control.

The current data suggest that, in a sample of currently incarcerated males, exposure to violence adversely impacts some components of harm learning, but not all. Participants with higher ETV scores showed an ability to develop accurate beliefs about others by objectively encoding their harm preferences. However, exposure to violence appeared to disrupt the formation of subjective, global impressions of other's moral character from observed harm behavior. Participants with higher ETV scores formed more positive and less uncertain impressions of harmful agents and more negative and less certain impressions of helpful agents. Moreover, these differences in subjective impressions associated with higher ETV scores led to maladaptive trust behavior, such that participants with higher ETV scores extended less trust than optimal when interacting with a good agent. Finally, the link between exposure to violence and maladaptive trusting behavior was mediated by the disturbances in impression formation. In turn, this led to significantly more violations in prison, suggesting that the effects of exposure to violence on real social behavior in prison is predicted by subjective impressions and trust behavior as measured by our task. On the whole, these findings raise the intriguing possibility that exposure to violence does not fundamentally disrupt all components of social learning, but instead may produce a problem with generating global subjective social impressions and translating those impressions into adaptive social decision-making.

Our findings are consistent with evidence that the ability to learn the value of information is cognitively and neurally distinct from the ability to use learned information to guide decision-making and behavior<sup>41,42</sup>. Cognitively, participants with higher ETV scores were able to learn harm preferences of different agents but formed subjective impressions that appeared to normalize beliefs about harm in the bad agent<sup>10,15</sup>, seeing that agent as more similar to the good agent. This dissociation between learning and using learned information is consistent with neural lesion studies show that the lateral orbitofrontal cortex (OFC) is associated with learning the value of stimuli, whereas the medial OFC is associated with translating stimulus value representations into decisions<sup>43</sup>. Although there are no imaging studies that directly examine exposure to violence, studies looking at the combination of physical abuse and adversity (e.g., seeing intimate-partner violence, bullying, physical and sexual abuse) note structural and functional abnormalities in the OFC<sup>44–46</sup>, and individuals (both incarcerated and non-incarcerated) prone to aggressive and antisocial behavior also display abnormal medial OFC structure and function<sup>47–50</sup>. Taken together, findings from

these parallel literatures suggest that diminished use of learned information may be a consequence of OFC dysfunction in individuals exposed to violence.

An especially unfortunate consequence of disrupted use of harm learning may be a pervasive inability to develop healthy social relationships with trustworthy individuals and a greater likelihood of placing trust in the wrong people. Consistent with findings in the present study showing the impact of subjective impressions on trust behavior, research from the fields of sociology, psychology, and economics highlight that individuals who reside in communities with high rates of crime and disorder experience mistrust in their interactions with strangers, prosocial members of their community, and institutions<sup>51–53</sup>. Justice-involved individuals tend to reside in these types of communities characterized by crime and disorder, where social interactions and systems that may provide pathways out of serious and persistent offending are absent, severely debilitated, or sparse<sup>51,54</sup>. Moreover, once caught within the justice system, it is likely that these types of interactions are reinforced through new exposures to violence and negative social interactions<sup>24</sup>. Combined, community context and justice-involvement translates to lower access to informal and formal resources, homebased learning, and chronic re-exposure to violence<sup>55,56</sup>. The resulting pattern is that some individuals, like justice-involved individuals, are more likely to live in communities of unrelenting social and economic deprivation. These environmental characteristics do not solely impact the incarcerated but spills over to other community members: those who are incarcerated are released and their behavior and experiences impact family members, social acquaintances, and strangers in their own communities. For that matter, the disproportionate presence of incarcerated individuals in disadvantaged communities is not seen as aberrant but is often just part of living in these communities<sup>57</sup>. Thus, the combination of environmental characteristics and disruptions in the cognitive processes at the individual level are critical for the development and maintenance of trust and may ensnare individuals in a trajectory that continually reinforces maladaptive social connections, ultimately limiting chances for economic stability<sup>58–60</sup> and psychosocial wellbeing<sup>5</sup>.

Before concluding, methodological and conceptual limitations should be noted. The present sample is limited to incarcerated offenders, and thus we do not know whether or how incarceration status may impact the relationship between exposure to violence and harm learning. However, it is important to note that all task main effects replicated previous findings in non-incarcerated samples. For instance, previous work using the same task has shown that people form less positive, more uncertain, and more volatile beliefs about the bad agent, relative to the good agent, and adjust their trust behavior according to the harm preferences of the agents<sup>18</sup>. We observe the same pattern of results in our sample of incarcerated individuals. Moreover, length of incarceration (see Supplementary Tables 4–6) and other correlates known to increase risk for incarceration did not impact the reported exposure to violence effects. Ultimately, being currently incarcerated is just one type of adverse outcome related to exposure to violence that should not be seen as excluding the importance of the lived experience of exposure to violence for these individuals<sup>61–65</sup>.

While it may be useful to replicate the findings in a sample of non-incarcerated individuals, this raises important experimental considerations. From a scientific perspective, using a sample with sufficient variability in ETV scores, and whose experience with exposure to violence has led to great personal cost, is essential. Notably, the distribution of ETV scores in our sample of incarcerated individuals covers the full range of the scale. Endeavors in samples typical of psychology research, such as university or

crowdsourced samples, often suffer from restricted range in ETV scores. Nonetheless, to test for generalizability, future research should replicate the present findings in a sample of non-incarcerated individuals whose ETV scores are reflective of a range of experiences.

A final consideration is that implementing shocks in the harmfulness learning task is not as extreme a behavior as what might be seen in the real world (e.g., sexual assault, murder) for individuals exposed to violence or involved in the justice system. Therefore, it is possible that the objective learning of other's harm preferences could be different with more extreme behaviors. Future research should continue to investigate components of learning in those exposed to violence and vary the stimuli used to assess learning that consider cultural and situational contexts.

The relationship between exposure to violence and negative life experiences is undeniable. However, an understanding of how this environment shapes cognition and behavior is less clear. The present study identifies a specific deficit in the ability of incarcerated individuals exposed to violence to adapt social behavior towards agents with distinguishable harm preferences. Continuing to identify and specify the processes that are altered by exposure to violence will be crucial for understanding how individuals experience, incorporate, and react to their particular social environment.

## Methods

**Participants.** The present sample included 119 males from a high-security correctional institution in Connecticut. We used a prescreen of institutional files and assessment materials to exclude justice-involved individuals who: were not between the ages of 18 and 75; scored below 70 on a brief measure of IQ (Shipley Institute of Living Scale<sup>66</sup>) performed below the fourth-grade level on a standardized measure of reading (Wide Range Achievement Test-III<sup>67</sup>) had diagnoses of schizophrenia, bipolar disorder, or psychosis, not otherwise specified; were currently taking psychotropic medication; or had a history of medical problems (e.g., uncorrectable auditory or visual deficits, head injury with loss of consciousness greater than 30 min, seizures, neurological disorders) that may impact their comprehension of the materials. The Yale University Human Investigation Committee approved the procedures used in the present study. The study complied with all relevant ethical regulations for work with human participants and all participants provided written informed consent. See Supplementary Table 1 for participant demographic information.

**Harmfulness learning task.** In the task, participants predicted a sequence of 50 choices made by each of two agents (100 choices total). For each choice, the agent chose between two options: more money for themselves plus more shocks for an anonymous other individual, or less money for themselves plus fewer shocks for the other individual (Fig. 1a). For each trial, participants received feedback about their accuracy. No a priori information was provided about the agents; thus, optimal behavior required participants to learn the agents' preferences over time. Participants predicted all choices for one agent at a time, and the order of agents was randomized across participants.

Following every third prediction, participants indicated their current impression of the agent's moral character on a continuous visual analogue scale rating from 0 (nasty) to 100 (nice)<sup>68</sup> and indicated how certain they were about their impression on a scale ranging from 0 (extremely uncertain) to 100 (extremely certain) (see Fig. 1a). Together, this provided us with a trajectory of participants' subjective impression ratings of each agent's moral character and how certain participants were about their characterization.

To manipulate harm preferences, we created one agent who was more averse to harming others (good agent) and one agent who was less averse (bad agent). This was operationalized as their exchange rate between money for themselves and shocks for the other individual, with the good agent requiring more money per shock inflicted than the bad agent (good agent = \$2.40 per shock; bad agent = \$0.43 per shock). The agents selected from identical choice sets; however, because the agents had different preferences towards harm, they behaved differently 50% of the time. For details on how the trial sequences were created and the agents' behavior simulated, see Supplementary Methods.

To incentivize participants to learn about the preferences of the agents, before beginning the task they were instructed to pay close attention to the deciders and learn about their behavior, because they will interact with each of them in a computerized trust game at the end of the study which could earn them reward points.

**Trust game.** After completing the harmfulness learning task, participants played a trust game with each of the agents. In the game, participants were endowed with 100 points that they could entrust with each agent. Any amount that they entrusted with the agent would be tripled, and the agent could then choose how much of the tripled amount to return to the participant. We instructed participants that the percent returned by each agent had been predetermined, and thus the agents were not playing actively. We set the returned amount to correspond to the agents' actual harm tendencies, such that the good agent behaved less selfishly than the bad agent and therefore returned a larger proportion of the entrusted points. The final number of points was tallied and the top five earners were added to a leader board that was on display to all study participants in the testing room (Note: The Connecticut Department of Correction did not allow researchers to pay justice-involved individuals). Of particular interest was the difference in amount entrusted with the good agent vs. the bad agent ( $\Delta$ entrust = amount entrusted with the good agent – amount entrusted with the bad agent).

**Exposure to Violence Scale.** The ETV scale<sup>29</sup> was used to measure lifetime exposure to violent events. The questionnaire consisted of 13 items, documenting the types of both experienced and observed violence (e.g., "Have you been hit, slapped, punched, or beaten up?" and "Have you seen someone else get attacked with a weapon, like a knife or bat?"). Participants were asked to respond to each item based on a dichotomous choice (*yes/no*). If *yes* was selected, participants indicated the number of times they experienced this situation in their lifetime. The two scales, experienced and observed, showed moderate overlap (Spearman's  $\rho$ ,  $\rho = 0.607$ ,  $p < 0.001$ ). Thus, we examined a total exposure to violence score using a sum of all 13 items. Internal consistency for ETV total score was 0.86. ETV scores were normally distributed (skewness:  $-0.605$ , kurtosis:  $-0.810$ ). Ninety-nine percent of the sample reported experiencing at least one exposure to violence in their lifetime and ~30% of the sample reported experiencing over nine (the median) different exposures to violence in their lifetime. Lifetime frequency of exposure to violence ranged from two times to 11,465 times (median = 88).

**General Trust Scale.** The General Trust scale<sup>31</sup> was used to measure general beliefs about the honesty and trustworthiness of others. Participants were asked to indicate to what extent they agree (1) or disagree (5) with six statements (e.g., "Most people are trustworthy"). The scores from each statement were averaged together to produce a continuous measure of generalized trust.

**Hare psychopathy checklist revised (PCL-R).** The PCL-R<sup>66</sup> used information gleaned from a life-history interview and a review of institutional files to score participants on the presence of 20 different items (e.g., superficial charm, shallow affect, impulsivity, criminal versatility). A score of 0, 1, or 2 was given for each item according to the degree to which a characteristic was present. PCL-R total scores ranged from 0 to 40. The reliability and validity of the PCL-R has been well established<sup>38,69</sup>. Inter-rater reliability for 24% of the sample was 0.991 (alpha).

**Antisocial personality disorder.** Participants were assessed for Antisocial Personality Disorder (APD) during a semi-structured diagnostic interview. The interview evaluated the age and frequency of engagement in behaviors outlined in the Diagnostic Statistical Manual-5<sup>39</sup> (DSM). A diagnosis of APD was given if there was evidence of conduct disorder (CD) prior to age 16 and sufficient adult antisocial symptoms (e.g., aggression, irresponsibility). Inter-rater reliability for 32% of the sample was 0.989 (Cohen's kappa).

**Childhood Trauma Questionnaire.** We used the Childhood Trauma Questionnaire (CTQ<sup>40</sup>), a 28-item questionnaire, to assess maltreatment experiences prior to age 18. It consisted of five clinical scales: emotional abuse, physical abuse, emotional neglect, and sexual abuse. Items were rated on a 5-point Likert-type scale with response options ranging from *Never True* to *Very Often True*. For the present study, the total score was examined. For this sample, the total score demonstrated good internal consistency (Cronbach's  $\alpha = 0.824$ ).

**Computational model.** We compared three different computational models to describe how participants learned the agents' preferences and predicted their choices. First, we fit a Hierarchical Gaussian Filter model<sup>30,70</sup> (HGF), which identified participant-specific parameters to describe each participant's learning process. Beliefs about an agent's harm preference were updated using a Bayesian reinforcement learning algorithm, with precision-weighted prediction errors driving belief updating at the different levels of the hierarchical model. Second, we fit a Rescorla Wagner model, in which beliefs were updated by prediction errors with a fixed learning rate. Third, we fit a modified Rescorla Wagner model, in which beliefs were updated by prediction errors with separate fixed learning rates for helpful and harmful outcomes. All model fitting was implemented using the HGF toolbox (<https://tnu.ethz.ch/tapas>). For a full list of priors, see Supplementary Table 2. For further details about each model see Supplementary Table 3.

As in previous studies<sup>18</sup>, formal model comparison indicated that the HGF model outperformed the two alternative Rescorla Wagner models in our sample of participants (see Supplementary Note 1 for details). The HGF model generated a

trial-wise sequence of belief estimates about each agent's harmfulness (i.e., the exchange rate between money and pain, latent variable,  $\mu$ ); a trial-wise sequence of uncertainties on those beliefs (latent variable,  $\sigma$ ); and a global estimate of belief volatility (parameter,  $\omega$ ) that describes the rate at which beliefs evolve over time (Fig. 1b). Belief volatility is set in log space and is monotonically related to belief uncertainty (i.e., more uncertain beliefs are more volatile<sup>30</sup>; for example, a change in  $\omega$  from  $-3.5$  to  $-4.0$  corresponds to a 20% decrease in the average variance of posterior beliefs,  $\sigma$ ). In short, the model describes trial-wise updating of beliefs about the agent's preferences towards harm, which approximates Bayes optimality (in an individualized sense given differences in  $\omega$ ) and determines the participant's estimate of the probability that an agent will harm.

**Statistical analyses.** All data analysis was completed in Matlab (Mathworks) and PASW Statistics 24 (SPSS/IBM). All statistical tests were two-sided. We used robust linear regression models with a bisquare weighting function to analyze the z-scored trial-by-trial rating data (impression and certainty ratings). We used nonparametric statistical tests that do not make any assumptions about the underlying distributions of variables (e.g., Spearman's  $\rho$  and signed rank tests). To investigate whether the relationship between ETV score and differences in social behavior were mediated by differences in participants final harm judgments of the agents we used the PROCESS macro for SPSS<sup>34</sup>.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. Source data for Figs. 2 and 3 are provided with the paper.

## Code availability

All relevant Matlab code are available from the corresponding author upon request.

Received: 17 August 2018 Accepted: 8 April 2019

Published online: 26 April 2019

## References

- Finkelhor, D., Turner, H. A., Shattuck, A. & Hamby, S. L. Prevalence of childhood exposure to violence, crime, and abuse: results from the National Survey of children's exposure to violence. *JAMA Pediatr.* **169**, 746–754 (2015).
- Finkelhor, D., Turner, H. A., Shattuck, A. & Hamby, S. L. Violence, crime, and abuse exposure in a National Sample of Children and Youth: an update. *JAMA Pediatr.* **167**, 614–621 (2013).
- Delaney-Black, V. et al. Violence exposure, trauma, and IQ and/or reading deficits among urban children. *Arch. Pediatr. Adolesc. Med.* **156**, 280–285 (2002).
- Bailey, B. N. et al. Somatic complaints in children and community violence exposure. *J. Dev. Behav. Pediatr.* **26**, 341 (2005).
- Moffitt, T. E. & Klaus-Grawe 2012 Think Tank. Childhood exposure to violence and lifelong health: Clinical intervention science and stress-biology research join forces. *Dev. Psychopathol.* **25**, 1619–1634 (2013).
- Baskin, D. & Sommers, I. Trajectories of exposure to community violence and mental health symptoms among serious adolescent offenders. *Crim. Justice Behav.* **42**, 587–609 (2015).
- Fowler, P. J., Tompsett, C. J., Braciszewski, J. M., Jacques-Tiura, A. J. & Baltes, B. B. Community violence: a meta-analysis on the effect of exposure and mental health outcomes of children and adolescents. *Dev. Psychopathol.* **21**, 227–259 (2009).
- Guo, X. et al. Exposure to violence reduces empathetic responses to other's pain. *Brain Cogn.* **82**, 187–191 (2013).
- DuRant, R. H., Pendergrast, R. A. & Cadenhead, C. Exposure to violence and victimization and fighting behavior by urban black adolescents. *J. Adolesc. Health* **15**, 311–318 (1994).
- Guerra, N. G., Huesmann, L. R. & Spindler, A. Community violence exposure, social cognition, and aggression among urban elementary school children. *Child Dev.* **74**, 1561–1576 (2003).
- Hawkins, J. D. et al. Predictors of youth violence. juvenile justice bulletin. <https://eric.ed.gov/?id=ED440196> (2000).
- Albert Bandura. Social learning theory of aggression. *J. Commun.* **28**, 12–29 (1978).
- Huesmann, L. R. & Kirwil, L. Why observing violence increases the risk of violent behavior by the observer. in *The Cambridge handbook of violent behavior and aggression* 545–570 (Cambridge University Press, New York, 2007). <https://doi.org/10.1017/CBO9780511816840.029>.
- Ng-Mak, D. S., Salzinger, S., Feldman, R. S. & Stueve, C. A. Pathologic adaptation to community violence among inner-city youth. *Am. J. Orthopsychiatry* **74**, 196–208 (2004).



15. Ng-Mak, D. S., Stueve, A., Salzinger, S. & Feldman, R. Normalization of violence among inner-city youth: a formulation for research. *Am. J. Orthopsychiatry* **72**, 92–101 (2002).
16. Haidt, J. & Joseph, C. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus* **133**, 55–66 (2004).
17. Schein, C. & Gray, K. The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personal. Soc. Psychol. Rev.* **22**, 32–70 (2018).
18. Siegel, J. Z., Mathys, C., Rutledge, R. B. & Crockett, M. J. Beliefs about bad people are volatile. *Nat. Hum. Behav.* **2**, 750 (2018).
19. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cogn. Neurosci.* **19**, 1508–1519 (2007).
20. Todorov, A., Pakrashi, M. & Oosterhof, N. N. Evaluating faces on trustworthiness after minimal time exposure. *Soc. Cogn.* **27**, 813–833 (2009).
21. Stanley, D. A., Sokol-Hessner, P., Banaji, M. R. & Phelps, E. A. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl Acad. Sci. USA* **108**, 7710–7715 (2011).
22. Alexander, R. *The Biology of Moral Systems (Foundations of Human Behavior)*. (Aldine Transaction, New York, 1987). <https://doi.org/10.1086/293057>.
23. Gintis, H. Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179 (2000).
24. Boxer, P., Middlemass, K. & Delorenzo, T. Exposure to violent crime during incarceration: effects on psychological adjustment following release. *Crim. Justice Behav.* **36**, 793–807 (2009).
25. Baskin, D. & Sommers, I. Exposure to community violence and trajectories of violent offending. *Youth Violence Juv. Justice* **12**, 367–385 (2014).
26. Finkelhor, D., Turner, H., Ormrod, R. & Hamby, S. L. Trends in childhood violence and abuse exposure: evidence from 2 national surveys. *Arch. Pediatr. Adolesc. Med.* **164**, 238–242 (2010).
27. Fitzpatrick, K. M. & Boldizar, J. P. The prevalence and consequences of exposure to violence among African-American youth. *J. Am. Acad. Child Adolesc. Psychiatry* **32**, 424–430 (1993).
28. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
29. Selner-O'Hagan, M. B., Kindlon, D. J., Buka, S. L., Raudenbush, S. W. & Earls, F. J. Assessing exposure to violence in urban youth. *J. Child Psychol. Psychiatry* **39**, 215–224 (1998).
30. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5**, 39 (2011).
31. Yamagishi, T. & Yamagishi, M. Trust and commitment in the United States and Japan. *Motiv. Emot.* **18**, 129–166 (1994).
32. Rand, D. G., Ohtsuki, H. & Nowak, M. A. Direct reciprocity with costly punishment: generous tit-for-tat prevails. *J. Theor. Biol.* **256**, 45–57 (2009).
33. Delgado, M. R., Frank, R. H. & Phelps, E. A. Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* **8**, 1611–1618 (2005).
34. Hayes, A. F. PROCESS: a versatile computational tool for observed variable mediation, moderation, and conditional process modeling. (2012).
35. Baskin-Sommers, A. R. et al. The impact of psychopathology, race, and environmental context on violent offending in a male adolescent sample. *Personal. Disord.* **7**, 354–362 (2016).
36. Baskin-Sommers, A. R. & Baskin, D. Psychopathic traits mediate the relationship between exposure to violence and violent juvenile offending. *J. Psychopathol. Behav. Assess.* **38**, 341–349 (2016).
37. Kimonis, E. R., Frick, P. J., Munoz, L. C. & Aucoin, K. J. Callous-unemotional traits and the emotional processing of distress cues in detained boys: Testing the moderating role of aggression, exposure to community violence, and histories of abuse. *Dev. Psychopathol.* **20**, 569–589 (2008).
38. Hare, R. *Manual for the Revised Psychopathy Checklist 2* edn, (Toronto: Multi-Health System, 2003).
39. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)* (Washington, DC, 2013).
40. Bernstein, D. P. et al. Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse. Negl.* **27**, 169–190 (2003).
41. Noonan, M. P. et al. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proc. Natl Acad. Sci. USA* **201012246**. <https://doi.org/10.1073/pnas.1012246107> (2010).
42. Zapparoli, L. et al. Dissecting the neurofunctional bases of intentional action. *Proc. Natl Acad. Sci.* **115**, 7440–7445 (2018).
43. Noonan, M. P., Kolling, N., Walton, M. E. & Rushworth, M. F. S. Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement. *Eur. J. Neurosci.* **35**, 997–1010 (2012).
44. De Brito, S. A. et al. Reduced orbitofrontal and temporal grey matter in a community sample of maltreated children. *J. Child Psychol. Psychiatry* **54**, 105–112 (2013).
45. McCrory, E., Brito, D., A. S. & Viding, E. The impact of childhood maltreatment: a review of neurobiological and genetic factors. *Front. Psychiatry* **2**. <https://doi.org/10.3389/fpsy.2011.00048> (2011).
46. McCrory, E., Brito, S. A. D. & Viding, E. Research review: The neurobiology and genetics of maltreatment and adversity. *J. Child Psychol. Psychiatry* **51**, 1079–1095 (2010).
47. Blair, R. J. R. The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain Cogn.* **55**, 198–208 (2004).
48. Blair, R. J. R. Neurocognitive models of aggression, the antisocial personality disorders, and psychopathy. *J. Neurol. Neurosurg. Psychiatry* **71**, 727–731 (2001).
49. Buckholtz, J. W. Social norms, self-control, and the value of antisocial behavior. *Curr. Opin. Behav. Sci.* **3**, 122–129 (2015).
50. Nelson, R. J. & Trainor, B. C. Neural mechanisms of aggression. *Nat. Rev. Neurosci.* **8**, 536–546 (2007).
51. Anderson, E. The Code of the Streets. *Monthy. Atlantic* **273**, 81–94 (1994).
52. Besbris, M., Faber, J. W., Rich, P. & Sharkey, P. Effect of neighborhood stigma on economic transactions. *Proc. Natl Acad. Sci. USA*. 201414139. <https://doi.org/10.1073/pnas.1414139112> (2015).
53. Raudenbush, D. “I Stay by Myself”: social support, distrust, and selective solidarity among the urban poor. *Sociol Forum*, **31**, 1018–1039. <https://doi.org/10.1111/sofc.12294> (2016).
54. Wilson, W. J. *More Than Just Race: Being Black and Poor in the Inner City (Issues of Our Time)*. (W. W. Norton & Company, New York, 2009).
55. Buka, S. L., Stichick, T. L., Birdthistle, I. & Earls, F. J. Youth exposure to violence: prevalence, risks, and consequences. *Am. J. Orthopsychiatry* **71**, 298–310 (2001).
56. Hetey, R. C. & Eberhardt, J. L. The numbers don't speak for themselves: racial disparities and the persistence of inequality in the criminal justice system. *Curr. Dir. Psychol. Sci.* **27**, 183–187 (2018).
57. Goffman, A. On the run: wanted men in a Philadelphia ghetto. *Am. Sociol. Rev.* **74**, 339–357 (2009).
58. Fehr, E. On the economics and biology of trust. *J. Eur. Econ. Assoc.* **7**, 235–266 (2009).
59. Knack, S. & Keefer, P. Does social capital have an economic payoff? A cross-country investigation. *Q. J. Econ.* **112**, 1251–1288 (1997).
60. LaPorta, R., Lopez-de-Silanes, F., Shleifer, A. & Vishny, R. W. *Trust in Large Organizations*. (National Bureau of Economic Research, 1996). <https://doi.org/10.3386/w5864>.
61. Casciano, R. & Massey, D. S. Neighborhoods, employment, and welfare use: assessing the influence of neighborhood socioeconomic composition. *Soc. Sci. Res.* **37**, 544–558 (2008).
62. Garbarino, J. & Sherman, D. High-risk neighborhoods and high-risk families: The human ecology of child maltreatment. *Child Dev.* **51**, 188–198 (1980).
63. Goffman, A. *On the Run: Fugitive Life in an American City*. (Picador, New York, 2015).
64. Monahan, K. C., King, K. M., Shulman, E. P., Cauffman, E. & Chassin, L. The effects of violence exposure on the development of impulse control and future orientation across adolescence and early adulthood: time-specific and generalized effects in a sample of juvenile offenders. *Dev. Psychopathol.* **27**, 1267–1283 (2015).
65. Tangney, J. P., Stuewig, J. & Mashek, D. J. Moral emotions and moral behavior. *Annu. Rev. Psychol.* **58**, 345–372 (2007).
66. Zachary, R. A. *Shipley Institute of Living Scale*. (Western Psychological Services (WPS), Los Angeles, CA, 1991).
67. Wilkinson, G. S. *WRAT-3: wide range achievement test administration manual*. (Wide Range, Inc., Wilmington, DE, 1993).
68. Lapsley, D. K. & Lasky, B. Prototypic moral character. *Identity* **1**, 345–363 (2001).
69. Hare, R. D. et al. The revised psychopathy checklist: reliability and factor structure. *Psychol. Assess. J. Consult. Clin. Psychol.* **2**, 338–341 (1990).
70. Mathys, C. D. et al. Uncertainty in perception and the hierarchical gaussian filter. *Front. Hum. Neurosci.* **8**, 825 (2014).

### Acknowledgements

We thank those affiliated with the Connecticut Department of Correction, particularly Warden Scott Erfe and Dr. Patrick Hynes for their continued support of this research; and the research assistants who helped collect these data. This work was supported by a Clarendon and Wellcome Trust Society and Ethics award (104980/Z/14/Z), a Wellcome Trust ISSF award (204826/Z/16/Z), and the Academy of Medical Sciences (SBF001/1008).

### Author contributions

J.Z.S., M.J.C. and A.B-S designed the experiment. A.B-S. and S.E. collected the data. J.Z.S. analyzed the data. J.Z.S., A.B.S. and M.J.C. wrote the article with S.E. providing critical revisions.



**Additional information**

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-09962-9>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Journal peer review information:** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019