PREPRINT. In press at the Oxford Handbook of Evolution and the Emotions (2022). Eds. Laith Al-Shawaf and Todd K. Shackelford

The Neutralization Theory of Hatred

Aaron Sell, PhD

Department of Criminology and Psychology

Heidelberg University

Coltan Scrivner

Department of Comparative Human Development

Institute for Mind and Biology

The University of Chicago

Mitchell Landers

Department of Psychology

Center for Early Childhood Research

The University of Chicago

Anthony C. Lopez

School of Politics, Philosophy and Public Affairs

Washington State University

Abstract

We argue herein that, while often conceptualized as an extreme form of anger, hatred is a human emotion distinct from anger, with unique triggers, conceptual orientations, and terminating conditions. An examination of the social conditions of our species' evolutionary history reveals that hatred evolved to address its own distinct adaptive problem: individuals whose existence was -- on balance -- costly to the hater. Because a well-designed system for solving this problem would have been tailored toward neutralizing those costs, we call this hypothesis 'the neutralization theory of hatred.' This theory places the features of hatred within a functional framework. Specifically, we argue that hatred is triggered by cues that an individual's existence causes fitness decrements for the hater. Cognitively, hatred orients the mind so as to view costs heaped onto the hated person as benefits to the hater -- thus motivating spiteful behavior -- and can be characterized as maintaining a negative intrinsic welfare tradeoff parameter toward the hated person. Behaviorally, hatred can motivate either avoidance or a predatory style cost infliction strategy that is designed to weaken, incapacitate, or terminate the target. Hatred can be a dangerous emotion, and we believe a more thorough understanding of its evolved function is crucial for developing strategies that help mitigate its costs to society at large.

Introduction

On March 16th 1984, Leon Gary Plauché ambushed and killed Jeff Doucet at the Baton Rouge Metropolitan Airport while Doucet was being transported to jail by the police. Leon spoke on a pay phone in the airport while waiting, and when the handcuffed Doucet was led by, Leon turned and fired one shot into Doucet's head. Local news captured the event on videotape.

Doucet had been taken into custody for kidnapping and molesting Plauché's son. But, strange as it is to say, Plauché did not seem to be angry with Doucet. Notably, Plauché shows no evidence of rage. His face -- as best it can be made out from the recording -- is calm, his mouth closed. His body is still before he fires. He utters no vocalizations, no yells, no insults, no cries, even after shooting. He does not pace back and forth; indeed, he surrenders calmly and even places the phone receiver he had been talking into back on the hook within a second of having fired a bullet into Doucet's skull. None of these behaviors is consistent with the empirical evidence of anger displays. Rather, we believe Leon Plauché was motivated by hatred -- and that his seemingly odd behavior starkly illustrates the functional distinctiveness of hatred.

In this chapter, we will describe the neutralization theory of hatred. We propose that this theory can explain the major features of hatred, including its triggering conditions, its effect on internal regulatory variables, and its behavioral strategies. The theory holds that hatred evolved via natural selection in order to address a specific selection pressure: the existence of individuals whose well-being imposed a net fitness cost on you. In simple parlance, some people are bad for you. In most species, the evolved solution to this adaptive problem is to heap costs upon the target in an economical way so as to diminish their ability to harm you. This is often done by killing the target (e.g. siblicide in various bird species). In humans, however, this lethal response occurs only in a minority of cases. Rather, human hatred appears to make use of a mixed bag of strategies to minimize the costs emanating from the hated target, including: i) information warfare to diminish the target's social power, ii) low level surreptitious cost infliction to diminish the target's health, well-being, power, and to incentivize social distance, iii) actual predatory-style aggression - potentially lethal - to diminish their power, and iv) avoidance of the target to minimize the costs emanating from them.

It is important to note that this chapter serves as a philosophical examination of the evolution of hatred at the start of a new theory. Very few of the posited design features of this hypothesized adaptation have been subjected to rigorous empirical testing. Enterprising researchers will find many hypotheses worth exploring.

Theoretical approach

How can we know the function of an emotion?

This chapter takes an adaptationist approach (Cosmides & Tooby, 2000, Tooby & Cosmides, 1992; Williams, 1966) that argues that natural selection generates phenotypic design that is geared toward solving problems of reproduction. According to this framework, the function of an adaptation is proven to the extent that researchers show a close alignment between the nature of a selection pressure and the design of the hypothesized adaptation. In particular, adaptations must be shown to be efficient, economical, and well-designed to have solved the problems of reproduction faced by our ancestors. Thus, to prove the function of a given emotion, one must not only clearly state the hypothesized selection pressure that gave rise to it (i.e. how did this emotion effectively replicate the genes that gave rise to it in past environments?), but also must demonstrate how each design feature of the adaptation functionally addresses that selection pressure.

Thus, the adaptationist program requires a rigorous exploration of the basic features of any putative adaptation. Unfortunately, basic features are often invisible because of a phenomenon called "instinct blindness" (Cosmides & Tooby, 1994) that leads humans to underestimate their own complicated nature because it is so intuitive to them. This was illustrated most eloquently by the oft-quoted William James in his Principles of Psychology (1890):

> "To the metaphysician alone can such questions occur as: Why do we smile, when pleased, and not scowl? Why are we unable to talk to a crowd as we talk to a single friend? Why does a particular maiden turn our wits so upside-down? The common man can only say, "Of course we smile, of course our heart palpitates at the sight of the crowd, of course we love the maiden, that beautiful soul clad in that perfect form, so palpably and flagrantly made from all eternity to be loved!"
>
> "And so, probably, does each animal feel about the particular things it tends to do in presence of particular objects. They, too, are a priori syntheses. To the lion it is the lioness which is made to be loved; to the bear, the shebear. To the broody hen the notion would probably seem monstrous that there should be a creature in the world to whom a nestful of eggs was not the utterly fascinating and precious and never-to-be-too-much-sat-upon object which it is to her. Thus we may be sure that, however mysterious some animals' instincts may appear to us, our instincts will appear no less mysterious to them." (pg. 387)

He goes on to say, "*A priori*, there is no reason to suppose that any sensation might not in some animal cause any emotion and any impulse. To us it seems unnatural that an odor should directly excite anger or fear; or a color, lust." We linger on this point only to establish the importance of explaining what seems obvious about our emotions. If a man were to abuse your child, it seems most obvious that you would feel hatred. But a scientific explanation of this fact is not obvious! In the case of Leon Gary Plauché, such hatred brought no obvious benefit to his son; furthermore, Plauché himself only narrowly avoided spending the rest of his life in prison. If there were – ancestrally – a reproductive benefit to the hatred program that existed in Plauché's brain, it is not obviously evidenced by his shooting Doucet. One is tempted to posit that the "function" of the emotion is the justice it created, but this is exactly the kind of faulty reasoning that William James warned us about. Evolution does not select a gene because of "justice;" nature is simple mathematics – a gene spreads if, on balance, it increases its proportional frequency in the next generation and for no other reason.

One way to avoid instinct blindness is to reason as an engineer: given a problem, what would a well-designed solution look like? For instance, suppose you were a software engineer tasked with designing a robot to avoid physical danger. Would that program look like human fear? By focusing on the selection pressure rather than the adaptation, one can avoid some of the pitfalls of instinct blindness.

Here, we focus on the selection pressure we believe resulted in the emotion of hatred: individuals whose future existence is a net fitness cost to you.

Engineering a solution to the existence of toxic individuals

The existence of others affects your fitness (Aktipis et al., 2018; Hamilton, 1964; Tooby & Cosmides, 1996; Trivers, 1971). This selection pressure can be seen clearly from a simple thought experiment. A given individual – let us call him Leon – will leave a certain finite, quantifiable number of his genes in 100 years. If Leon were to use magic powers to "vanish" another individual from his social group, that number of future genes either increases, decreases, or stays the same. Following others (Petersen et al., 2010; Tooby & Cosmides, 1996), we refer to this delta value measuring the actual change in reproductive success as the "association value" (AV) of the individual such that AVxy represents the impact of X's existence on Y's reproductive success. In this formulation, association value is an objective indicator (like Hamilton's *r*) that refers to the actual effect of another's existence on one's fitness. We do not presuppose that individual organisms will have perfect knowledge of association value, just as they do not have perfect knowledge of *r* (Lieberman et al., 2007).

Thus, there exists at any time and for every individual a subset of others whose existence affects their fitness negatively (i.e., individuals with negative association value; for convenience, we refer to such people as "toxic"). Such individuals impose net fitness costs and thus serve as a selection pressure. To the extent that one can mitigate this damage, remove those individuals, or otherwise shape the environment to decrease the fitness costs of the toxic person, one will reproduce more, passing on to future generations the genes that gave rise to mechanisms producing those fitness-enhancing behaviors. Given this logic, humans should possess an adaptation that functions to: i) identify toxic individuals (i.e., those with negative AV to oneself) and ii) act in ways that minimize the fitness costs coming from these individuals.

1. Identifying toxic individuals

How could one know who these toxic individuals are? In the thought experiment above, we eliminated a person and waited 100 years to see the impact. Such an experiment would be – of course – impossible. Rather, animals and humans presumably evolved to detect cues that an individual's future AV will be negative and respond to those – admittedly imperfect – cues.

In the swallow-tailed kites (a bird native to northern Guatemala), mothers typically lay two eggs. The first hatchling, however, would need to share food and space with their clutch-mate sibling. Therefore, the existence of this second sibling is, presumably, a fitness cost to the first. The solution that evolved was simple – the first hatchling pecks at the skull of its sibling until it cracks and ejects it from the nest (Gerhardt et al., 1997). In this way, the first hatchling secures a monopoly on parental investment and typically proceeds to reproduce more – and this difference was, on average, enough to make up for the copies of its genes in its siblings that can no longer reproduce.

In the swallow-tailed kite, the cue of this toxic individual is relatively straightforward: it is the other chick in the nest. But for complicated social species such as our own, with debt, mutual dependence, status competition, intergroup conflict, shifting alliances, large-scale group cooperation, mate poaching, and so on, an accurate cue detector will be more difficult to instantiate.

The future is difficult to predict in most cases, but we can store evidence from the past. Because the best predictor of future behavior is past behavior (Epstein, 1979; 1980; Ouellette & Wood, 1998), one can reasonably conclude that a person who has imposed very large fitness costs on you will be more likely to impose similar costs in the future. Therefore, if we were to engineer an adaptation that functions to identify toxic individuals, it should at least respond to individuals who have imposed substantial costs on the individual (without corresponding benefits). Because association value is about the net effect of the person's future existence, small repeated costs would also predict low AV – possibly even more so than one large cost.

Furthermore, humans have – presumably for this sort of reason – evolved the ability to run metacognitive counterfactuals as a form of artificial time travel (Epstude & Roese, 2008). We are capable of estimating what would have happened if we were to change one aspect of a person or situation and predict the future based on that difference. This is key to our ability to blame and credit individuals for outcomes (Martin & Cushman, 2016). This same ability allows us to estimate what our circumstances would be like if a given person did not exist. Such a hypothetical enables us to identify individuals whose existence is bad for us (e.g. if Jessie weren't here, maybe Rick Springfield would have his girl; if Bin Laden hadn't existed, many of our friends and family would still be alive). Note, however, that in the second example, removing a person *after* they have imposed massive fitness costs would not necessarily be selected for (e.g. killing Bin Laden didn't raise the dead, nor did killing Doucet un-molest Leon's son), unless those past behaviors predicted future costs as well. From the point of view of a well-engineered hatred system, toxic individuals should ideally be identified early, before massive fitness consequences occur.

One way to identify toxic individuals before their existence imposes costs is to learn who the toxic individuals are from others – a process akin to mate copying (Gouda-Vossos, 2018). Only instead of learning who is a desirable mate, we learn who has negative association value (AV). This strategy relies upon our fellow humans to relay aspects of the toxic individual to us with high fidelity and comes with an additional inaccuracy – a person's association value will differ from one person to the next (e.g., Jessie's girl presumably benefits from Jessie's continued existence even if Rick does not). Therefore, this method of negative AV detection is probably less valid than personal experience unless the person you are learning from is similar to you in ways that predict the target has similar association values for both of you (e.g., an enemy of my child is usually an enemy of mine as well).

Finally, this AV detection mechanism may make use of other specialized systems that identify targets with negative AV. For example, envy may identify individuals whose continued existence deprives you of a share of resources or stands in the way of your optimal mate choice (this volume). Anger identifies individuals who treat your welfare with insufficient respect, in ways that will lead to future cost infliction (Sell, 2006; Sell et al., 2017; this volume). And so forth.

2. Minimize the fitness consequences coming from the toxic individual

Once individuals with low association value have been identified, the system must implement a change of behaviors or strategies to reduce the costs of that individual's existence.

The most theoretically clear solution is to cease their existence by killing them. This is a solution readily seen in the animal kingdom. For example, ground squirrels compete for food with prairie dogs, so the prairie dogs kill the ground squirrel's infants and leave them for scavenger birds (Hoogland & Brown, 2016). Of course, prairie dogs have a size and strength advantage over the infants they kill, making this behavior relatively low cost. Killing conspecifics can be much more costly because of the costs of fighting, the possibility of retaliation by friends and family of the deceased, and the – often negative -- social and reputational consequences that arise by demonstrating a willingness to kill individuals when it is in one's own interest. Furthermore, due to its permanent nature, the killing of toxic individuals makes it impossible to recoup cooperative benefits if one has miscalculated the association value of the person one has killed.

Nonetheless, the fact that killing the target is a permanent and often complete solution to the problem of a negative association value, and that it has been repeatedly selected for in other animals, suggests that killing a toxic individual should be part of human nature's toolkit of evolved responses, even if circumstances frequently make this option less practical.

Alternative strategies for dealing with toxic individuals are difficult to predict *a priori* without an understanding of how that individual depresses one's fitness interests. However, generally speaking, if a person's existence is depressing your fitness, the situation would usually have improved to the extent that this person's influence over the social world was diminished. Lessening this person's influence would be a particularly good solution if the negative fitness consequences stemmed from this person actively pursuing their own interests, e.g., in cases of resource competition, mate competition, status blocking, and so on. In those circumstances, lowering the toxic person's health, wellbeing, reputation, and status, would result in improved fitness because the toxic individual would be less capable of pursuing their own interests effectively. In short, toxic individuals should provoke in people a desire to harm them in cost-effective ways. Finally, if the toxic individual is depressing one's welfare via interactions with that person, then avoiding the toxic individual will be a potentially cost-effective means of reducing the damage done.

In sum, a simple information-processing analysis of the problem of negative association values in past environments leads to the prediction that natural selection should have designed a mechanism that functions to identify individuals whose existence in the future is costly to you, then enacts a suite of cognitive and behavioral procedures designed to minimize the negative fitness consequences of the target's existence. It is our contention that this simple analysis has identified the major features of the emotion of hatred. We call this the neutralization theory of hatred (see also, Sell & Lopez, 2020).

The neutralization theory of hatred

According to the neutralization theory, hatred evolved in order to neutralize the effects of individuals with negative association values, i.e. individuals whose existence is costly to you. In short, hatred responds to cues that, ancestrally, predicted that a person's continued existence and wellbeing was a net fitness loss to you. Once triggered by these varied cues, hatred calibrates the individual to treat the target differently. In particular, hatred leads to a negative intrinsic welfare tradeoff ratio (iWTR; an internal index that determines when a person will tradeoff on their own welfare to benefit another[1]; Delton, 2016; Sell, 2006, 2011; Sell et al., 2017; Tooby et al., 2008). A negative WTR means that the hateful person will spitefully accept costs in order to impose costs (or avoid benefiting) the target of hate. An intrinsic negative WTR is experienced as a lack of empathy, a desire to see the individual suffer (i.e. sadism when costs are inflicted by the hater, schadenfreude when they are inflicted by other means), and a preoccupation with thoughts of imposing costs on the target. Finally, hatred prudently enacts behavioral strategies that include a predatory-style of aggression, information warfare and ally recruitment, and avoidance of the target. Hatred, unlike anger, does not have ready terminating conditions that shut it off (see Sell, this volume). It is predicted to maintain itself as an orientation toward the target until their association value becomes positive, though the behavioral strategies of hatred are not designed to bring about this endpoint.

Triggers of hatred

Our functional analysis revealed four predicted triggers of hatred:

i) directly experiencing costs from that individual
ii) hypothetical reasoning about how one's life would be different if that person did not exist or was diminished in power,
iii) socially learning who others find toxic – with increased certainty put on the opinions of individuals who are similar to us or have shared interests, and
iv) the outputs of other specialized mechanisms that identify individuals with negative association value (e.g. other emotion systems)

Key to each one, and our central prediction, is that hatred is triggered by cues of a negative association value. In short, the existence of the target predicts future costs. Importantly – and as always – we refer to ancestral conditions in which modern genes were selected (Tooby & Cosmides, 1992), not to rational analyses of modern circumstances. For example, it was unlikely that Doucet was going to molest Plauché's child again: Doucet was in police custody and being led to jail. But police and prison are modern inventions.

Note that cost infliction itself is not a sufficient trigger to know that a person's net sum future impact will likely be negative. Instead, hatred should be particularly activated by costs that predict large future costs. These include:

1. Extreme costs that demonstrate a low welfare tradeoff ratio (WTR). Recall that WTRs are internal regulatory variables that indicate the extent to which a person will

[1] "Intrinsic" in this context refers to one's WTR in conditions when the target cannot effectively bargain for their own interests. Thus, having a high intrinsic WTR means one cares about the target's welfare even if they will never know about the tradeoff. For example, I have a high intrinsic WTR toward my family and care about them even in their absence, but I have a high "monitored" WTR toward my boss, whose opinion and welfare is extremely important to me *when* he is in the room.

sacrifice their own welfare to benefit yours or vice versa. They are typically calibrated to .4 to .7 for friends (Delton & Robertson, 2016). A person's WTR is revealed by the kinds of costs they are willing to impose on another in order to benefit themselves (Sell et al., 2017). A low WTR is revealed by imposing large costs on another for relatively trivial benefits (e.g., using your hand knit scarf to clean ketchup off of their face; burdening your child with painful sexual memories and trauma for a fleeting sexual experience).

Extreme costs alone may be insufficient if they do not reveal an extremely low WTR. For example, if Ted is inattentive and hits your child with his car, it may have devastating effects on your welfare, but this may be insufficient to provoke intense hatred. Now, compare this to Ted seeing your child playing in the street and slamming the gas because he thinks that's funny. This second scenario indicates a much more serious future threat to your fitness, because it reveals his stunningly low WTR toward you and your child. The evolutionary mechanisms needed to distinguish intentional harm from unintentional harm are beginning to be mapped (Martin & Cushman, 2017; Sell et al., 2017), but more work here will be useful for understanding hatred as well. Because intentional harms reveal a much lower WTR toward their target and are thus more predictive of future costs, we predict they will generate more intense hatred. However, it is important to note that hatred does *not* require a low WTR to be triggered. For example, if a woman's husband is sexually attracted to a young Irish woman, his wife may hate her even if she evinces perfectly acceptable levels of WTR. Indeed, the Irish woman could lavish respect and care upon the married woman and still be hated by her.

Finally, we should reiterate that repeated small costs can lead to hatred, even in the absence of a low WTR. For example, as argued before (Sell, 2012), cases of elder abuse and child abuse appear to result from the persistent negative effect of having to – at great expense of time, energy, and money – care for another individual's needs.

2. The ability to reason hypothetically about a person's non-existence or diminished power should lead to estimates of a negative association value and trigger hatred. The mapping of our ability to engage in counterfactual reasoning is beyond the scope of this paper (though see Martin & Cushman, 2016), but the fact that we can reason about how our life would be without a person's existence or wellbeing is sufficient to provide a cue of another's association value. When these hypotheticals are run, and we determine that we would be better off without the toxic person, hatred should be triggered. This is particularly clear in cases of envy or jealousy, where the target of hatred has not demonstrated a low WTR or in other ways evinced low moral character or a willingness to harm others. The fact remains, however, that this individual has resources, or a mate, or territory, that might be yours if they were gone.

3. There would have been a selection pressure to identify toxic individuals as early as possible (e.g. Plauche probably wished he knew that Doucet was a pedophile

much earlier than he did). One way of discovering such individuals earlier is to copy the information from others who may have already experienced the costs that emanate from the hated target. For this reason, we expect for hatred to spread socially such that individuals will copy hatred toward targets under some circumstances. Those circumstances likely include the following: i) you are more likely to copy the hatred of your loved ones and peers because if your loved one hates a person, this person's existence is probably toxic for you as well given the relationship between love and shared interests[2], ii) hatred is more likely to be copied when it is more widespread because this gives some converging evidence that the target is toxic to a large number of people, iii) hatred is more likely to be copied when the cause of cost imposition threatens the individual copying the hate of the individual who hates, e.g. parents may be particularly likely to copy Plauche's hatred of Doucet.

Unfortunately, the social learning of hatred suggests that it can become contagious. An error in perception can lead one person to hate another, which is then copied, and can create a snowball effect. Of particular concern is the fact that individuals who defend the hated person are – in perception – preventing the mob from neutralizing this toxic person, and thus are becoming costly themselves. The mob then lowers their estimate of the defender's association value and often hates them as well.

4. Finally, hatred is predicted to make use of calculations from other emotional systems (and possibly other systems more widely) to identify individuals whose existence is fitness suppressing. We highlight the following examples:

a. Anger.

Anger is designed to identify people whose welfare tradeoff ratios are below the appropriate negotiated level as perceived by the angry individual (this volume). In short, anger identifies those who do not value you sufficiently. Such a calculation means that the individual will be imposing more costs than they otherwise would. Note, however, that in most cases of anger the target is not hated. On the contrary, they are often loved (Averill, 1983; see Sell, 2011). This is because a "lower than it should be" WTR can still translate into an overall beneficial relationship. Indeed, anger is most common between family members or friends who maintain positive association values with each other, but still find room to negotiate over welfare tradeoff ratios.

That said, a low WTR will lead someone to impose large costs for relatively trivial benefits and does portend future costs that are not necessarily proportionate to the costs already imposed. In short, if someone were willing to knock your ice cream out of your hand for the laughs, what else would they be willing to do? Such a person – if

[2] We consider love to be the opposite of hatred. It identifies individuals whose existence causes positive fitness outcomes for us. It responds to cues that are usually the opposite of those of hatred, and motivates a very high intrinsic WTR. It also triggers fantasies of benefiting and sacrificing for the individual rather than at their expense.

they do not possess other compensating traits – would presumably be a net fitness suppressor for you, and would trigger hatred.

Anger and Hatred can both respond to targets who exhibit a low WTR. A full contrast of these emotions is beyond the scope of this paper, but we note that anger and hatred function distinctly in that anger attempts to recalibrate and bargain with a target, while hatred attempts to neutralize them. These - and other distinctions - are worth more empirical scrutiny.

b. Envy.

We consider envy an underexplored emotion – at least from the perspective most able to produce clear thinking about its function, namely evolutionary psychology (see Ramachandran & Jalal, 2017; Sznycer et al., 2017). Like Sznycer et al., we take envy to be an adaptation that identifies individuals who hold resources or status that would further our reproductive interests. This gives an incentive to deprive them of that status or power. In short, unless they possess offsetting traits, their existence is a cost to us, and we would be better off if they were to suffer a deprivation of life, status, or resources. Hatred, thus, can be triggered from envy. We consider the longstanding demonization of the wealthy and middle-man minorities to be, in part, a consequence of this emotion (see Sowell, 2016). While this form of envy is generally not functional in modern market economies, envy and hatred evolved in small scale economies with limited resources shared between small numbers of individuals.

c. Jealousy.

Mate competition is arguably as strong a selection pressure as resource competition (Darwin, 1871; Buss, 2005), and because of its competitive nature an individual can benefit by eliminating a toxic rival. Thus, the preconditions are met for hating one's romantic rival in a mate competition – again, presuming no compensating traits. Importantly, this explanation predicts hatred at one's rival, but not necessarily toward one's mate who may be thinking of straying. We consider Daly & Wilson's arguments about mate infidelity, spousal killings, and domestic violence as part of a mate guarding adaptation to be the best explanation of this phenomena (Daly & Wilson, 1988; Wilson & Daly, 1992). We will note – puzzling though it may be – that spousal killers frequently describe both hatred and love for their victims (Chimbos, 1978). The co-existence of seemingly opposite adaptations that are characterized by self-sacrifice, care, high intrinsic WTR on one hand, and spite, aggression, and negative intrinsic WTR on the other hand, remains to be explored.

d. Fear.

Individuals capable of imposing great costs on you are – of course – a danger. To the extent that this danger becomes likely, and – importantly – that the source of this danger does not provide useful benefits to you that outweigh this risk, then the feared person or persons are predicted to be hated, i.e. one would be better off without them. The rise in hate crimes toward Muslims after the 9/11 attacks may be an example of this (Disha et al., 2011).

e. Disgust.

While disgust serves multiple purposes (see Lieberman et al., 2007; this volume), one feature of disgust is that it identifies individuals who are potential disease vectors. Disgust triggers avoidance, but it can also establish that the target's existence is a net harm to you. As a result, hatred may be triggered by those who are "disgusting". While removing pathogen vectors could clearly be selected for, we also note that there is a long history of attacks on people who engage in other behavior that can trigger disgust, e.g. individuals engaging in deviant sexual practices, eating unusual foods, and so on. We note, of course, that compensatory factors that upregulate one's intrinsic WTR will counteract this such that hatred is not reliably activated toward one's child when they get the sniffles. Indeed, introspection suggests that love appears to deactivate disgust.

f. Shame.

Shame is believed to have evolved in order to slow or stop the spread of negative information about oneself (Sznycer et al., 2012; this volume). As such, one feature of shame is to identify the vectors of that negative information – the person who has this information and may spread it to others. Such a person's existence is harmful and would lead to a lower association value as a result. Should that value be negative, we predict that the shamed person should hate the bearer of negative information, even if the person has done nothing with that information.[3]

g. Hatred.

One of the effects of hatred is to heap costs upon the target. This means that a hateful person will likely have a negative association value toward their victim. In other words, if someone hates you, they will lie about you, look for costs to put on you, and fantasize about harming you. As a result, your life is likely to be worse off for their existence. Thus, hatred should be reciprocal. This has important implications for how hatred should express itself (see below).

Interestingly, this creates a perverse – but empirically verified prediction (Schopler & Compere, 1971) – which is that we should hate those that we have unjustly harmed. If you harm a person – you are presumably triggering hatred in them – which means that they are now an enemy who will likely work against you in the future. Thus, their continued existence is bad for you, triggering hatred.

[3] The journalist Christopher Hitchens relates an anecdote about Saddam Hussein killing his translator who was present (and translating) when UN officials spoke down to the dictator. His explanation was that Saddam did not allow anyone to live who witnessed him feel shame.

Finally, we should say that this list is likely not exhaustive. There may be cases where a target is hated merely because they have an incentive to harm you; e.g., a non-offending pedophile is still a person who potentially *wants* to molest your children.

Computational structure of hatred

According to the neutralization theory, the most significant effect of hatred is to set a negative intrinsic WTR toward the hated target. The more negative, the more "hated" the target is. Recall that welfare tradeoff ratios set the accepted discount rate on another's welfare when making decisions that impact you both. For example, a WTR toward .7 to my friend will cause me to impose costs of 10 on them if I benefit 8 or more, but not if I benefit only 6 or less (i.e., the decision rule is "take the self-beneficial action whenever the benefit to the self is more than the cost to the other times the WTR," or $B_x > C_y * WTR_{xy}$). A negative WTR means that one will take any benefit no matter how much it hurts the target (if Bx is a benefit, it will always be higher than a negative number). For example, an intrinsic WTR of -.5 calibrates the hateful person to exploit the target for benefits, to accept costs in order to hurt the target (if the cost is half the damage to the hated enemy or less), deny themselves some benefits because the target would benefit as well (if the hated target would benefit twice as much or more), and so on.

Herman Melville perfectly illustrated an extreme negative WTR at the conclusion of Moby Dick, when Captain Ahab uses his last breath to spit at the whale. The WTR logic is as follows: Ahab willingly gave up his last breath of air (presumably a weighty benefit as it was all he had left; let's say Bx = 100), in order to impose a trivial cost on the whale (spitting on an ocean-soaked mammal with thick skin; let's say Cy = 1). Thus Ahab's WTR toward the whale is revealed to be less than or equal to -100, an intense amount of hate that licenses extraordinarily damaging and spiteful behavior on Ahab's part.

The consequence of a negative WTR is that one should be aware and searching for opportunities to impose costs on the target. Such cost infliction is incentivized in the same way that we are incentivized to look for opportunities to help people we value intrinsically. In this way, hatred causes a desire to see that individual hurt whether we are causing their hurt or not (Rempel et al., 2019).

Welfare tradeoff ratios are believed to be used in many downstream cognitive systems (see Sell et al., 2017). For example, WTRs appear to govern memory such that higher WTRs lead us to pay more attention to the target and remember more information about them. Forgetting about a person (or an aspect of that person) is thus a trigger of anger because it reveals a low WTR (Sell, 2014). For hatred, the negative WTR presumably causes a similar increase in memory fidelity and for the same reasons. We need to know information about those that we value highly so that we can make choices that benefit them (e.g. I need to remember that my child has an allergy). Similarly, we need to know if our hated enemy has an allergy as well, so that we can make choices that harm them. For this reason, we predict that it is the magnitude of the WTR rather than its valence that increases memory.

Similarly, the recalibrational theory of anger predicts (see this volume) that holding a high WTR toward someone implies that their interests must be considered frequently. For example, when making a decision about whether to move, a woman presumably weighs the likely welfare impact the move will have on those toward whom she has a high WTR (e.g. her husband, her friends, her children, her family). But she will not likely consider the interests of those she has a low intrinsic WTR toward, e.g., her mail carrier, her colleague from HR, her ex-boyfriend. The interests of these individuals are not likely to be considered at all because the low WTR discounts those interests to the point where they would not sway the decision.[4]A highly negative WTR, however, should have the same effect! It is important to calculate the impact of one's decisions on toxic individuals who are imposing costs on you. In this way, a hated person (e.g. WTR = -1.0) is as important as a loved one (e.g. WTR = 1.0) when making one's decisions.

Other effects of high WTRs appear to reverse when the WTR is negative. For example, we enjoy spending time with those that we have high intrinsic WTRs toward, while we tend to avoid those we hate (Aumer & Bahn, 2016). With high intrinsic WTRs, we often experience vicarious enjoyment of happiness (e.g., my wife's smile when she looked at our daughter for the first time still makes me happy) and pain at their pain (e.g., the actual birthing process). With negative intrinsic WTRs, these effects appear reversed such that the pain of our hated enemies is enjoyable (e.g. Thomas Aquinas suspected one of the pleasures of heaven is that we can watch the torture of the damned), and the happiness of our enemies is experienced as suffering.

Behavioral strategies of hatred

Our functional analysis of the "toxic individual" selection pressure suggests three kinds of behavioral strategies for neutralizing the target:

- i) killing the target
- ii) weakening the target to limit their power and influence
- iii) avoiding the target

Killing a target, requires aggression, and aggression is one of the most reliable behavioral tendencies triggered by hatred. However, aggression can be done in different ways with different functions (Sell & Lopez, 2020; Wrangham, 2018). For example, anger triggers bargaining style aggression (see Sell, 2011) designed to force compliance or recalibrate welfare tradeoff ratios. According to the neutralization

[4] This results in a delightful trigger of anger in which a person is angry that their interests were not consulted, even if the decision was ultimately satisfactory. For example, a woman was angry at her husband for pulling into a restaurant that she wanted to go to, because he didn't ask her where she wanted to go (Sell 2014).

theory, the function of hatred-based aggression is to impose costs efficiently on the target in order to weaken them, diminish their physical or social power, or potentially kill them (Rempel et al., 2019). Thus the style of aggression activated by hatred is predicted to be "predatory" in nature.

Predatory aggression

We define predatory-style aggression as aggression used to inflict damage in the most efficient way possible - minimizing risk and maximizing impact, e.g., a lion stalking a gazelle, or a kite killing its sibling. It is characterized mostly by the features that are notably absent: i) no signaling, ii) no escalation, iii) no monitoring for surrender or submission, iv) continued aggression upon the target's submission, v) no interrogations of the target's motive or reasoning, and willful violations of the implicit rules of combat (see Sell & Lopez, 2000; Romero et al., 2014). Instead, predatory aggression should be characterized by:

i) Deception in order to minimize the chance for the victim to prepare. This is presumably why hatred does not have a corresponding facial expression,e.g., the anger face exaggerates cues of physical strength to bargain with the target (Marsh et al., 2005; Reed et al., 2014; Sell et al., 2014), but intense hatred appears to have no discernible reliable facial expression for the same reason the lion does not roar at the gazelle.

ii) Rapid deployment of most costly aggression. Predatory aggression is designed to inflict costs, not demonstrate fighting skill, and so the usual pattern of conspecific ritualized aggression wherein two animals fight for dominance has no purpose in predatory aggression. The kinds of aggression should be the most costly for the victim (constrained by the risks to the attacker, of course). For this reason, hatred-based human aggression should not make use of the usual rituals of aggression (e.g., pushing and shoving, staring contests, threats, and so on; see Sell, 2011).

iii) Aggression should be timed to victim vulnerability. Because hatred-based aggression is often surreptitious, and because retaliation, flight, and self-defense will often inflate the costs of a second attack, a hateful person should choose their first attack judiciously - timed to when it is most cost-effective. This style of aggression is again evident in the stalking behavior of predators who time their aggression to when prey is most vulnerable.

iv) As a corollary to point three above, signs of submission or fear will serve as evidence that the victim is not in a good position to fight back or defend themselves. As such, these responses should have an excited effect on the predatory aggression, as it indicates that this is a judicious time to attack a helpless victim.

v) Temporary increases in formidability (such as that provided by being in a group of like-minded people) should increase the probability that hatred will give rise to predatory aggression.

vi) Predatory aggression should be more likely than other kinds of aggression to be lethal. While hatred rarely leads to homicide (at least in the modern world, Pinker, 2012), the hatred adaptation was forged in an ancestral world with much more aggression. We cannot know the frequency with which intense hatred led to homicide, but we do note that research on homicidal fantasies shows that they are abundant and common in the modern world (Buss, 2006). The neutralization theory of hatred predicts that many (possibly most) of these fantasies are test runs – i.e. hypotheticals computed to learn the feasibility and practicality of terminating a hated other. Note that the function of the fantasy is to gather information; it is not as a final check before a behavioral strategy is immediately deployed. By analogy, just because I look in the cookie jar does not mean I'm going to cheat on my diet…. I'm just seeing what's there in case a decision needs to be made in the future.

Despite its utility, killing those you hate has several potent limitations, including: i) it can be impractical to carry out given the target's fighting ability or social position, ii) it may invite retaliation, iii) it cannot be undone - errors of judgment are permanent, and iv) it has potentially harmful reputational costs for the killer. Presumably for these reasons, natural selection has equipped hatred with alternative strategies that can also be effective at limiting the power and influence of a toxic person: information warfare.

Information warfare

A person's power is often determined by the status, prestige and concern shown them by others. If such a person uses that power in ways that go against your interests -- to the extent that their association value is negative for you -- then diminishing their social power can help ameliorate the damage done by the person. In short, one can "damage" a hated other and diminish their power and influence by recalibrating the status-seeking machinery in the minds of other humans in the social group.

We can predict how this is done by understanding the status-setting systems themselves. Borrowing from research on welfare tradeoff ratios and the recalibrational theory of anger (Sell, 2011; Sell et al., 2017; Tooby et al., 2008) and from direct work on status itself (Durkee et al., 2020), one can postulate that status-setting mechanisms should grant status to those who are capable of defending their own interests (e.g. fighting ability, coalitional strength), producing benefits for others (e.g. hunting skill, holding useful knowledge), and being inclined to benefit those in their group (e.g. loyalty, reciprocity). As such, we can predict that hateful people should attempt to spread information about the hated targets that minimize that target's value in the eyes of others: e.g., they are poor cooperators either from effort or ability; they are weak; they are promiscuous backstabbing cowards; and so on. The functional goal of this information warfare is to instill lower others' WTR toward the target, preferably to the point of engendering hatred toward them. By doing this, the hateful person can mobilize other people's hatred mechanisms and deprive the target of allies, friends, social power, and -- at times -- their life.

Crucially, there is no particular need for this negative information to be truthful (see also Petersen et al., 2020). Given that gossip and character assassination rarely allow for the victim to respond, great gains can be had against a target provided there is no one to counter the negative information. Again, the contagious nature of hatred makes this feature of the adaptation dangerous from a societal perspective. An innocent person, tarred with hateful gossip, will become a bad investment for defenders because those defenders will be seen as helping maintain a toxic individual. The mob will then lower their approximations of the defender's association value, and frequently hate them as well. This can create a perverse incentive for third parties who are now incentivized to hate a target who has no genuine toxic effects merely to avoid the appearance of defending them.

Attentional direction and Information gathering

Emotions frequently direct attention at certain aspects of the environment that predict which of the multiple strategies available to the emotion will be most effective (Cosmides & Tooby, 2000). Of course, hatred directs attention to the target, such that the appearance of a hated other will often distract from other tasks and emotions.

Hatred appears to focus attention on the hated target (Aumer & Bahn, 2015). For example, it would be difficult to concentrate on anything else if you were seated next to your sister's rapist. The hated target is important to attend to for the same reasons that a loved one is: one's decisions need to mold to the welfare of that other. It is important,therefore,to know what the person does and does not like, who their allies are, what debts they have, what secrets they hold, which individuals they are attracted to, which individuals they hate, and so on. While it is pleasurable to learn about someone you love, there is an odd tendency to both not enjoy but also feel compelled to learn about someone you hate. The phenomena of "hate following" people on social media, for example, involves individuals paying attention to the words and opinions of individuals they hate. Importantly, this phenomena is not a well-intentioned desire to understand another's perspective, but rather is a hunt for information that can be weaponized against them. Indeed, hatred shows an active aversion to understanding the perspective of the target. It suggests not just a disinterest in the target's defense, but a claim that no defense should be considered. For example, after the 9/11 attacks on the United States by Al Qaeda, the actor Richard Gere spoke in public and suggested that America attempt to "understand" why the terrorists did this and was roundly booed for his suggestion. Tactically, understanding the motives of one's enemies may be useful, but curiously hatred (at least intense hatred) appears to negate this - at least over some facets of motivation. We consider the most likely explanation of this phenomena to be that understanding motives and desires of an enemy will lead to negotiations over those conflicts (see Halperin et al., 2011), and that negotiations are incompatible with the function of hatred which is to nullify an enemy rather than appease one.

The selection pressure for this feature of hatred -- specifically the aversion to learning the motives and explanations for a hated target's behavior -- is hypothesized to be this: if a hated target is allowed to offer explanations, caveats, or apologies to the larger social group, this will diminish the ability of a hateful person to recruit allies against that target. This is because association values will differ from person to person (e.g., Doucet's existence may be less negative to someone who does not have children). The ability to negotiate one's toxicity is one tool a hated person can use to diffuse hatred; e.g., via upregulating WTRs to compensate as is typically done in apologies - "I know it can't be easy to live with someone like me… I'll be more conscientious in the future", or simply bestowing benefits to "cancel out" their negative effects. A person whose hatred is based on an extremely potent negative association value, may not want the target to be able to bargain at all. This final point may explain why figures who are hated are also silenced by the larger society.

Avoidance

Presumably the fitness costs of some individuals with negative association values can be blunted by merely not being near them; e.g., a colleague who never lets you get back to work; an ex-boyfriend who tries to shame you with questions. To the extent that avoiding these people reduced their fitness suppressing effects, natural selection would have selected avoidance as a feature of hatred. Indeed, hatred does appear to motivate avoidance of a target, unless – of course – one is intent on aggression (see Aumer & Bahn, 2016).

Terminating conditions for hatred

What circumstances should lead hatred to deactivate? According to the selection pressure posited here, hatred should deactivate when the target's association value becomes zero or positive. This can occur for a number of reasons. We speculate on common cases here:

1. A misperception of association value is corrected

Hatred is activated by internal estimates of a target's association value. Those estimates are necessarily imperfect. If hatred is activated via a misperceived negative association value that later is corrected, hatred should deactivate and guilt should be activated to repair any damage done by hatred (this volume). This will sometimes happen upon re-evaluation of a target's actions, e.g., the police arrested the wrong man for killing your mother; the guy who kept pulling your pig-tails is actually flirting with you not bullying you; the stranger who is spying on your Facebook page turns out to be your long-lost brother.

2. The target recalibrates their WTR and this results in a positive association value

To reiterate, WTR is an internal regulatory variable that functions primarily to determine which self-interested actions to take and which altruistic actions to take (Delton & Robertson, 2016; Sell et al., 2017; Tooby et al., 2008). It stores (in colloquial terms) the degree of respect or regard one has for another. Raising another's low WTR is the primary function of anger (Sell, 2006; 2011; Sell et al., 2017; this volume), which contains distinct behavioral strategies and triggers. However, a low WTR will lead a person to impose costs on another – and deny them benefits -- and as such will (all else equal) lead to a lower association value. This is presumably why anger and hatred often activate together (i.e., anger bargains for better treatment while hatred neutralizes the target's power).

If anger successfully bargains for better treatment and the target apologizes and raises their WTR, then it is possible that the estimated association value for that target also becomes positive and hatred deactivates as well. Note,as mentioned earlier, that hatred can be activated even when the target has a high WTR toward you (e.g., the obsessive ex who is still in love with you has a high WTR toward you; the man who married the woman you were in love with could value you a lot).

3. Shifting alliance structures turn a hated enemy into an ally

Modern politics is replete with examples of enemies becoming allies and vice versa, often with a parallel shifting in the minds of the citizenry (e.g., the American movie Rambo 3 awkwardly ends with a tribute to the Taliban). The nature of these shifts are beyond the scope of the current paper, but given their regular occurrence ancestrally, we can assume that association value estimators should recalibrate upon new discoveries of alliance; e.g., the bully who teases you nonetheless defends you from a genuinely lethal threat.

4. New avenues of cooperation turn an enemy into a potential cooperator

Hatred, particularly mutual hatred, is costly. If there exists an opportunity to rekindle a cooperative relationship that would revert negative associations positive, this would be a potent selection pressure (McCullough, 2008). This will be particularly true when a change in circumstance or social patterns allows for new cooperation. Having a stake in someone else's welfare could be a potent tool for defusing hatred.

5. The costs of hatred outweigh the benefits

Hatred is costly. It can motivate spiteful behavior, trigger retaliation, squander attention and resources, and lead others to return hatred on you. If the function of hatred is to neutralize a toxic individual, but hatred fails at doing this because the target cannot be eliminated, cannot be diminished in power, cannot be warred against by coalitions spurred on by information warfare, and cannot be avoided, then the costs of monitoring and spiteful actions will be net costs on the hateful person. Under these circumstances, nature would select for hatred to deactivate rather than waste the effort on ineffective strategies. This conclusion depends on there having been frequent cases (ancestrally) where a hated person could not be cost-effectively neutralized. We are agnostic on this point, but it is a reasonable assumption that the strategies deployed by hatred should self-evaluate their success such that if – for example – an incident of predatory aggression worked effectively, then the hateful person may be more likely to continue that strategy. Or, if an avoidance strategy fails because the bully seeks out her victim, the victim may be more likely to switch strategies to aggression. If all strategies fail, it is likely that hatred will deactivate or remain dormant until circumstances change. This possibility makes an interesting prediction: a sudden resurgence of hatred should occur when a powerful hated target demonstrates a new weakness.

Conclusion

In conclusion, the neutralization theory of hatred appears to explain many of the major features of human hatred as the expression of an evolved adaptation that functions to neutralize individuals whose future existence will likely impose costs on the hateful person. This adaptation comes online in response to triggers that ancestrally predicted the existence of such a person. These triggers appear to include evidence of a low WTR (a trigger shared with anger), the outcomes of hypotheticals that reveal how one' life would improve without the person, social learning and "hate copying", and outputs from other evolved emotion systems that flag individuals whose continued existence is detrimental. Once activated, hatred coordinates a suite of cognitive responses including: i) recalibrate one's WTR toward the target to be negative, incentivizing spiteful behavior, ii) focusing attention on the target, iii) disengaging empathy for the target, and iv) frequent consideration of the target's welfare when making decisions. Hatred also activates a series of behavioral strategies designed to eliminate the target or minimize their power to negatively affect the hateful person's welfare. Predatory aggression is the most serious of these strategies, typically deployed after homicidal fantasies test the feasibility and practicality of the behavior. More commonly, a kind of informational warfare is deployed both to gather allies against the target but also diminish their social power. Finally, a strategy of avoidance may be pursued.

If correct, future research should be able to map out additional features of this complex adaptation which will further distinguish it from the anger system and other emotional systems. Such research should also be able to identify strategies for diminishing the negative effects of hatred at the societal level.

REFERENCES

Aumer, K., & Bahn, A. C. K. (2016). Hate in intimate relationships as a self-protective emotion. In *The psychology of love and hate in intimate relationships* (pp. 131-151). Springer, Cham.

Averill, J. R. (1983). Studies on anger and aggression: implications for theories of emotion. *American psychologist*, *38*(11), 1145.

Buss, D. M. (2005). *The dangerous passion: Why jealousy is a necessary as love and sex*. Odile Jacob.

Buss, D. M. (2006). *The murderer next door: Why the mind is designed to kill*. Penguin.

Chimbos, P. D. (1978). *Marital violence: A study of interspouse homicide*. San Francisco: R & E Research Associates.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. Cognition, 50(1-3), 41-77.

Cosmides, L. & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions, 2nd Edition*. (pp. 91-115.) NY: Guilford.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London, UK: Murray.

Daly M., & Wilson M. (1988). *Homicide*. Transaction Books.

Delton, AW & TE Robertson (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. Current Opinion in Psychology, 7, 12-16.

Disha, I., Cavendish, J. C., & King, R. D. (2011). Historical events and spaces of hate: Hate crimes against Arabs and Muslims in post-9/11 America. *Social problems*, *58*(1), 21-46.

Durkee, P. K., Lukaszewski, A. W., & Buss, D. M. (2020). Psychological foundations of human status allocation. *Proceedings of the National Academy of Sciences*, *117*(35), 21235-21241.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of personality and social psychology*, *37*(7), 1097. https://doi.org/10.1037/0022-3514.37.7.1097

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American psychologist*, *35*(9), 790. https://doi.org/10.1037/0003-066X.35.9.790

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and social psychology review*, *12*(2), 168-192.

Gerhardt, R. P., Gerhardt, D. M., & Vasquez, M. A. (1997). Siblicide in swallow-tailed kites. *The Wilson Bulletin*, 112-120.

Gouda-Vossos, A., Nakagawa, S., Dixson, B. J., & Brooks, R. C. (2018). Mate choice copying in humans: a systematic review and meta-analysis. *Adaptive Human Behavior and Physiology*, *4*(4), 364-386.

Halperin, E., Russell, A. G., Dweck, C. S., & Gross, J. J. (2011). Anger, hatred, and the quest for peace: Anger can be constructive in the absence of hatred. *Journal of Conflict Resolution*, *55*(2), 274-291.

Hoogland, J. L., & Brown, C. R. (2016). Prairie dogs increase fitness by killing interspecific competitors. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1827), 20160144.

James, W. (1890). *The principles of psychology* (Vol. 1). New York : H. Holt and Company.

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature*, *445*(7129), 727-731.

Marsh A. A., Adams R. B., & Kleck R. E. (2005) Why do fear and anger look the way they do? Form and social function in facial expressions. *Personality and Social Psychology Bulletin, 31*(1), 73–86.

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133-143.

McCullough, M. (2008). *Beyond revenge: The evolution of the forgiveness instinct*. John Wiley & Sons.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*(1), 54–74. https://doi.org/10.1037/0033-2909.124.1.54

Petersen, M. B., Osmundsen, M., & Tooby, J. (2020). The Evolutionary Psychology of Conflict and the Functions of Falsehood. *PsyArXiv. August*, *29*.

Petersen, M.B., Sell, A., Tooby, J., and Cosmides, L. (2010). Evolutionary Psychology and Criminal Justice: A Recalibrational Theory of Punishment and Reconciliation. In Høgh-Olesen, H. (Ed.) *Human Morality and Sociality*, Palgrave Macmillan.

Pinker, S. (2012). *The better angels of our nature: Why violence has declined*. Penguin Group USA.

Ramachandran, V. S., & Jalal, B. (2017). The evolutionary psychology of envy and jealousy. *Frontiers in psychology*, *8*, 1619.

Reed, L. I., DeScioli, P., & Pinker, S. A. (2014). The commitment function of angry facial expressions. *Psychological Science*, *25*(8), 1511-1517.

Rempel, J. K., Burris, C. T., & Fathi, D. (2019). Hate: Evidence for a motivational conceptualization. *Motivation and emotion*, *43*(1), 179-190.

Romero, G. A., Pham, M. N., & Goetz, A. T. (2014). The implicit rules of combat. *Human nature*, *25*(4), 496-516.

Schopler, J., & Compere, J. S. (1971). Effects of being kind or harsh to another on liking. *Journal of Personality and Social Psychology*, *20*(2), 155.

Sell, A. (2011) 'The recalibrational theory and violent anger' *Aggression and Violent Behavior*, *16,* 381-389.

Sell, A. (2012) 'Revenge can be more fully understood by making distinctions between anger and hatred', Commentary on McCullough, Kurzan & Tabak's 'Cognitive Systems for Revenge and Forgiveness', *Behavioral and brain sciences,* 36(1) 36-37.

Sell, A. (2014). Twelve triggers of anger and how they invalidate all major theories of anger and aggression. *International Society for Research on Aggression*, Georgia State University, July 19th.

Sell, A., Cosmides L. & Tooby, J. (2014) 'The human anger face evolved to enhance cues of strength*', Evolution and Human Behavior*, 35(5), 425-429.

Sell, A. & A.C. Lopez (2020). Emotional underpinnings of war: An evolutionary analysis of anger and hatred. In Ireland, C., Ireland J., Lewis, M. & Lopez A. (Eds.). *The International Handbook on Collective Violence: Current Issues and Perspectives*, Routledge.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., Rascanu, R., Sugiyama, L., Tooby, J. & L. Cosmides (2017) 'The grammar of anger: Mapping the computational architecture of a recalibrational emotion', *Cognition*, 168, 110-128.

Sell, A., Tooby, J., & Cosmides, L. (2009) 'Formidability and the logic of human anger', *Proceedings of the National Academy of Science*, *106(35),* 15073-78.

Sherif, M. (1954). *Intergroup conflict and group relations*. Wesleyan University Press.

Sowell, T. (2016). *Wealth, poverty and politics*. Hachette UK.

Sznycer, D., Seal, M. F. L., Sell, A., Lim, J., Porat, R., Shalvi, S., ... & Tooby, J. (2017). Support for redistribution is shaped by compassion, envy, and self-interest, but not a taste for fairness. *Proceedings of the National Academy of Sciences*, *114*(31), 8420-8425.

Sznycer, D., Takemura, K., Delton, A. W., Sato, K., Robertson, T., Cosmides, L., & Tooby, J. (2012). Cross-cultural differences and similarities in proneness to shame: An adaptationist and ecological approach. *Evolutionary Psychology*, *10*(2), 147470491201000213.

Tooby, J. & L. Cosmides (1996). Friendship and the Banker's paradox: Other pathways to the evolution of adaptations for altruism. *Proceedings of the British Academy, 88*:119-43.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. *The adapted mind: Evolutionary psychology and the generation of culture*, 19-136; Oxford University Press.

Tooby, J., Cosmides, L., Sell, A., Lieberman, D. & Sznycer, D. (2008) 'Internal regulatory variables and the design of human motivation: A computational and evolutionary approach', In Elliot, Andrew J. (2008) *Handbook of approach and avoidance motivation.* (pp. 251-271). New York, NY, US: Psychology Press.

Williams, G. C. (1966) *Adaptation and natural selection: A critique of some current evolutionary thought*, Princeton university press.

Wilson, M. I., & Daly, M. (1992). Who kills whom in spouse killings? On the exceptional sex ratio of spousal homicides in the United States. *Criminology*, *30*(2), 189-216.

Wrangham, R. W. (2018). Two types of aggression in human evolution. *Proceedings of the National Academy of Sciences*, *115*(2), 245-253.