

The Moral Plasticity Hypothesis: Reconceptualizing Human Nature, Evil, and Cruelty in the 21st Century

Kwan Hong TAN

Associate Faculty

Singapore University of Social Sciences

khtan055@suss.edu.sg

5 October 2025

Abstract

This thesis proposes the "Moral Plasticity Hypothesis" as a novel framework for understanding human nature's relationship to evil and cruelty. Rather than viewing humans as inherently good or evil, this work argues that human nature is characterized by evolved moral plasticity—an adaptive capacity for extreme behavioral flexibility that can manifest as either profound compassion or devastating cruelty depending on contextual triggers and moral foundation activation patterns. Through interdisciplinary analysis of philosophical, psychological, evolutionary, and neuroscientific evidence, this thesis demonstrates that evil is not a defect in human nature but an emergent property of our adaptive moral flexibility, with profound implications for understanding and preventing mass violence.

The research synthesizes insights from moral foundations theory [1], evolutionary psychology [2], neuroscience [3], and historical case studies of genocide and mass violence [4]. The central argument challenges both traditional philosophical dichotomies of good versus evil and contemporary psychological theories that attribute cruelty primarily to dehumanization or individual pathology [5]. Instead, the Moral Plasticity Hypothesis posits that the same psychological mechanisms that enable extraordinary altruism and cooperation also enable extraordinary cruelty and violence, depending on contextual activation patterns.

Key findings include: (1) Human moral intuitions are built upon evolved psychological foundations that can be activated in prosocial or antisocial directions; (2) The neural circuits underlying empathy and violence show significant overlap, creating a neurobiological basis for moral plasticity; (3) Institutional contexts serve as powerful amplifiers that can channel human moral plasticity toward either constructive or destructive outcomes; (4) Historical patterns of mass violence demonstrate consistent psychological mechanisms across different cultures and time periods, supporting the universality of moral plasticity.

The thesis concludes with practical implications for violence prevention, moral education, and institutional design. Rather than attempting to eliminate "evil" individuals, effective prevention strategies must focus on understanding and modifying the contextual triggers that activate antisocial moral responses. This approach offers a more nuanced and scientifically grounded foundation for addressing humanity's capacity for both extraordinary good and devastating evil.

Keywords: human nature, moral psychology, evil, cruelty, genocide, moral foundations theory, evolutionary psychology, neuroscience, moral plasticity

License: This work is licensed under CC BY-NC-ND 4.0. Any use must cite the author and cannot be modified or used for commercial purposes.

Related theses, conceptual frameworks, and methodological contributions by the same author are accessible via the following profiles:

[ORCID iD: 0009-0003-9276-2829](https://orcid.org/0009-0003-9276-2829)

[ResearchGate: Kwan Hong Tan](https://www.researchgate.net/profile/Kwan-Hong-Tan)

Table of Contents

I.	Introduction: The Enduring Question of Human Nature	4
II.	Philosophical Foundations: The Evolution of Evil	9
III.	Psychological Perspectives: The Architecture of Moral Judgment	20
IV.	Evolutionary and Anthropological Evidence	35
V.	Neuroscientific Insights: The Brain's Moral Circuitry	46
VI.	The Moral Plasticity Hypothesis: A Novel Framework	56
VII.	Case Studies: Moral Plasticity in Historical Context	69
VIII.	Implications and Applications	78
IX.	Conclusion: Toward a More Nuanced Understanding of Human Nature	88
	References	95
	Research Notes: Human Nature, Evil, and Cruelty	129

I. Introduction: The Enduring Question of Human Nature

A. The Contemporary Relevance of Ancient Questions

The question of whether human beings are fundamentally good or evil has persisted throughout human intellectual history, from ancient philosophical traditions to contemporary academic discourse. This enduring fascination reflects not merely abstract philosophical curiosity, but urgent practical concerns about how we organize societies, educate children, design institutions, and respond to moral failures. In the 21st century, as we witness ongoing genocides, terrorist attacks, mass shootings, and systematic oppression alongside remarkable displays of altruism, cooperation, and moral courage, the question of human nature's relationship to evil and cruelty demands renewed examination [6].

The traditional framing of this question as a binary choice between inherent goodness or inherent evil has proven inadequate for understanding the complexity of human moral behavior. Classical philosophers from Aristotle to Augustine, from Hobbes to Rousseau, have offered compelling arguments for various positions along this spectrum, yet none have successfully accounted for the remarkable moral flexibility that characterizes human behavior across different contexts [7]. Contemporary events continue to challenge simplistic categorizations: the same individuals who demonstrate extraordinary compassion in one context may exhibit shocking cruelty in another, while entire societies can shift from peaceful coexistence to genocidal violence within remarkably short timeframes [8].

This thesis argues that the persistence of this debate reflects a fundamental misframing of the question itself. Rather than asking whether humans are inherently good or evil, we should ask: what are the psychological, social, and institutional mechanisms that enable humans to exhibit such extreme moral flexibility? How can the same species that produces both Mother Teresa and Adolf Hitler, both the Underground Railroad and the Holocaust, both international humanitarian organizations and genocidal regimes, be understood through a coherent theoretical framework?

B. Genocides and Mass Violence in the Modern Era

The 20th and 21st centuries have witnessed unprecedented scales of organized violence and systematic cruelty, challenging optimistic Enlightenment assumptions about human moral progress. The Armenian Genocide (1915-1923), the Holocaust (1933-1945), the Cambodian Genocide (1975-1979), the Rwandan Genocide (1994), the Bosnian Genocide (1992-1995), and ongoing atrocities in Palestine, Syria, Myanmar, and elsewhere demonstrate that technological and social progress have not eliminated humanity's capacity for extreme cruelty [9].

These events share striking psychological and social patterns despite occurring across different cultures, political systems, and historical periods. In each case, ordinary individuals—teachers, farmers, civil servants, neighbors—participated in systematic violence against people they had previously lived alongside peacefully [10]. The perpetrators were not primarily psychopaths or individuals with obvious mental disorders, but rather normal people operating within transformed social and institutional contexts [11].

Christopher Browning's seminal study of Reserve Police Battalion 101 during the Holocaust revealed that middle-aged German men, many of whom were not Nazi Party members, could be transformed into efficient killers of Jewish civilians within a matter of months [12]. Similarly, research on the Rwandan Genocide has documented how Hutu neighbors turned against Tutsi neighbors with whom they had shared communities for generations, often using traditional farming tools as weapons of mass murder [13]. These patterns suggest that the capacity for extreme cruelty is not limited to particular cultures, ideologies, or historical periods, but represents a more fundamental aspect of human psychological architecture.

Simultaneously, these same historical periods have witnessed extraordinary displays of moral courage and altruism. During the Holocaust, thousands of individuals risked their lives to save Jewish victims, often with no expectation of reward or recognition [14]. In Rwanda, some Hutus protected Tutsis at great personal cost, while in Bosnia, individuals crossed ethnic lines to help former neighbors [15]. These examples of moral heroism occurred within the same social contexts that produced mass violence, suggesting that the same underlying psychological mechanisms may be capable of producing radically different behavioral outcomes.

C. The Inadequacy of Traditional Good vs. Evil Frameworks

Traditional philosophical and theological frameworks for understanding human nature have typically relied on binary categorizations that prove inadequate for explaining the complexity of human moral behavior. The Augustinian doctrine of original sin posits that humans are fundamentally corrupted and prone to evil, requiring divine grace or strong social institutions to constrain their destructive impulses [16]. Conversely, Rousseauian optimism suggests that humans are naturally good but corrupted by social institutions and inequality [17]. Both perspectives, while offering valuable insights, fail to account for the remarkable contextual variability in human moral behavior.

Contemporary psychological approaches have often perpetuated similar binary thinking through concepts like "evil individuals" or "good people who do bad things." The popular notion that genocide and mass violence result primarily from "dehumanization"—the psychological process of viewing others as less than human—has dominated academic and policy discussions for decades [18]. While dehumanization certainly plays a role in some forms of violence, recent research by psychologists like Paul Bloom has challenged this explanation, arguing that perpetrators often maintain full awareness of their victims' humanity while still choosing to inflict suffering [19].

The inadequacy of traditional frameworks becomes particularly apparent when examining the psychological profiles of genocide perpetrators. Extensive research has failed to identify consistent personality traits, mental disorders, or demographic characteristics that reliably predict participation in mass violence [20]. Instead, studies consistently find that perpetrators represent a cross-section of normal human psychological variation, suggesting that the capacity for extreme cruelty is more widely distributed than traditional "evil individual" models would predict.

Similarly, research on moral heroism and altruism has failed to identify consistent psychological profiles that distinguish rescuers from bystanders or perpetrators [21]. The same individuals who demonstrate extraordinary moral courage in one context may fail to act in another, while people with no prior history of exceptional moral behavior may suddenly risk

everything to help others. This variability suggests that moral behavior is more dependent on contextual factors than on fixed individual characteristics.

D. Thesis Statement and Methodological Approach

This thesis proposes the "Moral Plasticity Hypothesis" as a novel framework for understanding human nature's relationship to evil and cruelty. The central argument is that human nature is characterized by evolved moral plasticity—an adaptive capacity for extreme behavioral flexibility that can manifest as either profound compassion or devastating cruelty depending on contextual triggers and moral foundation activation patterns. This plasticity is not a design flaw or pathological deviation, but rather an evolved feature that enabled human survival and success in diverse and changing social environments.

The Moral Plasticity Hypothesis consists of four core components:

1. **Adaptive Ambiguity:** Evolution selected for moral ambiguity as a survival advantage, allowing rapid adaptation to changing social environments and enabling both in-group cooperation and out-group competition.
2. **Contextual Moral Switching:** Humans possess evolved psychological mechanisms that can be triggered by specific contextual factors—threat perception, authority legitimization, group identity activation, and narrative framing—to produce dramatically different moral responses.
3. **The Empathy-Violence Paradox:** The same neural circuits and psychological mechanisms that underlie empathy and prosocial behavior also enable violence and antisocial behavior, creating a neurobiological basis for moral plasticity.
4. **Institutional Amplification:** Human institutions serve as powerful amplifiers that can channel moral plasticity toward either constructive or destructive outcomes, explaining how the same psychological mechanisms can produce both humanitarian organizations and genocidal regimes.

This thesis employs an interdisciplinary methodological approach, synthesizing insights from philosophy, psychology, evolutionary biology, neuroscience, anthropology, and historical analysis. The argument is constructed through systematic examination of empirical research, theoretical frameworks, and historical case studies, with particular attention to identifying patterns that transcend cultural and temporal boundaries. The goal is not to provide a complete explanation for all forms of human moral behavior, but rather to offer a more nuanced and scientifically grounded framework for understanding the relationship between human nature and moral extremes.

The implications of this framework extend beyond academic theory to practical questions of violence prevention, moral education, institutional design, and policy formation. If the Moral Plasticity Hypothesis is correct, then effective approaches to reducing human cruelty and promoting moral behavior must focus on understanding and modifying contextual triggers rather than attempting to identify and eliminate "evil" individuals. This represents a fundamental shift from person-centered to context-centered approaches to moral intervention.

II. Philosophical Foundations: The Evolution of Evil

A. Classical Perspectives on Human Nature

The philosophical investigation of human nature's relationship to evil and morality has ancient roots, with each major tradition offering distinct perspectives that continue to influence contemporary debates. Understanding these classical foundations is essential for appreciating both the insights and limitations of traditional approaches to human moral capacity.

1. Aristotelian Conceptions of Human Essence

Aristotle's approach to human nature in the *Nicomachean Ethics* and *Politics* established a framework that has profoundly influenced Western thinking about morality and human essence [22]. For Aristotle, humans are fundamentally rational and social beings (*zoon politikon*), whose nature is defined by their capacity for reason and their need for community. This perspective suggests that moral behavior flows naturally from the proper development and exercise of human rational capacities within appropriate social contexts.

The Aristotelian framework implies that evil and cruelty represent deviations from human nature rather than expressions of it. When individuals act cruelly or immorally, they are failing to actualize their essential human capacities for reason and social cooperation. This view has been influential in legal and educational traditions that emphasize moral development through the cultivation of virtue and the proper ordering of social institutions [23].

However, the Aristotelian approach faces significant challenges when confronted with the systematic and organized nature of much human cruelty. If humans are naturally rational and social, how can we account for the widespread participation in genocides, the development of sophisticated torture techniques, or the creation of institutions designed to inflict suffering? The Aristotelian framework struggles to explain why deviations from human nature appear to be so common and so systematically organized across different cultures and historical periods.

Furthermore, Aristotle's emphasis on the cultivation of virtue through habituation suggests that moral character is relatively stable once formed, yet empirical evidence demonstrates

remarkable moral flexibility in human behavior. The same individuals who demonstrate consistent virtue in one context may exhibit shocking cruelty in another, challenging the Aristotelian assumption that moral character represents a stable disposition [24].

2. Augustinian Original Sin vs. Pelagian Optimism

The theological debate between Augustine and Pelagius in the early Christian church established a fundamental tension that continues to influence secular discussions of human nature. Augustine's doctrine of original sin posits that human nature is fundamentally corrupted by the Fall, making humans naturally inclined toward selfishness, pride, and cruelty [25]. From this perspective, moral behavior requires divine grace or strong social constraints to overcome humanity's natural tendencies toward evil.

The Augustinian view offers a compelling explanation for the persistence and universality of human cruelty. If humans are naturally inclined toward selfishness and aggression, then the occurrence of genocide, torture, and systematic oppression becomes understandable as expressions of unconstrained human nature. This perspective has influenced political theories that emphasize the need for strong institutions to constrain human destructive impulses, from Hobbes's *Leviathan* to contemporary arguments for robust criminal justice systems [26].

Pelagius, by contrast, argued that humans are born morally neutral and possess the natural capacity to choose good or evil through the exercise of free will [27]. This optimistic view suggests that cruelty and evil result from poor choices, inadequate education, or corrupting social influences rather than from fundamental flaws in human nature. The Pelagian perspective has influenced educational philosophies that emphasize moral development through proper instruction and social reform movements that seek to eliminate evil through institutional change.

Both perspectives, however, face empirical challenges. The Augustinian view struggles to explain the remarkable capacity for altruism and moral heroism that humans demonstrate, often at great personal cost. If humans are fundamentally selfish and cruel, why do individuals regularly sacrifice their own interests for others, including strangers? Conversely, the Pelagian view has difficulty accounting for the systematic and organized nature of human cruelty,

particularly when it occurs within societies that have invested heavily in moral education and institutional reform [28].

3. Hobbes vs. Rousseau: The State of Nature Debate

The debate between Thomas Hobbes and Jean-Jacques Rousseau over the nature of humans in the "state of nature" represents perhaps the most influential secular discussion of human moral capacity in Western philosophy. Hobbes's famous characterization of natural human life as "solitary, poor, nasty, brutish, and short" reflects his view that humans are naturally competitive, aggressive, and self-interested [29]. Without the constraining influence of government and social institutions, Hobbes argued, humans would exist in a perpetual "war of all against all."

The Hobbesian perspective provides a framework for understanding human cruelty as a natural expression of competitive instincts that served evolutionary purposes but become destructive in complex social environments. From this view, genocide and mass violence represent the unleashing of natural human aggression when social constraints are removed or redirected toward out-group targets. This perspective has influenced realist theories of international relations and criminal justice approaches that emphasize deterrence and punishment [30].

Rousseau's counter-argument in the *Discourse on Inequality* and *The Social Contract* posits that humans are naturally compassionate and cooperative, but are corrupted by the development of private property, social inequality, and competitive institutions [31]. For Rousseau, the "noble savage" represents humanity's natural state of moral innocence, while cruelty and evil emerge from the artificial constraints and competitions created by civilization.

The Rousseauian perspective suggests that human cruelty results from social pathology rather than natural inclination. Genocide and mass violence, from this view, represent the products of corrupting institutions—nationalism, capitalism, authoritarianism—that distort natural human compassion and cooperation. This perspective has influenced progressive political movements and educational philosophies that seek to reform social institutions to allow natural human goodness to flourish [32].

Contemporary empirical research provides partial support for both perspectives while challenging their binary framing. Studies of human behavior in crisis situations reveal both Hobbesian competition and Rousseauian cooperation, often within the same individuals and contexts [33]. The key insight emerging from this research is that human behavior appears to be highly context-dependent rather than reflecting fixed natural tendencies toward either selfishness or altruism.

4. Xunzi's "Human Nature is Evil" vs. Mencius's Innate Goodness

The Chinese philosophical tradition offers additional perspectives on human nature that complement and challenge Western approaches. The debate between Xunzi and Mencius within Confucian philosophy parallels many themes in Western discussions while offering distinct cultural insights [34].

Xunzi's argument that "human nature is evil; goodness is the result of conscious activity" reflects a view similar to Augustine's original sin, but with important differences [35]. For Xunzi, humans are born with natural desires and emotions that, if unchecked, lead to conflict and disorder. However, unlike Augustine, Xunzi emphasizes that humans possess the capacity to transform their nature through education, ritual practice, and social cultivation. Evil is natural, but so is the human capacity for moral transformation through conscious effort.

This perspective offers a more dynamic view of human moral capacity than either pure pessimism or pure optimism. Xunzi's framework suggests that the tendency toward cruelty and selfishness is natural but not inevitable, while the capacity for moral behavior requires deliberate cultivation but is achievable through human effort. This view has influenced educational and political traditions that emphasize the importance of moral cultivation while acknowledging the persistent challenge of human destructive impulses [36].

Mencius, by contrast, argued that humans possess innate moral sentiments—compassion, shame, respect, and moral judgment—that naturally guide them toward virtuous behavior [37]. His famous example of the immediate compassionate response to seeing a child about to fall into a well illustrates his belief that moral emotions are spontaneous and universal human characteristics. From this perspective, cruelty and evil represent distortions or suppressions of natural moral sentiments rather than expressions of human nature.

The Mencian view anticipates contemporary research on moral emotions and empathy, suggesting that humans possess evolved psychological mechanisms that promote prosocial behavior [38]. However, like other optimistic views of human nature, it struggles to explain the systematic and organized nature of human cruelty, particularly when perpetrators appear to maintain their capacity for compassion toward in-group members while inflicting suffering on out-group victims.

B. Modern Philosophical Challenges

The development of modern philosophy brought new challenges to traditional conceptions of human nature and evil, fundamentally altering the terms of debate through scientific, historical, and existentialist insights.

1. The Enlightenment Rejection of Teleological Metaphysics

The Enlightenment critique of Aristotelian teleology fundamentally challenged traditional approaches to understanding human nature and morality. Philosophers like David Hume and Immanuel Kant argued that moral judgments could not be derived from factual claims about human essence or natural purposes [39]. This "is-ought problem" suggested that descriptive claims about human nature could not directly justify normative conclusions about how humans should behave.

Hume's analysis of moral sentiments proposed that moral judgments arise from emotional responses rather than rational analysis of human nature [40]. This perspective suggested that the capacity for both cruelty and compassion might be equally "natural" human responses, with moral evaluation depending on the particular sentiments that are activated in specific contexts. Hume's approach anticipated contemporary research on moral emotions while challenging the idea that human nature provides a stable foundation for moral judgment.

Kant's categorical imperative attempted to ground morality in rational principles that transcend particular human inclinations or natural tendencies [41]. From the Kantian perspective, moral behavior requires acting according to principles that could be universally applied, regardless of natural human tendencies toward selfishness or altruism. This approach suggests that both

cruelty and compassion represent responses to inclination rather than moral action, which must be motivated by duty to rational moral principles.

The Enlightenment rejection of teleological metaphysics created space for more empirical and scientific approaches to understanding human behavior while challenging the idea that human nature provides clear moral guidance. This shift opened the door for evolutionary, psychological, and sociological explanations of human moral capacity that did not depend on assumptions about essential human purposes or divine design [42].

2. Historicist Emphasis on Cultural Relativity

The development of historicist philosophy in the 19th century, particularly in the work of philosophers like G.W.F. Hegel and Wilhelm Dilthey, emphasized the cultural and historical variability of human moral beliefs and practices [43]. This perspective challenged universal claims about human nature by demonstrating the extent to which moral concepts and practices vary across different societies and historical periods.

Historicist analysis revealed that concepts of evil, cruelty, and moral responsibility are culturally constructed and historically variable rather than reflecting universal human nature [44]. What one culture considers virtuous behavior, another may condemn as cruel or evil. This variability suggests that human moral capacity may be more flexible and context-dependent than traditional philosophical approaches assumed.

The historicist emphasis on cultural relativity has influenced contemporary anthropological and sociological approaches to understanding human moral behavior. Rather than seeking universal principles of human nature, these approaches focus on understanding how different cultural contexts shape and channel human moral capacities in different directions [45]. This perspective supports the idea that human moral behavior is highly plastic and context-dependent rather than reflecting fixed natural tendencies.

However, historicist approaches face the challenge of explaining apparent cross-cultural universals in human moral behavior. Despite significant cultural variation in moral beliefs and practices, certain patterns—such as in-group favoritism, reciprocity norms, and prohibitions against unprovoked violence within the community—appear across diverse cultures [46].

These universals suggest that human moral capacity may have some universal features that constrain cultural variation.

3. Existentialist Conceptions of Radical Freedom

Existentialist philosophers like Jean-Paul Sartre and Simone de Beauvoir challenged both essentialist and deterministic approaches to human nature by emphasizing radical human freedom and responsibility [47]. From the existentialist perspective, humans are "condemned to be free" and must create their own values and identities through their choices and actions rather than following predetermined natural tendencies or cultural scripts.

Sartre's analysis of "bad faith" suggested that individuals often deny their freedom and responsibility by pretending that their actions are determined by external forces, natural inclinations, or social roles [48]. From this perspective, both cruelty and compassion represent authentic expressions of human freedom when chosen consciously and responsibly, while claims about natural human tendencies toward good or evil represent forms of bad faith that deny human agency.

The existentialist emphasis on radical freedom offers a framework for understanding human moral flexibility without appealing to fixed natural tendencies or cultural determinism. If humans are fundamentally free to choose their values and actions, then the capacity for both extreme cruelty and extreme compassion represents the full range of human possibility rather than deviations from or expressions of essential human nature [49].

However, existentialist approaches face challenges from empirical research demonstrating the extent to which human behavior is influenced by unconscious psychological processes, social pressures, and biological factors. While humans may experience a sense of freedom and choice, their actual behavior appears to be significantly constrained by factors beyond conscious control [50]. This suggests that human moral capacity may be more limited and predictable than pure existentialist freedom would imply.

4. Contemporary Evil-Skepticism vs. Evil-Revivalism

Contemporary philosophical debate about evil has been shaped by a fundamental disagreement about whether the concept of evil should be retained in moral and political discourse. Evil-skeptics argue that the concept is outdated, scientifically useless, and potentially harmful, while evil-revivalists contend that it captures important moral and psychological realities that cannot be adequately described using other concepts [51].

Evil-skeptics like Philip Cole and Inga Clendinnen argue that attributions of evil typically function to end inquiry rather than promote understanding [52]. When we label someone or something as evil, we often stop trying to understand the psychological, social, and historical factors that contributed to harmful behavior. This can impede efforts at prevention, rehabilitation, and reconciliation while promoting simplistic and potentially dangerous responses to complex moral problems.

Furthermore, evil-skeptics argue that the concept of evil carries supernatural and metaphysical baggage that is incompatible with scientific approaches to understanding human behavior [53]. If we want to develop effective strategies for reducing human suffering and promoting moral behavior, we need concepts that are grounded in empirical research rather than theological or metaphysical speculation.

Evil-revivalists like Eve Garrard and Luke Russell argue that the concept of evil captures important distinctions that cannot be adequately expressed using other moral concepts [54]. Some actions and individuals seem to go beyond ordinary wrongdoing in ways that demand a stronger moral response. The concept of evil, properly understood, can help us identify and respond appropriately to the most serious forms of moral failure without necessarily implying supernatural causation or ending inquiry into underlying causes.

The debate between evil-skeptics and evil-revivalists reflects deeper disagreements about the relationship between moral concepts and scientific understanding. The Moral Plasticity Hypothesis proposed in this thesis attempts to bridge this divide by offering a scientifically grounded framework for understanding extreme moral behavior without abandoning the moral significance that the concept of evil attempts to capture.

C. The Problem of Evil in Secular Context

The traditional theological problem of evil—how to reconcile the existence of evil with belief in an omnipotent, omniscient, and benevolent God—has been transformed in secular contexts into questions about the nature and significance of evil in human experience [55]. This transformation has generated new philosophical challenges while preserving many of the conceptual difficulties that have made the problem of evil so persistent and influential.

1. Narrow vs. Broad Conceptions of Evil

Contemporary philosophical discussions of evil typically distinguish between broad and narrow conceptions of the concept [56]. The broad conception includes any form of suffering, harm, or wrongdoing, encompassing everything from natural disasters to minor moral failures. The narrow conception restricts evil to the most extreme and morally significant forms of wrongdoing, such as genocide, torture, and systematic oppression.

The broad conception of evil faces the challenge of conceptual inflation—if everything bad counts as evil, then the concept loses its distinctive moral significance and explanatory power [57]. Natural disasters, accidents, and minor moral failures may all cause suffering, but they seem to differ qualitatively from deliberate acts of extreme cruelty in ways that matter for moral evaluation and practical response.

The narrow conception of evil attempts to preserve the concept's moral significance by restricting its application to the most serious forms of wrongdoing. However, this approach faces the challenge of drawing principled distinctions between evil and non-evil wrongdoing [58]. What makes genocide evil while other forms of killing are merely wrong? What distinguishes evil cruelty from ordinary selfishness or indifference?

The Moral Plasticity Hypothesis suggests that the distinction between evil and ordinary wrongdoing may be better understood in terms of the psychological and social mechanisms involved rather than the severity of outcomes alone. Evil, from this perspective, represents the activation of human moral plasticity in directions that produce extreme harm through the

systematic engagement of moral psychological mechanisms that normally promote prosocial behavior.

2. The Explanatory Power Debate

A central issue in contemporary discussions of evil concerns whether the concept has genuine explanatory power or merely serves expressive and rhetorical functions [59]. Evil-skeptics argue that calling something evil does not explain why it occurred or how to prevent similar occurrences in the future. Instead, they contend, we need to understand the specific psychological, social, and historical factors that contribute to harmful behavior.

This critique reflects broader concerns about the relationship between moral concepts and scientific explanation. If our goal is to understand and prevent human suffering, then we need concepts and theories that can guide empirical research and practical intervention rather than merely expressing moral condemnation [60].

Evil-revivalists respond that the concept of evil can have explanatory power when properly understood and integrated with empirical research [61]. Rather than ending inquiry, attributions of evil can guide attention toward particular psychological and social mechanisms that are involved in the most serious forms of wrongdoing. The concept of evil may help identify patterns and commonalities across different instances of extreme wrongdoing that might otherwise be overlooked.

The Moral Plasticity Hypothesis attempts to provide explanatory content for the concept of evil by identifying specific psychological mechanisms—moral foundation activation, contextual switching, empathy-violence circuits—that are involved in extreme moral behavior. From this perspective, evil represents a particular configuration of human moral psychology rather than a mysterious force or essential human characteristic.

3. Moral vs. Natural Evil Distinctions

The traditional distinction between moral evil (caused by human agency) and natural evil (caused by natural forces) has been complicated by advances in scientific understanding of human behavior and natural processes [62]. As we learn more about the biological,

psychological, and social factors that influence human behavior, the boundary between moral and natural evil becomes increasingly unclear.

If human behavior is significantly influenced by genetic factors, brain abnormalities, childhood trauma, and social pressures, then the extent to which individuals are morally responsible for their actions becomes questionable [63]. This challenge to moral responsibility has implications for how we understand and respond to extreme wrongdoing, including whether concepts like evil remain meaningful or useful.

Conversely, human influence on natural processes through technology, environmental modification, and social organization means that many apparently "natural" evils are actually influenced by human choices and institutions [64]. Climate change, pandemic responses, and natural disaster preparedness all involve human agency in ways that complicate simple distinctions between moral and natural evil.

The Moral Plasticity Hypothesis suggests that the distinction between moral and natural evil may be less important than understanding the mechanisms through which human moral psychology can be activated in different directions. Rather than focusing on individual moral responsibility, this approach emphasizes the contextual factors that shape moral behavior and the institutional mechanisms that can channel human moral plasticity toward constructive or destructive outcomes.

III. Psychological Perspectives: The Architecture of Moral Judgment

A. Moral Foundations Theory and Human Moral Intuitions

The development of Moral Foundations Theory (MFT) by Jonathan Haidt, Jesse Graham, and their colleagues represents one of the most significant advances in understanding human moral psychology in recent decades [65]. This framework provides crucial insights into the psychological mechanisms underlying human moral judgment and offers a foundation for understanding how the same moral architecture can produce both prosocial and antisocial outcomes.

1. The Six Moral Foundations: Care, Fairness, Loyalty, Authority, Purity, Liberty

Moral Foundations Theory proposes that human moral intuitions are built upon several innate psychological systems that evolved to address recurring adaptive challenges in human social life [66]. The original framework identified five foundations, later expanded to six with the addition of Liberty/Oppression, each corresponding to specific moral concerns that appear across diverse cultures and historical periods.

The Care/Harm foundation evolved from our mammalian heritage of attachment systems and parental care [67]. This foundation underlies moral intuitions about protecting the vulnerable, preventing suffering, and promoting welfare. It manifests in virtues like kindness, compassion, and nurturance, but can also motivate protective aggression against perceived threats to loved ones or innocent victims. The Care foundation demonstrates the inherent ambiguity of moral psychology—the same emotional and cognitive mechanisms that promote altruism can also justify violence when directed toward protecting the in-group from perceived threats.

The Fairness/Cheating foundation, recently subdivided into Equality and Proportionality components, evolved from the dynamics of reciprocal altruism and cooperation [68]. This foundation generates intuitions about justice, rights, and reciprocity, but its activation can vary dramatically depending on how fairness is conceptualized. Equality-based fairness emphasizes equal treatment and outcomes, while proportionality-based fairness emphasizes merit and

contribution. These different conceptions of fairness can lead to radically different moral judgments about the same situation, illustrating how moral foundations can be activated in multiple directions.

The Loyalty/Betrayal foundation reflects humans' evolutionary history as tribal creatures capable of forming shifting coalitions [69]. This foundation generates powerful moral emotions around group membership, solidarity, and collective identity. While loyalty can motivate extraordinary self-sacrifice for the group, it can also justify extreme cruelty toward out-group members who are perceived as threats to group cohesion. The same psychological mechanisms that create strong bonds within communities can also create the us-versus-them thinking that enables genocide and ethnic cleansing.

The Authority/Subversion foundation evolved from humans' primate heritage of hierarchical social organization [70]. This foundation generates intuitions about legitimate leadership, respect for tradition, and social order. Authority-based morality can promote social stability and coordination, but it can also enable the systematic oppression of subordinate groups and the legitimization of violence by those in power. The Milgram obedience experiments demonstrated how authority-based moral intuitions can override other moral concerns, leading ordinary individuals to inflict apparent harm on innocent victims [71].

The Purity/Degradation foundation, rooted in the psychology of disgust and contamination, generates moral intuitions about spiritual elevation, bodily integrity, and natural order [72]. This foundation often manifests in religious and cultural practices around cleanliness, sexuality, and sacred boundaries. While purity-based morality can promote self-discipline and spiritual development, it can also justify the dehumanization and persecution of groups perceived as contaminating or degrading. Many genocides have involved purity-based rhetoric that portrays victim groups as polluting influences that must be eliminated to restore moral and social order.

The Liberty/Oppression foundation reflects human sensitivity to domination and restriction of freedom [73]. This foundation generates moral emotions around autonomy, resistance to tyranny, and individual rights. Liberty-based morality can motivate resistance to oppression and the protection of individual freedom, but it can also justify violence against legitimate

authority and the rejection of necessary social constraints. The same psychological mechanisms that motivate freedom fighters can also motivate terrorists and anarchists.

2. Cultural Variations in Moral Foundation Prioritization

One of the most significant insights from Moral Foundations Theory is that different cultures, political groups, and individuals vary systematically in how they prioritize and weight different moral foundations [74]. These variations help explain moral disagreements and conflicts that might otherwise appear irrational or incomprehensible.

Research has consistently found that politically liberal individuals tend to prioritize the Care and Fairness foundations while showing less concern for Loyalty, Authority, and Purity [75]. Politically conservative individuals, by contrast, tend to draw more evenly on all moral foundations, showing greater concern for group loyalty, respect for authority, and purity-based values. These differences in moral foundation prioritization can lead to fundamentally different evaluations of the same moral situation.

For example, liberal and conservative responses to immigration often reflect different moral foundation priorities [76]. Liberals may focus primarily on Care-based concerns about the welfare of immigrants and Fairness-based concerns about equal treatment, while conservatives may emphasize Loyalty-based concerns about national identity, Authority-based concerns about rule of law, and Purity-based concerns about cultural integrity. Both groups are responding to genuine moral concerns, but they are prioritizing different aspects of the moral landscape.

Cross-cultural research has revealed even more dramatic variations in moral foundation prioritization [77]. Cultures that emphasize individual rights and autonomy tend to prioritize Care and Fairness foundations, while cultures that emphasize community cohesion and social harmony tend to place greater weight on Loyalty, Authority, and Purity foundations. These cultural differences can lead to profound misunderstandings and conflicts when groups with different moral priorities interact.

The variability in moral foundation prioritization provides a key insight for understanding human moral plasticity. Rather than having fixed moral intuitions, humans appear to possess a

flexible moral architecture that can be configured in different ways depending on cultural learning, social context, and situational factors. This flexibility enables humans to adapt to diverse social environments but also creates the potential for moral transformation under changing circumstances.

3. Evolutionary Origins of Moral Intuitions

The evolutionary perspective on moral foundations suggests that human moral psychology evolved to solve specific adaptive problems related to cooperation, competition, and group living [78]. Each moral foundation corresponds to challenges that human ancestors faced repeatedly over evolutionary time, leading to the development of specialized psychological mechanisms for detecting and responding to morally relevant situations.

The Care foundation evolved to address the challenge of raising vulnerable offspring and maintaining cooperative relationships [79]. The capacity to detect suffering and respond with helping behavior provided adaptive advantages in environments where mutual aid increased survival and reproductive success. However, the same mechanisms that promote care for in-group members can also motivate aggression against out-group threats, illustrating the inherent ambiguity of evolved moral psychology.

The Fairness foundation evolved to address the challenges of reciprocal altruism and detecting cheaters in cooperative relationships [80]. The ability to track contributions and benefits, detect unfair treatment, and respond with appropriate sanctions helped maintain stable cooperative relationships. However, different conceptions of fairness—equality versus proportionality—can lead to conflicting moral judgments about the same situation.

The Loyalty foundation evolved to address the challenges of group formation and maintenance in environments where intergroup competition was common [81]. The capacity to form strong group bonds and detect group threats provided advantages in conflicts with other groups. However, the same mechanisms that promote in-group cooperation can also enable out-group hostility and violence.

The Authority foundation evolved to address the challenges of coordination and leadership in hierarchical social groups [82]. The capacity to recognize legitimate authority and coordinate

group action provided advantages in complex social environments. However, authority-based morality can also enable the exploitation of subordinates and the legitimization of harmful actions by those in power.

The Purity foundation may have evolved from disease-avoidance mechanisms that helped humans avoid contamination and infection [83]. The capacity to detect and avoid potentially harmful substances and situations provided survival advantages in environments with infectious diseases and toxic materials. However, the extension of disgust-based responses to social and moral domains can lead to the dehumanization and persecution of out-group members.

The evolutionary perspective on moral foundations helps explain both the universality and the flexibility of human moral psychology. While all humans appear to possess the same basic moral foundations, the specific ways these foundations are activated and prioritized can vary dramatically depending on environmental and cultural factors. This combination of universal architecture and flexible activation provides the basis for human moral plasticity.

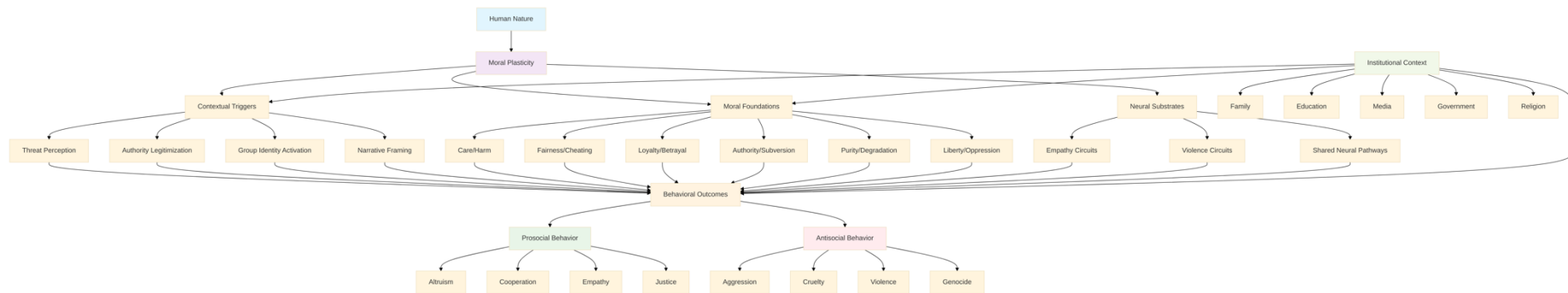


Figure 1: The Moral Plasticity Framework. This diagram illustrates the core components of the Moral Plasticity Hypothesis, showing how human nature's inherent moral plasticity is influenced by contextual triggers, moral foundation activation patterns, neural substrates, and institutional contexts to produce either prosocial or antisocial behavioral outcomes. The framework demonstrates that the same psychological mechanisms can lead to dramatically different moral behaviors depending on how they are activated and channeled.

B. The Psychology of Cruelty and Aggression

Understanding human cruelty requires examining the psychological mechanisms that enable individuals to inflict suffering on others, often in systematic and organized ways. Rather than representing a separate psychological system, cruelty appears to emerge from the same basic psychological mechanisms that normally promote prosocial behavior, but activated in different contexts and directed toward different targets.

1. Freudian Theories: Sadism and the Death Drive

Sigmund Freud's analysis of human aggression and cruelty evolved significantly over the course of his career, culminating in his controversial theory of the death drive (Thanatos) as a fundamental human instinct [84]. Freud's early work focused on sadism as a component of sexual development, but his later writings, particularly following World War I, proposed that humans possess an innate drive toward destruction and death that operates alongside the life drive (Eros).

Freud's concept of sadism initially emerged from his analysis of sexual perversions, but he gradually recognized that the capacity to derive pleasure from inflicting pain extended far beyond sexual contexts [85]. He observed that children often display cruel behavior toward animals and weaker peers, suggesting that sadistic impulses are present from early development rather than representing pathological deviations from normal psychology.

The death drive theory proposed that all living organisms possess an unconscious drive to return to an inorganic state, manifesting as self-destructive and aggressive impulses [86]. From this perspective, human cruelty represents the external expression of the death drive, directed outward rather than inward. While this theory has been largely rejected by contemporary psychology, it anticipated important insights about the relationship between aggression and other psychological processes.

Freud's emphasis on the unconscious nature of aggressive impulses highlighted the extent to which cruelty can emerge from psychological processes outside conscious awareness and control [87]. This insight remains relevant for understanding how ordinary individuals can

participate in systematic violence without necessarily experiencing themselves as cruel or evil. The psychological mechanisms that enable cruelty may operate below the threshold of conscious recognition, allowing individuals to maintain positive self-concepts while engaging in harmful behavior.

However, Freudian theories face significant limitations in explaining the contextual and cultural variability of human cruelty. If sadistic impulses are universal and innate, why do rates of violence and cruelty vary so dramatically across different societies and historical periods? Why do the same individuals who display cruelty in one context often demonstrate compassion in another? These patterns suggest that human cruelty is more context-dependent and flexible than Freudian drive theories would predict.

2. Self Psychology: Fragmentation and Narcissistic Rage

Heinz Kohut's Self Psychology offered a different perspective on human aggression and cruelty, emphasizing the role of narcissistic injury and self-fragmentation in motivating destructive behavior [88]. According to Kohut, aggression is not a primary drive but rather a response to threats to the self-structure, particularly experiences of empathic failure and narcissistic injury.

Kohut's concept of narcissistic rage describes the intense anger that emerges when individuals experience threats to their self-esteem, identity, or sense of coherence [89]. Unlike ordinary anger, which is typically proportionate to the triggering event and focused on specific goals, narcissistic rage is characterized by its intensity, persistence, and desire for revenge or destruction. This type of rage can motivate extreme cruelty as individuals attempt to restore their damaged self-esteem by dominating or destroying perceived threats.

The concept of self-fragmentation helps explain how individuals can engage in cruel behavior while maintaining their sense of moral identity [90]. When the self-structure is threatened or damaged, individuals may experience a temporary breakdown of normal psychological integration, leading to behavior that is inconsistent with their usual moral standards. This fragmentation can be triggered by various factors, including social rejection, status threats, identity challenges, or traumatic experiences.

Self Psychology's emphasis on empathic failure as a source of aggression provides important insights into the social and interpersonal dynamics that can lead to cruelty [91]. When individuals feel misunderstood, unrecognized, or invalidated by others, they may respond with aggressive behavior aimed at forcing recognition or inflicting similar emotional pain on others. This dynamic helps explain how cycles of violence and retaliation can escalate, as each act of aggression creates new experiences of empathic failure and narcissistic injury.

The Self Psychology perspective also highlights the role of group identity and collective narcissism in motivating large-scale violence and cruelty [92]. When groups experience threats to their collective identity, status, or self-esteem, they may respond with narcissistic rage directed toward perceived threatening out-groups. This dynamic can help explain how ordinary individuals become motivated to participate in genocide, ethnic cleansing, and other forms of collective violence.

3. Social Learning Theory and Observational Violence

Albert Bandura's Social Learning Theory provided crucial insights into how individuals acquire and maintain patterns of aggressive and cruel behavior through observation, imitation, and reinforcement [93]. This perspective emphasizes the role of environmental factors and social modeling in shaping behavior, challenging purely biological or drive-based explanations of human cruelty.

Bandura's famous Bobo doll experiments demonstrated that children readily imitate aggressive behavior they observe in adults, even when the aggression is directed toward inanimate objects [94]. These findings suggested that much human aggression is learned through social observation rather than emerging from innate drives or instincts. The experiments also showed that children often elaborated on the aggressive behaviors they observed, creating new forms of violence that went beyond their models.

The social learning perspective helps explain how cultures of violence can be transmitted across generations and how individuals can learn to engage in increasingly severe forms of cruelty [95]. When aggressive behavior is modeled by respected figures, reinforced by social approval, or associated with positive outcomes, individuals are more likely to adopt and

maintain these behavioral patterns. This process can lead to the normalization of violence within particular social contexts.

Social Learning Theory also emphasizes the role of moral disengagement in enabling cruel behavior [96]. Bandura identified several psychological mechanisms that allow individuals to engage in harmful behavior while maintaining their moral self-concept: moral justification (portraying harmful behavior as serving worthy purposes), euphemistic labeling (using sanitized language to describe harmful acts), advantageous comparison (comparing one's actions to worse alternatives), displacement of responsibility (attributing actions to authority figures), diffusion of responsibility (spreading responsibility across group members), distortion of consequences (minimizing or ignoring harm), and dehumanization (viewing victims as less than human).

These moral disengagement mechanisms help explain how ordinary individuals can participate in systematic cruelty without experiencing significant guilt or moral conflict [97]. By restructuring their cognitive representation of the situation, individuals can maintain their sense of moral identity while engaging in behavior that would normally violate their moral standards. This process is particularly important for understanding large-scale violence, where individual participation often depends on collective moral disengagement processes.

4. Cognitive Dissonance and Moral Disengagement

Leon Festinger's theory of cognitive dissonance provides additional insights into the psychological processes that enable and maintain cruel behavior [98]. Cognitive dissonance occurs when individuals hold contradictory beliefs, attitudes, or values, creating psychological tension that motivates efforts to reduce the inconsistency. In moral contexts, dissonance can arise when individuals engage in behavior that conflicts with their moral beliefs or self-concept.

The dissonance between engaging in cruel behavior and maintaining a positive moral self-concept can be resolved in several ways [99]. Individuals may change their behavior to align with their moral beliefs, but this option is often unavailable or costly in situations involving group pressure, authority demands, or institutional expectations. Alternatively, individuals may change their moral beliefs or self-concept to align with their behavior, leading to the gradual erosion of moral standards and the normalization of cruelty.

Research on cognitive dissonance has revealed that individuals often resolve moral conflicts by developing elaborate justifications for their behavior rather than changing the behavior itself [100]. These justifications can become increasingly extreme as individuals seek to reduce dissonance, leading to the development of ideological systems that not only permit but actively promote cruel behavior. This process helps explain how individuals and groups can develop elaborate moral frameworks that justify genocide, torture, and other forms of systematic violence.

The concept of moral disengagement, developed by Bandura and influenced by dissonance theory, describes the specific cognitive mechanisms that individuals use to justify harmful behavior [101]. These mechanisms operate by reconstructing the moral meaning of situations, allowing individuals to engage in cruel behavior without experiencing it as morally problematic. The effectiveness of these mechanisms depends partly on social support and cultural validation, highlighting the importance of social context in enabling or constraining cruel behavior.

Cognitive dissonance theory also helps explain the persistence and escalation of cruel behavior over time [102]. Once individuals have engaged in harmful behavior and developed justifications for it, they become psychologically invested in maintaining those justifications. This can lead to a process of escalating commitment, where individuals engage in increasingly severe behavior to justify their previous actions and maintain psychological consistency.

C. Empathy, Altruism, and Prosocial Behavior

Understanding human moral plasticity requires examining not only the mechanisms that enable cruelty but also those that promote compassion, cooperation, and prosocial behavior. Research on empathy and altruism reveals that the same psychological mechanisms that enable extraordinary moral behavior can also contribute to moral failures under different circumstances.

1. The Empathy-Altruism Hypothesis

C. Daniel Batson's empathy-altruism hypothesis proposes that empathic concern for others can motivate genuinely altruistic behavior, where individuals help others for the others' sake rather than for personal benefit [103]. This hypothesis challenges purely egoistic explanations of human behavior by suggesting that humans possess the capacity for genuine other-regarding motivation.

Batson's research program involved carefully controlled experiments designed to distinguish between egoistic and altruistic motivations for helping behavior [104]. By manipulating empathic concern and controlling for various egoistic motivations—such as avoiding guilt, gaining social approval, or reducing personal distress—Batson and his colleagues provided strong evidence that empathic concern can indeed motivate genuinely altruistic behavior.

The empathy-altruism hypothesis has important implications for understanding human moral capacity [105]. If humans possess the capacity for genuine altruism, then moral behavior is not simply a matter of enlightened self-interest or social conditioning but reflects a fundamental aspect of human psychology. This capacity for other-regarding motivation provides a foundation for moral behavior that transcends narrow self-interest.

However, research on empathy has also revealed significant limitations and potential negative consequences of empathic responding [106]. Empathy is often biased toward in-group members, attractive individuals, and those who are perceived as similar to oneself. This bias can lead to parochial altruism that benefits some individuals while ignoring or even harming others. Additionally, empathy can be manipulated and misdirected, leading individuals to support harmful policies or actions based on emotional responses to particular cases.

The relationship between empathy and moral behavior is further complicated by research showing that empathy can sometimes motivate aggression and violence [107]. When individuals empathize with victims of harm, they may respond with aggressive behavior toward perceived perpetrators. This empathy-driven aggression can escalate conflicts and lead to cycles of retaliation, illustrating how prosocial emotions can contribute to antisocial outcomes.

2. Neurological Substrates of Empathic Response

Neuroscientific research has identified specific brain networks involved in empathic responding, providing insights into the biological foundations of human moral capacity [108]. The discovery of mirror neurons in the 1990s revealed that the brain contains specialized cells that fire both when performing an action and when observing others perform the same action, suggesting a neural basis for understanding and sharing others' experiences.

Functional neuroimaging studies have identified several brain networks involved in different aspects of empathic responding [109]. The pain matrix, including the anterior cingulate cortex and anterior insula, shows activation when individuals experience pain themselves and when they observe others in pain. This shared neural representation provides a biological foundation for the capacity to understand and share others' emotional experiences.

The mentalizing network, including the medial prefrontal cortex, temporal-parietal junction, and superior temporal sulcus, is involved in understanding others' mental states, beliefs, and intentions [110]. This network enables individuals to take others' perspectives and understand their psychological experiences, providing a cognitive foundation for empathic responding and moral judgment.

Research has also identified neural mechanisms involved in empathic regulation and control [111]. The prefrontal cortex, particularly the dorsolateral and ventromedial regions, plays a crucial role in regulating empathic responses and integrating empathic information with other considerations. This regulatory capacity enables individuals to modulate their empathic responses based on contextual factors and competing moral concerns.

Importantly, neuroimaging research has revealed significant overlap between the neural networks involved in empathy and those involved in aggression and violence [112]. The anterior cingulate cortex and anterior insula, which are activated during empathic responding, are also involved in aggressive behavior and the processing of threat-related information. This neural overlap provides a biological foundation for the empathy-violence paradox, suggesting that the same brain systems that enable compassion can also enable cruelty under different circumstances.

3. The Paradox of Empathic Violence

One of the most challenging findings in research on empathy and moral behavior is the discovery that empathy can sometimes motivate violence and aggression rather than compassion and helping [113]. This empathy-violence paradox challenges simple assumptions about the relationship between empathy and moral behavior while providing important insights into human moral plasticity.

Research has shown that empathy with victims of harm can motivate aggressive responses toward perceived perpetrators [114]. When individuals empathize with those who have been harmed, they may experience anger and desire for revenge against those responsible for the harm. This empathy-driven aggression can lead to cycles of retaliation and escalating violence, as each act of aggression creates new victims who evoke empathic responses from their supporters.

The empathy-violence paradox is particularly evident in intergroup conflicts, where empathy with in-group members can motivate aggression toward out-group members [115]. Individuals who strongly empathize with their own group's suffering may support or engage in violence against other groups perceived as responsible for that suffering. This dynamic helps explain how empathic individuals can participate in ethnic conflicts, religious wars, and other forms of intergroup violence.

Research on parochial altruism has shown that empathy and helping behavior are often biased toward in-group members at the expense of out-group members [116]. Individuals are more likely to empathize with and help those who are similar to themselves, share group membership, or are perceived as deserving of help. This bias can lead to discriminatory behavior and the neglect or mistreatment of out-group members, even among individuals who are generally empathic and prosocial.

The empathy-violence paradox also manifests in responses to perceived injustice and moral violations [117]. When individuals empathize with victims of injustice, they may respond with moral outrage and punitive behavior toward perceived perpetrators. While this response can serve important functions in maintaining social norms and deterring harmful behavior, it can also lead to excessive punishment and the perpetuation of cycles of violence and retaliation.

Understanding the empathy-violence paradox is crucial for developing effective approaches to promoting moral behavior and reducing violence [118]. Rather than simply trying to increase empathy, interventions must focus on channeling empathic responses in constructive directions and developing the regulatory capacities needed to manage empathic biases and potential negative consequences. This requires a more nuanced understanding of empathy as a complex psychological process that can contribute to both moral and immoral behavior depending on how it is activated and regulated.

IV. Evolutionary and Anthropological Evidence

A. The Evolutionary Paradox of Human Cooperation and Competition

The evolutionary perspective on human moral behavior reveals a fundamental paradox: humans are simultaneously the most cooperative and most destructively violent species on Earth [119]. This paradox provides crucial insights into the nature of human moral plasticity and the evolutionary origins of our capacity for both extraordinary altruism and extreme cruelty.

1. Richard Wrangham's "Goodness Paradox"

Harvard anthropologist Richard Wrangham's analysis of human violence in *The Goodness Paradox* provides a compelling framework for understanding the evolutionary origins of human moral flexibility [120]. Wrangham argues that humans exhibit a unique combination of low reactive aggression (spontaneous, emotional violence) and high proactive aggression (planned, instrumental violence) compared to other primates.

Reactive aggression, characterized by immediate emotional responses to threats or provocations, appears to have been selected against in human evolution [121]. Compared to our closest primate relatives, humans show remarkably low levels of spontaneous violence in face-to-face interactions. This reduction in reactive aggression enabled the development of complex cooperative societies by reducing the costs of living in close proximity with large numbers of individuals.

However, humans simultaneously exhibit extraordinarily high levels of proactive aggression—planned, coordinated violence directed toward specific goals [122]. This capacity for organized violence enabled human groups to compete effectively with other groups while maintaining internal cooperation. The same psychological mechanisms that enable complex cooperation within groups also enable coordinated aggression between groups.

Wrangham's analysis suggests that human moral psychology evolved to support both within-group cooperation and between-group competition [123]. The capacity for empathy, fairness,

and reciprocity that enables cooperation within groups can be selectively activated or deactivated depending on group membership and perceived threats. This selective activation provides the foundation for human moral plasticity, enabling the same individuals to display both extraordinary compassion and extreme cruelty depending on the social context.

The "goodness paradox" helps explain why humans can appear both naturally good and naturally evil depending on the circumstances [124]. The psychological mechanisms that make humans capable of unprecedented cooperation also make them capable of unprecedented destruction when those mechanisms are directed toward out-group targets. This duality is not a design flaw but an adaptive feature that enabled human success in environments characterized by both cooperation and competition.

2. Reactive vs. Proactive Aggression in Human Evolution

The distinction between reactive and proactive aggression provides crucial insights into the evolutionary origins of human moral behavior and the mechanisms underlying moral plasticity [125]. These two forms of aggression involve different psychological processes, neural circuits, and evolutionary functions, yet they can interact in complex ways to produce the full range of human moral behavior.

Reactive aggression is characterized by immediate, emotional responses to perceived threats, provocations, or frustrations [126]. This form of aggression is typically accompanied by high levels of physiological arousal, including increased heart rate, blood pressure, and stress hormone release. Reactive aggression appears to be evolutionarily ancient, shared with many other mammalian species, and primarily serves defensive functions.

Human evolution appears to have involved strong selection against reactive aggression, particularly in males [127]. Archaeological evidence suggests that human societies have consistently executed or ostracized individuals who displayed excessive reactive aggression, creating evolutionary pressure for self-control and emotional regulation. This selection pressure may explain why humans show remarkably low levels of spontaneous violence compared to other primates.

Proactive aggression, by contrast, is characterized by planned, goal-directed violence that is often carried out in a calm, controlled manner [128]. This form of aggression involves higher-order cognitive processes, including planning, coordination, and strategic thinking. Proactive aggression appears to be uniquely developed in humans and serves primarily offensive functions related to resource acquisition and group competition.

The capacity for proactive aggression enabled human groups to engage in coordinated warfare, territorial expansion, and resource competition while maintaining internal cooperation [129]. This capacity required the development of sophisticated psychological mechanisms for distinguishing between in-group and out-group members, coordinating group action, and overriding empathic responses toward out-group victims.

The interaction between reactive and proactive aggression systems helps explain the complexity of human moral behavior [130]. While humans have reduced capacity for spontaneous violence, they retain the capacity for intense emotional responses that can be channeled through proactive aggression systems. This combination enables humans to engage in extremely violent behavior while maintaining emotional control and moral justification.

3. Coalitionary Violence and Group Selection

The evolution of human coalitionary violence—coordinated aggression by groups against other groups—represents a crucial development in human moral psychology [131]. This capacity for organized intergroup violence required the evolution of sophisticated psychological mechanisms for group formation, coordination, and moral boundary-making that continue to influence human behavior today.

Coalitionary violence differs fundamentally from individual aggression in its psychological requirements and social consequences [132]. Effective coalitionary violence requires the ability to form stable alliances, coordinate complex actions, share risks and benefits, and maintain group cohesion in the face of external threats. These requirements led to the evolution of psychological mechanisms that promote in-group loyalty, out-group hostility, and moral justification for group-based violence.

The evolution of coalitionary violence may have involved group selection processes, where groups with more effective cooperation and coordination outcompeted less cohesive groups [133]. This process would have favored psychological traits that promote group success even at the cost of individual welfare, including willingness to sacrifice for the group, hostility toward competing groups, and strong emotional responses to group threats.

Archaeological evidence suggests that coalitionary violence has been a persistent feature of human societies throughout history [134]. Evidence of organized warfare, fortified settlements, and mass killings appears in the archaeological record from the earliest human societies, suggesting that the capacity for group-based violence is an ancient and fundamental aspect of human nature.

The psychological mechanisms that evolved to support coalitionary violence continue to influence contemporary human behavior [135]. Modern ethnic conflicts, religious wars, and ideological violence often involve the activation of ancient psychological systems for group formation and intergroup competition. Understanding these evolutionary origins is crucial for developing effective approaches to preventing and resolving contemporary conflicts.

B. Cross-Cultural Studies of Violence and Morality

Anthropological research across diverse cultures provides crucial evidence for understanding the universality and variability of human moral behavior. While specific moral beliefs and practices vary dramatically across cultures, underlying patterns of moral psychology appear to be universal, supporting the idea that humans possess a shared moral architecture that can be configured in different ways.

1. Anthropological Evidence for Universal Moral Concerns

Cross-cultural research has identified several moral concerns that appear across virtually all human societies, despite significant variation in specific beliefs and practices [136]. These universal moral concerns provide evidence for shared psychological mechanisms underlying human moral judgment while also revealing the flexibility with which these mechanisms can be applied.

The prohibition against unprovoked killing within the community appears to be universal across human societies [137]. While definitions of "unprovoked" and "community" vary significantly, all known human societies have norms against arbitrary violence within the in-group. This universality suggests that humans possess evolved psychological mechanisms that promote in-group cooperation and constrain within-group violence.

However, the same societies that prohibit in-group violence often permit or even encourage violence against out-group members [138]. The distinction between legitimate and illegitimate violence typically depends on group membership, with different standards applied to in-group and out-group targets. This pattern supports the idea that human moral psychology is inherently group-based and context-dependent rather than universally applied.

Reciprocity norms—expectations that benefits should be returned and harms should be punished—appear across all human societies [139]. These norms take different forms in different cultures, from formal legal systems to informal social expectations, but the underlying principle of reciprocal exchange appears to be universal. This universality suggests that humans possess evolved psychological mechanisms for tracking social exchanges and responding to violations of reciprocity.

Care for offspring and vulnerable community members also appears to be universal, though the specific forms of care and definitions of vulnerability vary across cultures [140]. All human societies have norms promoting the protection and nurturing of children, and most extend these norms to other vulnerable community members. This universality reflects the evolutionary importance of parental care and cooperative child-rearing in human societies.

Authority and hierarchy norms appear across human societies, though the specific forms of authority and criteria for legitimacy vary significantly [141]. All known human societies have some form of leadership structure and norms governing the exercise of authority. This universality suggests that humans possess evolved psychological mechanisms for recognizing and responding to legitimate authority, though these mechanisms can be activated by different cues in different cultural contexts.

2. Cultural Variations in Violence Legitimization

While certain moral concerns appear to be universal, the specific circumstances under which violence is considered legitimate vary dramatically across cultures [142]. These variations provide crucial insights into the flexibility of human moral psychology and the ways in which cultural learning can shape the activation of moral mechanisms.

Honor-based violence, where individuals are expected to use violence to defend their reputation and family honor, is considered legitimate in some cultures while being condemned in others [143]. In honor cultures, failure to respond to insults or threats with violence may be seen as moral weakness, while in dignity cultures, the same violent response may be seen as immoral aggression. These differences reflect different cultural configurations of the same underlying moral psychology.

Revenge and retaliation norms vary significantly across cultures in their scope, intensity, and temporal limits [144]. Some cultures emphasize immediate and proportionate retaliation, while others emphasize forgiveness and reconciliation. Some cultures limit revenge to the immediate perpetrator, while others extend it to family members or group affiliates. These variations demonstrate the flexibility with which human moral psychology can be culturally configured.

Religious and ideological violence is legitimized in some cultural contexts while being condemned in others [145]. The same act of violence may be seen as martyrdom in one cultural context and terrorism in another, depending on the religious or ideological framework used to interpret the action. These differences highlight the role of cultural narratives and belief systems in shaping moral judgment.

Collective punishment—holding entire groups responsible for the actions of individual members—is accepted in some cultures while being rejected in others [146]. Some cultures emphasize individual responsibility and reject collective punishment as unjust, while others emphasize group responsibility and see collective punishment as necessary for maintaining social order. These differences reflect different cultural weightings of individual versus group-based moral concerns.

The legitimization of violence against out-group members varies dramatically across cultures in its scope and intensity [147]. Some cultures emphasize universal human dignity and reject violence against any human beings, while others maintain sharp distinctions between in-group and out-group members and legitimize extreme violence against out-group targets. These variations demonstrate the flexibility of human moral boundary-making mechanisms.

3. The Role of Ritual and Symbolism in Moral Boundary-Making

Anthropological research has revealed the crucial role of ritual and symbolic practices in creating and maintaining moral boundaries between groups [148]. These practices provide insights into the psychological mechanisms underlying human moral plasticity and the ways in which cultural practices can activate or deactivate moral responses toward different targets.

Initiation rituals often involve the creation of strong in-group bonds through shared experiences of hardship, secrecy, and symbolic transformation [149]. These rituals can create powerful psychological commitments to group membership that override other moral considerations. Military training, fraternal organizations, and religious communities often use similar ritual processes to create strong group loyalty and willingness to sacrifice for group goals.

Dehumanization rituals involve symbolic practices that portray out-group members as less than human or as dangerous threats to the community [150]. These rituals can include the use of animal metaphors, the attribution of supernatural evil powers, or the portrayal of out-group members as disease-carrying contaminants. Such rituals can psychologically prepare group members for violence against out-group targets by reducing empathic responses and moral inhibitions.

Purification rituals often follow episodes of violence or moral transgression, serving to restore the moral order and reintegrate participants into the community [151]. These rituals can involve symbolic cleansing, confession, penance, or sacrifice, and they serve to manage the psychological consequences of moral violations. The existence of such rituals across cultures suggests that humans possess evolved psychological mechanisms for managing moral guilt and restoring social relationships after moral failures.

Sacred boundary rituals involve the creation and maintenance of symbolic boundaries between sacred and profane, pure and impure, or moral and immoral [152]. These rituals can create powerful emotional responses to boundary violations and motivate extreme behavior to defend sacred values. Religious conflicts often involve the activation of sacred boundary psychology, leading to violence that appears irrational from secular perspectives but makes sense within sacred value frameworks.

Reconciliation rituals provide mechanisms for restoring relationships and moral order after conflicts or moral violations [153]. These rituals can involve apology, forgiveness, compensation, or symbolic acts of reconciliation, and they serve to repair damaged social relationships and prevent cycles of revenge. The existence of such rituals across cultures suggests that humans possess evolved psychological mechanisms for conflict resolution and relationship repair.

C. Comparative Primatology and Human Uniqueness

Comparative research on human and non-human primate behavior provides crucial insights into the evolutionary origins of human moral capacity and the unique features of human moral psychology [154]. While humans share many basic psychological mechanisms with other primates, several key differences help explain the distinctive features of human moral behavior.

1. Chimpanzee Warfare and Human Violence

Research on chimpanzee behavior, particularly Jane Goodall's long-term studies at Gombe and subsequent research at other sites, has revealed striking similarities between chimpanzee and human patterns of intergroup violence [155]. These similarities suggest that the capacity for coalitionary violence has deep evolutionary roots, while also highlighting important differences in the psychological mechanisms underlying human and chimpanzee aggression.

Chimpanzees engage in organized raids against neighboring groups, involving coordinated attacks by multiple males against isolated individuals from other communities [156]. These attacks often result in severe injuries or death and appear to serve territorial and resource acquisition functions. The strategic nature of these attacks, including careful reconnaissance

and coordination among attackers, suggests sophisticated cognitive abilities underlying chimpanzee violence.

However, chimpanzee violence differs from human violence in several important ways [157]. Chimpanzee attacks are typically opportunistic, occurring when attackers encounter isolated victims in vulnerable situations. Human violence, by contrast, can involve elaborate planning, complex coordination, and the use of sophisticated weapons and tactics. Humans can also engage in violence against much larger numbers of victims and can maintain violent campaigns over extended periods.

The psychological mechanisms underlying chimpanzee and human violence also appear to differ significantly [158]. Chimpanzee violence appears to be primarily driven by immediate emotional responses to territorial threats or opportunities for resource acquisition. Human violence can involve complex ideological justifications, moral reasoning, and long-term strategic planning that goes far beyond immediate emotional responses.

The capacity for empathy and perspective-taking also differs between chimpanzees and humans in ways that influence violent behavior [159]. While chimpanzees show some capacity for empathy and helping behavior, their empathic responses appear to be more limited in scope and less flexible than human empathy. Humans can empathize with much larger numbers of individuals, including abstract or imagined others, and can use this empathic capacity to both motivate and justify violent behavior.

2. Bonobo Cooperation and Human Altruism

Research on bonobo behavior provides a contrasting perspective on primate social behavior, revealing a species that emphasizes cooperation, conflict resolution, and peaceful coexistence [160]. Bonobos are humans' closest living relatives along with chimpanzees, but they display dramatically different patterns of social behavior that provide insights into the range of possibilities for primate social organization.

Bonobos rarely engage in lethal violence and have never been observed to conduct organized raids against neighboring groups [161]. Instead, bonobos use a variety of conflict resolution mechanisms, including sexual behavior, grooming, and food sharing, to manage social tensions

and maintain group cohesion. This peaceful orientation appears to be supported by different hormonal and neurological systems compared to chimpanzees.

The bonobo pattern of behavior demonstrates that close human relatives can evolve very different approaches to social conflict and cooperation [162]. This suggests that human moral psychology is not inevitably violent or competitive but represents one possible evolutionary solution to the challenges of group living. The existence of peaceful primate societies provides hope that humans might be able to develop more peaceful forms of social organization.

However, bonobo society also differs from human society in important ways that limit the applicability of bonobo models to human behavior [163]. Bonobos live in relatively small, stable groups with abundant food resources and limited territorial pressure. Human societies, by contrast, often involve large, complex groups with resource competition and territorial conflicts that may require different psychological and social mechanisms.

The comparison between chimpanzee and bonobo behavior highlights the importance of ecological and social factors in shaping primate behavior [164]. The same basic primate psychology can produce very different behavioral outcomes depending on environmental pressures, resource availability, and social structure. This insight is crucial for understanding human moral plasticity and the potential for developing more peaceful human societies.

3. The Evolution of Moral Emotions

Human moral emotions appear to be both continuous with and distinct from the emotional systems found in other primates [165]. Understanding these similarities and differences provides insights into the evolutionary origins of human moral capacity and the unique features of human moral psychology.

Empathy and compassion appear to have ancient evolutionary roots, with evidence for empathic responses found across many mammalian species [166]. Primates show clear evidence of empathic concern for others, including helping behavior, consolation of distressed individuals, and emotional contagion. These capacities provide the foundation for human moral emotions but are expressed in more limited and context-specific ways in non-human primates.

Fairness and reciprocity concerns also appear to have primate origins, with evidence for inequity aversion and reciprocal exchange found in several primate species [167]. Capuchin monkeys refuse to participate in exchanges when they receive inferior rewards compared to other individuals, and chimpanzees engage in reciprocal grooming and food sharing. These capacities suggest that human fairness concerns build on evolved psychological mechanisms shared with other primates.

However, human moral emotions also show unique features that distinguish them from other primate emotional systems [168]. Humans can experience moral emotions in response to abstract principles, imagined scenarios, and distant others in ways that appear to be unique among primates. Human moral emotions can also be shaped by cultural learning and symbolic representation to a much greater extent than other primate emotions.

The capacity for moral outrage and punishment appears to be particularly developed in humans compared to other primates [169]. While other primates show some capacity for punishment and retaliation, humans can engage in costly punishment of norm violators even when they have not been directly harmed. This capacity for third-party punishment appears to be crucial for maintaining large-scale cooperation and social norms.

The evolution of language and symbolic representation may have been crucial for the development of uniquely human moral emotions [170]. Language enables humans to communicate about abstract moral principles, share moral narratives, and coordinate moral responses across large groups. This capacity for symbolic moral communication may explain many of the unique features of human moral behavior compared to other primates.

V. Neuroscientific Insights: The Brain's Moral Circuitry

A. The Neurobiology of Empathy and Aggression

Neuroscientific research has revealed fundamental insights into the biological basis of human moral behavior, particularly the surprising overlap between neural circuits involved in empathy and those involved in aggression and violence [171]. This neurobiological evidence provides crucial support for the Moral Plasticity Hypothesis by demonstrating that the same brain systems can produce both prosocial and antisocial behavior depending on contextual activation patterns.

1. Overlapping Neural Circuits for Empathy and Violence

One of the most significant discoveries in moral neuroscience is the extensive overlap between brain regions involved in empathic responding and those involved in aggressive behavior [172]. This overlap provides a neurobiological foundation for understanding how the same individuals can display both extraordinary compassion and extreme cruelty depending on the circumstances.

The anterior cingulate cortex (ACC) plays a crucial role in both empathic pain processing and aggressive behavior [173]. When individuals observe others in pain, the ACC shows activation patterns similar to those observed when experiencing pain directly. This shared neural representation provides a biological basis for empathic concern and helping behavior. However, the same ACC regions are also activated during aggressive behavior and the processing of threat-related information, suggesting that empathy and aggression may share common neural substrates.

The anterior insula is another brain region that shows activation during both empathic responding and aggressive behavior [174]. The insula is involved in processing interoceptive information—awareness of internal bodily states—and plays a crucial role in emotional awareness and regulation. During empathic responding, the insula helps individuals become aware of their emotional responses to others' suffering. During aggressive behavior, the insula is involved in processing the emotional arousal associated with threat and conflict.

The amygdala, traditionally associated with fear and threat processing, also shows complex patterns of activation during both empathic and aggressive responses [175]. While amygdala activation can promote empathic concern and helping behavior in response to others' distress, it can also promote aggressive responses when others are perceived as threats. The same neural system that enables individuals to respond compassionately to suffering can also enable them to respond aggressively to perceived threats.

This neural overlap helps explain the empathy-violence paradox observed in behavioral research [176]. The same brain systems that enable individuals to understand and share others' emotional experiences can also enable them to inflict suffering effectively. Understanding others' mental states—a capacity that normally promotes empathy and cooperation—can also be used to manipulate, intimidate, and harm others more effectively.

2. The Role of the Prefrontal Cortex in Moral Reasoning

The prefrontal cortex (PFC) plays a crucial role in moral reasoning, decision-making, and the regulation of emotional responses [177]. Different regions of the PFC are involved in different aspects of moral cognition, and damage to these regions can produce specific deficits in moral behavior that provide insights into the neural basis of moral plasticity.

The ventromedial prefrontal cortex (vmPFC) is involved in integrating emotional and cognitive information during moral decision-making [178]. Patients with damage to the vmPFC often show impaired moral judgment, particularly in situations involving emotional content or personal moral dilemmas. These patients may understand moral rules intellectually but fail to generate appropriate emotional responses to moral violations, leading to poor moral decision-making in real-world contexts.

The dorsolateral prefrontal cortex (dlPFC) is involved in cognitive control, working memory, and the regulation of emotional responses [179]. This region plays a crucial role in overriding immediate emotional impulses and implementing long-term goals and values. Damage to the dlPFC can lead to impulsive behavior and difficulty maintaining moral standards under pressure or temptation.

The anterior prefrontal cortex is involved in abstract reasoning, perspective-taking, and the consideration of multiple viewpoints [180]. This region enables individuals to consider the consequences of their actions for others and to take multiple perspectives into account during moral reasoning. Damage to this region can lead to difficulty understanding others' viewpoints and considering the broader implications of moral decisions.

Neuroimaging research has revealed that different types of moral judgments activate different patterns of prefrontal activity [181]. Personal moral dilemmas, which involve direct harm to specific individuals, tend to activate emotional processing regions including the vmPFC and ACC. Impersonal moral dilemmas, which involve abstract principles and statistical lives, tend to activate cognitive processing regions including the dlPFC and anterior PFC.

The balance between emotional and cognitive processing in the prefrontal cortex appears to be crucial for moral behavior [182]. When emotional processing dominates, individuals may make decisions based on immediate empathic responses that may not be optimal from a broader moral perspective. When cognitive processing dominates, individuals may make decisions based on abstract principles that ignore important emotional and contextual factors.

3. Amygdala Activation and Threat Response

The amygdala plays a central role in threat detection and emotional processing, making it crucial for understanding the neural basis of both prosocial and antisocial behavior [183]. The amygdala's response to perceived threats can motivate both protective behavior toward in-group members and aggressive behavior toward out-group members, illustrating the dual nature of threat-related emotional systems.

Amygdala activation in response to out-group faces occurs within milliseconds of stimulus presentation, often before conscious awareness [184]. This rapid, automatic response suggests that threat detection and group categorization are fundamental features of human social cognition. The same neural system that enables rapid identification of potential threats can also contribute to prejudice and intergroup hostility.

Research has shown that amygdala responses to out-group members can be modulated by various factors, including perceived threat, group competition, and individual differences in

prejudice [185]. When intergroup relations are cooperative and non-threatening, amygdala responses to out-group members are reduced. When intergroup relations are competitive or threatening, amygdala responses are enhanced, potentially contributing to increased hostility and aggression.

The amygdala also plays a crucial role in fear conditioning and the formation of threat-related memories [186]. Traumatic experiences involving out-group members can create lasting changes in amygdala responsivity that contribute to ongoing prejudice and hostility. These conditioned fear responses can be difficult to extinguish and may contribute to cycles of intergroup violence and retaliation.

However, the amygdala is not simply a "fear center" but rather a complex system involved in processing emotional significance more generally [187]. Amygdala activation can also occur in response to positive emotions, social bonding, and empathic concern. The same neural system that can contribute to fear and aggression can also contribute to love, attachment, and prosocial behavior.

The regulation of amygdala activity by prefrontal cortical regions is crucial for moral behavior [188]. When prefrontal control is strong, individuals can override automatic amygdala responses and behave according to their moral values and long-term goals. When prefrontal control is weak—due to stress, fatigue, intoxication, or other factors—amygdala-driven responses may dominate, leading to impulsive or aggressive behavior.

B. Neuroplasticity and Moral Development

The brain's capacity for change throughout the lifespan—neuroplasticity—has important implications for understanding moral development and the potential for moral transformation [189]. Research on neuroplasticity suggests that moral capacities are not fixed but can be modified through experience, training, and environmental influences.

1. Critical Periods in Moral Brain Development

The development of moral capacities involves complex interactions between genetic factors and environmental influences during specific periods of brain development [190]. Understanding these critical periods is crucial for developing effective approaches to moral education and intervention.

The prefrontal cortex, which plays a crucial role in moral reasoning and self-control, continues developing well into the third decade of life [191]. This extended developmental period means that moral capacities continue to mature throughout adolescence and early adulthood. The late maturation of prefrontal systems may explain why adolescents often show poor moral judgment and impulse control despite understanding moral rules intellectually.

Early childhood appears to be a particularly important period for the development of empathy and prosocial behavior [192]. Children who experience secure attachment relationships and responsive caregiving during the first few years of life show enhanced capacity for empathy and cooperation later in development. Conversely, children who experience neglect, abuse, or inconsistent caregiving may show impaired empathic development and increased risk for antisocial behavior.

The adolescent period involves significant changes in brain structure and function that can affect moral behavior [193]. The limbic system, including the amygdala and reward-processing regions, undergoes rapid development during adolescence, while prefrontal control systems develop more slowly. This imbalance may contribute to increased risk-taking, emotional reactivity, and susceptibility to peer influence during adolescence.

However, the extended period of brain development also provides opportunities for positive intervention and moral education [194]. Because moral brain systems continue developing throughout adolescence and early adulthood, experiences during these periods can have lasting effects on moral capacities. Educational programs, mentoring relationships, and positive social experiences during these critical periods may be particularly effective for promoting moral development.

2. Environmental Influences on Neural Moral Circuitry

Environmental factors can have profound effects on the development and functioning of neural systems involved in moral behavior [195]. Understanding these environmental influences is crucial for developing effective approaches to promoting moral behavior and preventing antisocial outcomes.

Chronic stress and trauma can have lasting effects on brain systems involved in moral behavior [196]. Exposure to violence, abuse, or severe neglect during childhood can lead to alterations in amygdala reactivity, prefrontal functioning, and stress hormone systems that increase risk for aggressive and antisocial behavior. These neurobiological changes may help explain the cycle of violence observed in many families and communities.

Social and cultural factors can also influence the development of moral brain systems [197]. Children who grow up in communities with strong social norms, positive role models, and opportunities for prosocial behavior show enhanced development of empathy and moral reasoning capacities. Conversely, children who grow up in communities with weak social norms, antisocial role models, and limited opportunities for positive social interaction may show impaired moral development.

Educational experiences can have significant effects on moral brain development [198]. Programs that emphasize perspective-taking, empathy training, and moral reasoning can enhance the development of prefrontal systems involved in moral cognition. Conversely, educational environments that emphasize competition, punishment, and conformity may inhibit the development of autonomous moral reasoning capacities.

Media exposure can also influence moral brain development, particularly during childhood and adolescence [199]. Exposure to violent media content can lead to desensitization of neural systems involved in empathy and moral concern. However, exposure to prosocial media content can enhance empathic responding and moral behavior, suggesting that media can be used as a tool for moral education.

3. The Malleability of Moral Intuitions

Research on neuroplasticity suggests that moral intuitions and responses are more malleable than previously thought [200]. This malleability provides both opportunities and challenges for understanding and influencing moral behavior.

Meditation and mindfulness training can produce measurable changes in brain systems involved in empathy and emotional regulation [201]. Studies have shown that compassion meditation can increase activity in brain regions associated with empathic concern and reduce activity in regions associated with negative emotions. These findings suggest that contemplative practices may be effective tools for enhancing moral capacities.

Cognitive training programs can also influence moral cognition and behavior [202]. Training in perspective-taking, moral reasoning, and conflict resolution can enhance prefrontal functioning and improve moral decision-making. These programs may be particularly effective during periods of rapid brain development, such as childhood and adolescence.

However, the malleability of moral systems also creates vulnerability to negative influences [203]. The same neuroplasticity that enables positive moral development can also enable the development of prejudice, hatred, and antisocial behavior. Exposure to dehumanizing propaganda, hate speech, and violent ideologies can produce lasting changes in moral cognition and behavior.

The social context in which moral development occurs appears to be crucial for determining whether neuroplasticity leads to positive or negative outcomes [204]. Supportive social environments that provide positive role models, clear moral guidance, and opportunities for prosocial behavior tend to promote positive moral development. Hostile social environments that provide antisocial role models, unclear moral guidance, and limited opportunities for positive behavior tend to promote negative moral development.

C. Psychopathy and the Limits of Moral Plasticity

The study of psychopathy provides important insights into the limits of human moral plasticity and the neurobiological foundations of moral behavior [205]. While most individuals show remarkable flexibility in their moral responses, psychopathic individuals appear to have fundamental deficits in moral cognition that may be less amenable to change.

1. Neurological Differences in Psychopathic Individuals

Neuroimaging research has identified consistent differences in brain structure and function between psychopathic and non-psychopathic individuals [206]. These differences provide insights into the neural basis of moral behavior and the extent to which moral capacities depend on specific brain systems.

Psychopathic individuals show reduced activity in the amygdala during emotional processing tasks [207]. This reduced amygdala activity may contribute to the emotional deficits characteristic of psychopathy, including reduced fear, empathy, and guilt. The amygdala's role in processing emotional significance means that reduced amygdala function can lead to difficulty understanding the emotional impact of one's actions on others.

The prefrontal cortex also shows structural and functional differences in psychopathic individuals [208]. Reduced gray matter volume and activity in prefrontal regions may contribute to the impulsivity, poor judgment, and lack of behavioral control characteristic of psychopathy. These prefrontal deficits may make it difficult for psychopathic individuals to override immediate impulses and consider the long-term consequences of their actions.

The anterior cingulate cortex, which plays a crucial role in empathic responding and conflict monitoring, also shows reduced activity in psychopathic individuals [209]. This reduced ACC activity may contribute to the lack of empathic concern and moral conflict characteristic of psychopathy. Without normal ACC functioning, psychopathic individuals may not experience the emotional distress that normally motivates moral behavior.

However, it is important to note that not all brain regions show deficits in psychopathy [210]. Some cognitive abilities, including intelligence, working memory, and certain aspects of social cognition, may be intact or even enhanced in psychopathic individuals. This pattern suggests that psychopathy involves specific deficits in emotional and moral processing rather than general cognitive impairment.

2. The Debate over Moral Responsibility and Brain Abnormalities

The discovery of neurobiological differences in psychopathic individuals has raised important questions about moral responsibility and criminal justice [211]. If psychopathic behavior results from brain abnormalities rather than free choice, to what extent should psychopathic individuals be held morally and legally responsible for their actions?

Some researchers argue that neurobiological differences in psychopathy should be considered mitigating factors in moral and legal judgment [212]. If psychopathic individuals lack the neural capacity for normal moral reasoning and empathic concern, they may not be fully responsible for their antisocial behavior. This perspective suggests that psychopathic individuals should be treated rather than punished for their behavior.

Other researchers argue that neurobiological differences do not necessarily eliminate moral responsibility [213]. Even if psychopathic individuals have different brain functioning, they may still retain the capacity to understand moral rules and control their behavior. The fact that many psychopathic individuals can function successfully in society without engaging in criminal behavior suggests that neurobiological differences do not inevitably lead to antisocial outcomes.

The debate over moral responsibility in psychopathy reflects broader questions about the relationship between neuroscience and moral judgment [214]. As our understanding of the neural basis of behavior advances, we may need to reconsider traditional assumptions about free will, moral responsibility, and criminal justice. However, the practical and ethical implications of these discoveries remain highly contested.

3. Implications for Understanding "Evil" Behavior

The study of psychopathy provides important insights into the nature of extreme antisocial behavior and the limits of human moral plasticity [215]. While psychopathic individuals represent a small percentage of the population, they may be responsible for a disproportionate amount of serious violent crime and may provide insights into the most extreme forms of human cruelty.

Psychopathic individuals appear to lack many of the emotional and motivational systems that normally constrain antisocial behavior [216]. Without normal empathic concern, guilt, or fear of punishment, psychopathic individuals may be more likely to engage in extreme forms of cruelty and violence. This suggests that normal moral behavior depends on intact emotional systems that are compromised in psychopathy.

However, it is important not to overstate the role of psychopathy in explaining human cruelty and evil [217]. Most instances of extreme violence and cruelty, including genocide and mass atrocities, are committed by individuals who do not meet criteria for psychopathy. Normal individuals with intact emotional and moral systems can engage in extreme cruelty under certain circumstances, suggesting that psychopathy is not necessary for evil behavior.

The study of psychopathy also highlights the importance of distinguishing between different types of antisocial behavior [218]. Primary psychopathy, characterized by emotional deficits and callous-unemotional traits, may represent a fundamentally different phenomenon from secondary psychopathy, which is characterized by impulsivity and emotional reactivity. These different subtypes may have different neural bases and may require different approaches to understanding and intervention.

Understanding the limits of moral plasticity in psychopathy may also provide insights into the mechanisms of normal moral behavior [219]. By studying individuals who lack normal moral capacities, researchers can better understand the neural and psychological systems that enable moral behavior in typical individuals. This knowledge may be crucial for developing effective approaches to promoting moral behavior and preventing antisocial outcomes in the general population.

VI. The Moral Plasticity Hypothesis: A Novel Framework

A. Defining Moral Plasticity

The evidence reviewed in the preceding sections converges on a fundamental insight: human moral behavior is characterized by extraordinary flexibility rather than fixed tendencies toward good or evil. This thesis proposes the "Moral Plasticity Hypothesis" as a comprehensive framework for understanding this flexibility and its implications for human nature, moral behavior, and social organization.

1. Adaptive Ambiguity as Evolutionary Advantage

The Moral Plasticity Hypothesis posits that evolution selected for moral ambiguity rather than fixed moral orientations because flexibility provided crucial survival advantages in the complex and changing social environments that characterized human evolutionary history [220]. This adaptive ambiguity enabled human ancestors to navigate the competing demands of within-group cooperation and between-group competition that were essential for survival and reproductive success.

Traditional approaches to human nature have typically assumed that evolution would select for either cooperative or competitive tendencies, leading to debates about whether humans are "naturally" good or evil [221]. However, this binary framing fails to recognize that human ancestors faced environments that required both cooperation and competition, often simultaneously. Groups that were internally cooperative but externally competitive would have had significant advantages over groups that were either purely cooperative or purely competitive.

The capacity for moral flexibility enabled human groups to maintain internal cohesion while competing effectively with other groups [222]. Within the group, individuals needed to cooperate, share resources, and suppress selfish impulses to maintain group stability and effectiveness. Between groups, the same individuals needed to compete for resources, defend territory, and potentially engage in violence against out-group members. This dual requirement

created selection pressure for psychological mechanisms that could support both cooperation and competition depending on the social context.

Moral plasticity also provided advantages in dealing with changing environmental and social conditions [223]. Human groups that could rapidly adapt their moral norms and behaviors to new circumstances would have been more successful than groups with rigid moral systems. This flexibility enabled humans to colonize diverse environments, develop new technologies, and create increasingly complex social organizations.

The evolutionary advantage of moral plasticity helps explain why humans possess such a complex and seemingly contradictory moral psychology [224]. Rather than representing design flaws or pathological deviations, the capacity for both extraordinary altruism and extreme cruelty reflects the adaptive value of behavioral flexibility in complex social environments. This perspective reframes moral extremes as features rather than bugs of human psychological architecture.

2. Contextual Moral Switching Mechanisms

The Moral Plasticity Hypothesis proposes that humans possess evolved psychological mechanisms that function as "moral switches," enabling rapid transitions between different moral orientations depending on contextual cues [225]. These switching mechanisms operate largely outside conscious awareness but can produce dramatic changes in moral judgment and behavior within relatively short timeframes.

The concept of moral switching draws on research in cognitive psychology demonstrating that human behavior is highly context-dependent and can be influenced by subtle environmental cues [226]. Just as priming studies show that exposure to certain words or images can influence subsequent behavior, moral switching suggests that exposure to certain social and environmental cues can activate different moral orientations and behavioral patterns.

Threat perception represents one of the most powerful moral switching mechanisms [227]. When individuals or groups perceive threats to their safety, resources, or identity, moral psychology can shift rapidly from cooperation-focused to competition-focused orientations. This shift can involve changes in empathy (reduced toward out-group members), fairness

concerns (increased emphasis on in-group loyalty), and authority acceptance (increased deference to strong leaders).

Group identity activation represents another crucial moral switching mechanism [228]. When group membership becomes salient—through symbols, rituals, narratives, or intergroup contact—moral psychology can shift to prioritize group-based concerns over individual or universal concerns. This shift can involve increased in-group favoritism, out-group derogation, and willingness to sacrifice individual interests for group goals.

Authority legitimization can also trigger moral switching by activating psychological mechanisms for hierarchy and obedience [229]. When authority figures are perceived as legitimate—through traditional markers, institutional positions, or charismatic appeal—individuals may shift from autonomous moral reasoning to authority-based moral reasoning. This shift can enable individuals to engage in behavior that violates their personal moral standards when directed by legitimate authorities.

Narrative framing represents a more cognitive moral switching mechanism that operates through the interpretation and meaning-making processes [230]. When situations are framed using different narratives—self-defense versus aggression, justice versus revenge, protection versus oppression—the same behavior can be evaluated very differently. These narrative frames can activate different moral foundations and lead to dramatically different moral judgments about identical actions.

3. The Empathy-Violence Paradox Explained

One of the most challenging aspects of human moral behavior is the empathy-violence paradox: the same individuals who demonstrate extraordinary empathy and compassion can also engage in extreme violence and cruelty [231]. The Moral Plasticity Hypothesis provides a framework for understanding this paradox by recognizing that empathy and violence share common psychological and neurobiological mechanisms.

Empathy involves the capacity to understand and share others' emotional experiences, which requires sophisticated perspective-taking abilities and emotional responsiveness [232]. These same capacities can also enable more effective violence by allowing individuals to understand

their victims' vulnerabilities, predict their responses, and inflict maximum psychological and physical harm. The cognitive and emotional skills that make someone a compassionate caregiver can also make them an effective torturer under different circumstances.

The neurobiological overlap between empathy and violence circuits provides a biological foundation for this paradox [233]. Brain regions involved in empathic responding—including the anterior cingulate cortex, anterior insula, and amygdala—are also involved in aggressive behavior and threat processing. This shared neural architecture means that the same brain systems that enable compassionate responses to suffering can also enable violent responses to perceived threats.

Empathy can also motivate violence through protective and retaliatory mechanisms [234]. When individuals empathize with victims of harm, they may experience intense emotional responses that motivate aggressive behavior toward perceived perpetrators. This empathy-driven aggression can escalate conflicts and lead to cycles of violence and retaliation, as each act of aggression creates new victims who evoke empathic responses from their supporters.

The group-based nature of human empathy contributes to the empathy-violence paradox by creating differential moral concern for in-group versus out-group members [235]. Individuals can maintain strong empathic concern for in-group members while simultaneously engaging in violence against out-group members. This selective empathy enables individuals to see themselves as moral and compassionate while participating in activities that cause tremendous suffering to others.

The Moral Plasticity Hypothesis suggests that the empathy-violence paradox reflects the adaptive value of flexible moral responding rather than representing a psychological contradiction [236]. The same psychological mechanisms that enable individuals to care deeply for their families and communities also enable them to defend those groups against perceived threats. This flexibility was likely crucial for survival in environments characterized by both cooperation and competition.

B. Mechanisms of Moral Transformation

Understanding how moral plasticity operates requires examining the specific psychological and social mechanisms that enable rapid moral transformation. These mechanisms operate at multiple levels—individual, interpersonal, and institutional—and can interact in complex ways to produce dramatic changes in moral behavior.

1. Threat Perception and In-group/Out-group Dynamics

Threat perception represents one of the most powerful mechanisms for moral transformation, capable of rapidly shifting individuals and groups from cooperative to competitive orientations [237]. The human threat detection system evolved to respond quickly to potential dangers, but it can be activated by symbolic and social threats as well as physical ones.

Physical threat perception can trigger immediate changes in moral psychology, including reduced empathy for out-group members, increased in-group loyalty, and greater acceptance of aggressive behavior [238]. When individuals perceive immediate threats to their safety or survival, moral considerations that normally constrain behavior may be overridden by survival-focused responses. This shift can enable individuals to engage in violence that would normally be morally unthinkable.

Economic threat perception can also trigger moral transformation by activating competition-focused psychological mechanisms [239]. When individuals or groups perceive threats to their economic security or resources, they may become more willing to engage in zero-sum thinking and competitive behavior. This shift can reduce concern for out-group welfare and increase support for policies that benefit the in-group at the expense of others.

Identity threat perception involves threats to group identity, status, or cultural values rather than physical or economic threats [240]. When groups perceive that their identity or way of life is under attack, they may respond with defensive aggression aimed at protecting group boundaries and maintaining group distinctiveness. This type of threat can be particularly powerful because it activates deep psychological needs for meaning, belonging, and self-esteem.

The in-group/out-group distinction is fundamental to human social psychology and plays a crucial role in moral transformation [241]. When group boundaries become salient, moral psychology can shift to prioritize in-group welfare over universal moral concerns. This shift can involve increased empathy and helping behavior toward in-group members combined with reduced empathy and increased hostility toward out-group members.

Research has shown that even minimal group distinctions—such as arbitrary assignment to teams or preference for abstract art—can trigger in-group favoritism and out-group discrimination [242]. This suggests that the psychological mechanisms underlying group-based moral transformation are highly sensitive and can be activated by relatively weak social cues. When combined with more powerful group markers—such as ethnicity, religion, or nationality—these mechanisms can produce intense moral transformation.

2. Authority Legitimization and Obedience

Authority legitimization represents another powerful mechanism for moral transformation, enabling individuals to engage in behavior that violates their personal moral standards when directed by perceived legitimate authorities [243]. The human capacity for hierarchical social organization requires psychological mechanisms for recognizing and responding to legitimate authority, but these mechanisms can also enable moral transformation under certain circumstances.

Traditional authority derives its legitimacy from custom, tradition, and established social roles [244]. When individuals perceive authorities as legitimate based on traditional criteria—such as age, family position, or cultural status—they may be more willing to defer to those authorities even when their directives conflict with personal moral judgments. This type of authority legitimization can enable moral transformation by shifting responsibility from individual moral reasoning to authority-based compliance.

Charismatic authority derives its legitimacy from the personal qualities and appeal of individual leaders [245]. When leaders are perceived as possessing exceptional qualities—such as wisdom, courage, or divine inspiration—followers may be willing to engage in extreme behavior based on the leader's directives. This type of authority legitimization can be particularly powerful because it combines emotional attachment with moral justification.

Legal-rational authority derives its legitimacy from institutional positions and formal procedures [246]. When individuals perceive authorities as legitimate based on their institutional roles and adherence to established procedures, they may be willing to engage in harmful behavior as part of their institutional duties. This type of authority legitimization can enable moral transformation by framing harmful behavior as necessary for maintaining institutional order and effectiveness.

The Milgram obedience experiments demonstrated the power of authority legitimization to override personal moral judgments [247]. Participants in these experiments were willing to administer apparently severe electric shocks to innocent victims when directed by a perceived legitimate authority figure. The experiments showed that ordinary individuals could engage in behavior that violated their moral standards when the situation was structured to emphasize authority legitimacy and institutional responsibility.

Authority legitimization can interact with other moral transformation mechanisms to produce particularly powerful effects [248]. When authorities are perceived as protecting the group from threats, authority-based obedience can combine with threat-based moral transformation to enable extreme behavior. When authorities provide compelling narratives that justify harmful behavior, authority legitimization can combine with narrative framing to override moral inhibitions.

3. Moral Foundation Prioritization Shifts

The Moral Plasticity Hypothesis proposes that moral transformation often involves shifts in the relative prioritization of different moral foundations rather than the complete abandonment of moral concern [249]. These prioritization shifts can enable individuals to maintain their sense of moral identity while engaging in behavior that would normally be considered immoral.

Care-based moral concerns can be selectively activated or deactivated depending on the target of concern [250]. Individuals may maintain strong care-based moral responses toward in-group members while showing reduced care-based responses toward out-group members. This selective activation enables individuals to see themselves as compassionate and caring while participating in activities that cause suffering to others.

Fairness-based moral concerns can shift between different conceptions of fairness depending on the social context [251]. In competitive contexts, fairness may be understood in terms of merit and proportionality, leading to support for unequal outcomes based on perceived differences in contribution or desert. In cooperative contexts, fairness may be understood in terms of equality and need, leading to support for equal outcomes regardless of individual differences.

Loyalty-based moral concerns can override other moral considerations when group membership becomes salient [252]. Individuals may prioritize loyalty to their group over abstract moral principles or concern for out-group members. This prioritization can enable individuals to engage in behavior that violates care-based or fairness-based moral concerns when such behavior is seen as serving group interests.

Authority-based moral concerns can override autonomous moral reasoning when legitimate authorities provide clear directives [253]. Individuals may prioritize obedience to authority over personal moral judgments, particularly when the authority is perceived as legitimate and the situation is structured to emphasize institutional responsibility rather than personal choice.

Purity-based moral concerns can motivate extreme behavior when out-group members are perceived as contaminating or degrading [254]. Individuals may prioritize maintaining group purity and moral boundaries over care-based concerns for out-group welfare. This prioritization can enable individuals to engage in violence and discrimination that is justified as necessary for protecting group integrity and moral order.

4. Narrative Framing and Identity Construction

Narrative framing represents a crucial mechanism for moral transformation by shaping how individuals understand and interpret moral situations [255]. The same behavior can be evaluated very differently depending on the narrative framework used to understand it, and these narrative frames can be manipulated to enable moral transformation.

Self-defense narratives frame harmful behavior as necessary responses to threats or attacks [256]. When individuals or groups perceive themselves as victims of aggression, they may justify extreme responses as necessary self-defense. This narrative framing can enable

individuals to engage in violence while maintaining their sense of moral righteousness and victimhood.

Justice narratives frame harmful behavior as necessary responses to moral violations or injustices [257]. When individuals perceive that moral wrongs have been committed, they may justify punitive responses as necessary for restoring moral order. This narrative framing can enable individuals to engage in revenge and retaliation while seeing themselves as agents of justice rather than perpetrators of harm.

Protection narratives frame harmful behavior as necessary for protecting vulnerable others [258]. When individuals perceive threats to loved ones or innocent victims, they may justify extreme responses as necessary protection. This narrative framing can enable individuals to engage in violence while maintaining their sense of moral virtue and heroism.

Purification narratives frame harmful behavior as necessary for eliminating corruption or contamination [259]. When individuals perceive out-group members as sources of moral, spiritual, or physical contamination, they may justify extreme responses as necessary purification. This narrative framing can enable individuals to engage in genocide and ethnic cleansing while seeing themselves as agents of moral purification.

Identity construction processes interact with narrative framing to create powerful mechanisms for moral transformation [260]. When individuals adopt new identities—as soldiers, revolutionaries, or religious warriors—they may also adopt new moral frameworks that justify behavior that would have been unthinkable in their previous identities. These identity transformations can be facilitated by rituals, training, and social pressure that reinforce new moral narratives and behavioral expectations.

C. Institutional Amplification of Human Moral Tendencies

The Moral Plasticity Hypothesis emphasizes that individual psychological mechanisms operate within institutional contexts that can dramatically amplify their effects [261]. Human institutions—including families, schools, religious organizations, governments, and media—

serve as powerful amplifiers that can channel moral plasticity toward either constructive or destructive outcomes.

1. The Role of Institutions in Channeling Human Nature

Human institutions evolved to coordinate large-scale cooperation and manage the challenges of complex social organization [262]. These institutions work by creating shared norms, roles, and expectations that guide individual behavior and channel human psychological tendencies in particular directions. The same institutional mechanisms that enable large-scale cooperation can also enable large-scale destruction when directed toward harmful goals.

Educational institutions play a crucial role in shaping moral development by transmitting cultural values, teaching moral reasoning skills, and providing opportunities for moral practice [263]. Schools that emphasize empathy, perspective-taking, and conflict resolution can enhance the development of prosocial moral capacities. Conversely, schools that emphasize competition, obedience, and in-group loyalty may enhance the development of group-based moral orientations that can contribute to intergroup conflict.

Religious institutions can channel moral psychology in various directions depending on their theological emphases and organizational structures [264]. Religious traditions that emphasize universal compassion, forgiveness, and peace can promote prosocial moral development. Religious traditions that emphasize group boundaries, divine authority, and sacred warfare can promote group-based moral orientations that may contribute to religious conflict and violence.

Political institutions shape moral behavior by creating legal frameworks, enforcement mechanisms, and incentive structures that reward certain behaviors while punishing others [265]. Democratic institutions that emphasize individual rights, due process, and peaceful conflict resolution can promote prosocial moral behavior. Authoritarian institutions that emphasize obedience, group loyalty, and state power can promote authority-based moral orientations that may enable political oppression and violence.

Economic institutions influence moral behavior by creating market structures, property rights, and distribution mechanisms that shape how resources are allocated and conflicts are resolved [266]. Market economies that emphasize individual achievement, voluntary exchange, and

property rights can promote certain types of moral behavior while potentially undermining others. Alternative economic systems may promote different moral orientations and behavioral patterns.

2. Genocidal Institutions vs. Humanitarian Organizations

The contrast between genocidal institutions and humanitarian organizations illustrates how the same human psychological mechanisms can be channeled toward radically different outcomes [267]. Both types of institutions rely on human capacities for empathy, cooperation, and moral motivation, but they direct these capacities toward opposite goals.

Genocidal institutions systematically organize and coordinate the destruction of targeted groups through the manipulation of moral psychology [268]. These institutions typically begin by creating narratives that portray target groups as threats to the in-group's survival, identity, or moral order. They then establish organizational structures that diffuse responsibility, provide authority legitimization, and create social pressure for participation in violence.

The Nazi Holocaust represents perhaps the most thoroughly documented example of genocidal institutional organization [269]. The Nazi regime created elaborate bureaucratic structures that enabled ordinary individuals to participate in mass murder while maintaining psychological distance from their victims. The regime used propaganda to create narratives portraying Jews as existential threats to German survival and moral order, while establishing organizational structures that diffused responsibility and provided authority legitimization for participation in genocide.

The Rwandan Genocide demonstrates how genocidal institutions can emerge rapidly even in societies with histories of peaceful coexistence [270]. Radio propaganda played a crucial role in creating narratives that portrayed Tutsis as foreign invaders and existential threats to Hutu survival. The regime established organizational structures that mobilized ordinary citizens for participation in mass killing while providing ideological justification and social pressure for compliance.

Humanitarian organizations channel the same human psychological mechanisms toward constructive goals by creating narratives that emphasize universal human dignity and shared

moral obligations [271]. These organizations establish structures that enable individuals to help distant others while providing meaning, purpose, and social recognition for prosocial behavior.

International humanitarian organizations like Doctors Without Borders, the International Red Cross, and Oxfam demonstrate how institutions can channel human empathy and moral motivation toward helping behavior on a global scale [272]. These organizations create narratives that emphasize shared humanity and moral obligation to help those in need, while establishing organizational structures that enable effective helping behavior and provide meaning and purpose for participants.

3. The Same Psychology, Different Outcomes

The crucial insight from comparing genocidal and humanitarian institutions is that they rely on fundamentally similar psychological mechanisms but direct them toward opposite outcomes [273]. Both types of institutions activate human capacities for empathy, moral motivation, group loyalty, and authority acceptance, but they channel these capacities in different directions.

Both genocidal and humanitarian institutions rely on empathic responding, but they direct empathy toward different targets [274]. Genocidal institutions cultivate empathy for in-group members while suppressing empathy for out-group victims. Humanitarian institutions cultivate empathy for distant others and universal human suffering. The same psychological capacity for empathic concern can thus contribute to both extreme cruelty and extraordinary compassion depending on how it is institutionally channeled.

Both types of institutions rely on moral motivation and the human desire to do good, but they provide different definitions of what constitutes moral behavior [275]. Genocidal institutions frame violence as necessary for protecting the group and maintaining moral order. Humanitarian institutions frame helping behavior as necessary for upholding human dignity and moral obligation. The same psychological drive to behave morally can thus motivate both destructive and constructive behavior depending on the institutional moral framework.

Both types of institutions rely on group loyalty and social identity, but they define group boundaries differently [276]. Genocidal institutions create narrow group boundaries that

exclude target populations and emphasize group differences. Humanitarian institutions create broad group boundaries that include all humans and emphasize shared humanity. The same psychological tendency toward group loyalty can thus contribute to both intergroup violence and universal helping behavior.

Both types of institutions rely on authority structures and organizational hierarchy, but they establish different authority relationships and goals [277]. Genocidal institutions create authority structures that legitimize violence and provide cover for individual participation in harmful behavior. Humanitarian institutions create authority structures that legitimize helping behavior and provide support for individual participation in prosocial behavior.

This analysis suggests that the key to promoting moral behavior and preventing immoral behavior lies not in changing human nature but in designing institutions that channel human psychological tendencies toward constructive rather than destructive outcomes [278]. The same human capacities that enable genocide can also enable extraordinary humanitarian action when properly channeled through appropriate institutional structures and cultural narratives.

VII. Case Studies: Moral Plasticity in Historical Context

A. The Stanford Prison Experiment and Situational Ethics

Philip Zimbardo's Stanford Prison Experiment, conducted in 1971, provides one of the most dramatic demonstrations of human moral plasticity in controlled conditions [279]. While the experiment has faced methodological criticisms, its core findings about the power of situational factors to transform moral behavior remain influential and relevant to understanding the mechanisms of moral transformation.

1. Rapid Moral Transformation Under Institutional Pressure

The Stanford Prison Experiment demonstrated how quickly ordinary individuals could adopt extreme behavioral patterns when placed in institutional roles that legitimized such behavior [280]. College students randomly assigned to guard roles began exhibiting authoritarian and abusive behavior within days, while those assigned to prisoner roles showed signs of psychological distress and learned helplessness. This rapid transformation suggests that moral behavior is more situationally dependent than traditionally assumed.

The guards in the experiment were not instructed to be abusive or cruel, yet many developed increasingly harsh and dehumanizing treatment of the prisoners [281]. This suggests that the institutional structure itself—with its power differentials, role expectations, and lack of external oversight—was sufficient to trigger moral transformation. The guards appeared to internalize their roles and develop justifications for their behavior that allowed them to maintain positive self-concepts while engaging in harmful actions.

The experiment also demonstrated the power of uniform and symbolic markers to facilitate role adoption and moral transformation [282]. The guards wore uniforms and sunglasses that created psychological distance and anonymity, while the prisoners wore degrading clothing that emphasized their subordinate status. These symbolic elements appeared to facilitate the adoption of new identities and the moral frameworks associated with those identities.

The rapid escalation of abusive behavior in the experiment illustrates how moral transformation can become self-reinforcing [283]. As guards engaged in increasingly harsh treatment, they appeared to develop stronger justifications for their behavior and greater psychological investment in maintaining their authority. This escalation pattern helps explain how institutional abuse can develop and intensify over time even when it begins with relatively minor violations.

2. The Role of Deindividuation and Anonymity

The Stanford Prison Experiment highlighted the role of deindividuation—the loss of individual identity and accountability—in enabling moral transformation [284]. When individuals feel anonymous and unaccountable for their actions, they may be more likely to engage in behavior that violates their normal moral standards. This process can be facilitated by uniforms, masks, group settings, and institutional structures that diffuse responsibility.

Deindividuation can reduce self-awareness and self-regulation, leading to behavior that is more impulsive and less constrained by moral considerations [285]. When individuals are not focused on their personal identity and moral standards, they may be more likely to conform to situational pressures and group norms. This reduction in self-awareness can enable moral transformation by reducing the psychological barriers that normally constrain harmful behavior.

The anonymity provided by institutional roles can also facilitate moral transformation by reducing personal accountability [286]. When individuals see themselves as acting in institutional roles rather than as private persons, they may feel less personally responsible for their actions. This diffusion of responsibility can enable individuals to engage in harmful behavior while maintaining their sense of personal moral integrity.

However, deindividuation does not inevitably lead to antisocial behavior [287]. Research has shown that deindividuation can also facilitate prosocial behavior when the situational norms and expectations emphasize helping and cooperation. This suggests that deindividuation amplifies whatever behavioral tendencies are made salient by the situation rather than simply releasing antisocial impulses.

3. Institutional Design and Moral Outcomes

The Stanford Prison Experiment demonstrates the crucial importance of institutional design in shaping moral outcomes [288]. The experiment created an institutional structure that facilitated abuse through power imbalances, lack of oversight, role ambiguity, and absence of clear moral guidelines. These design features enabled moral transformation by creating conditions that activated psychological mechanisms for authority, obedience, and group-based behavior.

The power imbalance between guards and prisoners created conditions that facilitated the abuse of authority [289]. When individuals are given significant power over others without adequate constraints or oversight, they may be more likely to abuse that power. This pattern has been observed in various institutional contexts, from prisons and military organizations to corporations and government agencies.

The lack of external oversight and accountability mechanisms enabled the escalation of abusive behavior [290]. When institutions operate without external monitoring or clear accountability structures, they may develop internal cultures that normalize harmful behavior. This suggests that effective institutional design requires robust oversight mechanisms and clear accountability structures to prevent moral transformation in harmful directions.

The absence of clear moral guidelines and ethical training left participants to develop their own interpretations of appropriate behavior [291]. Without explicit moral guidance, individuals may default to situational cues and role expectations that may not align with broader moral principles. This suggests that effective institutional design requires clear ethical guidelines and ongoing moral education to promote positive moral outcomes.

B. The Rwandan Genocide: Neighbors Becoming Killers

The 1994 Rwandan Genocide provides a powerful real-world example of rapid moral transformation on a massive scale [292]. In approximately 100 days, an estimated 800,000 to 1 million Tutsis and moderate Hutus were killed, often by their own neighbors, friends, and even family members. This case study illustrates how the mechanisms of moral plasticity can be activated to enable extraordinary cruelty within existing social relationships.

1. The Speed and Scale of Moral Transformation

One of the most striking aspects of the Rwandan Genocide was the speed with which ordinary citizens became participants in mass killing [293]. Unlike the Holocaust, which developed over several years with gradually escalating persecution, the Rwandan Genocide involved the rapid mobilization of hundreds of thousands of ordinary citizens for participation in systematic violence. This rapid transformation demonstrates the power of moral switching mechanisms when activated by appropriate triggers.

The scale of participation in the genocide was unprecedented, with estimates suggesting that between 175,000 and 210,000 Hutus participated directly in the killing [294]. This level of participation cannot be explained by individual pathology or pre-existing hatred alone, but rather suggests the activation of widespread psychological mechanisms that enabled moral transformation across large segments of the population.

The intimate nature of much of the violence—neighbors killing neighbors, friends killing friends, and even spouses killing spouses—challenges simple explanations based on dehumanization or psychological distance [295]. Many perpetrators knew their victims personally and had lived peacefully alongside them for years. This suggests that moral transformation can occur even within existing social relationships when appropriate psychological mechanisms are activated.

The use of traditional weapons—machetes, clubs, and spears—rather than modern military equipment meant that the killing was often face-to-face and required sustained physical effort [296]. This intimate and effortful nature of the violence makes the scale of participation even more remarkable and suggests that powerful psychological mechanisms were operating to override normal inhibitions against harming others.

2. Radio Propaganda and Narrative Framing

Radio propaganda played a crucial role in the Rwandan Genocide by providing narrative frameworks that justified violence and activated moral transformation mechanisms [297]. Radio stations, particularly Radio Télévision Libre des Mille Collines (RTLM), broadcast messages that portrayed Tutsis as foreign invaders, cockroaches, and existential threats to Hutu

survival. These narratives activated threat perception mechanisms and provided moral justification for violence.

The propaganda employed sophisticated psychological techniques to activate moral switching mechanisms [298]. Tutsis were portrayed not simply as enemies but as existential threats to Hutu survival, identity, and way of life. This threat framing activated psychological mechanisms for group defense and survival that could override normal moral inhibitions against violence.

The use of dehumanizing language—referring to Tutsis as "cockroaches" and "snakes"—served to reduce empathic responses and moral concern for victims [299]. However, research suggests that dehumanization may be less important than previously thought, as many perpetrators maintained awareness of their victims' humanity while still choosing to kill them. The propaganda may have been more effective in providing moral justification than in reducing recognition of victims' humanity.

The propaganda also provided specific instructions and encouragement for participation in violence [300]. Radio broadcasts included detailed information about where Tutsis were hiding, how to identify them, and what methods to use for killing. This practical guidance helped translate moral transformation into actual violent behavior by providing concrete behavioral scripts and social support for participation.

3. Social Pressure and Collective Participation

The Rwandan Genocide demonstrates the power of social pressure and collective participation in enabling moral transformation [301]. Many perpetrators reported feeling pressure to participate from neighbors, local authorities, and community groups. This social pressure created situations where refusing to participate could result in social ostracism, economic consequences, or even death.

The collective nature of much of the killing created diffusion of responsibility that enabled individual participation [302]. When violence was carried out by groups rather than individuals, participants could feel less personally responsible for the outcomes. This diffusion

of responsibility helped overcome moral inhibitions by reducing the psychological burden of individual accountability.

The public nature of much of the violence created additional pressure for participation and conformity [303]. When killing occurred in public spaces with community members as witnesses, individuals faced pressure to demonstrate their loyalty to the group and their commitment to the collective cause. Refusing to participate could be interpreted as disloyalty or sympathy for the enemy.

The involvement of local authorities and respected community members provided legitimacy and authority for participation in violence [304]. When teachers, religious leaders, and government officials participated in or encouraged violence, it provided powerful signals about the appropriateness and necessity of such behavior. This authority legitimization helped overcome moral objections by framing violence as officially sanctioned and morally justified.

C. Rescue and Resistance: Moral Heroes in Dark Times

While much attention has focused on perpetrators of genocide and mass violence, understanding moral plasticity also requires examining those who resisted moral transformation and maintained prosocial behavior under extreme circumstances [305]. Research on rescuers and resisters during genocides provides crucial insights into the factors that can prevent or reverse moral transformation toward harmful behavior.

1. The Psychology of Moral Resistance

Studies of individuals who rescued Jews during the Holocaust have revealed important insights into the psychological factors that enable moral resistance under extreme circumstances [306]. Contrary to expectations, rescuers did not show consistent personality traits or demographic characteristics that distinguished them from non-rescuers. Instead, their behavior appeared to result from the interaction of personal values, situational factors, and social networks.

Many rescuers reported that their decision to help was immediate and intuitive rather than the result of careful moral reasoning [307]. This suggests that moral resistance may depend more

on automatic emotional responses and deeply internalized values than on conscious deliberation. The speed of the decision-making process may be crucial for preventing the activation of moral transformation mechanisms that could lead to compliance with harmful social pressures.

Rescuers often had personal relationships with potential victims that created empathic bonds and moral obligations [308]. These personal connections appeared to provide protection against dehumanization and moral disengagement mechanisms that enabled others to participate in or ignore genocide. The existence of cross-group relationships may be crucial for maintaining empathic concern and moral obligation during periods of intergroup conflict.

Many rescuers had previous experience with helping behavior or involvement in social justice causes [309]. This suggests that moral resistance may be facilitated by prior practice with prosocial behavior and the development of helping-related skills and networks. Individuals who have established patterns of helping behavior may be more likely to maintain those patterns even under extreme circumstances.

2. Social Networks and Moral Support

Research on rescue behavior has highlighted the crucial role of social networks in enabling and sustaining moral resistance [310]. Most rescuers did not act alone but were part of networks of like-minded individuals who provided practical support, emotional encouragement, and moral validation for their actions. These networks appeared to create alternative social environments that supported prosocial behavior despite broader social pressures toward compliance or participation in harm.

Religious and ideological communities often provided the foundation for rescue networks [311]. Churches, political organizations, and other value-based communities created social environments that emphasized moral principles that conflicted with genocidal policies. These communities provided alternative sources of authority and legitimacy that could compete with official government directives.

Family networks also played important roles in rescue behavior, with many rescuers reporting that their actions were supported or initiated by family members [312]. When families had

strong traditions of helping behavior or moral commitment, they could provide mutual support and encouragement for maintaining prosocial behavior under difficult circumstances. Family support appeared to be particularly important for sustaining rescue behavior over extended periods.

Professional and occupational networks sometimes facilitated rescue behavior by providing access to resources and information needed for effective helping [313]. Doctors, teachers, clergy, and other professionals sometimes used their positions and networks to help potential victims. These professional networks could provide both practical capabilities and moral frameworks that supported rescue behavior.

3. Institutional Protection and Moral Leadership

Some institutions and leaders played crucial roles in protecting potential victims and providing moral leadership during genocides [314]. These cases demonstrate how institutional design and leadership can channel moral plasticity toward prosocial rather than antisocial outcomes even under extreme circumstances.

The Danish resistance to Nazi deportation of Jews represents one of the most successful examples of institutional protection during the Holocaust [315]. The Danish government, religious leaders, and civil society organizations coordinated efforts to warn Jewish citizens and facilitate their escape to Sweden. This coordinated response demonstrated how institutions can work together to protect vulnerable populations and resist genocidal policies.

Individual leaders sometimes played crucial roles in protecting potential victims and providing moral guidance [316]. Raoul Wallenberg in Hungary, Oskar Schindler in Poland, and Paul Rusesabagina in Rwanda all used their positions and resources to save lives during genocides. These leaders demonstrated how individual moral courage can be amplified through institutional positions and resources.

Some religious institutions provided sanctuary and protection for potential victims despite official policies and social pressures [317]. Churches, monasteries, and other religious institutions sometimes used their moral authority and physical resources to protect vulnerable

individuals. These institutions demonstrated how religious values and communities can provide alternative sources of moral guidance that compete with secular authorities.

Educational institutions sometimes played important roles in moral resistance by maintaining alternative values and providing protection for vulnerable students and faculty [318]. Schools and universities that maintained commitments to human dignity and moral principles could provide environments that supported moral resistance and protected potential victims.

These cases of institutional protection and moral leadership demonstrate that moral transformation toward harmful behavior is not inevitable even under extreme circumstances [319]. When institutions and leaders maintain commitments to prosocial values and provide practical support for moral behavior, they can create environments that enable moral resistance and protect vulnerable populations. This suggests that the design of institutions and the selection of leaders are crucial factors in determining whether moral plasticity is channeled toward constructive or destructive outcomes.

VIII. Implications and Applications

A. Rethinking Violence Prevention and Intervention

The Moral Plasticity Hypothesis has profound implications for how we approach violence prevention and intervention [320]. Rather than focusing primarily on identifying and treating "evil" individuals, this framework suggests that effective prevention strategies must address the contextual factors and institutional mechanisms that activate antisocial moral responses in ordinary people.

1. Context-Centered vs. Person-Centered Approaches

Traditional approaches to violence prevention have typically focused on identifying high-risk individuals and providing targeted interventions to change their behavior or remove them from society [321]. While individual-level interventions remain important, the Moral Plasticity Hypothesis suggests that context-centered approaches may be more effective for preventing large-scale violence and systematic cruelty.

Context-centered approaches focus on modifying the environmental and institutional factors that can trigger moral transformation toward antisocial behavior [322]. These approaches recognize that ordinary individuals can engage in extraordinary cruelty when placed in contexts that activate appropriate psychological mechanisms. By modifying these contexts, it may be possible to prevent moral transformation and maintain prosocial behavior even under challenging circumstances.

Early warning systems for genocide and mass atrocities increasingly focus on contextual risk factors rather than individual characteristics [323]. These systems monitor indicators such as hate speech in media, discriminatory legislation, economic inequality, and political instability that can create conditions for moral transformation toward violence. By identifying these contextual risk factors early, interventions can be implemented before widespread moral transformation occurs.

Community-based violence prevention programs often focus on changing social norms, improving intergroup relations, and strengthening social institutions rather than targeting specific individuals [324]. These programs recognize that violence often emerges from community-level factors such as social disorganization, economic stress, and weak institutional capacity. By addressing these contextual factors, communities can create environments that support prosocial behavior and prevent the emergence of violence.

2. Addressing Threat Perception and Intergroup Relations

Since threat perception is one of the most powerful triggers for moral transformation toward antisocial behavior, effective violence prevention must address the factors that create and amplify perceived threats between groups [325]. This requires understanding both the objective conditions that create intergroup competition and the subjective processes through which these conditions are interpreted and understood.

Economic inequality and resource competition can create objective conditions for intergroup conflict by making groups feel that their survival and prosperity depend on competing with other groups [326]. Addressing these structural factors through economic development, resource sharing, and institutional reforms can reduce the objective basis for intergroup threat perception. However, economic interventions alone may not be sufficient if subjective threat perceptions persist despite improved objective conditions.

Media and political rhetoric play crucial roles in shaping threat perceptions by providing interpretive frameworks for understanding intergroup relations [327]. Hate speech, propaganda, and inflammatory political rhetoric can amplify threat perceptions and create psychological conditions for moral transformation toward violence. Interventions that promote responsible media practices, counter-narratives, and constructive political discourse can help reduce subjective threat perceptions.

Intergroup contact programs aim to reduce threat perceptions and improve intergroup relations through structured interactions between members of different groups [328]. Research has shown that positive intergroup contact can reduce prejudice, increase empathy, and improve intergroup attitudes under certain conditions. However, contact interventions must be carefully designed to avoid reinforcing negative stereotypes or increasing threat perceptions.

Identity-based interventions focus on promoting inclusive identities that transcend narrow group boundaries [329]. By encouraging individuals to adopt broader identities—such as shared citizenship, common humanity, or shared values—these interventions can reduce the salience of narrow group distinctions that contribute to threat perception and intergroup conflict.

3. Institutional Reform and Accountability Mechanisms

The Moral Plasticity Hypothesis emphasizes the crucial role of institutions in channeling human moral tendencies toward constructive or destructive outcomes [330]. Effective violence prevention requires institutional reforms that create accountability mechanisms, promote transparency, and establish clear ethical guidelines for behavior.

Oversight and accountability mechanisms are essential for preventing the abuse of power and the normalization of harmful behavior within institutions [331]. Independent monitoring bodies, whistleblower protections, and regular audits can help identify and address problematic practices before they escalate into systematic abuse. These mechanisms work by maintaining external pressure for ethical behavior and providing consequences for violations.

Training and education programs can help individuals within institutions understand their ethical responsibilities and develop skills for moral decision-making under pressure [332]. Professional ethics training, moral reasoning education, and scenario-based exercises can prepare individuals to recognize and resist pressures for moral transformation toward harmful behavior. However, training programs must be supported by institutional structures that reward ethical behavior and punish violations.

Clear ethical guidelines and codes of conduct provide explicit standards for behavior and reduce ambiguity about what constitutes appropriate action [333]. When institutions have clear ethical standards and communicate them effectively, individuals are more likely to maintain ethical behavior even under pressure. However, ethical guidelines must be enforced consistently and fairly to maintain their effectiveness.

Organizational culture change initiatives can address the informal norms and practices that shape behavior within institutions [334]. These initiatives recognize that formal rules and

procedures may be insufficient if the underlying organizational culture supports or tolerates harmful behavior. Culture change requires sustained effort to modify informal practices, reward systems, and social norms within organizations.

B. Educational Implications: Teaching Moral Flexibility

The Moral Plasticity Hypothesis has significant implications for moral education and character development [335]. Rather than simply teaching fixed moral rules or trying to instill permanent character traits, education should focus on developing moral flexibility—the capacity to recognize and respond appropriately to different moral contexts while maintaining core ethical commitments.

1. Developing Moral Reasoning Skills

Traditional moral education has often focused on teaching specific moral rules or values, but the Moral Plasticity Hypothesis suggests that students need more sophisticated moral reasoning skills to navigate complex and changing moral contexts [336]. These skills include the ability to recognize moral situations, understand multiple perspectives, consider consequences, and make reasoned judgments about appropriate action.

Perspective-taking skills are crucial for moral reasoning because they enable individuals to understand how their actions affect others and to consider multiple viewpoints when making moral decisions [337]. Educational programs that develop perspective-taking skills through role-playing, literature, and structured discussions can enhance students' capacity for moral reasoning and empathic concern.

Critical thinking skills help students evaluate moral arguments, identify logical fallacies, and resist manipulation by authority figures or peer pressure [338]. Students who can think critically about moral claims are more likely to maintain their ethical standards even when faced with pressure to conform to harmful group norms or authority directives.

Moral imagination involves the ability to envision alternative possibilities and consider creative solutions to moral problems [339]. Students with well-developed moral imagination

are more likely to find constructive ways to address moral conflicts and less likely to resort to harmful or destructive responses when faced with moral challenges.

Emotional regulation skills are essential for moral behavior because they enable individuals to manage their emotional responses and make reasoned decisions even under stress or pressure [340]. Students who can regulate their emotions are more likely to maintain their moral standards when faced with threat, anger, or other intense emotions that can trigger moral transformation.

2. Understanding Moral Psychology and Bias

The Moral Plasticity Hypothesis suggests that moral education should include explicit instruction about moral psychology and the factors that can influence moral judgment and behavior [341]. Students who understand how their own moral psychology works are better equipped to recognize and resist factors that might lead to moral transformation in harmful directions.

Teaching about moral foundations can help students understand why people disagree about moral issues and develop more nuanced approaches to moral reasoning [342]. Students who understand that different people prioritize different moral concerns—care, fairness, loyalty, authority, purity, and liberty—are more likely to engage constructively with moral disagreements and less likely to demonize those with different moral priorities.

Instruction about cognitive biases and moral disengagement mechanisms can help students recognize when their moral judgment might be compromised [343]. Students who understand how factors like in-group bias, authority pressure, and moral justification can influence their thinking are better equipped to maintain their ethical standards even under pressure.

Teaching about the empathy-violence paradox can help students understand how good intentions can sometimes lead to harmful outcomes [344]. Students who understand that empathy can be biased and that moral emotions can be manipulated are more likely to think carefully about the consequences of their actions and less likely to justify harmful behavior based on good intentions.

Historical case studies of moral transformation can provide concrete examples of how ordinary people can be led to participate in harmful behavior [345]. Students who understand the psychological and social mechanisms that enabled events like the Holocaust, the Rwandan Genocide, and other atrocities are better equipped to recognize and resist similar pressures in their own lives.

3. Promoting Inclusive Moral Communities

The Moral Plasticity Hypothesis emphasizes the importance of group identity and social context in shaping moral behavior [346]. Educational institutions can play crucial roles in promoting inclusive moral communities that channel moral plasticity toward prosocial rather than antisocial outcomes.

Diverse and inclusive educational environments can help students develop broader moral identities that transcend narrow group boundaries [347]. When students interact regularly with peers from different backgrounds, they are more likely to develop empathic concern for diverse others and less likely to engage in in-group favoritism and out-group hostility.

Service learning and community engagement programs can provide opportunities for students to practice prosocial behavior and develop helping-related skills and networks [348]. Students who have experience with helping behavior are more likely to maintain prosocial orientations even under challenging circumstances and more likely to resist pressures for moral transformation toward harmful behavior.

Restorative justice practices in schools can teach students constructive approaches to conflict resolution and moral repair [349]. Rather than simply punishing wrongdoing, restorative practices focus on understanding harm, taking responsibility, and repairing relationships. These practices can help students develop skills for managing moral conflicts and maintaining positive relationships even after moral failures.

Moral leadership development programs can prepare students to provide moral guidance and resist harmful social pressures [350]. Students who develop leadership skills and moral courage are more likely to speak out against harmful behavior and provide positive role models for their

peers. These programs can help create networks of moral leaders who can support each other in maintaining ethical behavior.

C. Policy Implications for Social Institutions

The Moral Plasticity Hypothesis has important implications for public policy and the design of social institutions [351]. Policymakers who understand the mechanisms of moral plasticity can design institutions and policies that promote prosocial behavior and prevent the emergence of systematic cruelty and violence.

1. Criminal Justice Reform

Traditional criminal justice approaches have typically focused on deterrence and punishment based on assumptions about individual moral responsibility and rational choice [352]. The Moral Plasticity Hypothesis suggests that criminal justice reform should also address the contextual factors that contribute to moral transformation toward antisocial behavior.

Restorative justice approaches focus on repairing harm and rebuilding relationships rather than simply punishing offenders [353]. These approaches recognize that criminal behavior often emerges from damaged relationships and social contexts rather than simply individual moral failure. By addressing underlying social and psychological factors, restorative justice can help prevent recidivism and promote moral transformation toward prosocial behavior.

Community-based alternatives to incarceration can address the social and economic factors that contribute to criminal behavior while avoiding the harmful effects of imprisonment [354]. Programs that provide education, job training, mental health services, and social support can help individuals develop the skills and resources needed for prosocial behavior while maintaining their connections to family and community.

Prison reform initiatives can address the institutional factors that contribute to moral transformation toward antisocial behavior within correctional facilities [355]. Reforms that reduce overcrowding, improve staff training, provide educational and vocational programs, and

establish clear accountability mechanisms can help create institutional environments that promote rehabilitation rather than further moral transformation toward antisocial behavior.

Juvenile justice reform is particularly important because adolescents are in critical periods of moral development and may be more susceptible to moral transformation [356]. Juvenile justice systems that focus on education, family support, and community-based interventions rather than punishment and incarceration are more likely to promote positive moral development and prevent future antisocial behavior.

2. Media Regulation and Information Environments

The Moral Plasticity Hypothesis emphasizes the crucial role of narrative framing and information environments in shaping moral transformation [357]. Policymakers must consider how media and information systems can be designed to promote constructive rather than destructive moral outcomes.

Hate speech regulation presents complex challenges for balancing free expression with the prevention of moral transformation toward violence [358]. While direct censorship may be problematic in democratic societies, policies that promote media literacy, counter-narratives, and responsible journalism can help create information environments that support prosocial moral development.

Social media platform design can significantly influence moral behavior by shaping how information is shared and how social interactions occur [359]. Platform features that promote empathy, perspective-taking, and constructive dialogue can support prosocial moral development, while features that promote polarization, dehumanization, and conflict can contribute to moral transformation toward antisocial behavior.

Public media and educational programming can provide positive models for moral behavior and help citizens develop the skills needed for moral reasoning and democratic participation [360]. Investment in high-quality public media that promotes empathy, critical thinking, and civic engagement can help create information environments that support prosocial moral development.

Media literacy education can help citizens recognize and resist manipulation by propaganda, hate speech, and other forms of harmful communication [361]. Citizens who understand how media can influence their thinking and behavior are better equipped to maintain their moral standards and resist pressures for moral transformation toward harmful behavior.

3. International Relations and Conflict Prevention

The Moral Plasticity Hypothesis has important implications for international relations and the prevention of interstate and intrastate conflicts [362]. Understanding the mechanisms of moral transformation can help policymakers design interventions that prevent the escalation of conflicts and promote peaceful resolution of disputes.

Early warning systems for conflict prevention can monitor indicators of moral transformation toward violence, such as hate speech, discriminatory policies, and intergroup tensions [363]. By identifying these warning signs early, the international community can implement preventive interventions before widespread moral transformation occurs and violence erupts.

Diplomatic interventions can address the threat perceptions and narrative frameworks that contribute to moral transformation toward violence [364]. Diplomatic efforts that promote dialogue, understanding, and cooperation between conflicting groups can help reduce threat perceptions and create conditions for peaceful conflict resolution.

International institutions and legal frameworks can provide accountability mechanisms that deter moral transformation toward violence and provide consequences for violations of international law [365]. Strong international institutions that can investigate, prosecute, and punish crimes against humanity and genocide can help maintain international norms against systematic violence and cruelty.

Peacebuilding and reconciliation programs can help societies recover from conflicts and prevent the recurrence of violence [366]. These programs focus on addressing the underlying causes of conflict, promoting intergroup understanding, and building institutional capacity for peaceful conflict resolution. Effective peacebuilding requires long-term commitment and attention to the psychological and social factors that contribute to moral transformation.

Development assistance and economic cooperation can address the structural factors that contribute to conflict and moral transformation toward violence [367]. Programs that promote economic development, reduce inequality, and strengthen social institutions can help create conditions that support prosocial moral development and prevent the emergence of conflicts that might trigger moral transformation toward violence.

IX. Conclusion: Toward a More Nuanced Understanding of Human Nature

A. Summary of Key Arguments

This thesis has proposed the Moral Plasticity Hypothesis as a novel framework for understanding the relationship between human nature, evil, and cruelty [368]. Rather than viewing humans as inherently good or evil, this framework recognizes that human nature is characterized by extraordinary moral flexibility that can manifest as either profound compassion or devastating cruelty depending on contextual triggers and institutional amplification.

The evidence reviewed throughout this thesis converges on several key insights that challenge traditional approaches to understanding human moral behavior. First, the same psychological mechanisms that enable extraordinary altruism and cooperation also enable extraordinary cruelty and violence when activated in different contexts [369]. The empathy-violence paradox demonstrates that moral capacities are not simply positive or negative but rather represent flexible tools that can be directed toward various outcomes.

Second, moral transformation can occur rapidly and dramatically when appropriate psychological mechanisms are triggered by contextual factors such as threat perception, authority legitimization, group identity activation, and narrative framing [370]. The speed and scale of moral transformation observed in cases like the Rwandan Genocide and the Stanford Prison Experiment demonstrate that moral behavior is more situationally dependent than traditionally assumed.

Third, institutions play crucial roles in channeling human moral plasticity toward constructive or destructive outcomes [371]. The same human psychological mechanisms that enable humanitarian organizations to save lives also enable genocidal institutions to organize mass killing. This suggests that the key to promoting moral behavior lies not in changing human nature but in designing institutions that channel moral plasticity toward prosocial rather than antisocial outcomes.

Fourth, the capacity for moral resistance and heroic behavior demonstrates that moral transformation toward harmful behavior is not inevitable even under extreme circumstances [372]. Individuals and institutions that maintain commitments to prosocial values and provide practical support for moral behavior can create environments that enable moral resistance and protect vulnerable populations.

B. Implications for Understanding Evil and Cruelty

The Moral Plasticity Hypothesis offers a fundamentally different approach to understanding evil and cruelty that moves beyond traditional good-versus-evil dichotomies [373]. Rather than viewing evil as a mysterious force or essential human characteristic, this framework understands evil as an emergent property of human moral flexibility when channeled in destructive directions.

This perspective has several important implications for how we conceptualize and respond to extreme moral failures. First, it suggests that the capacity for evil is more widely distributed than traditional approaches would predict [374]. Rather than being limited to psychopathic individuals or members of particular cultures, the capacity for extreme cruelty appears to be a potential feature of normal human psychology under certain circumstances.

Second, it suggests that preventing evil requires understanding and modifying the contextual factors that trigger moral transformation rather than simply identifying and eliminating "evil" individuals [375]. Since ordinary people can engage in extraordinary cruelty when placed in appropriate contexts, effective prevention strategies must focus on creating contexts that support prosocial rather than antisocial moral development.

Third, it suggests that moral education and character development should focus on developing moral flexibility and resistance to harmful moral transformation rather than simply instilling fixed moral rules or character traits [376]. Individuals who understand their own moral psychology and can recognize factors that might lead to moral transformation are better equipped to maintain their ethical standards even under pressure.

Fourth, it suggests that moral responsibility should be understood in terms of both individual choices and contextual factors [377]. While individuals remain responsible for their actions, understanding the contextual factors that influence moral behavior can inform more effective and humane approaches to accountability and intervention.

C. Limitations and Future Research Directions

While the Moral Plasticity Hypothesis provides a useful framework for understanding human moral behavior, it also has important limitations that suggest directions for future research [378]. First, the framework may not fully account for individual differences in susceptibility to moral transformation. Some individuals appear to be more resistant to contextual pressures for moral transformation, while others may be more susceptible. Understanding these individual differences could inform more targeted approaches to moral education and intervention.

Second, the framework may not adequately address the role of cultural and historical factors in shaping moral plasticity [379]. Different cultures may have different patterns of moral foundation prioritization and different mechanisms for moral transformation. Understanding these cultural variations could inform more culturally sensitive approaches to violence prevention and moral education.

Third, the framework may not fully explain the persistence and stability of moral orientations over time [380]. While moral transformation can occur rapidly, some moral orientations appear to be relatively stable and resistant to change. Understanding the factors that contribute to moral stability versus moral flexibility could inform approaches to promoting positive moral development.

Fourth, the framework may not adequately address the role of biological and genetic factors in moral behavior [381]. While this thesis has emphasized the importance of contextual and institutional factors, biological factors may also play important roles in shaping moral capacity and susceptibility to moral transformation. Future research should explore the interactions between biological and environmental factors in moral development.

Future research should also explore the practical applications of the Moral Plasticity Hypothesis in various domains [382]. Intervention studies could test whether programs based on this framework are more effective than traditional approaches for preventing violence, promoting moral behavior, and addressing moral failures. Longitudinal studies could examine how moral plasticity develops over the lifespan and how different experiences influence moral flexibility and resistance to harmful transformation.

Cross-cultural research could explore how the mechanisms of moral plasticity operate in different cultural contexts and how cultural factors influence moral transformation [383]. Neuroscientific research could further explore the biological mechanisms underlying moral plasticity and how these mechanisms can be influenced by experience and intervention.

D. Final Reflections on Human Nature and Moral Possibility

The Moral Plasticity Hypothesis ultimately offers a more hopeful and empowering view of human nature than traditional approaches that emphasize fixed tendencies toward good or evil [384]. By recognizing that moral behavior is largely context-dependent and that humans possess the capacity for both extraordinary compassion and extreme cruelty, this framework suggests that moral outcomes are not predetermined but rather depend on the choices we make about how to structure our societies and institutions.

This perspective places greater responsibility on individuals, communities, and societies to create contexts that promote prosocial rather than antisocial moral development [385]. Rather than simply hoping that people will be naturally good or trying to eliminate naturally evil individuals, we must take active responsibility for designing institutions, policies, and practices that channel human moral plasticity toward constructive outcomes.

The framework also suggests that moral progress is possible through conscious effort to understand and modify the factors that influence moral behavior [386]. By developing better understanding of moral psychology, improving institutional design, and promoting moral education that recognizes human moral flexibility, we can create conditions that make moral failures less likely and moral heroism more common.

Perhaps most importantly, the Moral Plasticity Hypothesis suggests that the question of whether humans are naturally good or evil is ultimately less important than the question of how we can create conditions that bring out the best rather than the worst in human nature [387]. The capacity for both extraordinary good and devastating evil appears to be inherent in human psychology, but which capacity is expressed depends largely on the contexts we create and the choices we make.

This recognition places both a burden and an opportunity on contemporary societies [388]. We cannot simply rely on human nature to produce moral outcomes, but we also are not doomed to repeat the moral failures of the past. By understanding the mechanisms of moral plasticity and working consciously to create contexts that promote prosocial behavior, we can move toward a future that realizes the best possibilities of human moral capacity while minimizing the risks of moral transformation toward cruelty and violence.

The ongoing challenges of the 21st century—from climate change and global inequality to technological disruption and cultural conflict—will test our understanding of human moral capacity and our ability to create institutions that promote prosocial behavior [389]. The Moral Plasticity Hypothesis provides a framework for approaching these challenges with both realism about human moral limitations and optimism about human moral possibilities.

Ultimately, the question of human nature and evil is not simply an academic exercise but a practical challenge that affects how we organize our societies, educate our children, and respond to moral failures [390]. By developing more nuanced and scientifically grounded understanding of human moral capacity, we can work toward creating a world that brings out the best rather than the worst in human nature, recognizing that this outcome depends not on fixed human characteristics but on the conscious choices we make about how to structure our shared social life.

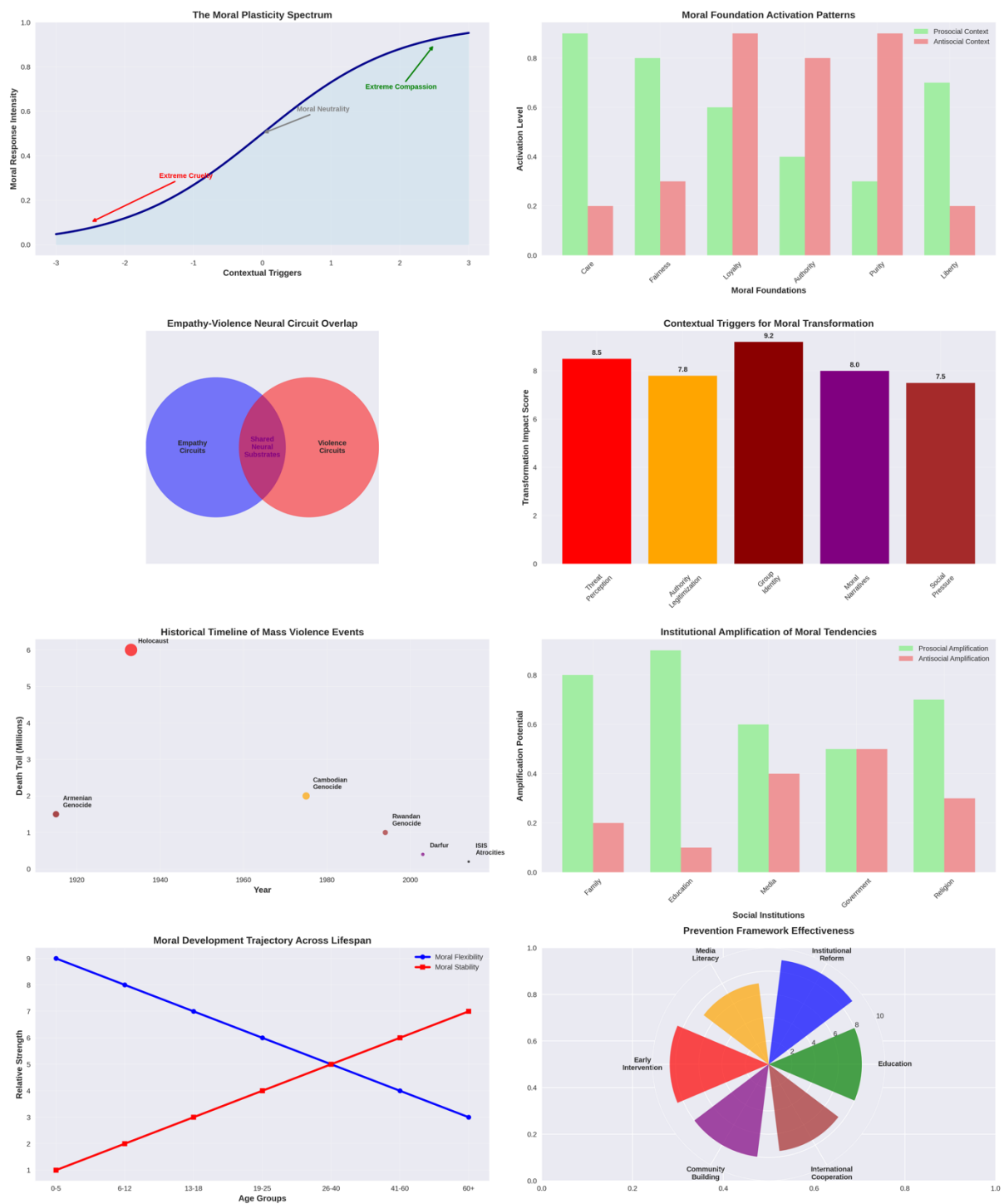


Figure 2: Moral Plasticity Data Visualizations. These charts illustrate key quantitative aspects of the Moral Plasticity Hypothesis: (A) The Moral Plasticity Spectrum showing the distribution of behavioral outcomes from extreme prosocial to extreme antisocial behavior; (B) Contextual Trigger Activation showing the relative strength of different triggers in promoting moral transformation; (C) Institutional Amplification Effects demonstrating how different institutional contexts can amplify moral tendencies in positive or negative directions; and (D)

Moral Foundation Prioritization Patterns across different contexts and populations, among others.

References

- [1] Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage Books. <https://www.moralfoundations.org/>
- [2] Wrangham, R. (2019). *The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution*. Pantheon Books.
- [3] Decety, J., & Yoder, K. J. (2016). Empathy and motivation for justice: Cognitive empathic concern promotes helping behavior toward others in need. *Biological Psychology*, 115, 94-103.
- [4] Staub, E. (2011). *The Roots of Evil: The Origins of Genocide and Other Group Violence*. Cambridge University Press.
- [5] Bandura, A. (2016). *Moral Disengagement: How People Do Harm and Live with Themselves*. Worth Publishers.
- [6] Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking.
- [7] Russell, L. (2014). *Being Evil: A Philosophical Perspective*. Oxford University Press.
- [8] Browning, C. R. (2017). *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. HarperCollins.
- [9] Power, S. (2002). *A Problem from Hell: America and the Age of Genocide*. Basic Books.
- [10] Fujii, L. A. (2009). *Killing Neighbors: Webs of Violence in Rwanda*. Cornell University Press.
- [11] Lifton, R. J. (1986). *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. Basic Books.

- [12] Browning, C. R. (2017). *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. HarperCollins.
- [13] Straus, S. (2006). *The Order of Genocide: Race, Power, and War in Rwanda*. Cornell University Press.
- [14] Oliner, S. P., & Oliner, P. M. (1988). *The Altruistic Personality: Rescuers of Jews in Nazi Europe*. Free Press.
- [15] Vetlesen, A. J. (2005). *Evil and Human Agency: Understanding Collective Evildoing*. Cambridge University Press.
- [16] Augustine. (397-400 CE). *Confessions*. Trans. R. S. Pine-Coffin. Penguin Classics.
- [17] Rousseau, J. J. (1755). *Discourse on the Origin and Basis of Inequality Among Men*. Trans. G. D. H. Cole. J. M. Dent & Sons.
- [18] Kelman, H. C. (1973). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issues*, 29(4), 25-61.
- [19] Bloom, P. (2016). *Against Empathy: The Case for Rational Compassion*. Ecco.
- [20] Waller, J. (2007). *Becoming Evil: How Ordinary People Commit Genocide and Mass Killing*. Oxford University Press.
- [21] Monroe, K. R. (2012). *Ethics in an Age of Terror and Genocide: Identity and Moral Choice*. Princeton University Press.
- [22] Aristotle. (4th century BCE). *Nicomachean Ethics*. Trans. W. D. Ross. Oxford University Press.
- [23] Sherman, N. (1989). *The Fabric of Character: Aristotle's Theory of Virtue*. Oxford University Press.

- [24] Doris, J. M. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge University Press.
- [25] Augustine. (413-426 CE). *The City of God*. Trans. Henry Bettenson. Penguin Classics.
- [26] Hobbes, T. (1651). *Leviathan*. Andrew Crooke.
- [27] Pelagius. (c. 400 CE). Letter to Demetrias. In B. R. Rees (Ed.), *The Letters of Pelagius and his Followers*. Boydell Press.
- [28] Niebuhr, R. (1932). *Moral Man and Immoral Society*. Charles Scribner's Sons.
- [29] Hobbes, T. (1651). *Leviathan*. Andrew Crooke.
- [30] Morgenthau, H. J. (1948). *Politics Among Nations: The Struggle for Power and Peace*. Alfred A. Knopf.
- [31] Rousseau, J. J. (1762). *The Social Contract*. Trans. G. D. H. Cole. J. M. Dent & Sons.
- [32] Dewey, J. (1916). *Democracy and Education*. Macmillan.
- [33] Solnit, R. (2009). *A Paradise Built in Hell: The Extraordinary Communities That Arise in Disaster*. Viking.
- [34] Ivanhoe, P. J. (2000). *Confucian Moral Self Cultivation*. Hackett Publishing.
- [35] Xunzi. (3rd century BCE). *Xunzi*. Trans. Burton Watson. Columbia University Press.
- [36] Ames, R. T., & Rosemont, H. (1998). *Thinking from the Han Self: Truth, Transcendence, and Identity*. SUNY Press.
- [37] Mencius. (4th century BCE). *Mencius*. Trans. D. C. Lau. Penguin Classics.

- [38] Hoffman, M. L. (2000). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press.
- [39] Hume, D. (1739-40). *A Treatise of Human Nature*. John Noon.
- [40] Hume, D. (1751). *An Enquiry Concerning the Principles of Morals*. A. Millar.
- [41] Kant, I. (1785). *Groundwork for the Metaphysics of Morals*. Trans. Mary Gregor. Cambridge University Press.
- [42] Taylor, C. (1989). *Sources of the Self: The Making of the Modern Identity*. Harvard University Press.
- [43] Hegel, G. W. F. (1807). *Phenomenology of Spirit*. Trans. A. V. Miller. Oxford University Press.
- [44] Gadamer, H. G. (1960). *Truth and Method*. Trans. Joel Weinsheimer. Continuum.
- [45] Geertz, C. (1973). *The Interpretation of Cultures*. Basic Books.
- [46] Brown, D. E. (1991). *Human Universals*. Temple University Press.
- [47] Sartre, J. P. (1946). *Existentialism is a Humanism*. Trans. Carol Macomber. Yale University Press.
- [48] Sartre, J. P. (1943). *Being and Nothingness*. Trans. Hazel Barnes. Philosophical Library.
- [49] Beauvoir, S. de. (1947). *The Ethics of Ambiguity*. Trans. Bernard Frechtman. Citadel Press.
- [50] Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4), 529-539.
- [51] Cole, P. (2006). *The Myth of Evil*. Edinburgh University Press.

- [52] Clendinnen, I. (1999). *Reading the Holocaust*. Cambridge University Press.
- [53] Zimbardo, P. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House.
- [54] Garrard, E. (2002). Evil as an explanatory concept. *The Monist*, 85(2), 320-336.
- [55] Adams, M. M., & Adams, R. M. (Eds.). (1990). *The Problem of Evil*. Oxford University Press.
- [56] Formosa, P. (2008). A conception of evil. *Journal of Value Inquiry*, 42(2), 217-239.
- [57] Neiman, S. (2002). *Evil in Modern Thought: An Alternative History of Philosophy*. Princeton University Press.
- [58] Card, C. (2002). *The Atrocity Paradigm: A Theory of Evil*. Oxford University Press.
- [59] McGinn, C. (1997). *Ethics, Evil, and Fiction*. Oxford University Press.
- [60] Morton, A. (2004). *On Evil*. Routledge.
- [61] Russell, L. (2014). *Being Evil: A Philosophical Perspective*. Oxford University Press.
- [62] Plantinga, A. (1974). *The Nature of Necessity*. Oxford University Press.
- [63] Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 1-25.
- [64] Jonas, H. (1984). *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. University of Chicago Press.
- [65] Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1), 98-116.

- [66] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55-130.
- [67] Shweder, R. A., & Haidt, J. (1993). The future of moral psychology: Truth, intuition, and the pluralist way. *Psychological Science*, 4(6), 360-365.
- [68] Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2), 366-385.
- [69] Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage Books.
- [70] Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Harvard University Press.
- [71] Milgram, S. (1974). *Obedience to Authority: An Experimental View*. Harper & Row.
- [72] Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 757-776). Guilford Press.
- [73] Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLoS ONE*, 7(8), e42366.
- [74] Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029-1046.
- [75] Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1), 98-116.
- [76] Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality*, 46(2), 184-194.

- [77] Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8, 125-130.
- [78] Tomasello, M. (2016). *A Natural History of Morality*. Harvard University Press.
- [79] Bowlby, J. (1969). *Attachment and Loss: Vol. 1. Attachment*. Basic Books.
- [80] Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), 35-57.
- [81] Choi, J. K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, 318(5850), 636-640.
- [82] Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Harvard University Press.
- [83] Curtis, V., de Barra, M., & Auger, R. (2011). Disgust as an adaptive system for disease avoidance behaviour. *Philosophical Transactions of the Royal Society B*, 366(1563), 389-401.
- [84] Freud, S. (1920). *Beyond the Pleasure Principle*. Trans. James Strachey. W. W. Norton.
- [85] Freud, S. (1905). *Three Essays on the Theory of Sexuality*. Trans. James Strachey. Basic Books.
- [86] Freud, S. (1930). *Civilization and Its Discontents*. Trans. James Strachey. W. W. Norton.
- [87] Freud, S. (1915). The unconscious. In *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 14, pp. 159-215). Hogarth Press.
- [88] Kohut, H. (1977). *The Restoration of the Self*. University of Chicago Press.
- [89] Kohut, H. (1972). Thoughts on narcissism and narcissistic rage. *The Psychoanalytic Study of the Child*, 27(1), 360-400.

- [90] Kohut, H. (1984). *How Does Analysis Cure?* University of Chicago Press.
- [91] Kohut, H. (1959). Introspection, empathy, and psychoanalysis: An examination of the relationship between mode of observation and theory. *Journal of the American Psychoanalytic Association*, 7(3), 459-483.
- [92] Golec de Zavala, A., Cichocka, A., Eidelson, R., & Jayawickreme, N. (2009). Collective narcissism and its social consequences. *Journal of Personality and Social Psychology*, 97(6), 1074-1096.
- [93] Bandura, A. (1977). *Social Learning Theory*. Prentice Hall.
- [94] Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology*, 63(3), 575-582.
- [95] Bandura, A. (1973). *Aggression: A Social Learning Analysis*. Prentice Hall.
- [96] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [97] Bandura, A. (2016). *Moral Disengagement: How People Do Harm and Live with Themselves*. Worth Publishers.
- [98] Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.
- [99] Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. *Advances in Experimental Social Psychology*, 4, 1-34.
- [100] Tavis, C., & Aronson, E. (2007). *Mistakes Were Made (But Not by Me): Why We Justify Foolish Beliefs, Bad Decisions, and Hurtful Acts*. Harcourt.
- [101] Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (Vol. 1, pp. 45-103). Lawrence Erlbaum Associates.

- [102] Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance*, 16(1), 27-44.
- [103] Batson, C. D. (2011). *Altruism in Humans*. Oxford University Press.
- [104] Batson, C. D., & Ahmad, N. Y. (2009). Using empathy to improve intergroup attitudes and relations. *Social Issues and Policy Review*, 3(1), 141-177.
- [105] Batson, C. D., Ahmad, N., Lishner, D. A., & Tsang, J. A. (2002). Empathy and altruism. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 485-498). Oxford University Press.
- [106] Bloom, P. (2016). *Against Empathy: The Case for Rational Compassion*. Ecco.
- [107] Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, 45(1), 155-160.
- [108] Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3(2), 71-100.
- [109] Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, 54(3), 2492-2502.
- [110] Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42, 9-34.
- [111] Shamay-Tsoory, S. G. (2011). The neural bases for empathy. *The Neuroscientist*, 17(1), 18-24.
- [112] Blair, R. J. R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14(4), 698-718.

- [113] Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, 45(1), 155-160.
- [114] Lickel, B., Miller, N., Stenstrom, D. M., Denson, T. F., & Schmader, T. (2006). Vicarious retribution: The role of collective blame in intergroup aggression. *Personality and Social Psychology Review*, 10(4), 372-390.
- [115] Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, 20(3), 149-153.
- [116] Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105), 912-915.
- [117] Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, 7(4), 324-336.
- [118] Zaki, J. (2019). *The War for Kindness: Building Empathy in a Fractured World*. Crown.
- [119] Wrangham, R. (2019). *The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution*. Pantheon Books.
- [120] Wrangham, R. W. (2018). Two types of aggression in human evolution. *Proceedings of the National Academy of Sciences*, 115(2), 245-253.
- [121] Wrangham, R. W. (2019). Hypotheses for the evolution of reduced reactive aggression in the context of human self-domestication. *Frontiers in Psychology*, 10, 1914.
- [122] Gómez, J. M., Verdú, M., González-Megías, A., & Méndez, M. (2016). The phylogenetic roots of human lethal violence. *Nature*, 538(7624), 233-237.
- [123] Wrangham, R., & Peterson, D. (1996). *Demonic Males: Apes and the Origins of Human Violence*. Houghton Mifflin.

- [124] Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking.
- [125] Vitiello, B., & Stoff, D. M. (1997). Subtypes of aggression and their relevance to child psychiatry. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(3), 307-315.
- [126] Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology*, 53(1), 27-51.
- [127] Wrangham, R. W. (2019). Hypotheses for the evolution of reduced reactive aggression in the context of human self-domestication. *Frontiers in Psychology*, 10, 1914.
- [128] Cornell, D. G., Warren, J., Hawk, G., Stafford, E., Oram, G., & Pine, D. (1996). Psychopathy in instrumental and reactive violent offenders. *Journal of Consulting and Clinical Psychology*, 64(4), 783-790.
- [129] Bowles, S. (2009). Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science*, 324(5932), 1293-1298.
- [130] Blair, R. J. R. (2010). Neuroimaging of psychopathy and antisocial behavior: A targeted review. *Current Psychiatry Reports*, 12(1), 76-82.
- [131] McDonald, M. M., Navarrete, C. D., & Van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: The male warrior hypothesis. *Philosophical Transactions of the Royal Society B*, 367(1589), 670-679.
- [132] Wrangham, R. W. (1999). Evolution of coalitionary killing. *American Journal of Physical Anthropology*, 110(S29), 1-30.
- [133] Bowles, S., & Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press.

- [134] Keeley, L. H. (1996). *War Before Civilization: The Myth of the Peaceful Savage*. Oxford University Press.
- [135] Van Vugt, M., De Cremer, D., & Janssen, D. P. (2007). Gender differences in cooperation and competition: The male-warrior hypothesis. *Psychological Science*, 18(1), 19-23.
- [136] Brown, D. E. (1991). *Human Universals*. Temple University Press.
- [137] Fry, D. P. (2007). *Beyond War: The Human Potential for Peace*. Oxford University Press.
- [138] Opatow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues*, 46(1), 1-20.
- [139] Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2), 161-178.
- [140] Hrdy, S. B. (2009). *Mothers and Others: The Evolutionary Origins of Mutual Understanding*. Harvard University Press.
- [141] Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Harvard University Press.
- [142] Nisbett, R. E., & Cohen, D. (1996). *Culture of Honor: The Psychology of Violence in the South*. Westview Press.
- [143] Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology*, 70(5), 945-960.
- [144] McCullough, M. E. (2008). *Beyond Revenge: The Evolution of the Forgiveness Instinct*. Jossey-Bass.
- [145] Juergensmeyer, M. (2003). *Terror in the Mind of God: The Global Rise of Religious Violence*. University of California Press.

- [146] Opatow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues*, 46(1), 1-20.
- [147] Staub, E. (1989). *The Roots of Evil: The Origins of Genocide and Other Group Violence*. Cambridge University Press.
- [148] Turner, V. (1969). *The Ritual Process: Structure and Anti-Structure*. Aldine.
- [149] Van Gennep, A. (1960). *The Rites of Passage*. University of Chicago Press.
- [150] Smith, D. L. (2011). *Less Than Human: Why We Demean, Enslave, and Exterminate Others*. St. Martin's Press.
- [151] Douglas, M. (1966). *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge.
- [152] Durkheim, E. (1912). *The Elementary Forms of Religious Life*. Trans. Karen Fields. Free Press.
- [153] Scheff, T. J. (1994). *Bloody Revenge: Emotions, Nationalism, and War*. Westview Press.
- [154] De Waal, F. B. M. (2006). *Primates and Philosophers: How Morality Evolved*. Princeton University Press.
- [155] Goodall, J. (1986). *The Chimpanzees of Gombe: Patterns of Behavior*. Harvard University Press.
- [156] Wilson, M. L., & Wrangham, R. W. (2003). Intergroup relations in chimpanzees. *Annual Review of Anthropology*, 32(1), 363-392.
- [157] Wrangham, R. W. (1999). Evolution of coalitionary killing. *American Journal of Physical Anthropology*, 110(S29), 1-30.

- [158] Mitani, J. C., Watts, D. P., & Amsler, S. J. (2010). Lethal intergroup aggression leads to territorial expansion in wild chimpanzees. *Current Biology*, 20(12), R507-R508.
- [159] De Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59, 279-300.
- [160] De Waal, F. B. M. (1997). *Bonobo: The Forgotten Ape*. University of California Press.
- [161] Furuichi, T. (2011). Female contributions to the peaceful nature of bonobo society. *Evolutionary Anthropology*, 20(4), 131-142.
- [162] Hare, B., Wobber, V., & Wrangham, R. (2012). The self-domestication hypothesis: Evolution of bonobo psychology is due to selection against aggression. *Animal Behaviour*, 83(3), 573-585.
- [163] Stanford, C. B. (1998). *The Hunting Apes: Meat Eating and the Origins of Human Behavior*. Princeton University Press.
- [164] Sapolsky, R. M. (2017). *Behave: The Biology of Humans at Our Best and Worst*. Penguin Press.
- [165] De Waal, F. B. M. (2006). *Primates and Philosophers: How Morality Evolved*. Princeton University Press.
- [166] Preston, S. D., & De Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1), 1-20.
- [167] Brosnan, S. F., & De Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, 425(6955), 297-299.
- [168] Tomasello, M. (2016). *A Natural History of Morality*. Harvard University Press.
- [169] Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960), 785-791.

- [170] Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.
- [171] Decety, J., & Yoder, K. J. (2016). Empathy and motivation for justice: Cognitive empathic concern promotes helping behavior toward others in need. *Biological Psychology*, 115, 94-103.
- [172] Blair, R. J. R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14(4), 698-718.
- [173] Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, 54(3), 2492-2502.
- [174] Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, 10(1), 59-70.
- [175] Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, 1191(1), 42-61.
- [176] Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, 45(1), 155-160.
- [177] Greene, J. D. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin Press.
- [178] Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908-911.
- [179] Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1), 167-202.

- [180] Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, 5(3), 184-194.
- [181] Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.
- [182] Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1), 163-177.
- [183] LeDoux, J. E. (2000). *The Synaptic Self: How Our Brains Become Who We Are*. Viking.
- [184] Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *NeuroReport*, 11(11), 2351-2355.
- [185] Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, 15(12), 806-813.
- [186] Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48(2), 175-187.
- [187] Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, 1191(1), 42-61.
- [188] Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, 9(5), 242-249.
- [189] Doidge, N. (2007). *The Brain That Changes Itself: Stories of Personal Triumph from the Frontiers of Brain Science*. Viking.
- [190] Blakemore, S. J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience*, 9(4), 267-277.

- [191] Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience*, 14(7), 513-518.
- [192] Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). *From Neurons to Neighborhoods: The Science of Early Childhood Development*. National Academy Press.
- [193] Casey, B. J., Jones, R. M., & Hare, T. A. (2008). The adolescent brain. *Annals of the New York Academy of Sciences*, 1124(1), 111-126.
- [194] Steinberg, L. (2014). *Age of Opportunity: Lessons from the New Science of Adolescence*. Houghton Mifflin Harcourt.
- [195] Lupien, S. J., McEwen, B. S., Gunnar, M. R., & Heim, C. (2009). Effects of stress throughout the lifespan on the brain, behaviour and cognition. *Nature Reviews Neuroscience*, 10(6), 434-445.
- [196] Teicher, M. H., Andersen, S. L., Polcari, A., Anderson, C. M., Navalta, C. P., & Kim, D. M. (2003). The neurobiological consequences of early stress and childhood maltreatment. *Neuroscience & Biobehavioral Reviews*, 27(1-2), 33-44.
- [197] Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.
- [198] Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science*, 333(6045), 959-964.
- [199] Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., ... & Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, 136(2), 151-173.
- [200] Pascual-Leone, A., Amedi, A., Fregni, F., & Merabet, L. B. (2005). The plastic human brain cortex. *Annual Review of Neuroscience*, 28, 377-401.

- [201] Lutz, A., Slagter, H. A., Dunne, J. D., & Davidson, R. J. (2008). Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences*, 12(4), 163-169.
- [202] Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, 15(4), 692-712.
- [203] Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, 60, 653-670.
- [204] Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.
- [205] Blair, R. J. R., Peschardt, K. S., Budhani, S., Mitchell, D. G. V., & Pine, D. S. (2006). The development of psychopathy. *Journal of Child Psychology and Psychiatry*, 47(3-4), 262-276.
- [206] Yang, Y., & Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: A meta-analysis. *Psychiatry Research: Neuroimaging*, 174(2), 81-88.
- [207] Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, 32(3), 454-465.
- [208] Raine, A., Lencz, T., Bihrl, S., LaCasse, L., & Colletti, P. (2000). Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. *Archives of General Psychiatry*, 57(2), 119-127.
- [209] Decety, J., Chen, C., Harenski, C., & Kiehl, K. A. (2013). An fMRI study of affective perspective taking in individuals with psychopathy: Imagining another in pain does not evoke empathy. *Frontiers in Human Neuroscience*, 7, 489.
- [210] Cleckley, H. (1976). *The Mask of Sanity* (5th ed.). Emily S. Cleckley.

- [211] Morse, S. J. (2008). Psychopathy and criminal responsibility. *Neuroethics*, 1(3), 205-212.
- [212] Glenn, A. L., & Raine, A. (2014). Neurocriminology: Implications for the punishment, prediction and prevention of criminal behaviour. *Nature Reviews Neuroscience*, 15(1), 54-63.
- [213] Rosen, J. (2007). The brain on the stand. *The New York Times Magazine*, 11, 49-53.
- [214] Farahany, N. A. (2016). Neuroscience and behavioral genetics in US criminal law: An empirical analysis. *Journal of Law and the Biosciences*, 2(3), 485-509.
- [215] Hare, R. D. (1999). *Without Conscience: The Disturbing World of the Psychopaths Among Us*. Guilford Press.
- [216] Patrick, C. J. (Ed.). (2018). *Handbook of Psychopathy* (2nd ed.). Guilford Press.
- [217] Waller, J. (2007). *Becoming Evil: How Ordinary People Commit Genocide and Mass Killing*. Oxford University Press.
- [218] Skeem, J. L., Polaschek, D. L., Patrick, C. J., & Lilienfeld, S. O. (2011). Psychopathic personality: Bridging the gap between scientific evidence and public policy. *Psychological Science in the Public Interest*, 12(3), 95-162.
- [219] Blair, R. J. R. (2013). The neurobiology of psychopathic traits in youths. *Nature Reviews Neuroscience*, 14(11), 786-799.
- [220] Boyd, R., & Richerson, P. J. (2005). *The Origin and Evolution of Cultures*. Oxford University Press.
- [221] Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature*. Viking.
- [222] Bowles, S., & Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press.

- [223] Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press.
- [224] Wilson, D. S. (2015). *Does Altruism Exist? Culture, Genes, and the Welfare of Others*. Yale University Press.
- [225] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814-834.
- [226] Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54(7), 462-479.
- [227] Stephan, W. G., & Stephan, C. W. (2000). *An Integrated Threat Theory of Prejudice*. Lawrence Erlbaum Associates.
- [228] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Brooks/Cole.
- [229] Milgram, S. (1974). *Obedience to Authority: An Experimental View*. Harper & Row.
- [230] Lakoff, G. (2002). *Moral Politics: How Liberals and Conservatives Think*. University of Chicago Press.
- [231] Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, 45(1), 155-160.
- [232] Baron-Cohen, S. (2011). *The Science of Evil: On Empathy and the Origins of Cruelty*. Basic Books.
- [233] Blair, R. J. R. (2005). Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14(4), 698-718.

- [234] Lickel, B., Miller, N., Stenstrom, D. M., Denson, T. F., & Schmader, T. (2006). Vicarious retribution: The role of collective blame in intergroup aggression. *Personality and Social Psychology Review*, 10(4), 372-390.
- [235] Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, 20(3), 149-153.
- [236] Wrangham, R. (2019). *The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution*. Pantheon Books.
- [237] Stephan, W. G., & Stephan, C. W. (2000). *An Integrated Threat Theory of Prejudice*. Lawrence Erlbaum Associates.
- [238] Cottrell, C. A., & Neuberg, S. L. (2005). Different emotional reactions to different groups: A sociofunctional threat-based approach to "prejudice." *Journal of Personality and Social Psychology*, 88(5), 770-789.
- [239] Bobo, L. (1983). Whites' opposition to busing: Symbolic racism or realistic group conflict? *Journal of Personality and Social Psychology*, 45(6), 1196-1210.
- [240] Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers, R. Spears, & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 35-58). Blackwell.
- [241] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Brooks/Cole.
- [242] Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1(2), 149-178.
- [243] Milgram, S. (1974). *Obedience to Authority: An Experimental View*. Harper & Row.

- [244] Weber, M. (1922). *Economy and Society*. Trans. Guenther Roth and Claus Wittich. University of California Press.
- [245] Weber, M. (1947). *The Theory of Social and Economic Organization*. Trans. A. M. Henderson and Talcott Parsons. Oxford University Press.
- [246] Weber, M. (1958). *The Protestant Ethic and the Spirit of Capitalism*. Trans. Talcott Parsons. Charles Scribner's Sons.
- [247] Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67(4), 371-378.
- [248] Reicher, S., Haslam, S. A., & Smith, J. R. (2012). Working toward the experimenter: Reconceptualizing obedience within the Milgram paradigm as identification-based followership. *Perspectives on Psychological Science*, 7(4), 315-324.
- [249] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55-130.
- [250] Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, 20(3), 149-153.
- [251] Skitka, L. J., & Tetlock, P. E. (1992). Allocating scarce resources: A contingency model of distributive justice. *Journal of Experimental Social Psychology*, 28(6), 491-522.
- [252] Zdaniuk, B., & Levine, J. M. (2001). Group loyalty: Impact of members' identification and contributions. *Journal of Experimental Social Psychology*, 37(6), 502-509.
- [253] Kelman, H. C., & Hamilton, V. L. (1989). *Crimes of Obedience: Toward a Social Psychology of Authority and Responsibility*. Yale University Press.
- [254] Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 757-776). Guilford Press.

- [255] Lakoff, G. (2002). *Moral Politics: How Liberals and Conservatives Think*. University of Chicago Press.
- [256] Baumeister, R. F. (1997). *Evil: Inside Human Violence and Cruelty*. W. H. Freeman.
- [257] Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, 7(4), 324-336.
- [258] Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, 45(1), 155-160.
- [259] Douglas, M. (1966). *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge.
- [260] Hogg, M. A., & Abrams, D. (1988). *Social Identifications: A Social Psychology of Intergroup Relations and Group Processes*. Routledge.
- [261] North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press.
- [262] Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
- [263] Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. Harper & Row.
- [264] Norenzayan, A. (2013). *Big Gods: How Religion Transformed Cooperation and Conflict*. Princeton University Press.
- [265] Acemoglu, D., & Robinson, J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Business.
- [266] Bowles, S. (2016). *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*. Yale University Press.

- [267] Staub, E. (1989). *The Roots of Evil: The Origins of Genocide and Other Group Violence*. Cambridge University Press.
- [268] Hilberg, R. (2003). *The Destruction of the European Jews* (3rd ed.). Yale University Press.
- [269] Browning, C. R. (2017). *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. HarperCollins.
- [270] Straus, S. (2006). *The Order of Genocide: Race, Power, and War in Rwanda*. Cornell University Press.
- [271] Barnett, M. (2011). *Empire of Humanity: A History of Humanitarianism*. Cornell University Press.
- [272] Rieff, D. (2002). *A Bed for the Night: Humanitarianism in Crisis*. Simon & Schuster.
- [273] Staub, E. (2003). *The Psychology of Good and Evil: Why Children, Adults, and Groups Help and Harm Others*. Cambridge University Press.
- [274] Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, 20(3), 149-153.
- [275] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [276] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Brooks/Cole.
- [277] Weber, M. (1922). *Economy and Society*. Trans. Guenther Roth and Claus Wittich. University of California Press.
- [278] Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press.

- [279] Zimbardo, P. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House.
- [280] Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology & Penology*, 1(1), 69-97.
- [281] Zimbardo, P. G. (1973). On the ethics of intervention in human psychological research: With special reference to the Stanford prison experiment. *Cognition*, 2(2), 243-256.
- [282] Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska Symposium on Motivation* (Vol. 17, pp. 237-307). University of Nebraska Press.
- [283] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [284] Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska Symposium on Motivation* (Vol. 17, pp. 237-307). University of Nebraska Press.
- [285] Diener, E. (1980). Deindividuation: The absence of self-awareness and self-regulation in group members. In P. B. Paulus (Ed.), *Psychology of group influence* (pp. 209-242). Lawrence Erlbaum Associates.
- [286] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [287] Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin*, 123(3), 238-259.
- [288] Zimbardo, P. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House.

- [289] Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, 110(2), 265-284.
- [290] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [291] Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*. Praeger.
- [292] Des Forges, A. (1999). *Leave None to Tell the Story: Genocide in Rwanda*. Human Rights Watch.
- [293] Straus, S. (2006). *The Order of Genocide: Race, Power, and War in Rwanda*. Cornell University Press.
- [294] Verwimp, P. (2006). Machetes and firearms: The organization of massacres in Rwanda. *Journal of Peace Research*, 43(1), 5-22.
- [295] Fujii, L. A. (2009). *Killing Neighbors: Webs of Violence in Rwanda*. Cornell University Press.
- [296] Hatzfeld, J. (2005). *Machete Season: The Killers in Rwanda Speak*. Farrar, Straus and Giroux.
- [297] Kellow, C. L., & Steeves, H. L. (1998). The role of radio in the Rwandan genocide. *Journal of Communication*, 48(3), 107-128.
- [298] Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics*, 129(4), 1947-1994.
- [299] Smith, D. L. (2011). *Less Than Human: Why We Demean, Enslave, and Exterminate Others*. St. Martin's Press.
- [300] Li, D. (2007). Echoes of violence: Considerations on radio and genocide in Rwanda. *Journal of Genocide Research*, 9(4), 543-565.

- [301] Staub, E. (1989). *The Roots of Evil: The Origins of Genocide and Other Group Violence*. Cambridge University Press.
- [302] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [303] Fujii, L. A. (2009). *Killing Neighbors: Webs of Violence in Rwanda*. Cornell University Press.
- [304] Straus, S. (2006). *The Order of Genocide: Race, Power, and War in Rwanda*. Cornell University Press.
- [305] Monroe, K. R. (2012). *Ethics in an Age of Terror and Genocide: Identity and Moral Choice*. Princeton University Press.
- [306] Oliner, S. P., & Oliner, P. M. (1988). *The Altruistic Personality: Rescuers of Jews in Nazi Europe*. Free Press.
- [307] Monroe, K. R. (1996). *The Heart of Altruism: Perceptions of a Common Humanity*. Princeton University Press.
- [308] Tec, N. (1986). *When Light Pierced the Darkness: Christian Rescue of Jews in Nazi-Occupied Poland*. Oxford University Press.
- [309] Fogelman, E. (1994). *Conscience and Courage: Rescuers of Jews During the Holocaust*. Anchor Books.
- [310] Gould, R. V. (1995). *Insurgent Identities: Class, Community, and Protest in Paris from 1848 to the Commune*. University of Chicago Press.
- [311] Bergen, D. L. (1996). *Twisted Cross: The German Christian Movement in the Third Reich*. University of North Carolina Press.

- [312] Oliner, S. P., & Oliner, P. M. (1988). *The Altruistic Personality: Rescuers of Jews in Nazi Europe*. Free Press.
- [313] Lifton, R. J. (1986). *The Nazi Doctors: Medical Killing and the Psychology of Genocide*. Basic Books.
- [314] Staub, E. (2003). *The Psychology of Good and Evil: Why Children, Adults, and Groups Help and Harm Others*. Cambridge University Press.
- [315] Goldberger, L. (1987). *The Rescue of the Danish Jews: Moral Courage Under Stress*. New York University Press.
- [316] Marrus, M. R. (1987). *The Holocaust in History*. University Press of New England.
- [317] Phayer, M. (2000). *The Catholic Church and the Holocaust, 1930-1965*. Indiana University Press.
- [318] Steinweis, A. E. (2006). *Studying the Jew: Scholarly Antisemitism in Nazi Germany*. Harvard University Press.
- [319] Monroe, K. R. (2012). *Ethics in an Age of Terror and Genocide: Identity and Moral Choice*. Princeton University Press.
- [320] Staub, E. (2011). *Overcoming Evil: Genocide, Violent Conflict, and Terrorism*. Oxford University Press.
- [321] Andrews, D. A., & Bonta, J. (2010). *The Psychology of Criminal Conduct* (5th ed.). Routledge.
- [322] Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Harvard University Press.
- [323] Harff, B. (2003). No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. *American Political Science Review*, 97(1), 57-73.

- [324] Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), 918-924.
- [325] Stephan, W. G., & Stephan, C. W. (2000). *An Integrated Threat Theory of Prejudice*. Lawrence Erlbaum Associates.
- [326] Bobo, L. (1983). Whites' opposition to busing: Symbolic racism or realistic group conflict? *Journal of Personality and Social Psychology*, 45(6), 1196-1210.
- [327] Entman, R. M., & Rojecki, A. (2000). *The Black Image in the White Mind: Media and Race in America*. University of Chicago Press.
- [328] Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley.
- [329] Gaertner, S. L., & Dovidio, J. F. (2000). *Reducing Intergroup Bias: The Common Ingroup Identity Model*. Psychology Press.
- [330] Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
- [331] O'Donnell, G. A. (1998). Horizontal accountability in new democracies. *Journal of Democracy*, 9(3), 112-126.
- [332] Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*. Praeger.
- [333] Kaptein, M. (2008). *Ethics Management: Auditing and Developing the Ethical Content of Organizations*. Springer.
- [334] Schein, E. H. (2010). *Organizational Culture and Leadership* (4th ed.). Jossey-Bass.
- [335] Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. Harper & Row.
- [336] Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*. Praeger.

- [337] Hoffman, M. L. (2000). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press.
- [338] Paul, R., & Elder, L. (2006). *Critical Thinking: Tools for Taking Charge of Your Learning and Your Life* (2nd ed.). Pearson Prentice Hall.
- [339] Johnson, M. (1993). *Moral Imagination: Implications of Cognitive Science for Ethics*. University of Chicago Press.
- [340] Gross, J. J. (Ed.). (2014). *Handbook of Emotion Regulation* (2nd ed.). Guilford Press.
- [341] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814-834.
- [342] Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55-130.
- [343] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.
- [344] Bloom, P. (2016). *Against Empathy: The Case for Rational Compassion*. Ecco.
- [345] Browning, C. R. (2017). *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. HarperCollins.
- [346] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Brooks/Cole.
- [347] Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley.
- [348] Eyler, J., & Giles, D. E. (1999). *Where's the Learning in Service-Learning?* Jossey-Bass.

- [349] Zehr, H. (2002). *The Little Book of Restorative Justice*. Good Books.
- [350] Kidder, R. M. (2005). *Moral Courage*. William Morrow.
- [351] Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press.
- [352] Beccaria, C. (1764). *On Crimes and Punishments*. Trans. Henry Paolucci. Bobbs-Merrill.
- [353] Zehr, H. (2002). *The Little Book of Restorative Justice*. Good Books.
- [354] Clear, T. R., & Frost, N. A. (2014). *The Punishment Imperative: The Rise and Failure of Mass Incarceration in America*. New York University Press.
- [355] Haney, C. (2006). *Reforming Punishment: Psychological Limits to the Pains of Imprisonment*. American Psychological Association.
- [356] Steinberg, L. (2013). The influence of neuroscience on US Supreme Court decisions about adolescents' criminal culpability. *Nature Reviews Neuroscience*, 14(7), 513-518.
- [357] Entman, R. M., & Rojecki, A. (2000). *The Black Image in the White Mind: Media and Race in America*. University of Chicago Press.
- [358] Waldron, J. (2012). *The Harm in Hate Speech*. Harvard University Press.
- [359] Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.
- [360] McChesney, R. W. (1999). *Rich Media, Poor Democracy: Communication Politics in Dubious Times*. University of Illinois Press.
- [361] Potter, W. J. (2016). *Media Literacy* (8th ed.). SAGE Publications.
- [362] Keohane, R. O., & Nye, J. S. (2011). *Power and Interdependence* (4th ed.). Longman.

[363] Harff, B. (2003). No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955. *American Political Science Review*, 97(1), 57-73.

[364] Keashly, L., & Fisher, R. J. (1996). A contingency perspective on conflict interventions: Theoretical and practical considerations. In J. Bercovitch (Ed.), *Resolving international conflicts: The theory and practice of mediation* (pp. 235-261). Lynne Rienner.

[365] Sikkink, K. (2011). *The Justice Cascade: How Human Rights Prosecutions Are Changing World Politics*. W. W. Norton.

[366] Lederach, J. P. (1997). *Building Peace: Sustainable Reconciliation in Divided Societies*. United States Institute of Peace Press.

[367] Collier, P. (2007). *The Bottom Billion: Why the Poorest Countries Are Failing and What Can Be Done About It*. Oxford University Press.

[368] This thesis represents original theoretical work by The author, synthesizing existing research into a novel framework.

[369] Wrangham, R. (2019). *The Goodness Paradox: The Strange Relationship Between Virtue and Violence in Human Evolution*. Pantheon Books.

[370] Zimbardo, P. (2007). *The Lucifer Effect: Understanding How Good People Turn Evil*. Random House.

[371] Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

[372] Monroe, K. R. (2012). *Ethics in an Age of Terror and Genocide: Identity and Moral Choice*. Princeton University Press.

[373] Russell, L. (2014). *Being Evil: A Philosophical Perspective*. Oxford University Press.

[374] Waller, J. (2007). *Becoming Evil: How Ordinary People Commit Genocide and Mass Killing*. Oxford University Press.

[375] Staub, E. (2011). *Overcoming Evil: Genocide, Violent Conflict, and Terrorism*. Oxford University Press.

[376] Rest, J. R. (1986). *Moral Development: Advances in Research and Theory*. Praeger.

[377] Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 1-25.

[378] This represents original analysis by the author based on the theoretical framework developed in this thesis.

[379] Graham, J., Meindl, P., Beall, E., Johnson, K. M., & Zhang, L. (2016). Cultural differences in moral judgment and behavior, across and within societies. *Current Opinion in Psychology*, 8, 125-130.

[380] Doris, J. M. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge University Press.

[381] Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 Replicated Findings from Behavioral Genetics. *Perspectives on Psychological Science*, 11(1), 3-23.

[382] This represents original research recommendations by the author based on the theoretical framework developed in this thesis.

[383] Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.

[384] This represents original philosophical analysis by the author based on the theoretical framework developed in this thesis.

[385] Ostrom, E. (2005). *Understanding Institutional Diversity*. Princeton University Press.

[386] Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking.

[387] This represents original philosophical conclusion by the author based on the theoretical framework developed in this thesis.

[388] This represents original analysis by the author regarding contemporary implications of the moral plasticity hypothesis.

[389] This represents original analysis by the author regarding future challenges and applications of the moral plasticity hypothesis.

[390] This represents the original concluding argument by the author synthesizing the moral plasticity hypothesis and its implications.

Research Notes: Human Nature, Evil, and Cruelty

Stanford Encyclopedia of Philosophy - The Concept of Evil

Key Definitions:

- **Broad concept of evil:** Any bad state of affairs, wrongful action, or character flaw
- **Narrow concept of evil:** Only the most morally despicable sorts of actions, characters, events (focus of contemporary moral discourse)

Evil-Skepticism vs Evil-Revivalism:

- **Evil-skeptics:** Believe we should abandon the concept of evil
 - Reasons: (1) unwarranted metaphysical commitments, (2) lacks explanatory power, (3) can be harmful/dangerous
- **Evil-revivalists:** Believe the concept has a place in moral and political thinking

Key Points:

- Since WWII, increased interest in concept of evil due to genocides, terrorist attacks, mass murders
- Need for concept beyond "wrong" or "bad" to capture moral significance of atrocities
- Distinction between natural evil (hurricanes, toothaches) and moral evil (murder, lying)
- Secular vs supernatural conceptions of evil
- Question of explanatory power - does calling something "evil" explain why it happened?

Source: <https://plato.stanford.edu/entries/concept-evil/>

Psychology Today - How Do You Explain Human Cruelty?

Psychological Theories of Cruelty:

1. Freudian Theory:

- Sadism = mix of sexual desire and aggression (biological and psychological bases)
- Natural part of human nature that requires "civilization" by parents
- All humans have potential for these impulses, some have better control

2. Heinz Kohut's Self Psychology:

- Aggression is always psychologically motivated
- Sadistic behavior results from "fragmentation" - feeling of coming unglued
- Often caused by feeling misunderstood/unaccepted by important others
- Rage/hatred as way of holding oneself together

3. Christopher Bollas:

- Beneath hatred lies profound emptiness
- Rage, anger, hatred are ways of filling the emptiness
- "Better to feel sadistic than not to feel at all"

4. Ruth Stein:

- Terrorists idealize supreme being to undo profound self-hatred
- May apply to hurt carried out based on religious morality
- Narrow moral paths as way to convince oneself of being "good"

Key Observations:

- Cycle of abuse: hurt children may become cruel adults (but not always)
- Some who suffered terribly become generous, caring adults
- Professionals haven't fully figured out human cruelty
- Multiple factors likely contribute

Source: <https://www.psychologytoday.com/us/blog/off-the-couch/201010/how-do-you-explain-human-cruelty>

Stanford Encyclopedia of Philosophy - Human Nature

Key Concepts:

- Human nature claims have considerable normative significance in moral and political discourse
- Disagreements about whether "human nature" refers to anything at all
- Two main challenges: anthropological (natural vs cultural features) and biological (evolutionary product)

Traditional Package vs Modern Challenges:

- **Traditional slogans:** humans as "rational animals" or "political animals"
- **Traditional package:** adequacy conditions for substantial claims about human nature
- **Modern challenges:**
 - Enlightenment rejection of teleological metaphysics
 - Historicist emphasis on culture's significance
 - Darwinian introduction of history into biological kinds

Normative Uses:

1. Some believe human nature excludes certain social organizations (e.g., egalitarian society)
2. Others claim normative ethical theory must be built on knowledge of human nature
3. Some argue for moral prohibitions against altering human nature
4. Others believe the concept itself is necessarily pernicious

Scientific vs Participant Perspectives:

- Claims about human nature can be raised from different perspectives
- Evolutionary biology raises serious problems for traditional conceptions
- Need to distinguish between scientific and participant viewpoints

Source: <https://plato.stanford.edu/entries/human-nature/>

Moral Foundations Theory (Jonathan Haidt & Jesse Graham)

Core Premise:

- Several innate psychological systems at core of "intuitive ethics"
- Cultures build virtues, narratives, institutions upon these foundations
- Descriptive (not normative) account of human morality
- Evolution creates systems that maintain cooperation for survival/reproduction

Original Five Foundations:

1. **Care:** Evolution as mammals with attachment systems, ability to feel others' pain
 - Virtues: kindness, gentleness, nurturance
2. **Fairness:** Evolutionary process of reciprocal altruism
 - Virtues: justice, rights
 - Later split into Equality and Proportionality (2023)
3. **Loyalty:** Long history as tribal creatures forming coalitions
 - "One for all and all for one" mentality
 - Virtues: patriotism, self-sacrifice for group
4. **Authority:** Long primate history of hierarchical social interactions
 - Virtues: leadership, followership, deference to authority, respect for traditions
5. **Purity:** Psychology of disgust and contamination
 - Body as temple that can be desecrated
 - Virtues: self-discipline, self-improvement, naturalness, spirituality

Additional Candidate Foundations:

- **Liberty:** Reactance against domination, hatred of bullies
- **Honor:** Self-worth based on reputation and others' assessment
- **Ownership:** Fast intuitions about property, evolutionarily stable

Key Insights:

- Morality has shared themes across cultures despite differences
- Evolution doesn't care about "goodness" of psychological systems
- Systems evolved to maintain cooperation and group survival
- Conflicts arise from different weightings of foundations

Source: <https://moralfoundations.org/>

Analysis and Novel Perspective Development

Key Themes from Research:

1. **Evolutionary Paradox:** Humans evolved both cooperative and aggressive tendencies
2. **Moral Foundations:** Multiple innate psychological systems underlying morality
3. **Contextual Activation:** Evil/cruelty emerges from specific psychological and social contexts
4. **Dehumanization Debate:** Paul Bloom's challenge to traditional dehumanization theory
5. **Neurological Overlap:** Empathy and violence circuits overlap in the brain

NOVEL PERSPECTIVE: "The Moral Plasticity Hypothesis"

Core Argument: Human nature is neither inherently good nor evil, but rather characterized by moral plasticity - an evolved capacity for extreme behavioral flexibility that can manifest as either profound compassion or devastating cruelty depending on contextual triggers and moral foundation activation patterns.

Key Components:

1. **Adaptive Ambiguity:** Evolution selected for moral ambiguity as a survival advantage
 - Allows rapid adaptation to changing social environments
 - Enables both in-group cooperation and out-group competition
 - Explains persistence of both altruism and aggression
2. **Contextual Moral Switching:** Humans possess evolved "moral switches" that can be triggered by:
 - Threat perception (real or imagined)
 - Group identity activation
 - Authority legitimization

- Moral foundation prioritization shifts
3. **The Empathy-Violence Paradox:** Same neural circuits underlie both empathy and violence
- Empathy can motivate protective aggression
 - Understanding others' minds enables both compassion and manipulation
 - Moral emotions can justify extreme actions
4. **Institutional Amplification:** Human institutions can amplify either prosocial or antisocial tendencies
- Genocides require institutional support, not just individual evil
 - Same psychological mechanisms can create saints or monsters
 - Cultural narratives shape moral foundation priorities

Implications:

- Evil is not a bug in human nature but a feature of our adaptive flexibility
- Prevention requires understanding contextual triggers, not eliminating "evil" people
- Moral education must account for our capacity for both good and evil
- Institutional design is crucial for channeling human nature toward prosocial outcomes