2 Mapping the Mind
Moral Psychology and International Humanitarian Law

In the early hours of October 3, 2015, American forces in Afghanistan accidentally bombed a trauma center run by the humanitarian organization, Doctors Without Borders (MSF), in Kunduz. Under IHL, hospitals are considered to be protected spaces, and so intentional attacks against them are war crimes. In the ensuing investigation of the Kunduz incident, US Central Command found that the bombing was the consequence of "a combination of human errors, compounded by process and equipment failures."[1] According to their report, on the night of October 2, 2015, Afghan forces planned to attack a Taliban-controlled site in Kunduz, and in the process, they requested air support from US Special Forces. When US forces "arrived…in the early morning on Oct. 3, 2015, they attempted to locate the Taliban-controlled target site. The Afghan forces provided the correct grid coordinates for the target site to the US Special Forces commander on the ground, who…relayed them to the aircrew through a Joint Terminal Attack Controller."[2] The aircrew "was initially unable to locate the target structure," but when they entered the "grid coordinates…the system directed" them to bomb "an open field." The crew then tried to "visually identify the target," deciding to strike what they thought was a Taliban structure, "but was actually the MSF" center. Forty-two people were killed.[3]

In its official report on the incident, the Pentagon decided that even though sixteen US military personnel engaged in conduct that warranted "appropriate administrative or disciplinary action," they did not commit any war crimes: "The Commander of U.S. Forces-Afghanistan concluded that certain personnel failed to comply with the law of armed conflict and rules of engagement. However, he did not conclude that these failures amounted to a war crime."[4] As the summary of the report stipulates, "the label 'war crimes' is typically reserved for *intentional* acts – *intentionally* targeting civilians or *intentionally* targeting protected

---

[1] United States Department of Defense (DOD) (2016b).   [2] DOD (2016b).
[3] DOD (2016b).   [4] DOD (2016b).

objects."[5] The bombing was deemed an accident because even though the crew intended to hit the target in question, it was working under the false impression that the hospital was in fact a lawful military target. As expected, the Pentagon report generated a firestorm of criticism – much of it from MSF. The critics charged that the bombing was, in fact, a war crime and that the US service personnel who conducted the strikes were let off with a slap on the wrists. Meinie Nicolai, the MSF President, claimed that "the threshold that must be crossed for this deadly incident to amount to a grave breach" of the laws of war "is not whether it was intentional or not."[6]

The MSF incident reveals several facts about IHL and its implications for battlefield conduct. Not only does it call attention to the principal dividing line between lawful attacks and war crimes – intentionally targeting civilians and civilian objects is one defining feature of a war crime – but in addition, it illustrates precisely what constitutes an intentional killing in the first place. According to the Pentagon report, an intentional attack on a civilian is one in which the perpetrator harms the person in question *and* knows that the person is in fact a civilian.[7] Moreover, the incident reveals that there is disagreement over what constitutes a grave breach of IHL. For MSF, the failure to take the proper precautions to spare civilian lives constitutes a grave breach of IHL, but for the US military, only an intentional attack on civilians is sufficiently weighty to be considered as a grave breach of IHL.[8] Yet, for civilians on the ground – and their loved ones who are condemned to suffer the aftereffects – these distinctions between intended and unintended, foreseen and unforeseen, may seem immaterial, for what matters to them are the effects of the military operation in question. As military lawyer W. Hays Parks aptly notes, "the distinction between intentional and unintentional injury or death is lost on the civilian who suffers that injury."[9] Why, then, does IHL make these distinctions, and what implications does this have for our ability to protect civilians through the force of law?

In this chapter, I argue that these kinds of judgments about what is just or unjust, legitimate or illegitimate, and lawful or unlawful, are driven largely by cognitive–emotional reactions rooted in the evolved structure of the human mind. More specifically, I defend two key points. First, I argue that evolved moral sentiments and cognitive heuristics have

---

[5] DOD (2016b), emphasis in original.   [6] Quoted in Doctor's Without Borders (2016).
[7] Still, Article 50 of Additional Protocol (I) holds that in cases of doubt, the person "shall be considered to be a civilian."
[8] See Article 147 of the Fourth Geneva Convention for a definition of "grave breaches."
[9] Quoted in Neer (2013, 190).

shaped not only modern IHL, but also the development of civilian protection norms in human societies more generally – both Western and Eastern. I claim that a more *naturalistic* account of moral cognition and emotion will enable IR theorists to better explain the cultural evolution of the laws of war across human societies and civilizations. I argue that cognitive and affective orientations bias the evolution of the rules of war toward instituting principles that are similar in kind to the modern principles of distinction and proportionality in IHL. Moral and legal principles that safeguard civilians are likely to emerge and remain durable because they are *affect-backed*, by which I mean that they are backed up by powerful emotional reactions.[10] Second, I argue that even though the laws of armed conflict do have *restrictive effects* that safeguard civilians from certain injustices, they also have *permissive effects* that enable states to kill civilians on a fairly large scale, albeit incidentally and often without intent. I argue that these permissive effects are rooted in the structure of moral beliefs and emotions, not just in the interests of powerful states, as realists and critical legal theorists argue. When states negotiate the laws of war in international diplomatic conferences, or when they develop strategical and tactical doctrines designed to ensure that their military practices are compliant with IHL, they do in many cases try to sincerely make sure that civilians are granted certain protections. However, the moral arguments and principles that they put forth create space for decision-making discretion.

Before I move on, three key points of clarification are in order. First, in saying that civilian protection norms are "durable," I mean to say that they are relatively long-lasting once they emerge, not that they are always effective in practice. Indeed, as recent research has shown, the effectiveness of civilian protection norms varies quite dramatically by case and over time.[11] That said, if they are to be thought of as norms at all, they should have some reasonable effect on behavior, even if it is less than full compliance. Second, in saying that civilian protection norms are "affect-backed," I do not mean to imply that affect by itself is the only factor that explains why certain norms are durable. To be sure, the most durable norms are those that benefit from a combination of affective, cognitive, strategic, and cultural considerations. My claim is simply that affective reactions tend to condition the initial emergence of civilian protection norms and they sustain the reasoning and argumentative efforts to

[10] Nichols (2002).
[11] Morrow (2007, 2014); Morrow and Jo (2006); Carpenter (2006); Bellamy (2012); Downes (2008); Valentino, Huth, and Croco (2006); Fazal and Greene (2015).

improve IHL after major atrocities – e.g., World War II and the Holocaust played a major role in International Committee of the Red Cross (ICRC) attempts to update and strengthen the Geneva Conventions. Third, although emotional responses often provide the *impetus* for efforts to create positive norms that protect innocent people in war,[12] my theory does not discount the importance of reason, deliberation, and argument. Indeed, as I clarify later, emotions, reason, and argument play an *interactive* role in the evolution of moral and legal norms for protecting noncombatants. Not only do reasoning and debate shape how people apply cognitive–emotional heuristics to novel situations that outstrip evolved programming, but so too they can get people to reflect on the negative implications of their evolved cognitive–emotional architecture. For example, the impulse toward revenge may be important for sustaining reciprocity, but, left unchecked, it can lead to extreme cruelty and violence, not to mention genocide. Although we cannot expunge our evolved cognitive and emotional architecture, we can use our capacities for reason and argument to create social institutions that control their worst excesses and redirect them toward rationally justifiable ends. Reason may be *informed* by the passions, but it is not their slave.

This chapter is organized as follows. In the following section, I briefly explain why existing theories of norm emergence and evolution in IR theory are lacking and why they must be integrated with a more naturalistic theory of moral beliefs, emotions, and arguments. Second, I argue that the laws of war, especially the principles of distinction and proportionality, are rooted in the cognitive and emotive wiring of the human mind. In particular, I argue that evolutionary pressures furnished the mind with capacities for perspective-taking and empathy, and that these capacities affected the development of the laws of war by encouraging people to accept limits on the treatment of foreign civilians. I then show that the core principles of IHL are rooted in emotional intuitions that govern how people judge the infliction of intentional harm. Intuitive precepts like the *intention/side-effect* distinction may have varying cultural expressions, but these precepts are largely shared rather than culture-specific schemas. Finally, I explain how this theory is consistent with cultural variation and normative change by discussing the relationship between individual moral intuitions and the thick social processes through which the law is created.

---

[12] For a similar argument applied to human rights norms, see Hunt (2007).

Norm Emergence and Cultural Evolution
in International Theory

In IR theory, there are three traditional approaches to explaining the design of international law. Realists and rationalists argue that states design international norms to advance their interests. While realists argue that powerful states create international laws to secure their political interests, rational choice theorists argue that they create institutions to overcome collective action problems and to enforce mutually beneficial agreements.[13] Finally, constructivists believe that international law is a social creation that reflects the shared values of international society.[14]

As noted in Chapter 1, there are three main limitations with existing scholarship on international norms. First, existing IR scholarship does not fully explain why states and nonstate actors create civilian protection rules in the first place, nor does it explain why civilizations that are otherwise very different nevertheless all converged on moral norms for protecting civilians in war. Second, existing IR scholarship does not account for the *content* of the laws of war, specifically the focus on intended/unintended harms. Finally, existing IR scholarship does not fully account for the *permissive effects* of the laws of war.

Realism: Power Politics and the Laws of Armed Conflict

Realists believe that the distribution of power influences the content of international norms. On their view, powerful states design international norms and institutions to maintain international order on their own terms. International law might "work," but it is designed to work in accord with the interests of powerful states. In international negotiations, states use Machiavellian stratagems to persuade other states to endorse specific policies, but in the end, such stratagems are intended to mask their goals and interests, which usually amount to maintaining or improving their military or economic power. Critical legal theorists generally accept this view. As Chris Jochnick and Roger Normand argue, states deliberately created permissive laws of war that privilege military necessity, and they did this to legitimate their military practices. For critical legal theorists, law is a technique of power in that it stifles potential criticisms: "law functions ideologically to both reinforce 'shared

---

[13] Mearsheimer (1994/1995); Krasner (1999); Keohane (1984); Koremenos, Lipson, and Snidal (2001); Morrow (2014).

[14] Wendt (1999); Brunnee and Toope (2001); Finnemore and Toope (2001); Finnemore (2003); Reus-Smit (2004); Tannenwald (2007); Hurd (1999).

values' and to impress upon people a sense of obligation to the existing order. More than simply supporting or deterring a particular act, law influences the public perception of an act by imbuing it with the psychic trappings of lawfulness."[15] Indeed, for critical theory, law and morality are both crucial forms of power: power works its effects not just through material coercion, but by shaping how people think – an idea that classical realists such as E. H. Carr generally endorsed.[16]

With respect to IHL, realism and critical theory provide some important insights. More so than any other states, powerful states possess the material, institutional, and ideological power to manipulate the law in ways that suit their interests. Yet, realism and critical theory have three key shortcomings. First, they do not explain why civilian protection norms have arisen in civilizations that are materially and culturally different. Power politics clearly shapes the evolution of the norms of war, but in the historical chapters that follow, I show that moral beliefs and emotions play a more important role than realism and critical theory recognize. Second, these theories do not explain *how* powerful states are able to persuade others to adopt particular norms. On the theory I develop here, some norms are attractive because they resonate with broadly accepted moral beliefs and emotions. Further, the archival evidence that I review in Chapters 6 and 7 shows that diplomats were often sincere in wanting to protect civilians from *intentional* killing, even though this meant creating permissive norms that allowed for unintentional fatalities. Finally, realism and critical theory underemphasize the restrictive effects of the laws of war. Even though the permissive effects of IHL are problematic, its restrictive effects do afford civilian populations significant legal protections.

Liberal Institutionalism: The Rational Design of the Laws of Armed Conflict

Rationalists argue that states design international institutions to overcome coordination and cooperation problems. With respect to IHL, rationalists argue that the laws of war are equilibrium institutions that create stable expectations about how wars should be fought. James Morrow argues that states create laws of war to restrain the untoward consequences of war. Furthermore, he argues that rational choice institutionalism best explains when states comply with or violate the law: states comply with the laws of war to stave off reciprocal enforcement

---

[15] Jochnick and Normand (1994, 57).   [16] Carr (1939); Babik (2013).

from their treaty partners. In short, the laws of war reflect "legal bright lines" that states create to avoid the inconveniences of anarchy, and states comply with the law, when they do, in the expectation that other states will reciprocate.[17]

Rationalism gives us a parsimonious way to explain why states develop international norms to limit the violence of war. However, it has three key shortcomings. First, it does not explain why human societies converge on similar rules of war. I do not disagree that the laws of war help resolve collective action problems, but the evidence shows that they do not always arise *for this particular reason*. Rationalism holds that states design the laws of war to secure their political interests – e.g., to protect their own citizens and soldiers. But in many cases, states have to be *persuaded* to endorse humanitarian laws of war, often by morally committed non-state actors.[18] For example, in Chapter 5, I examine how a medieval social movement known as the Peace of God led to the development of the principle of distinction in Western Europe. Second, rationalism does not explain the *content* of IHL. As I pointed out earlier, IHL reflects the *intention/side-effect* distinction, which holds that deliberate killings are worse than unintentional killings. My claim is that the content and the design structure of the laws of war track *emotional reactions*, not just rationalistic calculations.

In the chapters that follow, I provide two kinds of empirical evidence to back up this claim. First, some rules of IHL contradict the rationalist view that the laws of war are designed to facilitate reciprocity. As I discuss in Chapter 7, IHL does not permit reprisal killings when it comes to civilians, nor does it permit states to kill prisoners of war (POWs) or to torture them when their own soldiers are tortured. This suggests that the framers of IHL were concerned about civilian casualties and POW rights *as such*, not just protecting their own citizens from harm. Of course, this concern was not equally shared by all, and indeed the ICRC played the most significant role in outlawing reprisal as a tool of enforcement. Nevertheless, it does show that civilian protection rules are not just based on rational responses to collective action problems: they also have a strong moral dimension that is not reducible to the calculation of expected benefits and costs. Second, I discuss textual evidence and empirical cases that strongly suggest that rules against targeting civilians are based on gut-level emotional responses. As I discuss in greater detail later, the idea that civilians should not be intentionally targeted is based on gut-level emotional reactions to the thought of physically killing

---

[17] Morrow (2007, 561). Also see Posner (2003, 2010).
[18] Finnemore (1999); Price (1998); Petrova (2016).

someone in an *up-close-and-personal* way.[19] International norms concerning the use of military force closely track this idea. For example, IHL clearly prohibits the kind of personal killing campaigns that took place during the My Lai massacre in Vietnam, and it also prohibits the torture of POWs. Yet, as I discuss in Chapter 6, the 1907 Hague Conventions did not prohibit attacks on merchant ships on the high seas except at the beginning of war, and one reason for this was that attacks on merchant ships – which usually just involved confiscating property – were not perceived as being nearly as objectionable as mass killings of civilians. Mass killing campaigns are highly emotionally salient, and this is why international law singles them out as particularly egregious.

The problem is that rationalists generally ignore emotions. With respect to the laws of war, they believe that POW protections, civilian immunity norms, and rules outlawing perfidy arise out of rational decisions to resolve collective action problems. Emotions play no role in explaining the content of the laws of armed conflict. My claim is not that the laws of warfare are unresponsive to collective action problems, but rather it is that in order to explain the content of the laws of warfare, emotive response patterns need to be brought into consideration. Of course, I should point out that my theory of moral cognition and emotions is not completely at odds with rationalism. To be sure, the view that moral and social norms are based on evolved emotional responses actually provides a more compelling explanation of phenomena like diffuse reciprocity than does the view that they are based on rational cost–benefit calculations. Specific reciprocity implies that people will comply with social norms out of a concern for some direct and immediate benefit, but diffuse reciprocity implies that people comply with social norms even when there is no immediate benefit, but rather there is a long-term benefit associated with cooperation. Rationalists argue that in the case of diffuse reciprocity, people hope to reap the benefits of long-term cooperation. The problem with this view is that people vary in the degree to which they are willing and able to take on short-term costs to achieve long-term benefits. Indeed, people often care more about the present than they do about the future.[20] The theory of moral cognition and emotion that I put forth here helps to explain why people will often comply with norms that only bring long-term benefits. Decision making is often much more intuitive than the rationalist view of *homo economicus* would have us believe.[21] Hence, far from causing people to act

---

[19] Greene et al. (2001).   [20] Pinker (2011, 592–611).
[21] Pouliot (2010, 11–51); Hopf (2010); Bowles and Gintis (2011); Holmes and Traven (2015).

irrationally, moral intuitions can function as heuristic shortcuts that make it easier to comply with societal norms that, in the long run, bring diffuse benefits.

Finally, rationalism does not provide a compelling explanation of the permissive effects of IHL. Rational design theorists do recognize that states design laws that allow for decision-making discretion and sovereign prerogative.[22] But the permissive effects of IHL actually conflict with the idea that states rationally design international laws and institutions. Not only are the laws of armed conflict *less-than-fully-rational* in the sense that rational actors concerned to protect their own self-interests would probably create a more efficient set of norms, but the permissive effects of the laws of war can actually threaten the integrity of the regime. For example, consider the rationalist view that states create international laws to operate as legal bright lines that clearly delineate acceptable from unacceptable behavior. IHL creates a legal bright line with intended attacks, but it technically permits collateral damage. This principle makes intuitive sense, but it can be difficult to apply in practice, specifically because it can be easily misinterpreted and manipulated by enemy forces. For example, during Operation Rolling Thunder in Vietnam, US decision-makers directly prohibited intentional attacks on civilian targets, and in fact, during discussions over escalating the targets to include attacks on petroleum, oil, and lubricant facilities in 1966, Robert McNamara noted that the "final decision for or against will be influenced by extent they can be carried out without significant civilian casualties."[23] Still, inadvertent civilian deaths generated significant criticisms of US foreign policy in Vietnam.[24] But the criticism, in this case, was not just that US bombings caused *inadvertent* deaths; rather, it was that the United States *intended* to kill civilians, and thus violated the law. By causing a high number of incidental civilian casualties – both during Operation Rolling Thunder and during the Linebacker campaigns in 1972 – the United States took measures that unwittingly led to accusations that it had targeted civilians intentionally.[25]

The problem for rationalism is that the *intention/side-effect* distinction can lead to a lot of false positives, which could threaten the integrity of the regime. When unintended civilian deaths are high, the intuitive assumption on the part of those on the receiving end will likely be that enemy forces have deliberately targeted their civilian population. And even though the laws of war do not sanction reprisals, there will likely be a strong desire to exact vengeance on the enemy by harming their

---

[22] Abbott and Snidal (2000); Koremenos, Lipson, and Snidal (2001).
[23] DOD (2011, 124).   [24] DOD (2011, 177).   [25] Bellamy (2012, 173–182).

civilians, which can potentially unravel the civilian protection regime. Morrow says that this is a problem of noise,[26] but presumably rational states would have been able to predict and address this problem, perhaps by instituting casualty thresholds or by reducing the role that intentions play in the norms of war.[27] As a matter of fact, this would be quite reasonable, given that decisions to comply with or violate the laws of war are usually made behind closed doors where intentions are kept secret. When individual soldiers kill civilians, the military justice system is designed to signal to other states that the civilian deaths in question were not a result of an explicit policy. However, when large numbers of civilians are killed in airstrikes, and there is no clear evidence one way or the other as to whether the state intended to target them, enemy officials may draw dire conclusions, even at the risk of making a mistake. In short, even though states might have good reasons to make a distinction between intended and unintended killings (e.g., intentional killings signal that a state cannot be trusted to uphold the law), rational states would have created a far more consequentialist framework of IHL than states in fact created. Of course, this is not to say that IHL is irrational, but rather it is to say that a more plausible way to interpret IHL is that it is an outgrowth of cognitive heuristics and emotions, not just rational cost–benefit calculations.

Constructivism: Cultural Evolution and the Social Construction of the Laws of Armed Conflict

Constructivism holds that individual agents shape cultural norms, and that cultural norms, in turn, shape the identities and interests of individual agents. Agents and structures are mutually constituted.[28] As Martha Finnemore and Kathryn Sikkink argue in their classic article on international norms and political change, cultural norms emerge out of the activities of individual norm entrepreneurs who deliberately seek to change domestic and international institutions.[29] On the other hand, cultural norms work their effects on social practices by influencing how people think. As Ian Hurd argues, states comply with international norms because they see them as legitimate, and they come to see them as legitimate through a process of internalization.[30] For most

---

[26] Morrow (2007, 2014).
[27] The notion of crimes against humanity does institute a kind of casualty threshold: mass killings have to be part of a systematic policy to be regarded as crimes against humanity. However, there are no casualty thresholds when it comes to the principles of distinction and proportionality. See Dill (2015, 87).
[28] Wendt (1999).   [29] Finnemore and Sikkink (1998).   [30] Hurd (1999).

constructivists, the mind is like a blank slate that "imports" cultural ideas from the external social environment and uses them as mental models for how one should behave. Human societies may have some norms in common, but constructivists argue that we do not need to root theories of cultural evolution in assumptions about evolved cognition and sentiments. Similarities in cultural norms do not imply that there are similarities in the cognitive micro-processes upon which cultures supervene. Cultural norms and ideas are shared mental representations that can arise and spread in a wide variety of ways – e.g., through material constraints,[31] adaptive success in a competitive environment,[32] and imitation and social learning.[33] In addition, recent research in cultural anthropology and cognitive neuroscience suggests that the brain evolved for culture,[34] so it is not clear why IR scholars need to go down to the cognitive, or neural, level to explain similarities in cultural norms. Constructivist scholarship in IR has produced powerful insights into the dynamics of norm construction and diffusion, and this work gets by with fairly minimal assumptions about cognitive psychology. Why should this work turn to evolved cognition and emotion? There are three reasons.

First, all social science theories make assumptions about agency, and these assumptions should be consistent with the facts of social psychology and cognitive neuroscience, at least the well-tested and replicated facts. Recent findings in cognitive psychology and neuroscience show that the mind uses intuitive heuristics to cast judgment on a wide range of issues, from economic decision-making, to moral judgments and the law, to decisions to go to war.[35] Intuitive heuristics are fast, automatic, and inaccessible to consciousness, and as a result, they bias individual decision-making in ways that people do not directly control. If individual actors have some control over the construction of international law and culture, as many constructivists think they do, then we should expect intuitive moral judgments to seep into the content of international law. Therefore, building a theory of international law from a more naturalistic foundation will place IR on a firmer scientific basis. This does not mean that all aspects of international culture and law can be reduced to neuroscience, and indeed findings from neuroscience indicate that a mind-culture interactionist approach is the most defensible perspective

---

[31] Snyder (2002).   [32] Florini (1996).
[33] Wendt (1999); Richerson and Boyd (2005).   [34] Downey and Lende (2012).
[35] Holmes and Traven (2015); Holmes (2015); Kahneman (2011); Greene (2013); Sunstein (2009); Rosen (2005).

on mental content and morality. There are some basic principles that all cultures share – among them principles for regulating intentional harms – but there are also some significant cultural differences. I am not a strict moral universalist, if by this one means that all moral principles are universally shared.[36]

Second, as Sikkink suggests in *The Justice Cascade*, theories of norm emergence usually make assumptions about the motivations of social agents – especially norm entrepreneurs. Cultural ideas are macro-level representations, but they usually arise through the dedicated action of motivated individuals. To inject new ideas into the public sphere, individual norm entrepreneurs need to make them sink in, and this requires rhetorical action that appeals to audience emotions and moral beliefs. Finnemore argues that empathy is an important tool in the argumentative repertoire of norm entrepreneurs, and Neta Crawford shows how political activists used empathy to increase respect for the humanity of slaves and colonial subjects.[37] Empathy needs to be engaged to motivate people *to want to alter* the normative landscape of the international system. As a result, we need a theory of agency that builds on assumptions about the fundamental cognitive–emotional repertoire of individuals. But constructivists adopt an ambiguous model of agency. Cultural diffusion theorists hold that agent intentions and preferences are socially constructed all the way down. However, their explanations of the *process* of norm emergence often build upon assumptions about basic human emotions and beliefs. As an example, consider Sikkink's discussion of the norm of individual criminal accountability for human rights abuses in *The Justice Cascade*. In this book, Sikkink defends a theory that she refers to as *agentic constructivism*, which focuses on how agents promote new ideas and practices *even though* their interests are socially constructed. Yet, throughout the book, she is at pains to explain precisely where the concern for justice comes from, and in the end, she intimates that it may have something to do with evolved moral sentiments and beliefs. As she puts it, "core human rights norms [resonate] so profoundly in the world in part because of [an] intrinsic appeal."[38]

My proposal is that we cannot have it both ways. If we wish to develop an agent-centric theory of cultural evolution, one that is consistent with the fact that individual agents play a major role in the development of cultural norms, we should not assume that agent preferences are socially constructed all the way down *while also* thinking that the process of norm emergence is dependent on pan-cultural moral capacities such as

---

[36] For a useful discussion of different types of universalism, see Brown (1991).
[37] Finnemore (2003); Crawford (2002).   [38] Sikkink (2011, 262).

empathy and the desire for justice. My perspective is that we should start from fundamental moral capacities and intuitive heuristics, and then move our way up to macro-level cultural representations and institutions. Cultural institutions do influence how we think, but my view is that cultural institutions build upon shared cognitive rules and emotions.[39] I do not deny that much of international relations is socially constructed, but in my view, a deeper appreciation of the moral aspects of human nature can advance our knowledge of world politics. To some extent, constructivists do recognize the need to build theories of international norms on assumptions about human psychology. Not only do Finnemore and Sikkink think that empathy is important for explaining the evolution of international norms, but in addition, Alexander Wendt suggests that theories of agency should be based on material factors such as basic human drives.[40] However, Wendt argues that when it comes to things like moral ideas and culture, "biology matters relatively little."[41] My suggestion is not that we should reject the constructivist view of agency. Rather, it is that constructivists should adopt a *thicker* account of human nature – one that recognizes that some moral ideas and sentiments are substantially innate, or programmed prior to experience.[42]

In response, many constructivists might concede that the mind and the brain partly influence social interaction. This is part of the *rump materialism* that individuals bring to social interaction.[43] However, the hardware of the brain and nervous system is only an *input* into social interaction. It is the purely social aspects of human interaction – discussion, argumentation, diplomatic practices, and so forth – that do all the important explanatory work in shaping the process of norm emergence. In cases where states largely agree about the rules of war, then the kind of cognitive and emotional capacities that I explore in the rest of this chapter can help explain cross-cultural similarities. But what about cases where people disagree? When people disagree about the rules of war, they either abide by different standards, or they get together to work out their disagreements. In the latter case, the operative causal mechanism is diplomatic talk, or what Jürgen Habermas refers to as the power of the better argument.[44] Moral intuitions are an input of sorts, but they don't explain norm creation.

---

[39] Boyer and Petersen (2012).   [40] Wendt (1999, 130–135).
[41] Wendt (1999, 133); also see Crawford (2009).   [42] Haidt and Joseph (2007).
[43] Wendt (1999, 130). I am indebted to one of the anonymous reviewers for raising this criticism.
[44] Habermas (1990); Risse (2000); Mitzen (2005); Linklater (2005, 2007); Johnstone (2003).

Although a fuller response to this argument can only be had once my theory is on the table, a few brief comments are in order. First, I do not discount the causal significance of communicative reasoning. Not only is communicative reasoning a key mechanism whereby shared moral intuitions become institutionalized into positive international law, but it also helps people work out political and moral disagreements. With respect to moral disagreement, communicative reasoning performs three key functions in the theory. First, it provides the means through which people apply cognitive–emotional heuristics to novel situations that outstrip their evolved programming. Human beings did not evolve to think about the ethics of strategic bombing, nuclear weapons use, or cluster munitions, so when they reason through these issues, they have to find a way to make their evolved intuitions applicable to a novel domain. Communicative reasoning aids this process. Second, communicative reasoning – particularly the kind in which lawyers routinely engage – enables people to make vague intuitions more precise, and therefore more applicable to practical situations. Conscious reasoning enabled the participants of the 1899 and 1907 Hague Peace Conferences to converge on a definition of *lawful military targets* that effectively made the *intention/side-effect* distinction more applicable to the domain of naval warfare. Finally, although communicative reasoning can sometimes override our moral intuitions, much of what goes on in diplomacy is an attempt to frame arguments in ways that others will accept, which ordinarily requires giving reasons that pump different intuitions. This suggests that adopting a naturalistic view of moral cognition and emotion will enable constructivists to better explain what goes on in the thick social process of international diplomacy.

Third, building on the theory of evolved moral cognition and emotions will not only enable IR scholars to better explain patterns of norm emergence and cross-cultural similarities, but it will also heighten the explanatory power of existing theories of institutional development and design. In the chapters that follow, I show how building on the naturalistic theory of moral cognition and emotions that I elucidate in the following section can improve constructivist and rationalist theories of norms. More specifically, I argue that it will help IR theorists better account for the cross-cultural *timing of emergence* and the *durability* of civilian protection rules. Not only does the constructivist view that *all* moral beliefs are rooted in culture make it difficult to explain how different societies can converge on similar norms of war, but it also makes it harder to understand how norms emerge, spread, and remain durable *within* particular civilizations. Perhaps one might argue that contextual factors – whether material or ideational – make it more likely

that some societies will converge on civilian protection norms. However, I argue that incorporating a more naturalistic account of moral cognition and emotion will enable IR scholars to better understand *how* sociological and contextual variables generate convergence on civilian protection rules: they create support for stronger norms by working through the mind, thereby engaging shared moral sentiments. For example, innovations in technology such as photography made it possible for anti-war activists during the Vietnam War to build broad support for condemning the use of napalm by the US military.[45] I conjecture that *affect-backed* norms are more likely to arise and persist than *affect-neutral* norms. Civilian protection rules are prevalent because they *fit* with cognitive–emotional heuristics that govern how people think about the ethics of killing in war. As such, while it may be true that cultural similarities do not in and of themselves imply that political scientists need to "go cognitive" or "go neural," going cognitive and neural can improve theories of cultural evolution in ways that go beyond explaining transcultural similarities.

Theory of Mind: A Naturalistic Theory of Moral Cognition and Emotion

Human cultures are striking in their variety. Anyone who has stepped foot in a foreign culture or tried to speak with someone who speaks another language is all-too-familiar with this self-evident fact. And yet, human beings are not all that different from one another. They all learn how to speak languages,[46] they all care for their offspring (most of the time), and most of them believe that it is wrong to lie, cheat, steal, and kill in most circumstances.[47] The social and political systems within which people live are highly diverse, but the elementary moral and social rules that people use to get by are strikingly similar: do not free ride on the hard work of others, do not harm innocent people for no reason, shun or punish those who cheat on social agreements, share common resources with friends and family, etc. How is it possible to account for the immense richness and diversity of human cultures while also explaining the emergence of common norms? Since human behavior is motivated by psychological mechanisms, what must the mind be like such that human cultures have these immense differences and similarities?

There are two broad responses to these questions. One view posits that individuals are born with general learning mechanisms that enable them

---

[45] Neer (2013).   [46] Chomsky (1965, 1986, 2007, 2009); Jackendoff (2007).
[47] Brown (1991).

to pick up on relevant features in their material and social environment. With respect to moral rules, these mechanisms enable people to internalize the kinds of norms prevalent in their society. People internalize rules that prohibit inflicting harms on innocent persons through variable levels of punitive enforcement on the part of authority figures (e.g., parents and teachers), by imitating the behavior of others, and so on. When it comes to moral and social judgments, the mind is, for all intents and purposes, a blank slate. There is no innate moral faculty, and there are no innate cognitive biases or deontic rules. Some rules may be similar across cultures, but this can be explained by reference to the fact that most societies confront similar challenges and that people have general learning mechanisms for internalizing moral norms.[48] Moreover, emerging work on neuroplasticity suggests that neural networks are sensitive to cultural variations and that the social brain is overwhelmingly flexible. In this view, moral norms are mere social conventions, not expressions of broadly universal principles.[49]

Another view, which is the one that I adopt here, holds that moral intuitions are based on domain-specific learning mechanisms that constrain how humans internalize social norms. On this view, the mind is composed of a set of computational processing systems called "modules," each of which is devoted to specific tasks and capacities. Evolutionary psychologists believe that these domain-specific mechanisms evolved in response to recurrent situations in the ancestral environment of the species.[50] This is true for a wide variety of psychological mechanisms, including those that facilitate fight-or-flight reactions and those that facilitate coalitional affiliations.[51] The modular approach to mental functioning explains how certain capacities, e.g., the capacity to detect cheaters or the ability to understand the intentions of others[52] are programmed in the mind. Jonathan Haidt and Craig Joseph describe cognitive modules as "evolved…processing system[s] that [were] designed to handle problems or opportunities that presented themselves…in the ancestral environment of [the] species. Modules are little bits of input/output programming, ways of enabling fast and automatic responses to specific environmental triggers."[53] Modules are like equalizers on a stereo system that provide a "first draft" of the mind, which is then calibrated through cultural learning and experience.[54]

---

[48] Prinz (2007).   [49] Schwartz and Begley (2002); Crawford (2009, 277).
[50] Cosmides and Tooby (1994, 86).
[51] Haidt (2001); Lopez, McDermott, and Peterson (2011); Joyce (2006); Tooby and Cosmides (2010).
[52] Holmes (2013, 2018); Hall and Yarhi-Milo (2012).
[53] Haidt and Joseph (2004, 59–60).
[54] Haidt and Joseph (2007, 368); Graham and Haidt (2012); Haidt (2012).

Psychologists and neuroscientists disagree over the extent to which the mind and the brain are modular,[55] and even whether it is modular at all.[56] Among those who defend some version of the modularity thesis, there is disagreement over whether moral cognition is manifested in a single computational module,[57] in several modules,[58] or whether it is instead manifested in a combination of automated mechanisms and higher-cognitive thought processes.[59] Moreover, there is significant disagreement over how to incorporate each element of moral cognition, e.g., empathy, perspective-taking, moral emotions, conscience, condemnation, and culture.[60] One important debate has to do with the question of whether moral judgments are grounded in rationality and reasoning processes, or whether they are grounded in emotional reactions.[61] Here, I split the difference in these debates by claiming that there are three interrelated elements of moral cognition and emotion. First, people have innate abilities for perspective-taking and empathy. Perspective-taking and empathy are both crucial for moral and social development: individuals who lack the ability to empathize with or to see what the world is like from someone else's perspective will have difficulty understanding what the world is like from a moral point of view. Furthermore, perspective-taking and empathic concern play a crucial role in setting the dials of our moral equalizers (or fixing the parameters of our moral grammars) to recognize the moral interests of those outside our immediate circle of moral concern. Perspective-taking and empathy also induce people to act on behalf of others. This is the so-called empathy–altruism hypothesis: perspective-taking causes empathy, which in turn causes altruism.[62]

Second, while I am agnostic on the issue of whether there is a single computational module or a complex set of them,[63] a growing body of evidence in cognitive psychology and neuroscience suggests that the brain is constituted by a set of automated mechanisms and deliberative reasoning processes and that these processes compete for control over moral judgments. This is the so-called dual-process theory of the mind.

---

[55] Fodor (1983); Chomsky (2007); Cosmides and Tooby (1994); Sperber (1996); Buller and Hardcastle (2000).
[56] Prinz (2007).   [57] Mikhail (2011).
[58] Fiske (1991); Rai and Fiske (2011); Haidt (2001, 2012); Haidt and Joseph (2007); Shweder et al. (1997).
[59] Greene et al. (2001); Greene (2013); Kahneman (2011); Patterson, Rothstein, and Barbey (2012).
[60] Haidt (2001); DeScioli and Kurzban (2009, 2013).
[61] Kant (1998); Kohlberg (1981); Habermas (1990); Mikhail (2011); Haidt (2001); Jeffery (2016).
[62] Batson (2011).
[63] Mikhail (2011); Haidt (2012); Fiske (1991); Rai and Fiske (2011).

People often have quick intuitive reactions to emotional situations (e.g., the My Lai massacre or 9/11), and these reactions generate implicit moral judgments. However, people also have the ability to reflect on moral dilemmas. The automated mechanisms that govern our moral judgments are often highly emotional, but we can usefully think of them as operating in accordance with something like a grammar of moral principles. Here, I am suggesting that morality is *like* language in the sense that moral judgments are shaped by complex computational processes that operate below conscious thought and that we can construct rules, or grammars, to characterize how the process works. Although some theorists advocate a stronger interpretation of the linguistic analogy according to which moral rules are "engraved in the mind"[64] in ways that precede emotional reaction patterns,[65] here I argue that emotional responses play a more direct role in moral judgment. The analogy with language is used to help characterize the operations of the automated mechanism. It also provides a useful template for explaining how moral cognition is innate: just as children are "born knowing" the "constraints" on "human languages,"[66] so too they are born knowing the constraints on human moral systems. Finally, in line with the dual-process approach to moral cognition, I argue that reason and emotion are both causally implicated in moral judgment, motivation, and behavior. Deliberative reasoning shapes how people respond to novel situations, and it also has the ability to recalibrate our moral equalizers through internal reflection and communicative interaction.

To defend this theory of moral psychology, in the next subsection I define moral cognition and emotion, and I explain why moral intuitions conferred an adaptive advantage on individuals in the ancestral environment. In the second subsection, I explain how perspective-taking and empathy shape moral judgment, thinking, and behavior. Furthermore, I explain how they can induce people to condemn military operations that kill civilians. Next, I explore the grammar of moral cognition. I show that moral judgments are governed by subtle asymmetries in how people evaluate intentions: intended harms are perceived as morally worse than unintended side-effect harms. To explain how cultural differences are consistent with cultural similarities, I briefly explore how moral judgments are similar in kind to linguistic judgments. Although there are many differences between ethics and language, the idea that moral cognition is shaped, at least in part, by a shared grammar of principles helps square the knot between moral universalism and cultural

---

[64] Mikhail (2011, 101).   [65] Huebner, Dwyer, and Hauser (2009); Carchidi (2020).
[66] Fodor (1983, 4).

relativism. I argue that this account of moral psychology has three observable implications for the evolution of civilian protection rules. First, I argue that emotional dispositions bias the evolution of the laws of war. Specifically, I argue that perspective-taking and empathy help calibrate our moral intuitions so that they apply to a wider range of individuals. By contrast, moral arguments for mass killings, genocide, or ethnic cleansing often attempt to draw distinctions between "us" and "them." That is, they tend to involve dehumanized perception.[67] Second, I argue that civilian protection rules are relatively durable once they arise because they are *affect-backed*. Third, I argue that in face-to-face diplomacy, states use emotion-laden moral arguments to design the restrictive and permissive restraints of IHL.

Programming the Mind: Moral Judgments, Evolution, and Innateness

There is a burgeoning consensus in the study of moral psychology that evolution played an important role in the development of moral systems, though precisely how and why is a matter for debate. One prominent theory is that morality facilitated cooperation in the ancestral environment, and it did this through cognitive and emotional propensities focused on altruism.[68] Genetic predispositions that cause altruistic behavior conferred an adaptive advantage on individuals with those genes. Theorists who favor kin selection believe that people are altruistic because altruism favors the spread of altruism-causing genes to close relatives. Other scholars think that reciprocal altruism or group/multi-level selection mechanisms are better for explaining altruistic behavior toward non-kin. Reciprocal altruism involves taking on costs to oneself to benefit distantly related others under the expectation that these costs will be reciprocated.[69] Group/multi-level selection models hold that people evolved to be altruistic because altruism favored the groups in which they evolved.[70] Other scholars argue that altruism-centric theories ignore crucial facts about morality, such as the tendency for people to moralize issues that have very little to do with altruism. For example, some Muslims believe that it is morally wrong to draw pictures of Allah and the Prophet Muhammad, and Western just war theorists such as Hugo Grotius believed that it is wrong to destroy sacred objects in war.[71]

---

[67] Harris and Fiske (2011).
[68] Trivers (1971); Joyce (2006); Monroe (1996); Bowles and Gintis (2011); Greene (2013).
[69] Trivers (1971, 35).   [70] Bowles and Gintis (2011); Haidt (2012).
[71] Grotius (1925, 751–754).

Likewise, secular liberals believe that racism and sexism are not just morally wrong, but repulsive. In short, moral judgments often reflect concerns over purity and impurity, and these have very little to do with altruistic behavior. Here, I make the case that the moral capacities and intuitions that are relevant for the laws of war are strongly related to evolutionary concerns for altruism, but that moral judgments facilitate social cohesion in ways that go beyond motivating other-regarding behavior.

Moral cognition has four characteristics that are relevant for explaining why it might have evolved. First, moral judgments are *categorical* in the sense that they implicate beliefs about what must be done. Some moral philosophers believe that moral principles are hypothetical imperatives that require people to act in certain ways *depending on* whether they entertain particular objectives: e.g., "thou shalt not steal" is not a categorical imperative, but rather it depends on entertaining the goal of respecting property rights. If this view is right, moral principles are like rules of etiquette.[72] Though morality may in fact be a system of hypothetical imperatives, when people are in the grips of a moral judgment, there is a felt sense that the action must be done of necessity. Second, moral judgments are *objective* in the sense that when people make them, they believe that their judgments depict *extra-mental* features of reality, either natural reality or supernatural reality. Our moral beliefs may in the end all turn out to be false,[73] but at the time of judgment, most people take them to be descriptive, not just prescriptive. The judgment that "slavery is immoral" just seems to be right, and it seems to be right in the descriptive sense that it depicts something about the nature of reality.

Third, moral judgments are *action-orienting* in two senses. First, moral judgments usually have an intrinsic, but not always decisive, motivational force. It is analytically possible for someone to judge that an action is right without having the internal desire to comply with this judgment,[74] so moral judgments are not *necessarily* motivational.[75] However, in most cases, moral judgments, especially those that are infused with emotional content, have strong motivational power. Second, moral judgments are often designed to influence the behavior of others, often through explicit

---

[72] Foot (1972).   [73] Mackie (1977).   [74] Svavarsdottir (1999).

[75] Building on the work of Antonio Damasio (1994), philosopher Adina Roskies (2003) argues that ethical judgments are often decoupled from desires and motivations: people who suffer damage to the ventromedial cortex usually lack the desire to follow through on their moral judgments. However, this does not preclude their ability to make such judgments. People who suffer damage to these brain areas are able to make ethical judgments, but they are severely deficient in their ability to act on such judgments.

or implicit persuasion.[76] It is logically possible, but out of the ordinary, to judge that "killing civilians is wrong," yet to have no motivation to act on this judgment and to be unconcerned about whether other people agree. Finally, moral judgments derive from two dissociable, yet interrelated systems: an automatic system of moral intuitions (System 1) which generates emotive responses and action-orientations; and a slower, more deliberative system of reasoning that focuses on whether specific moral claims ought to be accepted (System 2).[77] While many scholars characterize the relationship between System 1 and System 2 as unidirectional, with System 1 being causally prior to System 2, later on in this chapter I suggest that it makes more sense to regard them as interactive. When people confront moral problems and dilemmas, their immediate responses are heavily informed by System 1's emotional intuitions. Furthermore, research on motivated reasoning shows that people implicitly use System 2 reasoning to justify or reinforce their intuitive reactions. As Haidt puts it, people are like intuitive lawyers arguing a moral case, not intuitive scientists seeking moral truth.[78] However, after spelling out the theory of moral cognition and emotion in more detail, I argue that we should regard the relationship between System 1 and System 2 as more interactive. Not only can individual reflection help people think through complex moral dilemmas where their intuitions are ambiguous, but individual and communicative reasoning can help to refine and develop their intuitive reactions.

To understand the evolutionary importance of morality, we should ask why evolution might have favored a cognitive system with features like this. In short, a cognitive system that produces judgments, motivations, and actions with these features would have been exceptionally well-suited to resolve recurrent problems in the ancestral environment of the species. One such problem has to do with the ability to create and coordinate social alliances and coalitions.[79] If we assume that the ability to maintain social alliances was necessary for the ancestral environment, early humans not only had to be able to motivate cooperative behavior, but they also had to be able to detect cheaters, to make credible commitments to punish transgressors, to persuade other people to accept social rules when disagreements arose, and to be able to evaluate the claims of others. Empathic abilities, along with System 1's intuitional heuristics,

---

[76] Stevenson (1937); Habermas (1984); Crawford (2002).
[77] Greene (2013); Kahneman (2011); Haidt (2012); Mercier and Sperber (2011).
[78] Haidt (2001, 820–822).
[79] Fiske (1991); Haidt and Joseph (2004); Tooby and Cosmides (2010); Lopez, McDermott, and Peterson (2011); Gat (2013); DeScioli and Kurzban (2009, 2013); Greene (2013).

can generate cooperative action orientations when the benefits from cooperation are neither immediate nor palpable. For example, as Rousseau once said, "compassion is a natural sentiment" that "carries us without thinking to the aid" of other people.[80] Because sympathy leads us to do what we have good evolutionary reasons to do – i.e., to altruistic-ally aid our conspecifics – a capacity like this would have been very useful in the ancestral environment. Thus, the categorical, objective, and action-orienting features of moral thinking enable individuals to manage coalitional affiliations. System 2 reasoning capacities focus on making persuasive arguments to others and on assessing their arguments in turn. They enable people to make better decisions, but they also enable them to shape how others think and behave through persuasion.

In the next several subsections, I explain in more detail how perspective-taking and empathy played an important role in the evolution of morality, and I explain how they affect judgments about the ethics of killing in war. In contrast to recent critics of empathy,[81] I argue that perspective-taking and empathic concern are important for explaining how individuals come to place greater value on the lives of other people, especially those outside their immediate circle of concern. Empathy may have evolved specifically for intragroup social life, but it can be recruited to help expand the bounds of the self. Second, I sketch out the grammar of intuitive moral rules that shape how people think about the ethics of killing in war, and I pay specific attention to what I call the *intention/side-effect* distinction. In each subsection, I outline the different moral capaci-ties, I highlight their evolutionary significance, and then I explain how they each contributed to the evolution of the norms of warfare.

Perspective-Taking, Empathic Concern, and the Ethics of Killing in War

Perspective-taking and empathy play a prominent role in moral reasoning, influencing not only how people make moral judgments, but also the degree to which they act on behalf of others, how they justify their moral claims, and how they define in-group/out-group identities. Perspective-taking and empathy are important for social life because they help people understand the mental states and experiences of others, and empathy, in particular, provides immediate motivating reasons for people to help others in need. Because social coordination and cooper-ation require that people be able to understand or predict how others will

---

[80] Rousseau (1988a, 29); Churchland (2011).   [81] Bloom (2016); Prinz (2011).

act, and because they require prosocial motivation, it is plausible to conclude that perspective-taking and empathy were important for promoting social cooperation in the ancestral environment. Indeed, primatologist Frans de Waal has suggested that "empathy evolved in animals as the main proximate mechanism for directed altruism," which is, in turn, important for survival.[82] In IR theory, several scholars have begun to explore the effects of perspective-taking and empathy – ranging from experimental studies of how perspective-taking influences militant assertiveness and internationalism,[83] to studies of how empathy affects patterns of interconnectedness and conflict.[84] Yet, despite the increased interest in emotions research in IR,[85] the effects of perspective-taking and empathy are far from clear. Part of the problem has to do with the fact that psychological variables are unobservable, and thus are difficult to measure accurately. Another problem is the so-called level-of-analysis problem: since empathy is an *individual* mental state, it is hard to see how it scales up to impact states, diplomacy, and international norms.[86] Finally, in recent years some commentators have begun to question the value of empathy, claiming that not only is it unnecessary for morality, but that it can sometimes have pernicious effects.[87]

After briefly outlining how I conceptualize the notions of perspective-taking and empathy, in this subsection, I provide a defense of empathy. In particular, I defend three points. First, I argue that, when appropriately targeted, perspective-taking and empathy can help *reduce* in-group biases that lead people to dehumanize others. Second, I suggest that a plausible extension of the empathy–altruism hypothesis is that empathy can *increase* moral condemnation of tactics that harm civilians. Third, I argue that although "empathy skeptics" such as Paul Bloom (2016) and Jesse Prinz (2011) are right in saying that empathy can sometimes lead people astray, they ignore the degree to which properly targeted forms of empathy can increase the extent to which people morally value the lives and well-being of others. Empathy can sometimes lead people to overlook impartial considerations of justice – empathizing with the *one* can lead people to overlook the interests of the *many* – but it can also encourage people to take cognizance of interests that they might otherwise totally ignore.

---

[82] De Waal (2008, 282).   [83] Kertzer and Renshon (2015).
[84] Linklater (2007, 2010); Head (2012, 2016); Holmes and Yarhi-Milo (2017).
[85] Bleiker and Hutchison (2008); Mercer (2010); Crawford (2000); Hall and Ross (2015); Jeffery (2016).
[86] For a compelling response to this problem, see Crawford (2014a) and Holmes (2018).
[87] Prinz (2011); Bloom (2016).

Perspective-Taking and Empathic Concern: A Conceptual Mapping.
Perspective-taking and empathy are strongly related, and in fact many
scholars see them as two components of an underlying psychological
construct. According to Mark Davis, "empathy in the broadest sense
refers to the reactions of one individual to the observed experiences of
another,"[88] a definition that reflects Adam Smith's idea of sympathy in
The Theory of Moral Sentiments. In this definition, empathy is similar to
perspective-taking; it is a *cognitive* capacity that involves taking on the
emotions or beliefs of other people, often in response to judgments about
what they may be thinking or feeling.[89] This definition of empathy is
fairly common in psychology and neuroscience, not to mention IR.[90] For
example, neuroscientists Tania Singer and Claus Lamm note that
empathy involves an isomorphic relationship between the thoughts or
feelings of one individual and another *as well as* a causal relationship
between those thoughts and feelings. For example, when children feel
fearful when their parents exhibit signs of fear or distress, they experience
a kind of empathy: not only do they experience the same emotion as their
parents, but their reactions are caused by their parents' reactions.
However, some scholars argue that there are important differences
between these two constructs. Whereas perspective-taking is a cognitive
or emotional *capacity*, empathy can also be understood as a *mental state* of
concern for others. For example, Jamil Zaki and Kevin Ochsner claim
that empathy often involves experience-sharing *and* prosocial concern.[91]

Given these conceptual disagreements, the most tractable way forward
is to define empathy and perspective-taking as two different constructs.
As Daniel Batson aptly points out, there is "no clear basis…for favoring
one" definition over another, and so in this situation, the best thing
to do is to "make clear the labeling scheme one is adopting, and use it
consistently."[92] To maintain definitional clarity, in this chapter I define
perspective-taking as having two basic components: (1) an ability to
understand the thoughts and feelings of other individuals and groups;
and (2) an ability to place oneself in the position of others – i.e., to take
on their thoughts, feelings, and intentions. Moreover, there are two
types of perspective-taking: egocentric and other-regarding. Egocentric
perspective-taking has to do with placing *myself* in the position of others.
When I think about what *I* would do if *I* were President of the United
States, I engage in egocentric perspective-taking. However, when I think
about what it might be like to be in the shoes of *President Joseph Biden*,
and to experience *his* thoughts and feelings, I experience other-regarding

---

[88] Davis (1983, 113).   [89] Davis (1983, 115).
[90] Singer and Lamm (2009); Singer and Leiberg (2009); Holmes and Yarhi-Milo (2017).
[91] Zaki and Ochsner (2012).   [92] Batson (2011, 20).

perspective-taking. Although perspective-taking can lead to empathy, here I follow Batson in defining empathy as *empathic concern*, which is an "other-oriented emotion elicited by and congruent with the perceived welfare of someone in need."[93] For example, if my daughter is in pain, and I too feel a sense of pain that is elicited by and congruent with her perceived welfare, I feel empathic concern.

Further, there are two main types of empathy. *Dispositional empathy* is a stable personality trait that makes some people more able to empathize with others. These people have an easier time thinking about what the world looks like from someone else's perspective, and they are much better at empathizing with and feeling compassionate toward others. By contrast, *state empathy* is a short-lived mental state of concern for the well-being and needs of others. In psychology experiments, researchers induce state empathy by presenting subjects with hypothetical stories of people in need and by asking them to take the perspective of the person in question. For example, Batson and his colleagues have successfully induced empathic concern by having research subjects read a fictitious story about "Katie Banks," a college student who has trouble caring for her family when her parents die.[94] The results show that when subjects are given perspective-taking instructions, they are more likely to report feeling empathic toward Katie. Hence, perspective-taking causes empathic concern.

The Empathy–Altruism Hypothesis and the Ethics of Killing in War. Perspective-taking and empathy have two important psychological effects that are relevant to the development of the laws of war. First off, encouraging people to take on the perspective of others increases feelings of empathic concern, and this, in turn, causes altruism – defined as other-regarding behavior done *for the sake* of improving the welfare of other people. This is the so-called *empathy–altruism hypothesis*, i.e., that empathic concern increases the tendency to care for and act on behalf of others in need.[95] According to Batson, empathic concern has two causal mechanisms: "(a) perceiving the other as in need[;] and (b) valuing the other's welfare."[96] Unfortunately, it can be difficult to encourage people to value the well-being of others. Imagine, for example, trying to get a neo-Nazi to feel empathetic toward Jewish people. Since neo-Nazis do not already value the well-being of Jewish people, it is likely quite difficult to get them to feel empathic toward Jewish people, and it is probably even harder to

[93] Batson (2011, 11–12).
[94] Batson, Early, and Salvarani (1997). For a similar approach, see Wood, James, and Ciardha (2014).
[95] Batson (2011).   [96] Batson (2011, 44).

get them to be altruistic toward them. This brings us to the second main effect of perspective-taking and empathy. Perspective-taking and empathy tend to reduce in-group/out-group biases. Since well-known cases of civilian victimization usually involve in-group/out-group biases – e.g., the Holocaust and the Rwandan genocide – this effect of perspective-taking is particularly important for understanding the evolution of civilian protection norms. In this subsection, I examine the empathy–altruism hypothesis and its importance for IHL, and in the following subsection, I look at the bias-reducing effects of perspective-taking and empathy.

Perspective-taking causes empathic concern through one of two causal pathways: first, by getting people to better understand the needs of others, and second, by getting people to weigh the needs or interests of others more heavily. One way that researchers induce empathic concern is by simply asking their subjects to take the perspective of others when they read through a hypothetical vignette. They can either ask them to engage in egocentric perspective-taking – "consider how you yourself would feel in this situation," or they can ask them to engage in other-regarding perspective-taking – "please consider how *X* feels in this situation." Interestingly enough, not only do different perspective-taking instructions lead to different behavioral effects, but they also implicate distinct areas of the brain.[97] In one study, Batson and his colleagues had participants read a vignette about a university student, Janet, who was having trouble making friends and was feeling generally very depressed. The findings showed that merely asking participants to take Janet's perspective, to think about how she feels, led to increased empathic concern and an increased desire to spend time with her. To control for the possibility that social desirability bias affected the results – i.e., that subjects did not want the researchers and their assistants to judge them for choosing not to spend time with Janet, Batson made sure that their responses were confidential. These findings show that empathic concern increases altruism irrespective of concerns about negative social evaluation.[98] Batson has also shown that inducing empathic concern toward *specific individuals* can increase empathy for *the groups* of which those individuals are a part. For example, in one study he and his colleagues asked subjects to listen to a fictional interview with "Julie," a woman with AIDS. In this interview, Julie "talked about her life since unexpectedly learning 3 months ago that she was HIV positive":[99]

[97] Batson, Early, and Salvarani (1997); Lamm, Batson, and Decety (2007).
[98] Fultz et al. (1986); Batson (2010).
[99] Batson, Polycarpou, Harmon-Jones, Imhoff, Mitchener, Bednar, Klein, and Highberger(1997, 108).

Well, as you can imagine, it's pretty terrifying. I mean, every time I cough or feel a bit run down, I wonder, is this it? Is this the beginning – you know – of the slide? Sometimes I feel pretty good, but in the back of my mind, it's always there. Any day I could take a turn for the worse. (pause) And I know that – at least right now – there's no escape. I know they're trying to find a cure – and I know that we all die. But it all seems so unfair. So horrible. Like a nightmare. (pause) I mean, I feel like I was just starting to live, and now instead I'm dying. …It can really get you down.[100]

As expected, the results showed that subjects in the perspective-taking condition were significantly more likely to express empathic concern for AIDS victims in general.[101] To see whether these findings are generalizable beyond sparking empathic concern for young women with AIDS, Batson ran the study using homeless people and murderers as the stigmatized group, and found similar results.

A plausible extension of the empathy–altruism hypothesis is that empathic concern for war victims can increase moral condemnation of military tactics that kill civilians, and that perspective-taking and empathy should encourage people to push for stronger norms for protecting civilians in armed conflict – a conjecture that I analyze later in the historical cases in Parts II and III of this book. Although it is hard to discern the relative causal impact of empathic concern on the evolution of the laws of war, some research shows that there is a connection between wartime violence, empathic concern, and altruistic behavior. In a recent study, Alexandra Hartman and Benjamin Morse (2020) showed that experiences of wartime violence increase the tendency to empathize with others, and that this, in turn, can "motivate helping behavior across group boundaries."[102] They show that Liberians who experienced wartime violence during the Liberian civil war were more likely to host refugees from the 2010–2011 crisis in the neighboring Ivory Coast. Furthermore, they showed that "violence-affected individuals" hosted "a higher *proportion* of non-coethnic and Muslim refugees," a result that points toward the bias-reducing effects of empathy.[103] Other research has shown that empathy is strongly correlated with support for humanitarian policies designed to aid civilian war victims. In one study, Sabrina Pagano and Yuen Huo (2007) showed that individuals who felt "higher levels of empathy" toward the Iraqi people were more likely "to support political actions to improve" their well-being in the aftermath of

---

[100] Batson, Polycarpou, Harmon-Jones, Imhoff, Mitchener, Bednar, Klein, and Highberger (1997, 108).
[101] Batson, Early, and Salvarani (1997a, 110); Pinker (2011, 586–587).
[102] Hartman and Morse (2020, 731).   [103] Hartman and Morse (2020, 733).

the 2003 Iraq invasion.[104] Anecdotal evidence also suggests a strong connection between perspective-taking, empathy, and moral judgments about harming civilians. For example, the publication of photographic images of the effects of war on civilians in Vietnam (e.g., the photo of Kim Phuc running from a napalm strike in South Vietnam in 1972 and the photos of the My Lai massacre in 1968) indicates that photography had a significant influence on public criticism of US foreign policy in Vietnam.[105] It is one thing to read about civilian suffering, yet quite another to see for oneself what US military operations mean for civilians on the ground.

To show how empathic concern influenced the evolution of the laws of war, I look for three key pieces of evidence that – taken together – plausibly indicate that empathic concern shaped the development of civilian protection norms. First, I look for information suggesting that compassion influenced the authors of specific texts and influenced the framing of political arguments in diplomatic negotiations. Specifically, I look for words and phrases that have a strong emotional content – e.g., Vattel's comment in *The Law of Nations* that "a man of exalted soul no longer feels any emotions but those of compassion towards a conquered enemy who has submitted to his arms."[106] Second, I look at how emotionally salient events such as major wars and social processes such as state formation and interdependence led people to place positive moral value on the lives and well-being of others. If the empathy–altruism hypothesis is correct, political events and processes that induce widespread empathy should encourage the development of stronger rules for protecting noncombatants. Third, although the focus of this book is on the *development* of civilian protection norms rather than their *effectiveness* in practice, in the historical cases in Parts II and III, I show that when people positively value the lives and well-being of others, and when they see them as relevantly similar to themselves, they are more inclined to observe restraints. However, when the conditions for empathy are absent, and in-group/out-group distinctions are heightened, people are less likely to observe restraints.

Why do perspective-taking and empathic concern have these effects on altruism and moral judgments? One possible explanation for the connection between perspective-taking, altruism, and morality is that perspective-taking can almost literally cause people to feel how others feel. Indeed, research in cognitive neuroscience shows that some of the same areas of the brain that process pain are also operative when people

---

[104] Pagano and Huo (2007, 245).   [105] Neer (2013).   [106] Vattel (2008, 553).

witness pain in other people. Empathy does not just involve feeling and understanding the mental states of others, but it also involves *feeling very similar experiences*. In one study, researchers showed that the brain areas responsible for processing pain, namely, "the 'pain matrix'—bilateral [anterior insula] AI, the rostral [anterior cingulate cortex] (ACC), brain stem, and cerebellum—were activated when subjects experienced pain themselves as well as when they saw a signal indicating that their loved one had experienced pain."[107] This is important for several reasons. First, while some commentators might view these results as confirming the idea that people only act altruistically to reduce their own discomfort, people nevertheless feel discomfort when seeing loved ones in pain. Second, they imply that, at the level of neurophysiology, there may be less of a difference than most people think between the pain that we experience and the pain that we see in other people. The Self and the Other may be closer than many people think. Finally, they help to explain, at the level of neurophysiology, why salient events like major wars, genocides, or humanitarian catastrophes can motivate many people to press for stronger norms for protecting others: because on some level they literally feel the pain of others.

Perspective-Taking, Empathy, and In-Group/Out-Group Biases. Not only do perspective-taking and empathic concern lead people to be more altruistic, but they also reduce in-group/out-group biases. In one study, Margaret Shih and her colleagues showed that perspective-taking reduces explicit and implicit in-group biases. Shih had subjects watch a clip of *The Joy Luck Club*, a movie that depicts the struggles of an Asian-American character, June. To induce empathy, the participants were asked "to imagine how June feels about what is happening while they were watching the clip. In the control condition, participants were instructed to watch the clip as a newspaper reviewer."[108] After watching the video, participants were asked to finish a computer task to measure in-group biases. Specifically, they were asked to "evaluate good and bad adjectives after being subliminally exposed to in-group pronouns (e.g., 'us') or…out-group pronouns (e.g., 'them') before each word," and "they measured how quickly participants were able to make the judgments."[109] Since in-group and out-group pronouns have diverging valences, subjects displayed in-group bias if they were faster at judging

---

[107] Singer and Leiberg (2009, 976); Singer et al. (2004).
[108] Shih, Stotzer, and Gutierrez (2013, 81).
[109] Shih, Stotzer, and Gutierrez (2013, 81).

good words after being subliminally exposed to in-group pronouns and faster at judging bad words after being exposed to out-group pronouns.

Interestingly, Shih found that subjects in the empathy condition showed no in-group biases. There was no statistically significant difference between the reaction times when subjects judged the valence of adjectives upon being primed with in-group pronouns or out-group pronouns.[110] In another study, research subjects were asked to evaluate a fictitious college application from a white American and an Asian American, both of whom were males. In the experiment, the subjects were asked to watch the brief excerpt of *The Joy Luck Club*, and those in the treatment condition were asked to put themselves in June's shoes and to imagine how she feels. Subjects were then asked to evaluate two college applications – one from a white American student and one from an Asian student. More specifically, they were asked which of the two applications they were more likely to accept. While there were no statistically significant differences between the groups in their decision to admit the student (the applicants were academically very strong), subjects in the perspective-taking condition revealed a much stronger liking for the Asian applicant than did the subjects in the control condition.[111]

These findings are especially significant in light of the fact that some neuroscience research shows that people have a hard time empathizing with members of relevant out-groups, and in some cases, they actually take pleasure in witnessing out-group pain.[112] This is particularly true for groups that are highly stigmatized.[113] Shih's work shows that perspective-taking can help mitigate in-group biases and out-group differentiation. Similarly, in the study I discussed earlier, Batson found that perspective-taking can generate empathic concern even for the members of stigmatized groups, e.g., AIDS victims, homeless people, and even criminals.[114] These findings show that getting people to empathize with particular *individuals* can help to engender empathic concern for an *entire group*. As philosopher Peter Singer suggests, perspective-taking helps expand our circle of concern, in the process making our judgments more impartial between Self and Other.[115] Furthermore, research by Marcus Holmes and Keren Yarhi-Milo shows that perspective-taking and

[110] Shih, Stotzer, and Gutierrez (2013, 81).
[111] Shih, Wang, Bucher, and Stotzer (2009).   [112] Cikara, Botvinick, and Fiske (2011).
[113] Harris and Fiske (2011).
[114] Batson, Polycarpou, Harmon-Jones, Imhoff, Mitchener, Bednar, Klein, and Highberger (1997).
[115] Singer (1981). For a similar argument, see Kohlberg (1981); Habermas (1990); and Linklater (2007).

empathy can help states overcome longstanding tensions. Empirically, they show that empathy played a key role in helping the United States, Israel, and Egypt sign the 1978 Camp David Accords, especially through President Jimmy Carter's ability to empathize with the Egyptian President Anwar Sadat and Israeli Prime Minister Menachem Begin. This suggests that even in cases where perspective-taking and empathy cannot reduce in-group/out-group *differences*, they can help overcome intergroup *tensions* and *conflict*.[116]

Against Empathy Skepticism. The scholarship reviewed earlier seems to indicate that empathy is one of the greatest forces for good in the modern world. However, in recent years some psychologists and philosophers have raised compelling arguments against the importance of empathy – a hodgepodge of ideas that I will refer to under the collective heading of "empathy skepticism." In *Against Empathy*, Bloom raises several problems with empathy, most of which have to do with what he calls the *spotlight* nature of empathy. Empathy is particularistic in its gaze and parochial in its chief objects of concern. As a result, it leads people to act in ways that compromise important moral principles:

Empathy is a spotlight focusing on certain people in the here and now. This makes us care more about them, but it leaves us insensitive to the long-term consequences of our acts and blind as well to the suffering of those we do not or cannot empathize with. Empathy is biased, pushing us in the direction of parochialism and racism. It is shortsighted, motivating actions that might make things better in the short term but lead to tragic results in the future. It is innumerate, favoring the one over the many.[117]

With the exception of Bloom's claim that empathy pushes us "in the direction of parochialism and racism," I largely accept most of what he says about the limits of empathy. If we define "empathy" as he does – i.e., that it is an affective capacity that enables people to feel what others feel[118] – then, yes, there are some problems with empathy. Not only is it innumerate, but it can cause us to ignore things like fairness and impartiality. Nevertheless, I argue that as long as empathy is appropriately guided by our rational capacities, it can be a powerful instrument for motivating prosocial concern.

Bloom lays out three key problems with empathy. First, he argues that in the moral domain, empathy is not only overrated, but that it often clashes with moral concerns like fairness and justice. In one study,

---

[116] Holmes and Yarhi-Milo (2017).   [117] Bloom (2016, 9).
[118] Bloom (2016, 3–4, 16–17, 39–41, 70).

Batson and his colleagues told subjects about Sheri Summers – a ten-year-old girl with "a fatal disease [who] was waiting in line for treatment that would relieve her pain. Subjects were told that they could move her to the front of the line. When simply asked what to do, they acknowledged that she had to wait because other more needy children were ahead of her. But if they were first asked to imagine what she felt, they tended to choose to move her up, putting her ahead of children who were presumably more deserving."[119] In this case, empathic concern for Sheri Summers led most people to unjustifiably choose her interests over the interests of others, thereby sacrificing rules of justice. Second, empathy is innumerate. It directs our attention to the here-and-now suffering of particular people. This is the *identifiable victim effect*, which Thomas Schelling aptly describes as follows:

Let a six-year old girl with brown hair need thousands of dollars for an operation that will prolong her life until Christmas, and the post office will be swamped with nickels and dimes to save her. But let it be reported that without a sales tax the hospital facilities of Massachusetts will deteriorate and cause a barely perceptible increase in preventable deaths – not many will drop a tear or reach for their checkbooks.[120]

Bloom raises two problems with the identifiable victim effect. First, it "sets up a perverse situation in which the suffering of one can matter more than the suffering of a thousand."[121] Furthermore, he argues that it is psychologically impossible for us to empathize with large numbers of individuals. Imagine, for instance, trying to empathize with the many millions who have suffered as a result of the Syrian civil war. In Bloom's view, this is simply beyond our cognitive and emotional capacities.

Finally, Bloom argues that empathy is heavily biased. In his view, we tend to feel empathy for those who are closest to us (family members, friends, tribal affiliates, and so forth). Citing cases like the Sandy Hook massacre, Bloom notes that "we are fascinated by the plight of young children, particularly those who look like us and come from our community. In general, we care most about people who are similar to us – in attitude, in language, in appearance – and we will always care most of all about events that pertain to us and people we love."[122] In evolutionary terms, this makes a lot of sense: those who are psychologically constituted "to favor…friends and family over strangers" will survive to pass on their genes at a higher rate than those who are not.[123] While he acknowledges

---

[119] Bloom (2016, 25).
[120] Quoted in Bloom (2016, 89). For more on the identifiable victim effect, see Small, Loewenstein, and Slovic (2007).
[121] Bloom (2016, 89).　　[122] Bloom (2016, 91).　　[123] Bloom (2016, 94).

that in-group biases are technically "separate from empathy," he contends that the "spotlight nature of empathy means that *it is vulnerable to them*."[124] Because empathy has a narrow, spotlight focus, it will always be focused on "what captures our attention," including in-group biases.[125]

These are all excellent points. However, they overlook the fact that perspective-taking and empathy not only help to temper in-group biases, but, in addition, they tend to encourage people to become more concerned about larger categories of individuals. And although Bloom is right to say that the spotlight nature of empathy can encourage people to ignore pressing concerns like fairness, empathy is sufficiently malleable that it can be directed by emotional framing and reason. In short, as long as they are properly targeted, perspective-taking and empathy will foster prosocial concern.

First, although Bloom is right to say that it is very difficult to empathize with a large number of individuals and that the spotlight nature of empathy can lead people to ignore things like fairness or justice, we can use the spotlight nature of empathy to make people more aware of broader moral concerns. Citing the work of Elaine Scarry, Bloom suggests that we are psychologically incapable of putting the interests of others on the same plane as those we love. To recognize the demands of fairness, Bloom suggests that we should *depersonalize ourselves and those we love*. For example, when deciding on who to hire or who to give an award, we should give equal weight to friends and strangers. But instead of putting strangers on the same moral plane as our friends and family, Bloom suggests that we find ways of *reducing* our personal biases – e.g., by using a blind review process where we do not know a job candidate's age, sex, race, or appearance – all factors that can influence hiring practices. In short, we should look at ourselves and other people behind a "veil of ignorance," to use Rawls' terminology.[126] Although I agree that justice and fairness sometimes require that we depersonalize ourselves and those we love, it all depends on the context. When in-group biases are so bad that they lead people to systematically violate the moral rights of others – e.g., during ethnic cleansing campaigns – justice requires that the victims be personalized through empathy. Moreover, even though empathy can sometimes encourage people to overlook considerations of fairness, the experiments that Bloom reviews all seem to show that, at the end of the day, empathy is incredibly malleable. For example, if we get people to empathize with Sheri Summers – the ten-year-old with the fatal disease – then people will empathize with her. However, if we induce

---

[124] Bloom (2016, 95).   [125] Bloom (2016, 95).   [126] Rawls (1999a).

people to empathize with the other children in line, then the results will likely go the other way. Also, by getting people to empathize with a *particular* individual, we can open their eyes to the moral interests of a *broader category* of individuals, as the research findings of Batson and Shih duly illustrate.

Similarly, as historian Lynn Hunt has shown, and as I discuss in more detail later, getting people to empathize with particular individuals – even *fictional* individuals – has generated *real* consequences for the promotion of social justice. According to Hunt, the reading of epistolary novels helped bring about the idea of human rights in the West.[127] Reading these novels helped Europeans empathize "across traditional social boundaries between nobles and commoners, masters and servants, men and women, perhaps even adults and children. As a consequence, they came to see others – people they did not know personally – as like them, as having the same kinds of inner emotions."[128] A similar process helped bring about the US Civil War and the destruction of slavery in the United States. Though the causes of the war were diverse, historians agree that *Uncle Tom's Cabin* played an important role in illustrating the moral horrors of slavery. As Civil War historian James McPherson points out, although it is difficult "to measure precisely the political influence of *Uncle Tom's Cabin*," it sold over two million copies in ten years, thereby making it one of the most noteworthy best sellers "of all time in proportion to population."[129] One main objective of the book was to make the "whole nation feel what an accursed thing slavery is,"[130] and it largely succeeded. Even contemporaries agreed. When Abraham Lincoln met with Harriet Beecher Stowe in 1862, "he reportedly greeted her with the words: 'So, you're the little woman who wrote the book that made this great war.'"[131] What made Stowe's book such a success was the degree to which it helped people empathize with the plight of American slaves. Far from detracting from fairness or justice, empathic concern for the rights of slaves helped fuel the abolitionist movement in the United States.

Second, and relatedly, although I agree with Bloom's point that empathy can sometimes be parochial, it can sometimes reduce in-group biases, as the experiments by Shih reviewed earlier duly suggest. Perspective-taking and empathy do not get rid of in-group biases, but they can certainly help with reducing their most pernicious effects.

---

[127] Hunt (2007).   [128] Hunt (2007, 40).   [129] McPherson (1988, 89).
[130] This quote comes from a letter from Harriet Beecher Stowe's sister-in-law. See McPherson (1988, 89).
[131] McPherson (1988, 90).

In short, we should use empathy to help deal with in-group biases. In response, empathy skeptics such as Bloom think that this is asking too much. In a discussion of the Israeli–Palestinian conflict, Bloom remarks that

> some would argue that the solution is more empathy. For Israelis, then, empathy not just for their neighbors sitting in the café, but for suicide bombers who set off the bomb that maimed them. For the Palestinians, empathy not just for their brothers and sisters who had their homes crushed by tanks but for the soldiers driving the tanks.[132]

In Bloom's view, "this is a nice thought," but this is just "not how empathy works."[133] It is simply unrealistic to ask Israelis and Palestinians "to feel as much empathy for an enemy as for their own child."[134] I have two responses to these points. First, when people call for Israelis and Palestinians to empathize with each other, they do not seem to be calling for them to empathize with extremists. Instead, they are calling for greater empathy between the two communities.

Second, although empathic concern may have evolved to facilitate intra-group connections, particularly the connection between mothers and their offspring,[135] this does not mean that natural proclivities like empathy are inflexible to behavioral control and rational redirection. For example, marketing experts routinely appeal to base human instincts to sell products, e.g., appealing to desires for sex, status, and dominance to sell everything from soda products to SUVs. Marketing behavior is well outside the evolved functioning of psychological traits and dispositions, but it clearly works in selling cars and driving up company profits. If marketing experts can manipulate evolved human psychology to sell products that are bad for our health (e.g., cigarettes), then there is no reason why we cannot redirect our natural proclivity to empathize with individuals within our own social groups to empathize with wider social networks of people on the outside. Indeed, the literature I reviewed earlier shows that empathy helps to reduce in-group biases and in-group/out-group tensions. If this is true, then it means that as long as we can rationally redirect how empathy is used, we can make it apply to a wider network of individuals and groups outside our own immediate circles of concern. As I illustrate in more detail later, perspective-taking and empathic concern *have* helped to temper our naturally excessiveat-tachments to kith and kin, particularly when it comes to civilian

---

[132] Bloom (2016, 190).   [133] Bloom (2016, 190).   [134] Bloom (2016, 190).
[135] Churchland (2011).

protection norms, which, as I show in Parts II and III of this book, used to be far more parochial than they are at present.

The Universal Grammar of the Laws of War: Moral Intuitions about Intentional Harm

The research reviewed earlier shows that perspective-taking and empathy influence patterns of moral thinking, and that they plausibly encourage people to morally condemn civilian victimization. In the chapters that follow, I show that if these capacities are engaged, they lead political agents to create stronger legal norms for protecting civilians in armed conflict. Consistent with the empathy–altruism hypothesis, I show that if the social environment encourages people to take the perspective of war victims, or if it encourages them to place a higher value on noncombatant lives, it will lead people to push for stronger legal institutions for protecting civilians and other victims of war. For example, in the aftermath of World War II, empathic concern for the victims of the Nazi regime – a regime that committed moral crimes that are said to have shocked the conscience of humanity – strongly influenced support for creating better protections for the civilian populations in occupied territories. Archival documents show that US support for such provisions was genuine, and it was not motivated by concerns for reciprocity or power politics. Similar considerations apply to earlier developments in the law and ethics of war. A close analysis of the language that Francisco de Vitoria used to express his displeasure with Spanish imperial policies in the New World – which involved killing and enslaving indigenous peoples – suggests a degree of empathic concern for the plight of Native Americans. By his own account, when Vitoria learned of the "bloody massacres…of innocent individuals pillaged of their possessions and dominions," he concluded that there were reasons "for doubting the justice of what [had] been done."[136] It was knowledge of the "bloody massacres," not a rational deduction from first principles, that led Vitoria to question the morality of Spanish policy in the Americas. For these reasons, I argue that perspective-taking and empathy can induce people to create *restrictive* norms for protecting civilians against intentional attacks.

But in addition to perspective-taking and empathy, cognitive–emotional biases and heuristics strongly influence the content of the laws of war, particularly cognitive–emotional biases that focus on intentional harm. More specifically, moral reasoning is governed by subtle

---

[136] Vitoria (1991, 238).

Table 2.1. *A grammar of intuitive moral rules on intentional and unintentional harms*

Rule	Description
R1	1. It is permissible to use violence in cases of individual or collective self-defense.
	2. Do not harm individuals in a specific category C, where C is defined by culturally created membership rules.
R2	1. Intentional acts of violence have higher cognitive and emotional valence than unintentional acts of violence.
	2. Intentional harms are morally worse than unintended harms.
R3	1. It is impermissible to intentionally inflict harm on "innocent" people.
	2. "Innocent" people are those who do not exhibit an intent to violate moral or social norms, and they are not responsible for posing a threat to the lives of others.
R4	1. It is permissible to act in ways that risk incidentally harming other people.
	2. In this case, two conditions must be met: (1) the action in question must be necessary to achieve an otherwise justifiable end; and (2) the potential benefit of the action must be sufficient to outweigh the harms imposed on others.

asymmetries in how people evaluate intentional actions: intended harms are routinely judged as worse than unintended harms, a cognitive–emotional bias that I refer to as the *intention/side-effect* distinction. Deliberately shooting one civilian is a war crime, but bombing a military target that incidentally kills hundreds of civilians might not be, depending on the military value of the target in question. This cognitive–emotional bias not only shapes how political and military decision-makers think about the law and ethics of warfare, but it shapes moral reasoning in virtually all spheres of social life, and there are strong reasons for thinking that the *intention/side-effect* distinction is substantially innate, by which I mean that it is programmed prior to experience.[137] Although this *intention/side-effect* distinction does impose certain limitations on what states can do in armed conflicts, it constitutes the basis for the most significant permissive effects of IHL. Indeed, were it not for the fact that the law reflects a distinction between intended and unintended killing – or if there were more significant concerns for unintentional or accidental collateral damage – the accidental bombing of the MSF hospital in Kunduz might have been more widely criticized as a war crime. To streamline the discussion, Table 2.1 spells out the grammar of moral rules that people use to evaluate intentional harm – heuristics that I claim are not only substantially innate, but also affected the content of IHL.

---

[137] Haidt and Joseph (2007).

In this grammar of moral psychology, R1 serves to restrict the bounds of moral subjectivity. Moral principles only apply to individuals who are seen as full moral subjects, worthy of protection. It is a way of capturing the fact that moral cognition is sensitive to group distinctions. Not only do people happen to care more about those within their own sphere of concern, but they tend to think that such care is justifiable: treating friends as friends and enemies as enemies is usually considered to be a moral imperative. However, R1 has flexible parameters in that it can be applied to a narrower or wider group depending on the context. As I explained in the previous section, perspective-taking and empathy have been shown to reduce in-group biases. R1 also reflects concerns over collective self-defense, which is important because most cultures have norms that permit the use of violence for self-defense. Rules R2–R4 reflect general moral rules associated with intentional harm, battery, and the doctrine of double effect. The doctrine of double effect holds that when one action has two different effects – one "good" and one "bad" – we should evaluate the action by reference to a norm that permits the intended good effects, yet prohibits the intended bad effects. More specifically, the doctrine of double effect holds that it is impermissible to intentionally do something "bad," e.g., to abuse children, to steal, or to kill innocent people. But the doctrine of double effect permits people to cause something bad just in case it is an unintended side-effect of an otherwise justifiable action, e.g., defending one's home from a violent intruder, bombing an enemy target, or droning individual terrorists or terrorist cells. In situations like this, the moral benefits of the good effect must outweigh the moral costs of the bad effect. If the benefits outweigh the costs, the action is deemed permissible.

These rules help to characterize the automated mechanisms that make up System1 thinking, specifically with respect to moral judgments regarding the infliction of harm. As I mentioned earlier, moral thinking encompasses a wide range of issues, including concerns related to purity/impurity, justice and fairness, in-group/out-group relations, authority ranking, and physical and mental harm. Haidt argues that these five issues constitute the moral foundations of all ethical systems: they govern how people intuitively react to a range of external stimuli. For example, just as secular liberals react negatively to the infliction of harm on innocent people, so too religious conservatives react negatively to the desecration of sacred objects like churches, shrines, mosques, or synagogues. In other words, human moral systems focus on issues related to physical and mental harm, but they also focus on a wide range of issues that have little to do with harm. My emphasis on harm should not be taken to imply that these other issues are unimportant for morality, and

indeed I suggest that these other concerns often play an important role in how people think about the ethics of harm. For example, as is reflected in R1, in-group/out-group thinking has a significant impact on how people think about the morality of killing: restraints on killing are ordinarily much higher for people who are within our sphere of concern. In addition, concerns for purity and impurity can amplify in-group/out-group dynamics. As I illustrate with a discussion of ISIS in Chapter 4, the ISIS killings of Yazidis and Shia Muslims were motivated by a toxic mixture of in-group/out-group dynamics amplified by a concern for in-group purity. Thus, the intuitive foundations of morality can interact with each other to produce variations on the same basic moral concerns. My reason for emphasizing the harm mechanism is that it is the one that carries the most significance for the law and ethics of warfare.

Why should we believe that rules R1–R4 characterize the moral judgments that most people make with respect to deliberate harm, and why should we believe that they are substantially innate, or programmed prior to experience? There are three key reasons why. First, a wide range of studies show that rules like R1–R4 characterize the ethical judgments that many people make with respect to the use of violence, and some of these studies show that specific reaction patterns are rooted in specific neural networks in the brain. With respect to R1, decades of work on social identity theory shows that not only are people inclined to feel a stronger sense of obligation to those within their own sphere of concern, but they often base these Self–Other distinctions on seemingly minor group characteristics, and they often display strong in-group biases. For example, in a famous experiment, Henri Tajfel and Michael Billig found that research subjects establish social identities on the basis of seemingly minor traits, such as whether someone likes Kandinsky or Klee paintings.[138] Social psychologists and IR scholars have used these findings to explain why nationalism, race, ethnicity, and patriotism influence patterns of conflict and violence in international relations.[139] Research in neuroscience shows that Self–Other distinctions, as well as the likelihood of aggressing against the Other, are rooted in distinctive neural networks. In one study, Mina Cikara, Matthew Botvinick, and Susan Fiske examined how fans of rival baseball teams (the Red Sox and the Yankees) responded to the successes and failures of the other team. Interestingly, they found that reward-processing neural systems are correlated with self-reports of the "likelihood of harming the rival team's fans."[140] For example, when Red Sox fans see the Yankees make a bad play, they feel a

---

[138] Billig and Tajfel (1973).   [139] Brewer (1999); Gat (2013).
[140] Cikara, Botvinick, and Fiske (2011, 309).

sense of pleasure. What is more, the region of the brain that codes for this pleasure – the ventral striatum – is also correlated with the desire to harm a fan of the rival team.[141] These findings show that even at the neural level, Self–Other distinctions are key to understanding the propensity to inflict harm on others.

With respect to the law and ethics of armed conflict, these findings are important because they show that moral cognition and emotion are sensitive to in-group/out-group distinctions. Even though perspective-taking and empathy can moderate Self–Other distinctions, such distinctions are extremely important for understanding how people think and feel about the ethics of killing in war. In particular, they help explain why some societies and individuals betray a distinct lack of concern for the plight of foreign civilians. In the American Southwest, for example, the wars between white settlers and indigenous tribes were extraordinarily brutal, and both sides directly killed civilians.[142] Likewise, Spanish conquistadors killed indigenous people in what is now Latin America on a mass scale, often justifying their behavior by claiming that the natives were "less than human." Although some may argue that examples like these show that there really is no moral grammar, a more plausible interpretation is that these kinds of cases show that moral cognition is group-oriented. Psychological restraints on killing civilians are lower when there is a Self–Other distinction between political and military decision-makers, on the one hand, and civilians, on the other. Some IR scholars argue that decisions to target civilians are made primarily on the basis of cost–benefit calculations rather than considerations of race and social identity.[143] But some cases show that states treat co-ethnics very differently from others. For example, during the American Civil War, the Union initially adopted a conciliatory policy toward Confederate civilians. When it became apparent that this policy would not work, they eventually abandoned it in favor of scorched earth tactics. Yet, as Civil War historian Mark Grimsley (1995) points out, Union forces generally spared civilian lives. However, the United States was not similarly restrained in its treatment of native tribes, whom it later killed off and dislocated in mass numbers. Racial differences are the most likely explanation for the variation in treatment.

With respect to R2–R4, several studies have shown that most people place a higher valence on *intentional* as opposed to *unintentional* harms. Several studies have shown that judgments concerning the moral

---

[141] Cikara, Botvinick, and Fiske (2011, 309).
[142] Gwynne (2010); Kinsella (2011); Janda (1995).
[143] Downes (2008). For a compelling response, see Fazal and Greene (2015).

wrongness of an action are often correlated with judgments about the mental state of the agent who caused the harm, and in particular, people tend to focus on whether the agent intended to cause harm to the victim.[144] However, there are some variations between children and adults. Though adults tend to evaluate actions by reference to mental states, children at or younger than five years of age tend to rely more heavily on information regarding the consequences of an action. At six to seven years of age, however, they begin relying more directly on information about intentions. The most plausible explanation for this developmental variation is that by six to seven years of age, small children start to develop a theory of mind – that is, they start to understand and explain the actions of others by reference to their mental states, e.g., desires, motivations, and intentions.[145] In one study, the neuroscientists Liane Young and Rebecca Saxe showed that people vary in the degree to which they exculpate others for accidental harms, and more specifically they show that the propensity to exculpate is mediated by the right temporo parietal junction (RTPJ), a region of the brain recruited in belief attribution. Young and Saxe found that when subjects are asked to consider an accidental harm – e.g., putting poison in a colleague's drink under the false belief that the poison was sugar – those who experience "higher activation" in "the RTPJ are more likely to exculpate" others for harming their colleagues.[146] They are likely to exculpate because they believe that the agent is acting under a false belief that the poison is sugar. This finding shows that even in the case of accidental harms, assessments of mental states are critical for assigning blame and responsibility to moral agents.

The *intention/side-effect* distinction, which lies at the core of R2–R4 as well as the laws of armed conflict, is grounded in emotional gut reactions to the thought of intentional harm. Cognitive neuroscientists and social psychologists have shown that people often place a higher moral valence on intended harms as opposed to unintended, side-effect harms. Even though the consequences of a specific action may be similar (i.e., someone is harmed), most people are inclined to believe that intentional harms are morally worse than incidental harms or accidents. The neuroscientist Joshua Greene has shown that this tendency is grounded in two distinctive processes of moral judgment. Using functional magnetic resonance imaging (fMRI) scans, Greene shows that when research participants are asked to evaluate harms that involve personal force – e.g., pushing

---

[144] Cushman (2008); Baird and Astington (2004); Darley and Shultz (1990); Hauser et al. (2007); Mikhail (2007, 2011); Borg et al. (2006).
[145] Baird and Astington (2004); Young and Saxe (2009).
[146] Young and Saxe (2009, 2069).

someone off a footbridge to stop a trolley from running over and killing five people – the emotional areas of the brain "light up," and people tend to judge that it would be wrong to push the person off the bridge. However, when they are asked to evaluate actions that do not involve the use of personal force – e.g., pushing a switch that diverts the trolley from a path containing five innocent people onto a track that contains only one person – they are more likely to judge the action as morally right. Greene has shown that in this kind of case, the more rational areas of the brain associated with higher-order cognition are active, and consequently, most people are inclined to make the utilitarian judgment that it is morally right to flip the switch.[147] In other words, the emotional framing of an action has a strong effect on the degree to which people judge that it is the right thing to do. As Jon Mercer aptly suggests "justice is an emotional belief."[148]

Rules R2–R4 bear a direct relationship to the principles of distinction and proportionality in IHL, and the social psychological and neuroscience evidence reviewed earlier helps explain the content of the laws of war. This evidence suggests that people routinely judge intended killings as being worse than unintended killings. Thus, a plausible explanation for why the laws of war distinguish between intended and unintended killings is that when states created these laws they drew on these implicit cognitive-emotional biases. Indeed, later in this chapter I theorize that in diplomatic negotiations, states utilize these affective responses to facilitate what Todd Hall and Andrew Ross call the horizontal transmission of emotions, thus ensuring that individual-level emotions take on a collective dimension.[149] When emotional responses lead to the development of legal norms, they become institutionalized.[150] In addition, findings from cognitive neuroscience and psychology help explain how people justify or judge military tactics that lead to civilian casualties. For example, in the MSF case reviewed at the beginning of this chapter, the Pentagon's justification for US behavior was clearly based on the fact that the aircrew that fired on the hospital operated under the mistaken belief that the hospital was, in fact, a military target. The findings of Young and Saxe show that there are particular neural networks that underlie such exculpatory judgments. Further, the *intention/side-effect* distinction helps to explain why states often justify military policies that expose civilians to harm by using what Bruce Cronin refers to as the "collateral damage exception to IHL,"[151] i.e., that they did not intend civilian deaths.

---

[147] Greene et al. (2001); Greene (2013).   [148] Mercer (2010, 6–7).
[149] Hall and Ross (2015).   [150] Crawford (2014a); Holmes and Traven (2015).
[151] Cronin (2013, 175).

The second reason why we should conclude that R1–R4 are substantially innate is that rules like them would have been functional for promoting survival in the ancestral environment. In other words, there are strong reasons to believe that they conferred an adaptive advantage on individuals who made judgments in accordance with such rules and behaved accordingly. The reasons for why rules like R1 would have provided adaptive advantages have been well articulated in the literature on the evolution of morality,[152] so here we can be brief. R1 functions to promote in-group cohesion by enabling individuals to make moral distinctions between in-group and out-group members and by giving them reasons to act on behalf of kin and other conspecifics.[153] Presuming it is true that in-group relationships conferred an adaptive advantage on individuals in the social environment, one way of reinforcing such relationships is through moral judgments and obligations. People are more likely to act on behalf of in-group members if they think they have a moral obligation to do so and if they are likely to feel guilty about violating such obligations. Morality helps bind groups together, and it does this by helping people distinguish the ethical obligations that they have toward members of their own tribe from the ethical obligations that they have toward members of other, perhaps rival, tribes. Within the tribe, obligations of mutual respect prevail. However, between tribes, moral obligations are generally seen as far less dense. We owe more to our compatriots than we do to outsiders. Rules R2–R4 characterize the obligations that prevail among those who see each other as full moral subjects.

Within in-group relations, R2–R4 provide people with the cognitive and emotive repertoire to resolve a range of coordination and cooperation problems. The ability to distinguish intentional from accidental harms helps people detect cheaters: intentional harms against innocent persons are a clear and reliable signal that someone cannot be trusted. For example, if the United States intentionally violates a treaty obligation, this signals to the international community that the United States cannot be trusted. Because moral judgments express categorical, objective, and action-orienting prescriptions, an individual who *moralizes* the *intention/side-effect* distinction would be more likely to take it upon herself to implement a costly punishment when it is violated. Furthermore, since intentional harms signal that one cannot be trusted *and* since it can lead others to reciprocate, moralizing the distinction between intended and

---

[152] Haidt (2001, 2012); Joyce (2006); Tooby and Cosmides (2010); Bowles and Gintis (2011); Greene (2013); DeScioli and Kurzban (2009, 2013).

[153] Gat (2013); Lopez, McDermott, and Peterson (2011); Haidt (2012).

unintended harms would have made it easier for people to form more beneficial social relationships by encouraging them to implicitly monitor their own behavior. Indeed, Greene suggests that people likely evolved an action-monitoring system that prevents them from "being casually violent."[154] Also, since moral intuitions often generate emotional responses, they make it easier for people to credibly commit to follow through on costly punishments. When people are angered by a social infraction, for example, they may lash out irrationally, and this will, as Schelling would argue,[155] increase their credibility should they decide to make a punitive threat. In short, moral emotions and cognitive biases facilitate coalitional affiliations and help actors solve the kind of collective action problems that would have been prevalent in the ancestral environment.

Finally, research on moral and social development indicates that cognitive–emotional biases are substantially programmed prior to experience. In one study, Amrisha Vaish, Malinda Carpenter, and Michael Tomasello showed that small children, i.e., eighteen- to twenty-five-month-olds, expressed concern and exhibited helping behavior when they witnessed others experiencing a negative event, and that this effect held even when the person in question expressed no signs of harm. Vaish et al. exposed their subjects to a harmful or a neutral event (the control condition). In both conditions, one of the experimenters showed the child a favored object (a necklace, a belt, a picture, or ball of clay), and then showed the child subject how much they valued the object. In the harm condition, another experimenter took the object, and said, "in a mildly aggressive tone 'I'm going to take/tear/ break this now,' and [did so] mildly aggressively for 15 [seconds]."[156] In the neutral condition, the second experimenter said and did the same thing, only this time they acted less aggressively. Vaish et al. found that children in the harm condition were significantly more likely to show a sense of concern for the victimized experimenter. They were also more likely to try to help them. Vaish et al. argue that children do this through a process of affective perspective-taking.[157] These findings show that empathetic abilities stem from an ontogenetic process, not a general learning mechanism.

Other studies show that the ability to evaluate intentions arises very early in the process of human development.[158] In one study, five- to twelve-month-old research subjects were shown videos "depicting two

---

[154] Greene (2013, 226).
[155] Schelling (1960); Bowles and Gintis (2011, 186–194); Greene (2013, 40).
[156] Vaish, Carpenter, and Tomasello (2009, 536).
[157] Vaish, Carpenter, and Tomasello (2009, 534).
[158] Hamlin, Wynn, and Bloom (2007); Hamlin (2013).

objects (a square and a triangle), one engaging in helping behavior and one engaging in hindering behavior toward a third object (a ball) that attempted to climb a hill."[159] After being habituated to the helping behavior and hindering behavior in the first set of videos, the infants were subsequently shown two new test videos, one in which the ball approaches the helping object, and another in which the ball approaches the hindering object. The researchers found that the subjects looked longer at the test video in which the ball approached the helping object versus the test video in which it approached the hindering object. They concluded that the infant research subjects "preferred the test movie in which the ball approached the object that had previously helped it reach the top of the hill."[160]

In another study that used material objects for the characters, infants were shown a similar scene in which a climbing character was either helped or hindered by another character, and they were subsequently given the option of reaching for the helper or the hinderer, along with a neutral object.[161] When they were given the choice between selecting a helper versus a neutral character, the infants "systematically chose the helper," and when given the choice between selecting a neutral character or a hindering character, they likewise chose the neutral character.[162] Follow-up studies have shown that children also engage in acts of retribution against the antisocial characters.[163] Though preliminary, studies like this indicate that small children have an innate capacity to attribute mental states such as goals, intentions, and desires to other people, and they also suggest that small children have a rudimentary ability to attribute positive and negative intentions to others, an ability that is critical for their social development. Kiley Hamlin, Karen Wynn, and Paul Bloom conjecture that "the presence of social evaluation so early in infancy suggests that assessing individuals by the nature of their interactions with others is central to processing the social world, both evolutionarily and developmentally," and that, as a result, the "capacity for such evaluations can be seen as a biological adaptation."[164] Also, even though most of the trolley experiments reviewed earlier are conducted with adults, some of them have been conducted with children as young as four-years-old, and with very similar results.[165]

---

[159] Kuhlmeier, Wynn, and Bloom (2003, 402).
[160] Kuhlmeier, Wynn, and Bloom (2003, 405).
[161] Hamlin, Wynn, and Bloom (2007); Wynn (2007, 344).
[162] Hamlin, Wynn, and Bloom (2007, 558).   [163] Hamlin (2013).
[164] Hamlin, Wynn, and Bloom (2007, 558).   [165] Pellizzoni, Siegal, and Surian (2010).

Why should we believe that intuitive heuristics R1–R4 are preprogrammed, or innate, rather than learned? Simply put, it is hard to explain how people can possibly develop the moral abilities that they in fact develop by supposing that they simply use general learning mechanisms. The main argument for the idea that people have innate moral modules is that the environment to which they are exposed is *impoverished*: it does not contain the amount of information that people would need to have in order to make the judgments that they in fact make.[166] For example, the fact that children have certain social capacities from a very early age indicates that it is unlikely that these capacities are internalized from the social environment, which does not necessarily contain all of the relevant distinctions. In addition, the work of social psychologists Elliot Turiel and Judith Smetana shows that small children know how to distinguish between basic social conventions and moral rules, and that they make this distinction without any outside input. Though social conventions are perceived as contingent and context-dependent, moral rules are perceived as universal and applicable beyond a specific social circumstance.[167] For example, speaking out in class without first raising your hand violates a social convention, but stealing or homicide violates basic moral rules. When children are asked to judge behaviors such as "not saying grace before snack" or "not hanging one's coat in the designated place," they routinely judge them to be permissible when there are not explicit societal rules that prohibit them – which indicates that they are societal conventions.[168] However, when kids are asked to evaluate behaviors such as hitting another child, they tend to say that such actions are seriously impermissible even "in the absence of a rule," i.e., that they are moral transgressions.[169] Cross-cultural research on Korean children shows that they make very similar distinctions between social conventions and moral principles. When Korean children are asked to explain why they feel a particular action is impermissible, they tend to talk about transgressions "in terms of" individual "welfare, obligation, and fairness," suggesting that these abilities are cross-culturally prevalent.[170]

The fact that small children make these distinctions between social conventions and moral principles indicates that they believe there is a difference between the moral domain and the social domain. As a result, philosopher Richard Joyce argues that in the case of morality this "poverty of the stimulus" problem is especially stark: "The problem here isn't just that the environment doesn't offer *enough* data for a child to grasp the

---

[166] Mikhail (2007, 2011); Joyce (2006).   [167] Turiel (1983, 37).
[168] Smetana (1981, 1335–1336).   [169] Smetana (1981, 1335).
[170] Song, Smetana, and Kim (1987, 580–581).

necessary distinction[s]; it's that it is puzzling what there even *could be* in the environment – even a rich and varied environment – for a generalized learning mechanism to latch on to in order to develop the idea of a moral transgression."[171] Indeed, the fact that small children pass judgment on the intentions of agents is puzzling enough; the fact that they are able to draw fine distinctions between general social conventions that they learn from authority figures and moral transgressions indicates that there is a preprogrammed *moral faculty* that enables them to understand the difference between what is morally right and what is morally wrong.

Innate Biases, Individual Differences, and Cultural Variation: The Linguistic Analogy

Poverty of the stimulus arguments like this draw quite explicitly on the work of the linguist Noam Chomsky, so a brief excursus on this approach to human cognition will help to clarify certain aspects of my argument. In particular, although I interpret the linguistic analogy as simply that – an analogy – the theory of human cognition upon which this analogy builds provides a useful template for explaining how it is possible for System 1 moral intuitions to be relatively universal even though there are important differences between individuals and across cultures. As Chomsky described it, the goal of generative linguistics is to explain how it is possible for people to gain an extraordinarily rich and complex knowledge of language in light of the finite range of experiences and information to which people are ordinarily exposed. In his view, the social environment does not provide people with the amount of information that they would need to have in order to converge on knowledge of a particular language. Since Plato raised a similar problem in the *Meno*, Chomsky referred to this as "Plato's Problem."[172] To clarify Plato's Problem, consider the following facts. Assuming some degree of normal functioning, most human beings who reach a certain age become competent users of a particular language, whether English, Spanish, Mandarin, or Arabic. This means that the neural architecture of the brain must be such that it enables human beings to learn languages, to engage in abstract thinking, and so on. In contrast to the blank slate view of the mind put forth by behaviorists, Chomsky argued that knowledge of language is partially innate.[173]

According to Chomsky, the only way to account for linguistic knowledge is to focus on the internal cognitive structures and mechanisms that

---

[171] Joyce (2006, 137).   [172] Chomsky (1986).   [173] Chomsky (1959).

underlie complex linguistic abilities – a view that Chomsky refers to as "mentalism."[174] Prior to the Chomskyan revolution in linguistics, languages were conceived as externalized objects, or collections of utterances and written statements out there in the public domain. In this view, the grammar of a specific language was viewed as a set of rules from which one could precisely deduce all of the available sentences or utterances in the language. The problem with this understanding of language is that it is inconsistent with linguistic creativity: individuals can enumerate an infinite number of statements that outstrip any rules that one can inductively derive from the available utterances of a public language. Since it is possible to construct an infinite number of sentences in any language, we cannot view language as an external, publicly available object. Instead, we have to regard it as a component of the mind/brain of individuals. That is, we should regard language as an "internal" object, or as "some element of the mind of the person who knows [a] language."[175] In this view, the linguistic knowledge of a mature individual consists of two key elements: an intuitive knowledge of universal grammar, and an experientially derived knowledge of the grammar of a particular language.[176]

Although there are significant differences between language and morality, this view of the mind provides a useful framework for understanding how moral intuitions might be innate. As the philosopher Jerry Fodor describes it, the Chomskyan view holds that individuals are "born knowing certain facts about universal constraints on possible human languages," and in response to their experiences with the social and cultural environment, they are able to construct more mature linguistic grammars.[177] In the case of morality, the notion that some moral intuitions are innate – or "organized in advance of experience"[178] – helps explain how people come to develop a moral sense in the first place, and why certain intuitions emerge early in the process of human development and in uniform stages.[179] Finally, this approach helps to make sense of how individual and cultural differences are consistent with the notion that certain cognitive structures are substantially innate, and thus broadly universal. There are clear cultural differences in language across societies, but if the Chomskyan view is right, these differences are underwritten by a shared knowledge of universal grammar. With respect to morality, people start out with an intuitive toolkit of moral principles,

---

[174] See Chomsky (1965, 4) and Mikhail (2011, 19).   [175] Chomsky (1986, 22).
[176] Chomsky (1986, 24). For more on Chomskyan linguistics, see Chomsky (1965); Hornstein and Antony (2003); Lasnik (2005); Hornstein (2005).
[177] Fodor (1983, 4).   [178] Haidt and Joseph (2007, 380).   [179] See Kohlberg (1981).

and their cultural experiences help them develop a more specific moral system. There are some innate biases that encourage them to accept certain rules and to adopt certain ways of thinking, but their experiences set the parameters of the moral principles that they adopt. For example, perspective-taking and empathy help redefine the line between in-groups and out-groups in R1, and emotional experiences can encourage people to adopt a higher threshold for making the kind of proportional trade-offs called for in R4.

Haidt likens the relationship between innate moral intuitions and social interaction to the setting of a sound equalizer on a stereo system. In his view, we all come equipped with some basic *moral foundations* (a concern for harm/care, justice/fairness, in-group pride, purity, and social hierarchy), which are akin to dials we can use to manipulate how our music sounds (volume, treble, bass, etc.). Our experiences help to set the values of our moral equalizers:[180] while some people are particularly concerned about individual harm and justice, others are more concerned about in-group pride and hierarchy. With respect to the ethics of war, cultural experiences and traditions influence *who* one perceives as worthy of protection, and they also shape what kinds of moral values people seek in war – e.g., whether they fight to enforce the commands of an omniscient deity, or whether they fight for more secular objectives like national self-determination. Furthermore, although most people distinguish intended from unintended killings, the *extent* to which individuals are concerned about killing the innocent and protecting people from harm likely varies with personal experiences. Indeed, as the study by Hartman and Morse (2020) cited earlier suggests, experiences with violence generally encourage people to be more compassionate toward individuals in similar circumstances. Also, experimental research by Eric Uhlmann, David Pizarro, David Tannenbaum, and Peter Ditto (2009) suggests that political ideology plays a strong role in whether people are willing to support operations that kill civilians. Specifically, their results show that conservatives are more likely than liberals "to condone the killing of innocent civilians in a military attack when those civilians [are] Iraqis killed by Americans rather than Americans killed by Iraqis," a result that is not surprising in light of Haidt's finding that conservatives tend to be more "groupish" than liberals.[181] Rules R1–R4 should thus not be regarded as *exceptionless* moral universals, but rather as a set of *innate biases*.[182]

That said, one might wonder how useful this approach to human cognition is when it comes to explaining the evolution of the law and

---

[180] Graham and Haidt (2012); Haidt (2012).   [181] Uhlmann et al. (2009, 489).
[182] Sripada (2008).

ethics of war. After all, if the moral mind is like a stereo system that can be set high and low by the social and cultural environment, then the social, cultural, and individual variables that push our moral dials up and down are the key causal factors we should focus on, not the innate structures of the mind. In other words, evolution may have provided human beings with an innate knowledge of basic moral concerns, but this innate knowledge is of little use for explaining the content of the laws and ethics of war in particular circumstances. Here, we need to bring in sociological mechanisms and culture-specific values. I could not agree more, and indeed in the following section, I sketch out in more detail how the innate mind intersects with sociological mechanisms and culture to generate convergence on similar norms of war. That said, the theory of moral cognition and emotion sketched out in this chapter provides a useful template for explaining why very similar moral principles and concerns repeatedly emerge across societies. Individual and cultural differences can be understood as manifesting different dial settings on the same underlying principles. That said, when the social environment primes people to feel compassionate toward the victims of war, there is a broad tendency for civilizations to converge on more humanitarian norms.

The analogy with language breaks down when we consider the causal role of emotions. As I suggested earlier, emotions play a highly significant and independent causal role in shaping moral judgments. Research by Haidt and others has shown that when moral sentiments are primed, even unconsciously, this has a significant effect on how people evaluate the morality of a given action.[183] Furthermore, the research by Greene that I reviewed earlier shows that the emotional centers of the brain play a key role in shaping moral judgments.[184] Those who defend a strong interpretation of the linguistic analogy believe that cognitive appraisals, or "unconscious computations," condition emotional reaction patterns.[185] But in light of the fact that emotions play a clear role in shaping and amplifying moral judgments, I adopt a weaker version of the linguistic analogy according to which the rules set forth in Table 2.1 characterize the automated mechanisms that constitute System 1's moral judgments. While some of our System 1 moral intuitions may be purely cognitive, others are highly emotional. In addition, emotional experiences help set the parameters of the moral rules that people use to think about everyday moral decision-making. As I mentioned earlier, perspective-taking and empathy can shape how people define in-group/

---

[183] Haidt (2001, 2012); Schnall et al. (2008).   [184] Greene et al. (2001); Greene (2013).
[185] Mikhail (2007, 2011); Huebner, Dwyer, and Hauser (2009).

out-group distinctions. So too, feelings of guilt can encourage people to act more conscientiously in the future when they commit moral infractions. In short, while the linguistic analogy provides a useful way of thinking about how innate moral biases are consistent with individual differences and cultural diversity, there are some key differences between moral principles and linguistic principles. We can think of rules R1–R4 as like a moral grammar in the sense that they help to characterize the automated mechanisms that shape our moral judgments.

Reason and Emotion in the Making of Moral Judgments

As I mentioned earlier, a traditional line of debate in moral philosophy, one that goes back at least as far as Plato's *Republic*, has to do with the extent to which our moral judgments are guided by reason or by emotions. Whereas moral rationalists such as Plato and Kant believe that reason should govern the emotions, moral sentimentalists such as David Hume believe that reason is the slave of the passions. This debate has also influenced competing approaches to the study of moral cognition. Rationalists such as Lawrence Kohlberg (1981) believe that moral reasoning plays a decisive role in how individuals arrive at specific judgments and how they develop over time. Moral intuitionists such as Haidt (2012) turn this view on its head. Like Hume, moral intuitionists believe that reason is the slave of the passions. Instead of viewing human beings as intuitive scientists who are out on a search for moral truth, moral intuitionists view human beings as intuitive lawyers who are intent on using their reasoning skills to justify their preexisting intuitions – a view that is supported by experimental work on motivated reasoning.[186] For example, in one study, the legal scholar Dan Kahan showed that not only do liberals and conservatives disagree about the empirical validity of climate science, but that liberals and conservatives with better reasoning skills are more inclined to be ideologically polarized. Specifically, he showed that liberals and conservatives who performed well on the Cognitive Reflection Test were more ideological in how they evaluate information.[187] In the case of morality, Haidt argues that people typically feel a quick flash of emotions in response to a particular action – e.g., they feel "revulsion at the thought of incest" – and then, when they are asked to provide a "verbal justification" for their judgments, they search for

---

[186] Jost and Amodio (2012); Kahan (2013); Haidt (2012). For an IR application, see Jeffery (2016).

[187] This test measures the ability to engage in System 2 reasoning (Kahan 2013, 410).

reasons to justify their emotions in the same way that a lawyer searches for reasons to defend their clients.[188]

While I accept the view that intuitions and emotions inform the process of moral reasoning, I reject the claim that reason is their slave. There are two reasons why. First off, although evolution provided humans with an architecture of preprogrammed moral intuitions that helps them navigate the contours of everyday social life, people often confront situations for which their intuitions leave them ill-prepared. For example, consider the moral dilemmas that states sometimes face when deciding whether to impose economic sanctions on regimes that violate international law: imposing sanctions sends a strong signal that law-breakers will be held accountable, but sanctions may also have a detrimental impact on the civilian population. In cases like this, our moral intuitions about intentional or unintentional harm provide, at best, a simplistic guide to moral action. People need to reason through the various considerations to help them figure out the right thing to do. System 1 intuitions inform our thinking, but when these intuitions conflict, System 2 steps in to help us make a decision. Haidt claims that moral reasoning is only freed from the passions when intuitions are weak or when "processing capacity is high,"[189] but System 2 can also be freed from System 1 when System 1 intuitions are conflicting and roughly equal in strength. The strongest intuitions do usually win out, but System 2 still helps people think through the issues.

Second, reasoned reflection – either subjective or intersubjective – can help recalibrate our intuitions, possibly through neuroplasticity,[190] a neurological process that I describe in more detail in the following section. On the individual level, reasoning through a moral dilemma, or reflecting on a personal experience, can help people decide which of their moral intuitions to listen to, and it can also shape their moral outlook for the future. At the intersubjective level, the effects of reason are potentially even more pronounced. Moral positions that liberal Westerners now take for granted (e.g., that slavery is wrong) were once openly contested. If Crawford (2002) is correct, these intuitions evolved through a long-term process of political contestation and moral argument, a view that implies that collective reasoning can help people recalibrate their moral intuitions and develop institutional responses to fix their various drawbacks.[191] Far from simply operating at loggerheads, reason and emotion often work together to shape individual and collective moral judgments.

---

[188] Haidt (2001, 814).   [189] Haidt (2001, 819).   [190] Holmes and Traven (2015).
[191] Patterson, Rothstein, and Barbey (2012); Paxton and Greene (2010).

Culture, Cognition, and the Evolution of the Laws of War: How Mind and Culture Intersect

Taken together, there are strong reasons to believe that the moral brain is equipped with a set of computational systems that enable people to make fast, intuitive judgments when confronted with complex moral dilemmas, as well as a set of cognitive reasoning abilities that help them think through their options. We also have strong reasons to believe that these moral abilities and intuitive heuristics are substantially innate. However, recent research in cognitive neuroscience casts doubt on the idea that the brain is hardwired for specific functions, especially research on neuroplasticity. It is well known that the brains of small children are incredibly plastic, but more recent scholarship shows that neuroplasticity continues well into adulthood. In one study, Bogdan Draganski and his colleagues showed that research subjects who were asked to learn how to juggle showed increases in gray matter "in the mid-temporal area…and in the left posterior intraparietal sulcus," an area of the brain "associated with the processing and storage of complex visual motion."[192] Other experiments have shown that people who learn how to play music at an early age develop a larger corpus callosum – the area of the brain associated with interhemispheric communication.[193] In one experiment, Alvaro Pascual-Leone showed that having subjects merely think about learning a new task, such as playing the piano, led to increases in the size of the associated regions of the brain. Pascual-Leone concludes that the human "brain is highly plastic, and that plasticity represents evolution's invention to enable the nervous system to escape the restrictions of its own genome."[194] In response to findings like this, constructivist IR scholars might argue that a culture-centered account of moral psychology is more consistent with neuroplasticity than the account that I have defended here.[195]

In this section, I briefly explain how neuroplasticity works, and why it does not falsify my account of moral cognition and emotion. In light of the fact that the empirical evidence is too rough to arrive at any definitive conclusions, my claims about neuroplasticity are necessarily speculative. With that in mind, I argue that the phenomenon of neuroplasticity actually helps tie together some of the loose ends of the argument put forth in the preceding section. In particular, it helps to explain how mind and culture might intersect. Historical contingencies such as major wars and broad sociological processes such as the rise of economic interdependence across societies affect our ethical thinking by engaging our

---

[192] Draganski et al. (2004, 311).   [193] Muente, Altenmueller, and Jaenke(2002, 475).
[194] Pascual-Leone (2009, 150).   [195] Crawford (2009).

abilities for empathy, intuition, and higher-level moral reasoning. Since the human brain is plastic, this opens up the possibility that social interaction and cultural experiences intersect with our moral psychology by calibrating the parameters of our intuitional heuristics – "fixing" the dials of our moral equalizers. Neuroplasticity may be particularly important for explaining how empathy can lead people to expand their circle of concern: their experiences help alter their neural networks. This is why I say that there are innate *biases* rather than *exceptionless* universals. Next, I show how my emphasis on innate biases can help explain the durability of civilian protection rules once they emerge: the fact that these rules fit with strong affective biases explains why people are motivated to keep them in place and to strengthen them over time. Finally, I describe how emotional framing and logical arguments shape the restrictive and permissive effects of IHL in face-to-face diplomacy.

Is the Moral Brain Inflexible? Cultural Variation and the Cognitive Science of Neuroplasticity

Neuroplasticity refers to the ability of the brain to rewire itself in response to experimental cues, environmental factors, cultural particularities, and so forth. At the level of neural networks, it is manifested in the synaptic connections that form in response to environmental triggers, but it is also manifested in the ability of the brain to generate new neurons later on in life (neurogenesis). In the early stages of development, the brain is wired through the overproduction of synaptic connections. Psychiatrist Jeffrey Schwartz points out that "the factor that provides the developing brain with the right connections is, ironically, an early profusion of wrong connections…. About half the neurons that form in the fetal brain die before the baby is born: 200 billion neurons…are reduced to the 100 billion of a newborn as neurons that fail to form functional synapses vanish."[196] As the child grows and develops, its brain continually rewires itself in response to learning, environmental cues, and cultural transmission processes like imitation or teaching. In one set of studies, Schwartz demonstrated that psychotherapy can cause marked changes in the orbital frontal cortex of patients with obsessive–compulsive disorder (OCD), the location of the brain where OCD is localized.[197] Interestingly, he found that simply getting patients to recognize that their symptoms were the result of faulty brain wiring was sufficient to alter their brains.

---

[196] Schwartz and Begley (2002, 117).   [197] Schwartz and Begley (2002).

Some philosophers and psychologists argue that findings on neural plasticity problematize the claim that there are innate cognitive structures and abilities. There are several reasons for this. First off, it is mathematically impossible for genetic programming to be solely responsible for the development of synaptic connections. Philosophers David Buller and Valerie Hardcastle claim that "the total number of genes in human DNA is currently estimated at around 80,000," and "as much as 50% of these may be concerned with our brain," and "yet we have literally trillions of synaptic connections in our head," and "[t]here is no way even 40,000 genes could code for that exactly."[198] Moreover, they argue that genetic programming is not responsible for the development of "higher cognitive functions…for it appears that most of the genetic 'specification' for our brain concerns the more peripheral structures."[199]

However, the argument that I have advanced in this chapter does not presume that neural networks and structures are inflexible or determined entirely by genetic programming. Some aspects of moral cognition and emotion seem to be innate – particularly in light of the evidence concerning moral development that I reviewed earlier. Empathic abilities, the group-oriented nature of morality, and the *intention/side-effect* distinction all seem to be based upon relatively hard-wired neurological structures. But other aspects of moral psychology are certainly the result of enculturation and neural plasticity. Notice, however, that my argument is consistent with neuroplasticity and neurogenesis. My claim is that the mind comes equipped with cognitive and emotional biases that lead people to endorse particular norms in particular circumstances, not that the mind deterministically produces these norms.[200] Also, I presume that certain elements of moral cognition and emotion are thoroughly plastic: genetic programming does not determine who we empathize or identify with, and it leaves substantial room for the development of group-specific norms and cultural values (although human beings are strongly inclined to value life). As Aristotle once said, human beings are by nature social animals. But *which* society or *which* group we happen to identify with is a matter of experience.

Finally, research on neuroplasticity shows that attention and volition are causally important for neuroplasticity, at least in the later stages of adulthood.[201] This leaves room for the possibility that evolved moral intuitions interact with cultural learning in the following way: moral cognition and emotion are wired for a particular content ("do not intentionally kill members of group g"), but experiential learning shapes how

---

[198] Buller and Hardcastle (2000, 314).   [199] Buller and Hardcastle (2000, 314).
[200] Sripada (2008).   [201] Schwartz and Begley (2002).

the brain applies this rule in practice. Similarly, experiential learning may affect the extent to which people believe intended harms are morally worse than accidental harms. It may even be possible to train people to believe that intentionally harming innocent people is a good thing. However, in light of the evidence I reviewed earlier, this would require going against the grain of System 1. Since even small children manifest basic ethical concerns, helping behavior, and so forth, overriding these inclinations is difficult. So although it is possible for neural networks to rewire themselves, the cognitive science of neuroplasticity does not falsify my argument. Indeed, my argument depends on perspective-taking, empathy, and System 2 reasoning to calibrate System 1 intuitions: when the social environment encourages people to feel empathy for others or to reason from an impartial position, it may affect changes in how the parameters of their intuitive heuristics are set. Of course, this means that our moral modules change and develop over time, but not in the radical way that blank slate theorists would expect. Encouraging people to treat outsiders as similar in kind to themselves does not require a wholesale revamping of their moral architecture, it simply requires an increase in the value that people attach to the lives and welfare of outsiders. It requires the Self to humanize the Other and to observe some basic rules of restraint for how he/she is treated.

Perspective-Taking, Empathy, and the Emergence of Civilian Protection Norms

This implies that in order to understand how the civilian protection norms emerge, we need to understand how the circle of concern expands. In the historical case studies in Parts II and III of this book, I show that there is no *one* structural-systemic force, or "master variable," to use Wendt's terminology,[202] that leads people to perceive the Other as similar in kind to the Self, but rather there is a complex set of them. Indeed, this is one reason why I claim that moral cognition is crucial for understanding the emergence of civilian protection rules and why structural-systemic theories of norm emergence are inadequate: material and cultural forces are indeterminate. That said, the literature reviewed earlier provides some clues as to the kinds of thick social processes that will increase the extent to which people identify with or value the lives of others. First, social processes that encourage people to engage in routine perspective-taking should increase empathetic concern. For example, in her book *Inventing*

---

[202] Wendt (1999, 343).

Human Rights, the historian Lynn Hunt shows how print media and epistolary novels helped facilitate the rise of human rights ideas in Western Europe by eliciting empathy. These novels helped readers feel the experiences of the characters in an entirely new way. According to Hunt, they "made the point that all people are fundamentally similar because of their inner feelings," inner feelings that were accessible by the form in which the books were written.[203] In Chapters 3 and 5, I argue that this helps explain the initial emergence of civilian protection norms in ancient China and the West. In ancient China, the Warring States regimes were increasingly staffed by commoners who had a better understanding of the plight of ordinary people. This forced the leadership to take the interests of commoners into consideration in the making of state policies. In the West, the Peace of God movement promulgated civilian protection norms through increased social contact and perspective-taking in face-to-face meetings between clergymen and local lords.

Second, as the empathy–altruism hypothesis implies, empathetic experiences tend to induce people to take steps to improve the welfare of others. In Chapter 5, I show that emotional responses to the treatment of indigenous tribes in the Americas encouraged just war theorists such as Vitoria to extend moral recognition to the interests of native peoples. Similarly, in Chapters 6 and 7 I show how empathetic experiences and emotional framing helped motivate the creation and strengthening of contemporary IHL. In Chapter 6, I show how emotional framing helped influence the evolution of the concept of *lawful military targets* in IHL, and in Chapter 7 I describe how direct experiences with war victims led Henri Dunant to push for stronger positive norms for aiding wounded soldiers. In addition, I show that sympathetic reactions to World War II and the Holocaust encouraged states to agree to create stronger laws for safeguarding the civilian population against intentional killing.

Third, although this book focuses on explaining norm convergence, it also provides insight into the psychological factors that cause civilian victimization. Although states and nonstate actors certainly kill civilians for strategic or tactical reasons, moral cognition also plays an important role. First, people vary in their empathic accuracy scores. Some people score very high on dispositional empathic concern scales, but others score much lower. Some people score appallingly low. Though it is unclear what impact these latter outliers have on international politics and war, anecdotal cases like Adolf Hitler, Joseph Stalin, and Pol Pot suggest that when extremely depraved people get into positions of

---

[203] Hunt (2007, 39).

political power, the consequences for innocent civilians can be especially dire. Second, I argue that explicit attempts to dehumanize people and to magnify in-group/out-group differences can motivate civilian victimization. Just as perspective-taking can encourage the Self to humanize the Other, so too mass outbursts of interethnic violence are often precipitated by conscious attempts to prevent people from considering the inner lives of others. As social neuroscientists, Lasana Harris and Susan Fiske show, disgust reactions toward deviant groups or persons facilitates dehumanized perception, or "a failure to spontaneously consider the mind of another person."[204] These reactions in turn may facilitate practices such as torture or genocide, practices that are "generally not reserved for human beings because perceiving a person implies" adhering to "a number of moral rules and norms."[205] These findings help explain the rhetoric that perpetrators of mass killings often use: they frequently try to compare their victims to nonhuman animals, thereby eliciting disgust reactions that inhibit the tendency for their followers to consider the mental states of their victims.

Moral Psychology and the Durability of Civilian Protection Norms

Because they are backed by powerful moral emotions and beliefs, civilian protection norms are likely to remain relatively durable once they emerge. This point is important because it suggests that regardless of the specific historical mechanisms through which these norms arise, moral beliefs and emotions are still important for explaining how they evolve because they render certain norms more stable. By this, I mean that they are persistent standards of behavior that people use to evaluate state practices,[206] not that they are always effective at regulating behavior. Since norms can emerge for any number of reasons, my claim in this subsection is that *affect-backed* norms are more likely to survive through the process of cultural evolution than are *affect-neutral* norms. Here, I build on the epidemiological account of cultural transmission set forth by anthropologist Dan Sperber. The epidemiological account holds that human psychology plays an important role in the transmission of cultural ideas. In short, it holds that cultural items that *fit* with universal mental modules are more likely to emerge and stay salient with a specific population. The epidemiological account of norms is not intended to explain the origin of norms per se; rather it is intended to account for the survival of cultural ideas. Philosopher Shaun Nichols argues that since

---

[204] Harris and Fiske (2011, 175).   [205] Harris and Fiske (2011, 180).
[206] Crawford (2002, 86).

"affective mechanisms are regarded as universal denizens of human psychology," they play an important role in cultural transmission:[207]

Emotional items are typically accorded extra importance. Put crudely, we *care* more about information that is emotionally gripping for us. And it seems likely that information that we care more about will be more culturally viable.[208]

Civilian protection norms are backed up by two sets of emotions: (1) empathic concern; and (2) the emotional responses that underwrite the *intention/side-effect* distinction. Moreover, they are reinforced by a range of emotional reaction patterns, including disgust, revulsion, shame, and guilt. Individuals feel pangs of disgust and revulsion in response to mass atrocities such as the massacre at My Lai or the Rwandan genocide. These emotions are primarily directed at the actions of *others:* virtuous people feel disgusted or repulsed by such actions. Emotions of guilt or shame, on the other hand, are primarily directed at *ourselves*. We feel guilt when we fail to live up to our own normative standards. For example, soldiers who participated in mass killings may, later on, feel guilty about what they did. We feel shame when we fail to live up to the normative standards of our own moral community. In other words, shame is more akin to embarrassment,[209] and as such, it can be recruited to strengthen or reinforce existing norms, as well as to crack down on deviant behavior.[210] Here is one area where my view about the neurological foundations of moral beliefs and emotion intersects with constructivist scholarship on socialization and norm diffusion: neurologically based intuitions provide the material substrate for socialization processes such as "naming and shaming." However, the theory set forth here provides a deeper explanation for why certain norms arise again and again: they fit reasonably well with evolved cognitive structures and powerful moral emotions.

Of course, just because a norm or law is *affect-backed* does not necessarily mean that it will be stable across time. As Steven Pinker has shown,[211] revenge norms are *affect-backed*, but modern societies have found ways of tempering some of the more pernicious effects of revenge. The desire for revenge is at the heart of reciprocity, but, taken too far, it can lead to costly cycles of retaliation and disproportionate reprisals.

---

[207] Nichols (2002, 240).   [208] Nichols (2002, 240); Sperber (1996).
[209] See Prinz (2007, 76–79).
[210] Several IR scholars have examined the political and rhetorical significance of shame. See Finnemore and Sikkink (1998, 903); Keck and Sikkink (1998); Krebs and Jackson (2007); Petrova (2016); Adler-Nissen (2014); Price (1998).
[211] Pinker (2011, 529–547). I am indebted to one of the anonymous reviewers for raising this point.

Liberal philosophers as far back as John Locke have understood this, and indeed Locke believed that this was one of the inconveniences of the state of nature that encourages people to create a civil government to protect their rights. In Pinker's view, social institutions such as the rule of law help to calm "the impulse for revenge before it spirals into a destructive cycle."[212] While it may be difficult to completely expunge the desire for vengeance, cultural mores and social institutions can help to calibrate these impulses in various directions. In cultures of honor, the desire for revenge is revved up, and as a result, costly cycles of retaliation are more likely to arise in these cultures. However, in more liberal cultures, the desire for revenge is often calibrated toward what is minimally required for reciprocity. Here, the emotion is not expunged, just redirected.

This suggests that *affect-backed* norms are subject to cultural calibration. Consequently, one might wonder how this squares with the naturalistic theory of moral cognition and emotions I offer here. I have three responses. First, not all emotions are created equally. The moral emotions that underwrite the civilian protection regime are not only psychologically powerful, but they also have a strong negative valence when people are in their grips. John Hibbing, Kevin Smith, and John Alford (2014) describe the "negativity bias" as a psychological phenomenon that leads most people "to respond more strongly, to be more attentive, and to give more weight to negative elements of their environment."[213] The negativity bias affects individual life choices – e.g., people expend more resources to avoid threats than they do to achieve positive gains – but it also affects politics. People who have negative reactions to political events should be more likely to endorse institutional norms that reduce their likelihood. The negative reactions that people have to mass killings and genocides are a case in point: they tend to encourage humanitarian activists to push for stronger norms of war. This helps explain why rules that forbid deliberate killing are likely to be durable once they emerge.

Second, my argument does not preclude the idea that culture-specific norms and institutions can tailor our moral intuitions, nor does it preclude the idea that the sociopolitical environment can favor some of our intuitions and emotions over others. Certain events can amplify the harm settings of our moral equalizers, shaping how the parameters of our intuitive moral grammar are set. Other sorts of events can lead in the opposite direction. In his book *On Killing*, Lieutenant Colonel David Grossman explains how military training and conditioning can reduce

[212] Pinker (2011, 541).   [213] Hibbing, Smith, and Alford (2014, 303).

our psychological restraints on killing. One key mechanism involves creating a sense of cultural distance between soldiers and their victims: "It is so much easier to kill someone if they look distinctly different from you. If your propaganda machine can convince your soldiers that their opponents are not really human but are 'inferior forms of life,' then their natural resistance to killing their own species will be reduced."[214] Similarly, Pinker points out that the desire for revenge comes with a "dimmer switch" that is "most easily modulated when the perpetrator falls within our natural circle of empathy."[215] This suggests that the social environment can shift our intuitive reaction patterns, and that it does so via empathy.

Finally, affective reactions such as empathic concern motivate political activists, diplomats, and lawyers to *institutionalize* their cognitive–emotional biases into the content of international law, thereby reshaping the culture of international society in a way that will ultimately serve to reinforce the affective mechanisms that discourage civilian victimization. The process I am thinking of here is akin to niche construction in the animal kingdom where organisms shape their local environment in ways that favor the survival and propagation of their genes. Here, affective mechanisms, which are strongly connected to behavioral motivations, lead people to change the normative environment in a way that makes it more likely that the kinds of socialization processes that constructivists often emphasize work to cultivate moral beliefs that discourage civilian victimization. Through emotion-based persuasion and reasoned argumentation, activists, lawyers, and diplomats work to reconstruct the normative environment in ways that toggle the dimmer switch on our more violent, destructive tendencies and amplify the volume switch on our more empathic tendencies. As I mentioned at the outset of this chapter, the most durable norms are those that benefit from a combination of strategic, affective, cognitive, and cultural reinforcements. By motivating people to alter the legal landscape in ways that fit with basic moral intuitions, the affective mechanisms that I have examined in this chapter should lead to particularly durable international norms.

Moral Intuitions and Communicative Action: The Restrictive and Permissive Effects of International Humanitarian Law

Rationalists argue that states create international laws to resolve cooperation problems. As I noted earlier, in the case of the laws of war they

---

[214] Grossman (2009, 161).   [215] Pinker (2011, 541).

design laws that will enable them to better protect their own citizens and that will establish legal bright lines of acceptable and unacceptable behavior. By contrast, Habermasian constructivists claim that states design the law via the logic of arguments. Although I agree that states often rationally design international law, the empirical evidence shows that they do this by invoking System1 moral intuitions. In the empirical chapters in Parts II and III of this book, I show how empathic concern motivated human rights activists, diplomats, and lawyers to *sincerely* argue for developing stronger restrictions on the use of force against the civilian population. Moral intuitions affect the process of argumentation in two key ways. First, they affect the proposals that states make to advance their interests. Second, they affect the process of argumentative consensus. Regardless of its semantic content, framing a proposal in the right kind of way can often mean the difference between success and failure. If a proposal is framed to evoke a visceral, gut reaction, this will likely overwhelm the proposal's semantic content and pave the way for a successful argument.

Again, this does not mean that System 2 communicative reasoning is causally insignificant. System 2 reasoning enables diplomats and lawyers to identify logical gaps in existing conventions, and it also enables them to respond to technological innovations that render existing laws obsolete. Furthermore, it is System 2 reasoning that helps people see that the biases and distortions of System 1 are in fact biases and distortions. However, in my analysis of international conference diplomacy in Part III, I show that attempts to generate reasoned consensus almost always rely on the intuitional heuristics and emotional response patterns of System 1. As I have stressed throughout this chapter, and as I demonstrate empirically in the rest of the book, states and nonstate actors frequently utilize the *intention/side-effect* distinction to design the restrictive and permissive effects of IHL. Although states do sometimes use this distinction in a Machiavellian attempt to create permissive norms, they also seem to use it because they regard it as intuitively compelling. What is more, the frequent use of this precept in IHL proceedings indicates that even if a significant number of states use it out of a Machiavellian attempt to reduce the effect that IHL has on their operational strategies and tactics, the distinction is perceived as having some sort of intrinsic moral merit. Regardless of the motives for which states use it, this precept gave rise to the main restrictive and permissive effects of IHL. The principle of distinction prohibits states from intentionally targeting civilians or civilian objects (a restrictive effect). However, IHL allows states to expose civilians to incidental risks just in case they adhere to the principle of proportionality (a permissive effect).

While the bulk of this book focuses on how the principles of distinction and proportionality emerged, in the conclusion of the book I discuss two key problems with these legal principles. First, they enable plausible deniability.[216] Since intentions are not public, states and nonstate actors can directly target civilians or civilian locations, but deny that any civilian casualties that resulted from their tactics were intended. Second, the rule of proportionality, in particular, gives military decision-makers wide room for maneuver. Technically, states can justify killing almost any number of people as long as the expected gains are deemed sufficiently weighty to override the anticipated losses. And although *in bello* proportionality takes the consequences of strategies and tactics into account, like the principle of distinction, it focuses on the intentions of decision-makers. As Janina Dill notes,

Rather than giving us a standard to evaluate outcomes, the principle of proportionality merely adds a requirement to what it means to have the right state of mind in war: right intent in war not only means fixing one's thoughts on a military objective (distinction). The right state of mind includes the desire to "balance" the pursuit of military progress against the imperative of protecting civilians (proportionality).[217]

Although some tactics are clearly disproportionate – e.g., carpet-bombing an entire neighborhood to take out a sniper – there is a large and highly subjective gray area. In the conclusion, I argue that even if it sometimes makes sense to distinguish between intended and unintended killings when it comes to assessing culpability for a war crime, IHL currently allows far too much decision-making discretion with respect to unintended side-effect killings and the indirect effects of armed conflict, and that as a result, it conflicts with basic principles of equal consideration. In light of the argument I have made in this chapter, it is unrealistic to presume that the *intention/side-effect* distinction will be expunged from IHL. However, I argue that we have strong moral reasons to increase the weight of concern that our legal institutions place on the unintended and indirect effects of armed conflict.

Considerations on Method

This theory of moral cognition, emotions, and argumentation generates several hypotheses that I examine in the following chapters. First, I argue

---

[216] The concept of plausible deniability is most often used in reference to covert operations, but here I use it to describe the concern that military personnel may intend to kill civilians in certain situations, yet deny that they meant to do so.
[217] Dill (2015, 87).

that moral intuitions influence the creation of the norms of war in human societies. To evaluate this claim, in Part II of this book, I use discourse analytic methods to show that there is a shared grammar of moral discourse and that this grammar causally influences the development of humanitarian protection norms. Depending on the quality of the historical data, in each case, I trace the creation and evolution of the norms of warfare. However, since the empirical record is often incomplete, and thus it can be very difficult to know for sure why a particular norm arose in the first place, I also look at why particular norms survive the process of cultural evolution. Here, I argue that the *affect-backed* nature of civilian protection norms contributes to their cultural durability, which is the second hypothesis that I examine in this book. Thus, even if it is not possible to isolate the contingent historical events and mechanisms that facilitated the emergence of civilian immunity norms in a given culture, my theory is still subject to empirical verification. Third, I argue that in face-to-face negotiations, states and nonstate actors invoke emotionally resonant principles to try to institutionalize restrictive and permissive principles into the structure of the laws of war.

Since part of my goal is to tap into shared meanings, I conduct a discourse analysis of the major philosophical, legal, and military treaties in four culturally distinctive international societies: Warring States China, the early Islamic empire, medieval Europe, and modern international society. As cultural psychologists Richard Shweder and Nancy Much point out, one of the main objectives of discourse analysis is to develop a coherent account of what individual utterances and texts mean: "since speakers always mean and convey more than they say, meaning is revealed by making explicit the relationship between the said and the unsaid."[218] My hypothesis is that some of the implicit rules and meanings that underlie everyday moral discourse are *broadly* universal and that they are grounded in basic moral intuitions. Thus, my objective in using discourse analytical methodologies to uncover shared meanings is not simply to develop a coherent interpretation of culturally specific meanings, but rather it is to evaluate the conjecture that some moral ideas are broadly shared across cultures and that they influence the creation of humanitarian protection norms.

One potential limitation of this research strategy is that it involves using textual information to derive conclusions about unobservables – i.e., the mental states that motivate moral judgments. There are two key problems here, one of access and one of sincerity. The problem of access

---

[218] Shweder and Much (1991, 186).

has to do with the fact that mental states are inherently subjective, and as a result, they are only accessible from a first-person perspective. Political scientists can only access them indirectly, through written documents, speeches, and overt behaviors.[219] When it comes to interpreting the psychological states that motivated ancient Chinese theorists, early Islamic jurists, and medieval Christian theologians, the problem of access is even more acute. Indeed, since these cases are far back in history, it is very difficult to know for sure exactly what mental states motivated people to make specific judgments. As such, there is no easy way to know for certain whether the emotional intuitions reviewed in this chapter causally influenced the norms of war in these early cultures. When it comes to face-to-face negotiations, the problem of access is still a challenge, but here the key problem is that of sincerity. In face-to-face settings, people are usually positioning for a specific audience (or two or three). As the sociologist Erving Goffman might put it, they are trying to manage an impression.[220] This means that what people *say* may not necessarily reflect what they *mean*, nor even what they *really believe*.

The problems of access and sincerity bedevil all historically oriented scholarship. Not only do political and military historians routinely try to understand the reasons for why historical figures such as Roosevelt, Churchill, or Stalin acted as they did in particular circumstances, but so too historians of philosophy frequently disagree over how to interpret the arguments of Plato, Aristotle, and Kant. These problems are not unique to the approach I use in this book. However, while some IR scholars think that the problems of access and sincerity are so difficult to get around that we should abandon the effort and focus solely on overt behavior and publicly available rhetoric,[221] here I operate under the presumption that social scientists should make an effort to theorize about the mental states that drive political action. In what follows, I address the problems of access and sincerity in three ways.

First, in each case, I look for evidence that either confirms or disconfirms my interpretations. For example, in my analysis of Warring States China I look for evidence showing that the rules of war were shaped by moral considerations, but I also look for evidence that suggests that they were driven by power politics considerations. In addition, in each chapter I examine the kind of language that philosophers, lawyers, and diplomats used to express their arguments, and I focus specifically on whether it is *plausible* that emotion-based moral intuitions may have influenced their reasoning. For example, in my examination of Western just war doctrine,

---

[219] O'Mahoney (2015) refers to this as the "fundamental problem of reason attribution."
[220] Goffman (1959).   [221] Krebs and Jackson (2007).

theorists such as Vitoria and Vattel frequently use emotion-laden terms to make their arguments. In addition, Vitoria directly explains how moral considerations stemming from empathy led him to expand the scope of the law to apply to Native Americans. Where necessary, I also use secondary sources to back up my interpretations. Given the difficulty of interpreting old and ancient texts, this methodology is imperfect. Therefore, my interpretations should be read as *plausible conjectures* that attempt to make the most sense of available textual and historical evidence. In light of the plausible assumption that emotions affect word choice and persuasion, this methodology helps to provide indirect evidence for the claim that moral beliefs and emotions shape the development of the norms of warfare.

Second, to address the problem of sincerity in face-to-face settings where the likelihood of deception is high, I follow a principle of consilience.[222] This involves looking at multiple sources of evidence (including diplomatic cables, memos, letters, speeches, diary entries, and secret archives) to evaluate competing theoretical explanations. I focus specifically on comparing what people say in public versus what they say behind closed doors. For example, to show that even powerful states care about creating norms that safeguard civilians from deliberate attacks, I examine secret memos and reports on the development of the Geneva Conventions of 1949 and their Additional Protocols. Even though backstage discussions still take place before an audience, secrecy reduces the political incentives to deceive. At the very least, it provides a more holistic understanding of state intentions. Although the principle of consilience cannot provide us with certainty as to actual state intentions, it can help reduce uncertainty. To measure the influence of moral emotions on institutional design, I examine how differences in the emotive salience of legal arguments impacted persuasive success. For example, in Chapter 6 I show that differences in the emotive framing of legal arguments played a key role in the persuasive success, or lack thereof, of proposals to prohibit the use of asphyxiating gas, the dropping of projectiles from balloons, and immunity for merchant vessels on the high seas.

Third, I show that the historical evolution of civilian protection norms in each of these cases closely matches the theoretical expectations set forth in this chapter. In the cases of ancient China and the West, I show that historical events and sociological processes that required political actors to take the interests and well-being of noncombatants into account

---

[222] O'Mahoney (2015); Holmes (2018).

led them to develop similar *jus in bello* principles. Since civilian protection rules in early Islam were a byproduct of earlier cultural norms, in this case I examine why these norms might have survived the transition from pre-Islamic Arabia to Islamic Arabia. One problem with this approach is that there are several factors that could affect whether a culture develops civilian protection ideas. Political theorists or state officials could adopt civilian protection ideas because they believe that this is the most efficient way of conquering and subsequently pacifying enemy territory – a *winning of hearts and minds* logic – or they could adopt them out of a concern for safeguarding their own civilians – a *reciprocal enforcement* logic. In each of the case studies, I argue that these two logics would lead to significantly different civilian protection rules than the *moral psychology* logic I have laid out in this chapter. In the ancient China case, I argue that the winning of hearts and minds logic doesn't explain why Warring States theorists frequently viewed state administration and warfare in moral terms in the first place. After all, some Warring States theorists (i.e., the Legalists) adopted the *realpolitik* view that the goal of statecraft is to ensure order, not to extend universal love to the people, as Mozi argued. The fact that this option was available suggests that moral concerns such as compassion did influence those who advocated more restrictive limits on how the people should be treated. I make a similar argument with respect to the West, although in this case, I add that empathic concern arising out of face-to-face interactions also shaped the early evolution of civilian protection norms. In the case of early Islam, I argue that the reciprocal enforcement logic does not explain why Muslims endorsed civilian protection norms even when they had little to fear from reprisals. Finally, in modern IHL, I argue that the reciprocal enforcement logic does not explain why reprisal attacks on civilians are regarded as war crimes.

To supplement this methodological approach, I also explore whether the theory can account for divergent or "hard" cases. There are at least four kinds of divergent cases: (1) cases where a state or nonstate actor knowingly violates a norm; (2) cases where a state or nonstate actor apparently violates a norm, but denies that their actions amount to a moral or legal infraction; (3) cases where a state or nonstate actor exploits a permissive loophole in the norm; and (4) cases where there are real cultural differences in the underlying norms. Cases of Types (1)–(3) are fully compatible with my argument. With respect to Type (1) cases, my theory does not presume full compliance. Indeed, because political actors often have strategic or power-political motivations to kill civilians, we need laws and institutions to hold them accountable. Cases of Types (2) and (3) are also consistent with my argument. These kinds of cases

usually involve an attempt to engage in norm evasion. As Zoltán Búzás defines it, norm evasion involves compliance with the letter of the law, but not its purpose.[223] Most of the divergent cases in the cultures under consideration in Parts II and III are cases of Types (1), (2), or (3). For example, in Chapter 6 I argue that although many strategic bombing operations clearly violated the principle of distinction, the United States and Great Britain attempted to justify their actions as consistent with the laws of war, arguably rendering them Type (2) or Type (3) cases.

Type (4) cases are more complicated. As noted earlier, I do not think there are exceptionless cultural universals, and so there are likely to be exceptions to the claims I make in this book. That said, the theory set forth here does provide a useful starting point for explaining *why* specific cases are exceptional. For example, in Chapter 4 I examine how the theory can deal with the motivations of Islamist extremist groups like al-Qaeda and ISIS. In particular, I argue that social identity-based factors often play a critical role in radicalization. Furthermore, the political rhetoric of ISIS clearly represents a genocidal ideology that is rooted in in-group/out-group differences and a strong desire for political and cultural purity. ISIS is an exceptional case, but not one that undermines the theory, even though it may underscore the theory's outer limits. Furthermore, my theory is about cognitive–emotional *biases*, and, indirectly, sociohistorical *tendencies*. It does not take a "snapshot" view of history, but rather it focuses on a broad tendency for societies to converge on humanitarian norms. Exceptional cases like the Nazis or ISIS may simply be a *diversion* from a broader historical trend. Hence, the proper way to evaluate the theory is to look at the *process* whereby humanitarian norms develop, which is roughly the approach that I take in the empirical chapters of Parts II and III.

Why choose these particular cultures? An ideal research strategy would be to analyze the norms of war in every culture across human history. But since this is clearly impossible, the next best option is to examine a set of cultures that are broadly distributed in time, space, religion, and so on. The cases that I review in Part II were chosen to maximize cultural variation and to ensure sufficient geographical breadth: Warring States China, Islam, the early West, and modern IHL cover a large cross-section of human cultures, and they are well distributed geographically. Furthermore, they all have a written history, which makes it possible to examine the evolution of the rules of war over time. One problem with this is that it introduces a possible selection bias. Cultures that have a

---

[223] Buzas (2017, 862).

written record are likely to be very different from those that do not, particularly on dimensions that are relevant for explaining the development of the norms of war. For example, written cultures are probably more likely to have entrenched norms for ensuring cooperation. This is because cooperation is necessary for sustaining the kind of economic developments that enable societies to have written cultures to begin with. As a result, it is possible that legal norms for ensuring the protection of innocent people may be related to these other factors rather than to psychological propensities. Additional research on non-written cultures is certainly needed to address this problem in full. However, in the case studies in Part II, I address this problem by examining how ecological factors interacted with psychological biases to generate support for particular norms of war. This gets around the problem by accepting that ecological factors play an important role in the evolution of cultural ideas. However, ecological factors (e.g., those that separate written cultures from non-written cultures) are inert. They influence the design and evolution of social norms by altering incentive structures, encouraging people to cooperate, and fostering empathy. For example, societies that are connected by trade and commerce have an incentive to develop cooperative institutions for facilitating trade. This requires them to engage in routine perspective-taking, which can in turn promote empathic concern, a reduction in in-group biases, and so forth.[224] Ecological factors shape cultural norms by engaging our emotional capacities, cognitive predispositions, and individual reasoning abilities.

One additional puzzle is that it is very difficult to tease apart the causal arrows that connect moral intuitions with the norms of armed conflict. As constructivist work on socialization suggests, as the norms of war become more robust and institutionalized, our moral sentiments become more and more attuned to protecting innocent people in armed conflict situations. This means that rather than our moral intuitions affecting the development of the laws of war, the laws of war may be affecting our moral intuitions. Modern human beings may be more concerned about civilian immunity norms because they have internalized the laws of armed conflict. Though I agree that moral intuitions and IHL are "co-constituted,"[225] in the following chapters I address this issue by looking for empirical evidence that clarifies the temporal order of the chain of events that led to the evolution of the laws of war. For example, if the evidence shows that moral beliefs and emotional reactions arose before the development of the laws of war, then we have good evidence to

---

[224] Pinker (2011, 76–77).   [225] Wendt (1999).

conclude that moral intuitions influenced the creation of civilian protection norms, not the other way around. Of course, if the law is working properly, states and their officials should eventually internalize them so that compliance becomes taken-for-granted. Once norms have become taken-for-granted, violating them should not only be seen as unlawful, but in addition, it should lead to significant moral condemnation.

But even though there are strong interconnections between the positive laws of war, on the one hand, and our moral response patterns, on the other, in the chapters that follow I argue that, in terms of their origins, the laws of war are rooted in shared moral intuitions. In Part II, I show how the theory outlined in this chapter helps explain why civilizations that are otherwise very different nevertheless converged on very similar norms for protecting civilians. In Part III, I show how it can help account for the restrictive and permissive effects of IHL.