IDENTIFYING, MEASURING AND REGULATING THE PSYCHOLOGICAL BIASES THAT CONTRIBUTE TO POLITICAL VIOLENCE

Emile G. Bruneau and Rebecca Saxe

Brain and Cognitive Sciences Department

Massachusetts Institute of Technology

Members of groups involved in conflict face a number of tangible forces that drive conflict and inhibit reconciliation: competition for limited resources, a history of violence, differences in cultural and religious beliefs. Inter-group antagonism and political violence can clearly be motivated by such factors: a young man might be motivated to commit an act of violence against a member of the 'other' group because his relative was killed by them; because he believes that his land or resources are being stolen; because he sees his cultural or religious beliefs threatened by the other group.

Accompanying these socio-political factors are a suite of psychological factors that can also motivate hostility. The same young man could also be tipped towards violence, for example, by extreme empathy for the suffering of ingroup members, or lack of empathy for outgroup members; because he views the other side as untrustworthy or irrational; because he views their motivations as unworthy rationalizations rather than reasonable justifications. These psychological biases that can be just as potent as socio-political factors in driving conflict and preventing reconciliation.

In this paper, we will highlight some of the best-categorized psychological biases that may help drive inter-group hostility and prevent the resolution of intractable conflicts, how biases can potentially be reduced with positive interventions, and finally how promising new technology (e.g. functional

neuroimaging) might help us to better understand the cognitive underpinnings of unconscious psychological biases.

I. PSYCHOLOGICAL BIASES

Empathy

A loved one loses their parent to cancer; on television, an Israeli soldier is wounded and bloody; in the newspaper, a Palestinian mother cradles the body of her injured child. How do people react when others are in distress? Much of the time, we feel pain or sadness in response to another's suffering. A key component of this response is the suite of cognitive and affective capacities called empathy (Batson, 2009): people recognize emotional experiences in others, experience matched sensations and emotions, and are motivated to alleviate the others' suffering, which frequently results in helping behaviors.

Too often, though, we are likely to feel no pain, no sadness, and no motivation to help. Failures of empathy are especially likely if the sufferer is socially distant, for example, a member of a different social or cultural group. We often fail to detect such outgroup members' emotional experiences or perceive them in substantially distorted ways, and are only weakly, if at all, motivated to reduce their suffering. In fact, depending on the victim, we may feel secretly pleased about their misfortunes. Such failures of empathy lead to indifference, and may even facilitate further harm against outgroups. Examining failures of empathy at the intergroup level is particularly important because intergroup conflicts engender significantly more aggression than interpersonal interactions (Insko, Pinkley, Hoyle, Dalton, Hong et al., 1987). Although interpersonal morality prohibits people from harming others, engaging in violence on behalf of the ingroup is accepted in times of group conflict (Cohen, Montoya, & Insko, 2006). Here we take an interdisciplinary look—including affective, behavioral, physiological, and neural data—at intergroup empathic failures, and consider potential negative alternatives (i.e., Schadenfreude).

Intergroup Failures of Empathy

Empathy is generally recognized as a central component of the human condition; because it promotes prosocial behavior, it is an essential aspect of human social life. From a young age, typical people are affected by another's suffering: they 'step into the other person's shoes', 'feel their pain' and are motivated to help (Batson, 2009). One popular theory suggests that among typical people, empathic responses arise out of an automatic, universal mechanism in the human brain that detects another person's experience and activates a matching experience in the observer. In this view, shared neural circuits provide a direct functional bridge between first- and second-person experiences (Decety & Ickes, 2009). Seeing another human being in pain, observers thus *feel* the other's pain.

We know, however, that adults with normal empathic capacity also frequently fail to respond to another's suffering. This may be because people are less likely to detect and attend to another's suffering when the victim is distant in space, time, kinship, or across racial, political, or social group boundaries (Batson & Ahmad, 2009). Empathy is even fragile between minimal groups—groups in which the boundary is arbitrary (e.g., red team and blue team)—such that children randomly assigned to color teams show greater empathy for ingroup members than for outgroup members when those children are socially rejected (Masten, Gillen-O'Neel, & Brown, 2010). Recent studies are beginning to unpack the physiological and neural underpinnings of these empathic failures. For example, transcranial magnetic stimulation (TMS) was used to demonstrate that Black and White participants have a stronger visceral response when watching an ingroup member's hand (or even an artificially colored, purple, hand) being pricked by a pin, compared to the hand of an outgroup member (Avenanti, Sirigu, & Aglioti, 2010); and fMRI imaging has shown that in White and Asian participants, the shared neural circuit for pain is more active when viewing pictures of someone from one's own race, versus the other race, in pain (Xu, Zuo, Wang, & Han, 2009).

Thus, outgroup members—merely by virtue of who they are and not anything they have done—as compared to ingroup members elicit diminished perceptions of suffering, and fail to elicit equivalent physiological and affective empathic responses. More concerning is that these dampened empathic responses

are related to less *helping*. For example, people who attributed fewer uniquely human emotions (e.g., anguish, mourning) to opposite-race Katrina victims were also less willing to volunteer for relief efforts to help those victims (Cuddy, Rock, & Norton, 2007).

Competition and Schadenfreude

Social identity—'us' and 'them'—is most salient when groups are set in direct competition. Not surprisingly, intergroup competition strongly modulates empathic responding: distressed ingroup members typically elicit empathy (Batson & Ahmad, 2009), whereas competitive rivals' pain may even elicit pleasure, sometimes referred to as *Schadenfreude* (Smith, Powell, Combs, & Schurtz, 2009).

When individuals compete, brain regions associated with experiencing "reward" show positive activation when a competitor receives a painful electric shock (Singer, Seymour, O'Doherty, Stephan, Dolan, & Frith. 2006), or has rumors spread about them (Takahashi, Kato, Matsuura, Mobbs, Suhara, & Okubo, 2009). People also show activity in reward-related regions when they themselves have the opportunity to punish an uncooperative individual (de Quervain, Fischbacher, Treyer, Schellhammer, Schnyder et al., 2004).

Similar effects occur when the sufferer is not a direct competitor, but a member of a competitive group. Competitive outgroups may become targets of Schadenfreude following failures in intergroup competition, particularly if participants are reminded of their own group's inferiority prior to the outgroup's failure (Leach & Spears, 2009). In the context of a real-world sports rivalry, Red Sox and Yankees fans report feeling pleasure, and show activity in reward-related brain regions when they watch their rival fail against their favored baseball team, and also against a less competitive team in the same league (i.e., the Orioles). Attaching positive value to outgroup members' suffering may provide motivation for inflicting suffering. For example, people who show more reward-related activity when watching the rival team fail also report being more likely to actively harm the rival team's fans (Cikara, Botvinick, & Fiske, under review).

Competitive groups may also become targets of Schadenfreude simply by virtue of the stereotypes associated with their group. While people self-report

feeling neutral watching a high-status, competitive stranger (e.g., an investment banker) sit in gum on a park bench, they also smile (i.e., cheek muscle engagement, measured by facial electromyography), indicating the presence of positive affect (i.e., Schadenfreude), not just the absence of negative affect (i.e., feeling neutral) (Cikara and Fiske, under review). On a positive note, manipulating status and competition-relevant information can attenuate this reaction: people exhibit a more empathic response when the unfortunate target is perceived as having lowered-status or as cooperative (Cikara & Fiske, under review).

Schadenfreude is thus a powerful, and common, alternative to empathy, offering positive emotions and self-affirmation in the face of a competitive threat (Leach & Spears, 2008). The lure of Schadenfreude can even overpower self-interest: people feel pleasure at rivals' misfortunes, even when the misfortunes have negative implications for themselves and society more broadly (Combs, Powell, Schurtz, & Smith, 2009). For example, Democrats, especially those who strongly identified with their political party, reported considerable Schadenfreude as a result of reading an article that noted a mild economic downturn that occurred during a Republican administration. Schadenfreude may function as a signal of ingroup cohesion, in opposition to competitors. Demonstrating pleasure instead of empathy in response to someone's misfortune is a clear sign to both ingroup and outgroup members that one's interests are not aligned with the victim (Leach & Spears, 2009).

Paradoxically, people with the most empathy for members of their ingroup may thus experience the most Schadenfreude toward a threatening outgroup. When an outgroup is perceived as antagonistic, people respond less empathically to outgroup members, but also *more* empathically to ingroup members (Dovidio, Johnson, Gaertner, Pearson, Saguy et al., 2010). Agent-based simulations suggest that the motivation to help ingroup members, and hostility toward people from other ethnic or racial groups, may have co-evolved in humans: group survival is more likely when many members are willing to fight in inter-group wars and even sacrifice themselves to protect others in their group (Choi & Bowles, 2007). The most dramatic incidents of intergroup violence are consistent with these

suggestions: most suicide bombers are not psychopaths, but rather may experience 'perochial altruism', or high empathy selectively for their own group's suffering (Ginges & Atran, 2009).

These studies illustrate that empathy is quite easily over-ridden, and Schadenfreude is readily induced, even in mildly antagonistic groups (supporters of athletic teams) or arbitrary groups. Thus, absence of empathy, and presence of Scadenfreude, is likely to be highly prevalent in groups with a history of conflict or who are actively involved in hostilities (e.g. Israelis and Palestinians, Irish Catholics and Protestants, American military personnel and Afghani fighters). How completely empathy is suppressed (and Schadenfreude enhanced), and whether empathy failures are mediated by group membership (e.g. more prevalent for the empowered group) has not been directly investigated in members of real conflict groups. This is currently being investigated in our lab.

HIGHER-LEVEL PSYCHOLOGICAL BIASES

The combination of enhanced in-group empathy and a failure of out-group empathy may provide a 'hot', emotional motivation for political violence. At the same time, a group of 'cold', and seemingly more rational biases may also drive hostility.

In 2009, the Black Harvard professor Henry Louis Gates was arrested by a White Cambridge police officer outside of his home after a White passerby called the Cambridge police when she saw Henry Louis Gates forcing open his front door (which had been stuck). Many people assumed that the White woman would not have made the call, and the White police officer would not have made the arrest, had Henry Louis Gates not been Black. Other people assumed that Henry Louis Gates would not have acted confrontationally if the police officer were not White. Both the police office, and the professor, strongly denied any bias in himself, while endorsing the view that the other man was biased. In general, people have no problem acknowledging the existence of bias in human decision-making. But there is a 'bias blind spot' when they are reflecting on the influence of bias and self-interest in their own decisions: they report overwhelmingly that they themselves are more immune

to psychological biases than are others (Pronin and Ross, 2002; Ehrlinger et al., 2005). That is, humans are 'naïve realists', believing that they have an objective view of reality (Ross and Ward, 1995, 1996). This creates a problem when we encounter disagreement with another. Naive realism predicts that people first assume that the other person lacks the correct perspective on the issues – "If only they knew what I knew, they would agree with me". However, when simple exchange of information fails to resolve the disagreement, people quickly switch to the interpretation that the other person or group is inherently biased and irrational. For example, in a disagreement among students over academic policy, each side is more likely to ascribe 'valid' reasons over 'biasing' reasons for their own position, but 'biasing' reasons over 'valid' reasons for the student they disagree with (Pronin, Gilovich & Ross, 2004). This effect has also been demonstrated at the group level: when asked about their views of the conflict in the Middle East, Jewish and Arab American respondents each report that their own identities provide insights on the issues, while the others' identity confers bias (Ehrlinger, Gilovich & Ross, 2005).

The greater the divide in opinion, the more people assume that another's views are based on non-normative factors like bias and ideology. The perception of out-group bias is thus exacerbated by another psychological bias: partisans tend to over-estimate their disagreements with the other group. This 'false polarization bias' acts at the group level, amplifying disagreement between groups beyond the actual levels of disagreement, specifically for one's most strongly held views (Robinson, Keltner, Ward & Ross, 1995; Chambers, Baron & Inman, 2006).

The perception of out-group bias can fuel political violence. Perceiving the other as biased makes people less willing to cooperate or negotiate with the other side, and more inclined towards aggressive or competitive actions, like sanctions or shows of force (Kennedy and Pronin, 2008). This has been hypothesized to lead to a 'perception of bias-conflict spiral'. The first side sees the group differences as amplified, and differences in opinion are perceived as wider than they are; these differences in opinion accentuate the perception of the second side's views as biased and irrational; seeing the second side as biased leads the first side to choose conflict-escalating behaviors and reduce the tendency towards rational negotiation;

these actions reinforce the second side's perception of the first side as irrational and biased, thus continuing the cycle. Altogether, this spiral of psychological effects drives partisans towards more adversarial options such as political violence.

If naïve realism is a consequence of the human condition, and these psychological biases are present at the interpersonal as well as intergroup levels, is there any way to get past them? Although the vast majority of work on higher-level cognitive biases has been devoted to categorizing and describing them, the few studies that have attempted to ascertain how stable these biases are over time provide some tentative hope. For example, our own preliminary work has shown that, given the right intervention conditions, empathy biases and higher level cognitive biases can be altered between different cultural groups (Westerners and people in Arab/Muslim countries), and even groups embroiled in intractable conflict (Israelis and Palestinians).

II. CONFLICT RESOLUTION INTERVENTIONS

When two groups are in conflict, prejudice, discrimination and open hostility can thrive. Each group's perception of the other is characterized by failures of empathy and perceptions of bias. Conflict resolution and prejudice-reduction programs aim to turn this situation around by using several types of interventions: perspective-taking, role playing, simulation and positive intergroup contact. The general hypothesis of these programs is that improving attitudes for specific outgroup members can enhance attitudes towards the outgroup as a whole, thus engendering a willingness to help and reluctance to harm outgroup members.

In a handful of recent studies, such interventions have increased empathy for the outgroup. Heterosexual empathy towards homosexuals was enhanced following a guided simulation of exclusion and repression (Hodson et al., 2009). Chileans' empathy towards native Mapuche, and Bosnian Serbs' empathy towards Bosnian Muslims, was increased by perspective-taking (Cehajic et al., 2009). In an impressive large-scale field study, a radio drama in Rwanda depicting positive intergroup interactions increased empathy of Hutus towards Tutsis (Paluck, 2009). A conflict

resolution program in Sri Lanka demonstrated that the positive effects of interventions can be long-lasting: relative to control groups, Singhalese participants in a 4 day intergroup workshop expressed enhanced empathy towards Tamils, even a year after participating in the program (Malhotra and Liyanage, 2005). Another study conducted by our lab in the Middle East illustrated that positive effects from interventions can act very rapidly, improving attitudes of Israeli and Palestinian participants for each other even after a 20 minute interaction with an outgroup member. Furthermore, increased empathy can lead to improved attitudes towards, and willingness to help the outgroup (Hodson et al., 2009; Batson et al., 1997; Pettigrew and Tropp, 2008). For example, increasing empathy increased donations to an outgroup charity (Malhotra and Liyanage, 2005), and forgiveness for past atrocities (Cehajic et al., 2008).

While success is possible, interventions are not always beneficial: empathy, positive attitudes and helpful intentions toward an outgroup can also *decrease* following perspective-taking. For example, metastereotypes—thoughts about how one (as a majority group member) may be evaluated by an outgroup member—are activated when individuals empathize with an outgroup member in the context of an intergroup interaction. These thoughts have the deleterious effect of interrupting other-focused empathic responses that are required for prejudice reduction. Moreover, among relatively high-prejudice participants, empathy-induction can elicit overtly *negative* reactions to a nearby outgroup member (Vorauer & Sasaki, 2009).

Intergroup interventions can also fail for one of the groups involved. In fact, a meta-analysis of conflict resolution programs based on the 'Contact Hypothesis' found that although the programs generally improve attitudes of the majority group towards the minority group, they are ineffective for improving attitudes of minority group members towards the majority group. This raises the possibility that interventions may interact with group membership to produce asymmetric effects. Although this idea has received little attention, recent studies have supported this notion. For example, a more 'assimilationist' orientation more effectively predicts positive interracial orientations among majority group members, while 'integration'

representations are more effective at predicting positive interracial orientations among minority group members (Dovidio et al, 2000; van Oudenhoven et al., 1998; Verkuyten & Brug, 2004).

Our own preliminary work shows an asymmetric effect of intervention type on attitudes of Israelis and Palestinians towards each other. In a study conducted simultaneously in Tel Aviv and Ramallah, Israelis and Palestinians were exposed to a member of the other group in a surprise, online interaction in which they either wrote about 'one or two of the most difficult aspects of life in [their] country' ('perspective-giving'), or read what a member of the other group wrote about this topic, summarizing that view at the end ('perspective-taking'). We found that Israeli attitudes towards Palestinians significantly improved only in the perspective-taking condition, and Palestinian attitudes towards Israelis significantly improved only in the perspective-giving condition (Bruneau, Cohen and Saxe, unpublished).

Understanding the causes and contexts of interventions, and the short and long-term effects of interventions on both groups, is critical to better understanding the positive effects and unintended consequences of conflict resolution efforts. Unfortunately, well-controlled empirical studies of prejudice-reduction and conflict resolution programs remain rare, and relevant data are scarce (Paluck & Green, 2009). Since well-intended programs sometimes have no effect or even negative effects, or the effects are asymmetrical for the groups involved, it is particularly important that empirical evaluations of these programs match the pace of their creation. An essential step in this process is to develop evaluation tools that effectively and authentically evaluate attitudes and beliefs towards the 'outgroup'.

III. NEUROIMAGING AS AN EVALUATION TOOL

A recent evaluation of conflict resolution programs done by the U.N.-affiliated International Conflict Resolution Research group found that "…evaluation theory specific to conflict resolution has not kept up with the demand, leaving the field comparatively lagging in this endeavor" (Church and Shouldice, 2002). That is, there are few reliable measures of how or even whether these programs are working. In

order to evaluate the efficacy of conflict resolution programs or activities, it will be necessary to have measures that are sensitive, accurate and predictive of behavior. An ideal measure would identify who was affected by the program, how much, and in what ways. It would also accurately predict which changes are likely to transfer to 'real-life' outcomes, like curtailing inter-group hostility and enhancing positive attitudes towards the out-group.

Quantitative evaluation of conflict resolution programs will depend on developing sensitive and accurate measures of beliefs, attitudes, and emotions about the out-group and the conflict. Three classes of measures may be useful: (1) explicit questions that directly assess individual beliefs, attitudes and emotions about other groups; (2) implicit measures that generally examine associations with the out-group that are outside of conscious awareness and/or control; and (3) neuroimaging measures that examine brain responses to in-group and out-group stimuli.

The most common way to assess inter-group attitudes and beliefs is explicitly, though survey and self-report measures. Explicit surveys have been used extensively to assess White/Black attitudes. The most commonly used measure of attitudes and beliefs about Black Americans is the Modern Racism Scale, which consists of 6 or more questions that encompass both subtle racism (e.g. perceiving the media is biased against Black/White Americans) and overt racism (e.g. feeling opposed to interracial marriage) (McConahay, 1986). Another common explicit measure is the "feeling thermometer", on which participants rate how "warm" they feel towards the out-group, from 'cold/unfavorable' to 'warm/favorable' (Cairns et al., 2006).

Explicit measures are simple and convenient, but pose well-known methodological challenges because participants are motivated to present themselves in a positive light (Greenwald and Banaji, 1995). For example, White Americans who express positive attitudes and behavior intentions towards Black Americans nevertheless show impulsive avoidance of a Black confederate (Dovidio et al., 2002). Explicit measures also do not predict behavior when participants are unaware of behaviors that they exhibit. For example, school teachers who have an

explicit aversion to sexual discrimination nonetheless unwittingly treat boys and girls differently in the science and math classroom, calling on boys more often and giving them more time to respond (Jones and Wheatley, 2006). Therefore, whenever participants endorse a strong norm of equality and non-discrimination, stereotypes and subtle forms of bias may still affect behavior, but these effects will be hard to measure or predict with surveys and questionnaires.

An alternative approach to assessing inter-group attitudes is through implicit measures that tap physiological changes (e.g. heart rate, blood pressure and skin conductance) (Amodio et al., 2003; Guglielmi, 1999; Olsson et al., 2005) or response latency (Dovidio et al., 2002; Dovidio et al., 1997). The most widely used response latency measure is the implicit association test (IAT) (Greenwald et al., 1998). In this test, words belonging to four categories (for example, good words, bad words, Black American names and White American names) appear sequentially. The participant then sorts the words as quickly as possible into two compound categories (e.g. White names/good words vs. Black names/bad words). The IAT depends on the observation that participants can make accurate sorting decisions *faster* when the category pairing is congruent with their implicit associations (e.g. for White participants: White/good, Black/bad) than when the pairing is incongruent (e.g. White/bad, Black/good). IATs have been used to assess implicit bias towards groups, including those defined by race, gender and political partisanship (Aberson et al., 2004; Greenwald et al., 2003; Knutson et al., 2007; Phelps et al., 2000).

Implicit tests have the potential to reveal unconscious associations that are opaque to the person being assessed, and also have been shown to be less susceptible to cognitive control: even when participants are aware that the test is being used to assess bias, the effect remains (Kim, 2003). Implicit measures may therefore provide better predictors of behavior than explicit surveys, particularly when normative pressures to be non-prejudiced are high (Blanton et al., 2009; Dovidio et al., 2002; Fazio and Olson, 2003; Greenwald et al., 2005). There are limitations to standard implicit measures, however. First, the output is usually a single measure generalized to positivity or negativity, so multiple interacting processes could be confounded. Second, implicit tests usually measure associations

rather than more complex beliefs and emotions. Third, what exactly the IAT measures is still debated, particularly since the IAT has been shown to be influenced by priming effects and training (Foroni and Mayr, 2005; Kawakami et al., 2007).

A more recent method of measuring inter-group bias is neuroimaging. In principle, neuroimaging could provide a measure that is less susceptible to pragmatic control and more proximal to behavior. Measures of neural activity also have the potential to unconfound multiple interacting processes, providing a more comprehensive view of behavioral precursors. Over the past decade neuroimaging has been increasingly used to look at inter-group bias among White and Black Americans. For example, studies have reported increased activation in the amygdala and decreased activation in the fusiform face area (FFA) when participants view out-group vs. in-group faces (Cunningham et al., 2004; Golby et al., 2001; Hart et al., 2000; Phelps et al., 2000). In these groups, activity in the amygdala correlated positively with an IAT measure but not with explicit measures of out-group bias (Cunningham et al., 2004; Hart et al., 2000).

Thus, previous neuroimaging studies indicate that quantitative neural measures of out-group bias are possible, at least for aspects of face perception. However, the generalizeability to many geopolitical conflicts is uncertain. Members of actual conflict groups (e.g. Hutus and Tutsis in Rwanda, Tamils and Singhalese in Sri Lanka, Israelis and Arabs in the Middle East) are often physiognomically indistinguishable. Also, a growing social psychological literature indicates that the driving force behind the escalation and perpetuation of conflict often lies in higher-level cognitive processing about the thoughts, motivations and beliefs held by the out-group (Ehrlinger et al., 2005). As described above, people see those who hold other political views as either uninformed, biased or irrational, fueling a negative feedback cycle, away from rational negotiation and towards political violence (Kennedy and Pronin, 2008).

Given the role of complex psychological biases in the perpetuation of conflict, it therefore seems likely that the most useful neuroimaging measure predictive of positive behavioral change towards the out-group would be at a relatively high level

of cognition. An appropriate neuroimaging assessment tool should fulfill three conditions: the diagnostic brain activity should (1) occur in regions involved in psychological biases in reasoning, (2) reflect beliefs, attitudes or emotions towards the out-group, and so be correlated with measures of out-group negativity across groups and individuals, and (3) be a better predictor of long-term behavior than existing explicit or implicit measures.

As a first step towards the goal of evaluating conflict resolution interventions and programs with neuroimaging, we developed a task that was designed to examine neural activity while members of conflict groups were engaged in a higher-level reasoning task (Bruneau and Saxe, 2010). Arab and Israeli participants were presented with statements related to the conflict in the Middle East that were favorable to the in-group or the out-group, and evaluated the 'reasonableness' of each statement while in an fMRI scanner. As predicted, participants rated pro-in-group statements as reasonable and pro-out-group statements as unreasonable. We found one brain region (the precuneus) that was sensitive to emotionally-valenced reasoning. In this brain region, activity was correlated across individuals with explicit and implicit measures of negative attitudes towards the out-group (Figure 1).

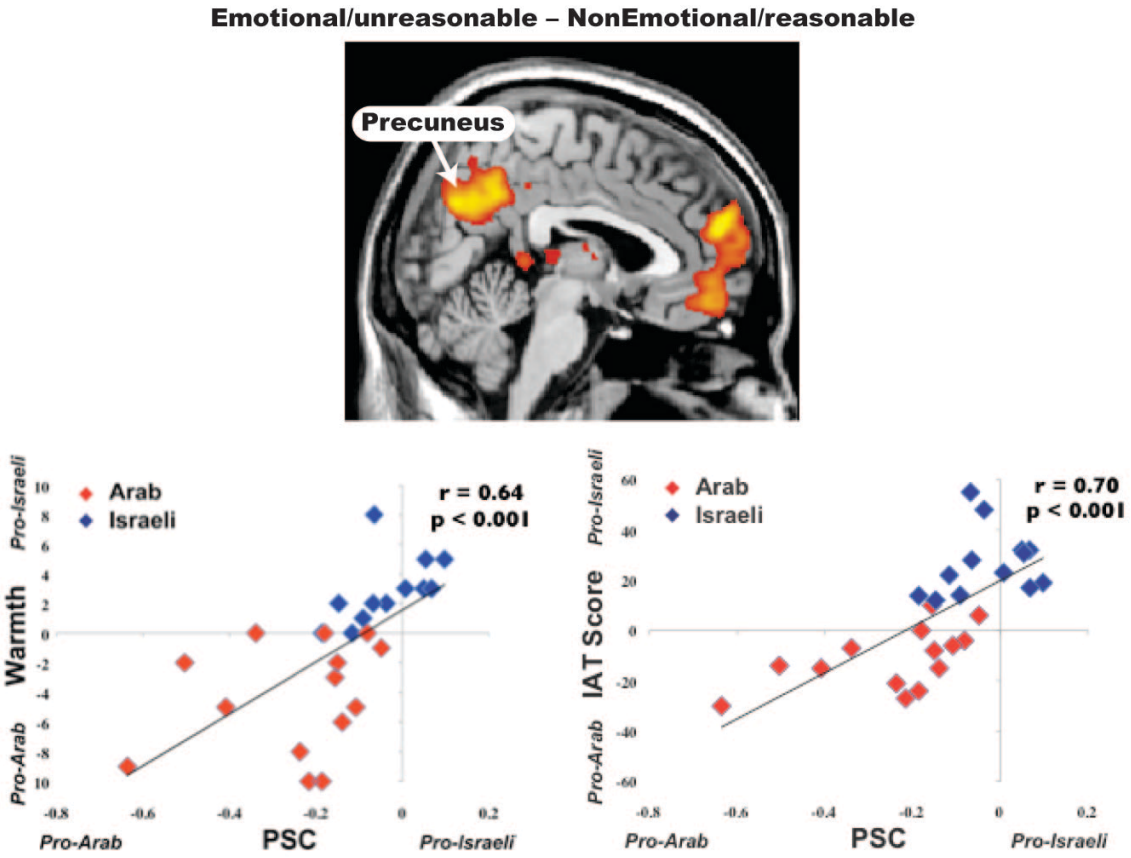

Figure 1. *Emotional and unreasonable control statements unrelated to the conflict in the Middle East (e.g. 'Hurricane Katrina was an act of God as punishment for the wickedness of the people in New Orleans…) generated activity in a number of brain regions. In one of these regions, the Precuneus, we found that activity for Pro-Arab statements (e.g. 'Israeli is effectively an Apartheid state…') was higher in Israeli participants, and activity for Pro-Israeli statements (e.g. 'Palestinians have wasted 60 years, in that time they could have created a modern state next to Israel, but instead they chose violence…) was higher in Palestinian participants. The difference in activity in the precuneus was predicted both by explicit measures of intergroup antipathy ('how warm or cold do feel towards Arabs' and 'how warm or cold do you feel towards Israelis'), and by implicit measures of negative outgroup associations.*

Program Evaluation.

The ultimate goal of our research program is to develop a neuroimaging measure that accurately predicts behavior in members of conflict groups. To determine how the present measure performed, it is useful to discuss how closely this task met our criteria of an ideal measure.

Criteria 1: activity is observed in brain regions associated with psychological bias

The more two people disagree about a political or moral issue, the more biased they perceive each other to be, and the less worthy of cooperative gestures (Kennedy and Pronin, 2008). Since conflict-escalating actions are driven by a perception of bias, it is important to develop cognitive and neural measures of these perceptions. Previous neuroimaging research had focused on measuring perception of out-group faces (Cunningham et al., 2004; Eberhardt, 2005; Golby et al., 2001; Hart et al., 2000). Our results suggest that fMRI can also be used to measure the neural correlates of high-level, cognitive components of bias toward the out-group. The present study was designed to find brain regions where activity was associated with emotion-laden cognition that involved judgments about bias-perception in members of conflict groups (Israelis and Arabs). Rather than images of faces, participants were presented with emotionally arousing statements. Also, participants were asked not whether they personally agreed with the statements, but how reasonable the statements were (i.e. whether anyone could reasonably agree with them). Activity in a particular brain region (the precuneus (PC)) was higher for statements (1) that were regarded as unreasonable by all participants, or (2) that specifically favored the outgroup. This activity was correlated, within and across participants, with judgments that pro-outgroup opinions were *unreasonable*. Thus, activity in the PC appears to be associated with one key aspect of psychological bias toward the out-group: the perception that their opinions and views are unreasonable, irrational, and biased.

Criteria 2: brain activity correlates with individual difference measures

A further characteristic of an ideal cognitive measure of conflict resolution programs is that it correlates with individual differences on behavioral measures. In the present study, we collected both explicit measures of out-group antipathy (warmth) and implicit measures of negative out-group associations (IAT). We found that responding on both of these measures correlated strongly with activity in the

PC. This provided the strongest internal support of our neuroimaging data as an accurate reflection of inter-group attitudes and beliefs.

Criteria 3: brain activity is a better predictor of behavioral change than either explicit or implicit measures

As a practical assessment tool of unconscious attitudes, neuroimaging of conflict resolution (i.e. "neuro-evaluation") shares a number of characteristics with neuro-marketing, which uses brain imaging to assess consumer preferences. Neuro-marketing assumes that consumer behavior is caused at least in part by subconscious motives that are undetectable by questionnaires or focus groups. Neuro-marketers aim to look "under the hood" at these motives, and thus hope to outperform surveys and focus groups in predicting subsequent consumer behavior (Fugate, 2008). Similarly, there is considerable evidence that the causes of inter-group behavior and attitudes are at least partially inaccessible to the participant themselves (and thus missed by standard explicit measures), and more differentiated than a simple positive-negative access (and thus missed by standard implicit measures; (Fazio and Olson, 2003; Greenwald and Banaji, 1995)). Like neuro-marketing, neuro-evaluation offers the chance to look "under the hood" at these causes of inter-group behavior.

Even more importantly, in both neuro-marketing and neuro-evaluation, actual behavior provides a ground-truth for comparing alternative behavioral and neural predictors. Neuro-marketers will be evaluated not by their ability to predict explicit or implicit attitudes in a focus group, but by their ability to predict actual buying behavior outside of the lab. Likewise, an ideal neural measure of inter-group hostility should not only be correlated with explicit attitudes and implicit associations, but specifically should outperform (cheaper and faster) behavioral assessments in predicting long-term pro-social and anti-social inter-group behavior outside of the lab (e.g. inter-group friendships, or voting for or participating in negotiations rather than violent conflict). In principle, we believe that neuro-evaluation could outperform behavioral assessments in exactly this way. For both

neuro-evaluation and neuro-marketing, however, this horizon remains a long way off. As a first step, future studies must include long-term measures of behavioral outcomes.

Concluding remarks:

As a practical method for determining higher-level psychological biases and evaluating conflict resolution programs, fMRI imaging has a number of hurdles that must be overcome: time on a scanner is expensive, access is limited, and the procedure is intrusive. However, neuroimaging can potentially circumvent the limitations of both explicit and implicit measures to provide a rich, complex quantitative measure that evades self-presentation pressures and is immediately proximal to behavior. It is therefore an important avenue of research. The results from our work make us cautiously optimistic that neuroimaging has the potential to live up to its theoretical promise of predicting behavioral change in members of conflict groups. Further studies involving explicit and implicit measures, neuroimaging and measures of behavior, preferably over time, will help to determine the true empirical utility of neuroimaging as an assessment tool.

SUMMARY

The psychological edifice erected between group members, often without their conscious awareness, combine with socio-political barriers to drive members of conflict groups towards aggressive intergroup behaviors and away from intergroup reconciliation. Crucially, group membership interacts with these psychological forces, potentially rendering uniform interventions less effective for one of the groups; in some conditions well-meaning interventions aimed at decreasing intergroup hostilities can even have an ironic effect. A better understanding of the strength of psychological factors on members of conflict groups, and evaluation of efforts to address these biases require evaluation tools. One promising tool is neuroimaging, which has been demonstrated here to provide a neural measure of higher-level inter-group bias that correlates strongly with both explicit and implicit measures of inter-group hostility. Using a combination of

explicit, implicit and neuroimaging measures, our current and future work aims to better understand the psychological factors driving conflict and hindering reconciliation in members of groups involved in intractable conflicts.

Acknowledgments:

The authors would like to thank Mina Cikara for her contributions to the section on empathy. Support for this work was provided by a gift from the Alliance of Civilizations, Media Fund, the Air Force Office of Scientific Research (managed through the Office of Naval Research), and a grant from the Wade Family Fund at MIT.