WORKING PAPER

CASCADE INSTITUTE

# Catastrophic Dehumanization

## A formal model

Thomas Homer-Dixon
Cascade Institute

November 2024

## CITATION

Homer-Dixon, Thomas, Catastrophic Dehumanization: A Formal Model (November 14, 2024). Available at SSRN: https://ssrn.com/abstract=5061588
or http://dx.doi.org/10.2139/ssrn.5061588

## ABSTRACT

A formal model draws on insights from conflict, critical-transition, and catastrophe theories to explain the causal mechanisms that produce sudden dehumanization in situations of extreme conflict, a phenomenon characterized by a sharp discontinuity between the psychological states of identification and non-identification with other people.

A set of discrete, plausible assumptions about the psychology of human conflict are stated as axioms, allowing them to be represented as mathematical functions. These functions conjoin to produce a bifurcated catastrophe “response surface” in a three-dimensional state space defined by the variables injustice, opportunity, and identity. The surface represents a finite set of possible psychological states and pathways of movement between these states.

Taken together, the model and surface provide a psychological theory of the nonlinear collapse of interpersonal identification. The model’s internal operations and alternative parameterizations are discussed and possible avenues for testing suggested.

# TABLE OF CONTENTS

## 1. INTRODUCTION[1]

Participants in violent conflict often dehumanize their opponents. Indeed, some form of dehumanization is arguably a near ubiquitous feature of the most brutal acts of human violence, such as saturation bombardment of civilian populations, terrorist attacks on urban centers, intense battlefield combat, and genocide.

People who dehumanize members of another group generally undergo three cognitive shifts, not necessarily in the following order. First, they de-individuate the members of the other group. Whereas previously they might have perceived the other group's members as individuals each with distinct and complex characteristics, histories, personalities, and goals—thus "according them identity," as the renowned social psychologist, Hebert Kelman wrote decades ago[2]—once dehumanized, these members of the other group are perceived as undifferentiated within their group. Second, they usually apply to these undifferentiated members a simplified and highly pejorative caricature or stereotype, often equating them with an unattractive species of animal.

Third, and most importantly, they deny the moral legitimacy of the other group's way of life, interests, actions, and even existence, and in the process deny its members the protection afforded by general principles of morality.  Esses et al. (2008: 6) write: "To the extent that a group is seen as . . . immoral, the group is likely to be seen as less than human and thus as less deserving of humane treatment. That is, because they do not share our humanity, the fate of members of such a group is less relevant to our own, and their interests may be ignored."

Together, these three cognitive shifts render the group's members as "unlike us" and, critically, put them beyond the boundary of the perceivers' community and thus beyond the range of the perceivers' responsibility and care. Kelman further writes:

> It is difficult to have compassion for those who lack identity and who are excluded from our community; their death does not move us in a personal way. Thus when a group of people is defined entirely in terms of a category to which they belong, and when this category is excluded from the human family, then the moral restraints against killing them are more readily overcome.

Or, as David Berreby (2005: 222) notes: "Killing someone who doesn't feel to us like a part of the human community does not, in fact, seem like killing a person."[3]

When people undergo these three cognitive shifts, and especially when they deny the moral legitimacy of another group's way of life, interests, and actions, they no longer *identify* with members of that group. But absence of identification does not entail absence of empathy. A person might not identify with another person while still being able to imagine—and feel—what it's like to be in the other's situation.[4]

The fact that empathy can be sustained in the absence of identification explains an oft-noted but nonetheless peculiar feature of some of the most vicious human conflicts. Participants sometimes treat members of the dehumanized group far more brutally than they would ever treat truly non-human beings, such as higher animals.[5] They will sometimes go out of their way to inflict the most horrible suffering imaginable, for instance by forcing parents to watch their children be tortured and killed. Such acts only make sense if empathy is sustained, because empathy allows conflict participants to know exactly how to inflict the greatest mental suffering on their opponents. Empathy allows them to access their opponents' minds, but loss of identification, which is the essence of dehumanization, ensures that these opponents no longer have the moral protection afforded by membership in a shared community, so they can be harmed at will.[6]

Dehumanization often happens quickly, even suddenly, with a sharply nonlinear shift or "flip" from one mental state to another. Under the right circumstances, people who are peaceable neighbors and even friends one day can become fierce enemies—each dehumanizing the other—soon thereafter. In our everyday lives, road rage may be the most common manifestation of this nonlinear phenomenon.

During the late evolution of hominids, this ability to flip from a mental state of identification to one of non-identification might have aided reproductive fitness. As hominids evolved a capacity to identify with others, a capacity that likely increased community cohesion and thus generally improved individual and group fitness, those able to suspend that capacity abruptly under specific circumstances—such as when under attack by outsiders—would have had a reproductive advantage over those who could not (Panksepp and Biven, 2012: 42-44).[7]

Scholars have developed insights into the main types of dehumanization (Haslam, 2006; Livingstone Smith, 2012) and into the societal conditions that contribute to and facilitate the phenomenon (Kelman, 1973; Esses et al., 2008; and Hagan and Rymond-Richmond, 2008). But the precise psychological mechanisms within the individual that produce dehumanization, especially the mechanisms that cause the flip from the mental state of identification to that of de-identification, are still not well understood.

I propose here for discussion, analysis, and testing one possible mechanism. In the next section, I show how my argument connects with dominant views on the psychology of human conflict, while introducing some general assumptions and caveats. In section 3, I highlight four hypothesized causal mechanisms at the model's core. In section 4, I then draw on conflict research, critical-transition theory, and complexity theory more generally to derive nine specific and plausible assumptions, stated as formal axioms, about the psychology of dehumanization within the individual. This section also provides precise definitions of all variables.

Section 5 presents the full formal model. It shows that the nine axioms, when represented as mathematical functions, can account for the sharp discontinuity or "flip" that often occurs between an individual's mental states of identification and non-identification with other people, a discontinuity that makes inaccessible intermediate values of identification between these states (Gilmore 1981, van der Maas 1992). The model's functions combine to generate a bifurcated catastrophe "response surface" in a three-dimensional phase space defined by the variables injustice, opportunity, and identity. The surface identifies a set of possible psychological states and pathways of movement between these states in a representative individual.

In section 6, I discuss the model's internal operation and the individual's possible decision logic at the point of the dehumanization catastrophe. I conclude with suggestions of possible avenues for testing the model. An appendix considers alternative model parameterizations and outlines some empirical questions these alternatives raise.

## 2. THEORETICAL FOUNDATIONS AND GENERAL ASSUMPTIONS

I offer this paper's formal model as a hypothesis to explain sudden dehumanization at the individual level, specifically dehumanization involving a sharp discontinuity between mental states of identification and non-identification. It addresses the following question: Given what we currently know about human conflict psychology generally and dehumanization specifically, what is a plausible and simple set of causal mechanisms that could produce such a discontinuity?

Although not a direct product of inductive empiricism, my proposed model nonetheless reflects intuitions developed over several decades of study of—and reflection on—the nature and dynamics of human conflict. Indeed, I start with several key variables and causal propositions that recur in the field of conflict studies. To explain the apparent dehumanization discontinuity, I use critical transition theory (Scheffer, 2009), which offers a powerful mathematical tool for representing and modeling discontinuous change. This tool guides my representation of the conflict variables and propositions as formal axioms. My model is thus consistent with both our existing knowledge of violent conflict and the mathematics of discontinuous change.

By representing my key assumptions about conflict and dehumanization mathematically, I show how their conjunction can explain the apparent sharp nonlinearity of dehumanization that conventional scholarship does not examine. Together, the assumptions, mathematical functions, and resulting catastrophe response surface offer a general psychological theory of the nonlinear collapse of interpersonal identification at the level of the individual.

Some may find the model too simple, especially in its underlying assumptions; others may find it too complex, especially in its number of intermediate variables. Any model of this kind necessarily strikes a somewhat arbitrary balance between, on the one hand, the criteria of transparency and explanatory power, both of which tend to encourage simplicity, and on the other, the criteria of theoretical grounding and empirical validity, both of which tend to encourage complexity. Within the present model, though, one can shift the balance in either direction, if desired, by relaxing for instance certain assumptions to allow for more complexity or stripping out some variables to highlight the core feedback that produces sharp nonlinearity.

Other mechanisms may cause nonlinear dehumanization, other than the one this model identifies. Dehumanization, as a general type of psychological phenomenon, almost certainly encompasses a number of subtypes. Each subtype is likely the product of one or more distinct generalized causal mechanisms involving specific sets of causal factors. In other words, dehumanization as a broad phenomenon is probably the equifinal outcome of multiple causal pathways, of which my model's pathway may be one.[8] I would argue, however, that this pathway is common—perhaps even dominant—and therefore critically important.

## 2.1 Theoretical Foundations: Injustice, Opportunity, and Identity

The model explicitly integrates three factors—injustice, opportunity, and identity—mentioned repeatedly in theories of human conflict across disciplines as diverse as psychology, economics, political science, international relations, and social psychology. Different disciplinary approaches to conflict, however, often emphasize one or other of these factors. For instance, economic or rational-choice theories tend to emphasize changes in perceived opportunity as a cause of conflict, while social psychological theories often focus on changes in group identity as either a cause or consequence of conflict.

A theoretical emphasis on one of these three factors often entails specific ontological commitments, especially to a particular conception of the essential nature of human beings. A focus on grievances arising from perceived injustice privileges a conception of human beings as *moral evaluators* of the fairness of their own and others' circumstances; a focus on perceived opportunity as a cause of conflict, in contrast, encourages conflict theorists to see human beings as *cost-benefit analyzers* akin to calculating machines; and lastly, a focus on identity leads theorists to see human beings primarily as recursive *identity constructors* both of their own and others' identities.

These ontological commitments anchor three quite distinct paradigmatic approaches to understanding human conflict that can seem, at first glance, largely incommensurable. But the different commitments are not intrinsically incompatible; in principle, for instance, human beings can be moral evaluators, cost-benefit analyzers, and identity constructors simultaneously. By explicitly integrating injustice, opportunity, and identity in the present model, I show that the ontological commitments commonly associated with these factors are complementary and that much theoretical and empirical leverage can be gained from exploiting these complementarities.

## 2.2 Integrating Hot and Cold Cognition

Of particular significance, in this regard, is the model's integration of "hot cognition" and "cold cognition" approaches to explaining the psychological dynamics that lead to conflict. Hot-cognition approaches generally focus on the causal role of emotions such as anger, fear, and greed as motives for actors to engage in conflict[9]; whereas cold-cognition approaches generally focus on the causal role of actors' non-emotional (or "rational") assessments of the costs and benefits of different strategies to achieve their goals, given their social situation's prevailing structural *opportunities*.[10] The present model bridges this gulf: it identifies specific causal mechanisms through which cost-benefit assessments produce emotional responses that in turn drive dehumanization.[11]

In the model, the key emotions are anger and fear.[12] Indeed, the extent to which an individual dehumanizes other people is a direct consequence of the sum of that individual's anger towards those other people and fear of them. These are what Hall and Ross (2015) call "high-intensity" affective responses to changes in situational variables of critical concern to an individual (Frijda, 1988). Although the model focuses on these two emotional pathways to dehumanization, in the real world other pathways—involving, for instance, contempt or disgust—likely operate too. Also, it is important to note that dehumanization need not be highly emotional. Haslam (2006) has distinguished between "animalistic" and "mechanistic" dehumanization, where the latter is characterized by "indifference" and "disregard." The present model describes a possible psychological mechanism producing what Haslam terms animalistic dehumanization.

## 2.3 General Assumptions: The Individual Psychological State and Processes

The model assumes that the individual whose psychological state is being modelled is representative—that is, they exhibit psychological characteristics close to the mean of the human population generally. In particular, the individual does not exhibit unusual psychological dispositions or traits, such as extreme aggressiveness, authoritarianism, or disgust sensitivity.

The model also assumes that the individual whose psychological state is being modelled identifies with or "humanizes" other people in his or her normal or rest mental state. People are naturally inclined to identify with others (Baron-Cohen, 2011). The mental state of dehumanization is therefore not a psychological starting point but rather a psychological outcome open to causal explanation.

## 2.4 General Assumptions: The Individual and the Surrounding Population

The model distinguishes between the individual whose psychological state is being modelled and the population of people surrounding that individual. It assumes this population's size substantially exceeds the maximum number of people that someone can know individually well enough to accurately anticipate their specific behaviors and maintain stable relations with them. The anthropologist and evolutionary psychologist Robin Dunbar has famously identified this limit for modern humans as 150 people, although researchers have questioned the empirical basis for this specific figure (Lindenfors, 2021). Nonetheless, that a limit exists—probably in the low- to mid-hundreds—is not in doubt; and when an individual interacts with a substantially larger number of people, they have to apply generalized categories and properties (often pejoratively referred to as "stereotypes") to at least some of those people (Berreby, 2005: 219-222).[13]

The model also assumes that the individual in question has had no prior interaction with any members of the surrounding population and therefore has no prior knowledge of those members, other than that they exist and are human beings. In other words, the model as presented here is ahistorical, although I will show later how memory and history could be represented within it.

Additionally, the surrounding population's members exhibit no easily detectable characteristics that could be the basis for simple categorization, and the individual being modeled cannot access a cultural schema or script (Petersen, 2002: 63-64) that shows how to categorize other members of the population and specifies a repertoire of possible actions towards those categories. So, in the individual's mind, at the outset of the psychological processes the model represents, the members of the surrounding population are wholly undifferentiated.

Together, these might be called the model's "tabula rasa" assumptions. Social-psychological research suggests that even in such circumstances, people are still highly susceptible to categorization that differentiates ingroup from outgroup and that they easily engage in discrimination that favors their ingroup after such categorization (Billig and Tajfel, 1973; Berreby, 2005: 205-211). Even then, though, their perceptions of the outgroup will generally not be highly pejorative—that is, they will not extend to dehumanizing the outgroup.

Finally, the individual being modeled is assumed to actively engage in a cognitive process of assessing his or her material and social circumstances and those of members of the surrounding population. The individual then draws on a previously internalized set of values and beliefs (discussed below) to evaluate these circumstances and then arrive at judgements about prevailing levels of injustice and opportunity.

For purposes of analytical tractability, the model assumes these judgments about injustice and opportunity are the sole determinants of the psychological process of dehumanization. Importantly, therefore, the model in its simplest form represents the individual as socially passive: it does not account for the effects on dehumanization processes—at both the individual and group levels—of reciprocal interaction from the individual to other members of the population and then back from those members to the individual.[14] Such interactions, of course, are potentially enormously important; they could, for instance, enhance group polarization and in turn further raise the likelihood and severity of dehumanization.[15] In section 5, I discuss one way these processes could be incorporated within a more complex and generalized form of the model.

## 2.5 Application of Catastrophe Mathematics

System modelers often call a marked discontinuity in a particular system's behavior a "catastrophe."[16] Two subfields of complexity theory—catastrophe theory and critical-transitions theory (Thom, 1975; Zeeman, 1976, 1977; Scheffer, 2009)—have developed mathematical approaches to understand and explain such discontinuities. Both approaches generate a graphical surface, often called a "response surface," in a specified state space. The surface represents the full set of possible system states and shows the values of the system's state variables at the points in the space where the discontinuity occurs.

The two approaches differ, however, in how their mathematical models describe and reproduce a discontinuity. The approach grounded in catastrophe theory, which I call "behavioralistic," generally specifies a single equation incorporating the system's independent and dependent variables, which are usually externally observable system behaviors.[17] This equation generates a response surface directly. Although the relationships between the independent and dependent variables may be causal, the precise mechanisms connecting them are not specified. The internal workings of the system are "black boxed."

In contrast, the approach grounded in critical-transitions theory, which I call "mechanistic," generally specifies a set of functions representing the system's internal causal dynamics or mechanisms. These functions mathematically describe the causal relationships between changes in the values of the system's independent variables, changes in the values of a set of precisely specified intermediate variables, and changes in the value of the specified dependent variable. The interaction of these functions produces a response surface indirectly.

The two approaches are not equivalent. In particular, the mechanistic approach can produce a response surface with a shape that a single-equation behavioralistic approach cannot easily generate. Also, while a mechanistic approach can generate response surfaces exhibiting sharp discontinuities, these surfaces may not have some of the features, such as hysteresis and singularities, of the catastrophe surfaces that behavioralistic approaches conventionally produce.

In the present analysis, I use critical-transition theory to ground a mechanistic model of the psychology of dehumanization.

## 3. THE MODEL'S CORE: FOUR CAUSAL MECHANISMS

At the core of the formal model proposed here are four hypothesized causal mechanisms.

The first mechanism is the dual causal effect of changes in perceived opportunity ("Opp"), which is a person's assessment of how much members of the surrounding population can exercise independent agency, on both "attributed responsibility" ("AR") and perceived "capacity to harm" ("CtH"). Changes in an individual's perception of the opportunity available to other population members will affect the degree to which that individual attributes responsibility for events in the surrounding environment to those other members (Opp→AR, see Figure 1). Also, changes in the individual's perception of opportunity will affect their assessment of other population members' capacity to harm them (Opp→CtH).

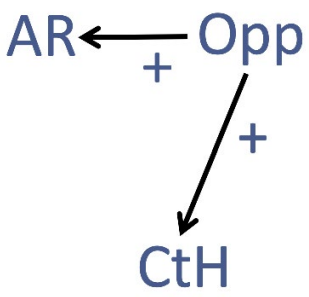



**Figure 1: The Dual Causal Effect of Perceived Opportunity on Attributed Responsibility and on Perceived Capacity to Harm**

The second hypothesized causal mechanism is the dual effect of changes in the product of perceived injustice ("InJ") and attributed responsibility—a product term here referred to as "attributed injustice"—on both an individual's level of anger towards other members of the population ("A") and their distrust ("DT") of those other members (see Figure 2).

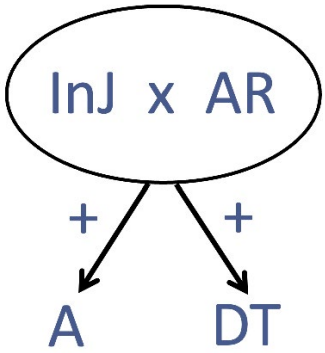



**Figure 2: The Dual Causal Effect of Attributed Injustice on Anger and on Distrust**

The third hypothesized causal mechanism is the critical feedback relationship between the product of distrust and capacity to harm—referred to as "expected harm"—and an individual's level of fear (see Figure 3).

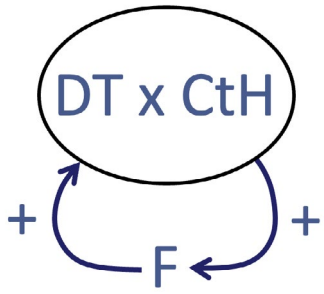



**Figure 3: The Positive (Self-Reinforcing) Feedback between Expected Harm and Fear**

The model proposes that a change in an individual's assessment of expected harm changes the individual's degree of fear ("F"); this change in fear then affects the individual's estimate of expected harm, by changing both the individual's degree of distrust of other members of the population and the individual's perception of those other members' capacity to harm.

And finally, the fourth key hypothesized causal mechanism is the additive impact of changes in a person's levels of anger and fear on that person's degree of identification ("I") with other members of the population. Identification declines in direct proportion to the sum of the individual's anger and fear (see Figure 4).

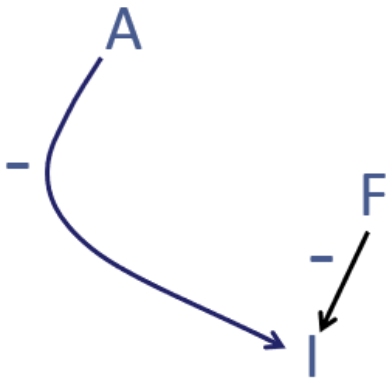



**Figure 4: The Additive Impact of Anger and Fear on Level of Identification**

These four mechanisms are integrated into the causal model shown in Figure 5. The following sections outline the empirical rationale for this model, represent the causal relationships in Figure 5 as axioms, implement the axioms mathematically, and show how the integrated set of mathematical functions produces a discontinuity between the mental states of identification and non-identification.

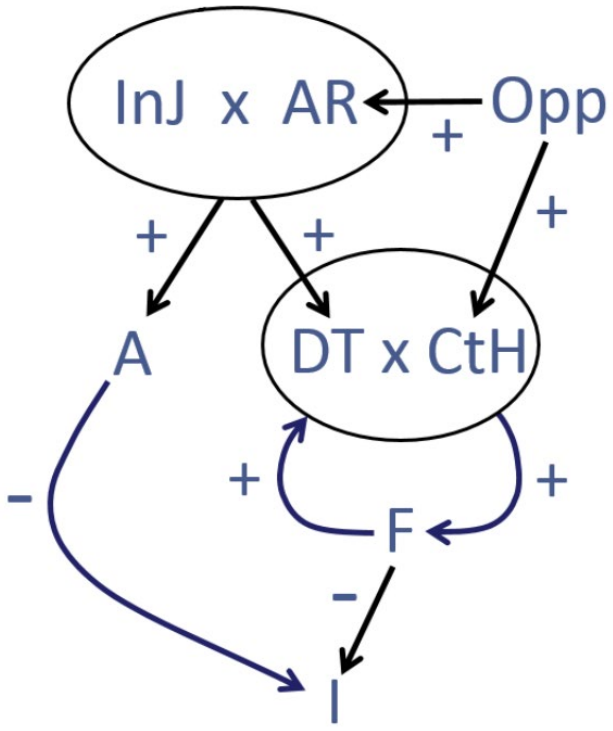



**Figure 5: The Core Model of Catastrophic Dehumanization**

## 4. SPECIFIC ASSUMPTIONS, AXIOMS, AND VARIABLE DEFINITIONS

The model of dehumanization proposed here implements mathematically a set of specific assumptions about the dehumanization process. For the purposes of the model, these assumptions serve as formal axioms. In this section, I state the axioms and provide definitions of key variables.

*Axiom 1: Change in identity (I) is a consequence of the interaction of perceived injustice (InJ) and perceived opportunity (Opp).*

Axiom 1 identifies one dependent variable, identity, and two independent variables, injustice and opportunity.

Dependent variable:

> **Identity (I)**: An individual's perception of the degree to which the population surrounding him or her is divided into an ingroup (of which the individual is a member) and an outgroup, and the degree to which, if the population is so divided, the individual regards the outgroup negatively.

Identity varies between inclusive/tolerant and exclusive/antagonistic.[18] At the inclusive/tolerant end of the scale, the individual perceives the ingroup boundary to encompass (or include) everyone in the surrounding population, including himself or herself.[19] In other words, the individual's referent for "we" is the entire population, and he or she shares the identity of the other population's members. The individual perceives no outgroup—or "they"—and can easily identify with any member of the surrounding population.

As the value of identity moves away from the inclusive/tolerant end of the scale towards the exclusive/antagonistic end, the individual begins to categorize people as either ingroup or outgroup members. In other words, he or she starts to exclude some members of the population from the ingroup. The individual also begins to link his or her personal identity to an emerging ingroup identity. Yet the outgroup is still only partially defined: the criteria for outgroup membership remain unclear, the ingroup-outgroup boundary is permeable, and the perceived identity of the outgroup is still fluid.

As movement from the inclusive/tolerant end continues, the distinction between ingroup and outgroup becomes sharper in the individual's mind. He or she believes the groups' membership criteria are largely unambiguous, perceives their identities to be clearer and more permanent, and explicitly designates the groups by "we" and "they." Yet the individual may still have a positive-to-neutral perception of the outgroup and its members. Also, he or she can still individuate outgroup members, when attention is focused on specific individuals, and still regards them as part of his or her moral community.

Once the individual has made the jump to the exclusive/antagonistic end of the scale, the individual perceives the population to be divided decisively into an ingroup and an outgroup. Each member of the population must be in one group or the other. The individual has a severely negative perception of the outgroup. He or she deindividuates and caricatures members of the outgroup and does not regard them as participants in his or her moral community. The individual does not identify with any outgroup members. Dehumanization has occurred.[20]

Independent variables:

> **Injustice (InJ)**: The individual's assessment of the degree of injustice of 1. the individual's own situation, 2. the situation of others in the population with whom the individual identifies, and/or 3. the actions of others in the population towards the individual or towards people with whom the individual identifies.

The individual's subjective norms of justice, discussed below, determine his or her expectations of what people deserve or what they ought to do, and the individual's normative assessment of a given situation or action is a function of the gap between what he or she expects based on these norms and what he or she perceives is actually happening.

> **Opportunity (Opp)**: The individual's assessment of the degree to which members of the surrounding population can act independently.

If opportunity is low, the individual perceives that other members of the population have limited latitude to act independently, because of severe structural constraints, low intrinsic capabilities, or both. If opportunity is high, he or she perceives that structural constraints are weak enough, and the other members' intrinsic capabilities are great enough, to allow those members wide latitude to act independently. Structural constraints include factors such as the presence of police officers, the rule of law, and normative sanctions against violence; intrinsic capabilities are influenced by the factors such as the availability of weapons and financial resources.

The model assumes that the individual perceives the degree of opportunity to be homogenous across all the surrounding population's members. However, its account of dehumanization can be adjusted to accommodate a perception of differential opportunity.

The remaining axioms specify the causal mechanisms leading from change in the independent variables to change in the dependent variable.

*Axiom 2: Change in opportunity (Opp) positively correlates with change in attributed responsibility (AR).*

Intervening variable:

> **Attributed responsibility (AR)**: The individual's assessment of the degree to which other members of the population are responsible for events in the surrounding environment.

The model proposes that change in an individual's perception of the opportunity of other members of the population has two simultaneous consequences. First, as specified in Axiom 2 above, it changes the degree to which the individual attributes responsibility for events in the surrounding environment to those other members of the population (Opp→AR). Second, as specified in Axiom 4 below, it affects the individual's assessment of other members' capacity to harm him or her (Opp→CtH); in this latter causal role, it therefore mediates the degree to which the individual's distrust of other members of the population (a variable defined below) makes him or her afraid. These two consequences are anchored in different temporal perceptions: the Opp→AR relationship arises from the individual's interpretation of events in the past or present, whereas the Opp→CtH relationship involves the individual's interpretation of possibilities in the future.[21]

With regards to the first of the two consequences (Opp→AR), perceived opportunity is assumed to affect attribution of responsibility because people generally understand that the capacity to act independently is a necessary condition for responsibility. Specifically, people generally will not hold another person responsible for an event, if they believe the person could not have acted independently to produce that event.[22]

The model permits experimentation with different functional forms representing the relationship between perceived opportunity and attribution of responsibility (Opp→AR). The simplest functional form is linear and proportional: as opportunity increases, attribution increases by the same ratio. But if, for instance, the modeler wants to represent an individual who is susceptible to the fundamental error of attribution, the Opp→AR function can be adjusted to produce a much more rapid and nonlinear increase in attribution. I discuss this matter further in this paper's appendix.

*Axiom 3: Change in the product of injustice and attributed responsibility (InJ x AR), referred to as **attributed injustice**, positively correlates with both change in anger (A) and distrust (DT).*

Intervening variables:

> **Anger (A)**: The individual's level of anger towards other members of the population in response to attributed injustice.
>
> **Distrust (DT)**: The individual's level of distrust of other members of the population.

Here is an example of this model's integration of "hot" and "cold" cognition approaches to explaining the psychology of conflict. The product of injustice and attributed responsibility (InJ x AR) has its effect on identity along two causal pathways. Along the first, the product term (InJ x AR), which is "attributed injustice," affects the individual's degree of anger, an emotion, which in turn negatively affects identity.[23] Along the second, the product term affects the individual's level of distrust, a non-emotional assessment of other actors' inclinations.

Trust is a mental attitude of one individual towards another. An individual's degree of trust in another is proportional to his or her estimate of the other's willingness to bear direct costs or to forgo benefits in order to protect that individual's interests. Distrust, therefore, is proportional to the individual's estimate of the other's willingness to seek benefit by acting against that individual's interests. Distrust is most extreme when the individual perceives this willingness to be high and at the same time perceives that the other's benefits are directly proportional to the individual's costs. An individual defines his or her interests within a local normative framework and moral economy that he or she may perceive to be institutionalized as the shared "rules of the game."

In this model, distrust is equivalent to a probability assessment. It captures the individual's answer to the question: What is the likelihood that other members of the population will harm me if they have the capacity to do so?[24]

*Axiom 4: Change in opportunity (Opp) positively correlates with change in capacity to harm (CtH).*

Intervening variable:

> **Capacity to harm (CtH)**: The individual's assessment of the capacity of other members of the population to inflict harm on him or her.

Capacity to harm is a joint function of the individual's assessment of the other members' capabilities and of his or her own capabilities. For instance, all things being equal, the higher the individual's assessment of his or her own defensive capability, the lower his or her assessment of other members' capacity to harm. The model assumes that the individual's defensive preparations remain low and constant up to any decision not to identify, so any change in capacity to harm (CtH) during this period is driven by perceived change in opportunity (Opp).

*Axiom 5: Change in the product of distrust and capacity to harm (DT x CtH), referred to as **expected harm (EH)**, positively correlates with change in fear (F).*

Intervening variable:

> **Fear (F)**: The individual's level of fear of other members of the population.

By influencing capacity to harm (CtH), opportunity (Opp) mediates the relationship between distrust (DT) and fear (F), because distrust of another person is insufficient by itself to produce fear of that person. *Distrust* must be accompanied by a judgment that the distrusted person has the capacity to inflict harm (a direct function of opportunity, by axiom 4). Even high levels of distrust produce little fear if the individual believes the distrusted actor(s) cannot actually inflict harm. In this model, therefore, distrust and capacity to harm are individually necessary and jointly sufficient for fear.

The model regards these two causes of fear as two elements of an expected utility calculation, where expected (dis)utility is the product of the probability of a given outcome and the estimated (dis)utility of that outcome, should it come to pass. In this case, the individual's estimate of expected harm is the product of the individual's estimate of the probability that other members of the population will cause him or her harm if they have the capacity to do so (indicated by the individual's level of distrust) and the individual's estimate of the capacity of those members to inflict harm (a function of the perceived opportunity of those other members).[25]

When the second component of the expected-harm calculation—the perceived capacity to harm and therefore the disutility of that harm, should it happen—remains small, the estimate of expected harm itself is likely to remain small and fear low. However, as opportunity increases and the perceived capacity to harm rises, this second component becomes proportionately more powerful.[26]

*Axiom 6: The positive correlation between expected harm (EH) and fear (F) is nonlinear and sigmoid in form.*

The model uses a Hill function to represent the EH→F relationship. This sigmoid curve has its steepest slope at intermediate values of expected harm. This functional form reflects an assumption that, as rising expected harm crosses a threshold value, the individual becomes more reactive or sensitive to this variable, with the result that his or her fear rises more rapidly.

*Axiom 7: Change in fear (F) positively correlates with change in expected harm (EH); this correlation is linear.*

*Axiom 8: The functions describing the EH→F and F→EH relationships intersect at three points.*

The conjunction of axioms 5 through 8 creates a particular kind of feedback relationship between expected harm (EH) and fear (F). I call this the "EH/F feedback." An increase in expected harm produces fear; this greater fear then raises the individual's estimate of expected harm, because it causes both greater distrust of other members of the population and a perception that their capacity to harm has increased.[28]

In the model, these causal relationships have specific functional forms that together generate three equilibria—one unstable and two stable—that in turn create the discontinuity between mental states of identification and non-identification (see Figure 8 below and the accompanying explanation). This is the mathematical heart of the formal model. The functional forms are derived from critical transition theory (Scheffer 2009). I discuss in the appendix the particular parameter settings needed to generate the shape of the EH→F sigmoid curve that the model requires to create the discontinuity. These parameter settings, like the model as a whole, should be regarded as hypotheses that need empirical confirmation.

*Axiom 9: Change in identity (I) is negatively correlated with changes in anger (A) and fear (F).*

The model assumes that the individual, in his or her rest state, identifies with or humanizes members of the surrounding population. So in the absence of anger (A) towards those other people or fear (F) of them, identity (I) remains at its maximum level. Identity declines in direct proportion to the sum of the individual's anger and fear.[29]

## 5. THE FULL MODEL OF CATASTROPHIC DEHUMANIZATION

Figure 6 is a causal diagram of the full catastrophic dehumanization model. The "core" model , which we have already seen in Figure 5, consists of everything inside the red dotted rectangle, specifically the model's independent, intermediate, and dependent variables and the causal relationships between these variables, as identified in axioms 1 through 9 above. The product term "InJ x AR" in the top oval is referred to as "attributed injustice"; while the product term "DT x CtH" in the bottom oval is called "expected harm" (EH). The four variables across the top—MSC, TofJ, ExCon, and EnCap—are the model's contextual inputs. Those inputs and the feedback loop between I and MSC on the left were not included in Figure 5.

The core model explains the relationship between an individual's perceptions of injustice and opportunity and the emotions anger and fear; change in the level of these emotions in turn affects the degree to which the individual identifies with other members of the population.

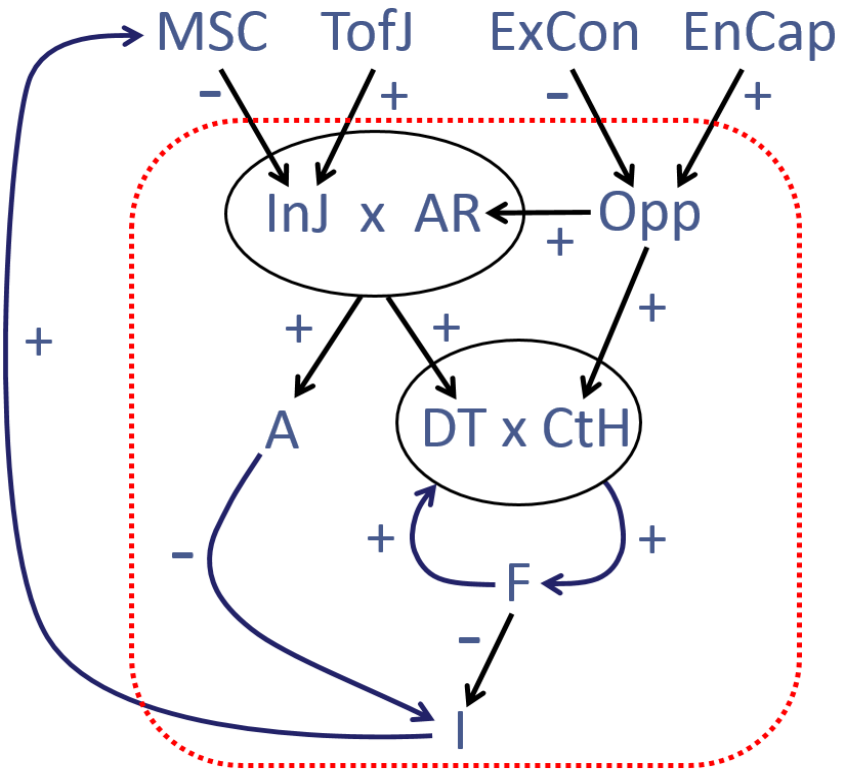



**Figure 6: The Full Model of Catastrophic Dehumanization**

## 5.1 Model Contextual Inputs

To complete the model conceptually, I propose four discrete inputs into the model's independent variables. These inputs are not included in the model's axioms or mathematical implementation.

**Material and social circumstances (MSC)**: The individual's assessment of the material and social environment surrounding him or her.

**Theory of justice (TofJ)**: The individual's normative framework, which may be incomplete or even incoherent, but which provides a basis for the individual's evaluation of the fairness of his/her or others' situations or actions.

**Exogenous constraint (ExCon)**: The individual's assessment of the degree of structural constraint, both material and social, affecting members of the surrounding population.

**Endogenous capability (EnCap)**: The individual's assessment of the intrinsic capability or power, both material and social, of members of the surrounding population.

These inputs constitute the broader material, political, institutional, cultural, and ideological context of the dehumanization process. Material conditions—ranging from the laws of thermodynamics to the availability of natural resources—act as ultimate constraints on this context's plasticity. Within these constraints, the political context of dehumanization can sometimes change fast, as when a sudden redistribution of social resources, such as an election that gives a particular group control of the state, markedly shifts power relations between individuals or between groups. Institutional, cultural, and ideological change, however, is generally slower. Political entrepreneurs or elites wishing to influence processes within the core model and ultimately change the value of identity (I) must exploit opportunities present within this material, social, and ideational context.[30]

Figure 6 shows proposed positive and negative correlations between the four inputs and the model's two independent variables, injustice (InJ) and opportunity (Opp). A deterioration in the individual's material and social circumstances (MSC), such as decreased food availability, is proposed to produce an increase in perceived injustice (InJ). The individual's theory of justice (TofJ) might be measured, in part, by his or her "justice sensitivity" (or "moral demandingness"), in which case a strengthening of this sensitivity will also lead to higher perceived injustice.[31] From the individual's point of view, perceived weakening of structural constraints facing members of the surrounding population (ExCon) raises those members' opportunity (Opp) as does perceived increases of those members' intrinsic power (EnCap).

Although the model as presented here is ahistorical, it need not be so. Culturally or ideologically embedded beliefs about, or memories of, past intergroup relations can be represented or encoded within the model's four external inputs, especially material and social circumstances (MSC) and theory of justice (TofJ), from where they can powerfully shape perceptions of injustice and opportunity. Particular group identities and their meanings, a history of intergroup hostility, or institutional ascriptions of rights or penalties to specific groups will influence, for instance, whether the modeled individual, in his or her normal or rest mental state, "humanizes" members of the surrounding population and, if not, which people in that population the individual dehumanizes.

## 5.2 The Identity Response Surface

Figure 7 shows one possible set of values of the variables injustice, opportunity, and identity within a coordinate space where each axis represents one of these variable's full range of possible values. The x axis from left to right represents injustice (InJ); the z axis from back to front represents opportunity (Opp); and the y axis from bottom to top represents Identity (I).

The set of values creates a surface in this coordinate space; this surface is bifurcated—or split into two parts. The core model always generates this bifurcated surface, but the particular positions and configurations of the two parts shown in Figure 7 reflect assumptions regarding the structure and parameters of the model's core functions (see "Appendix: Implementation Assumptions").

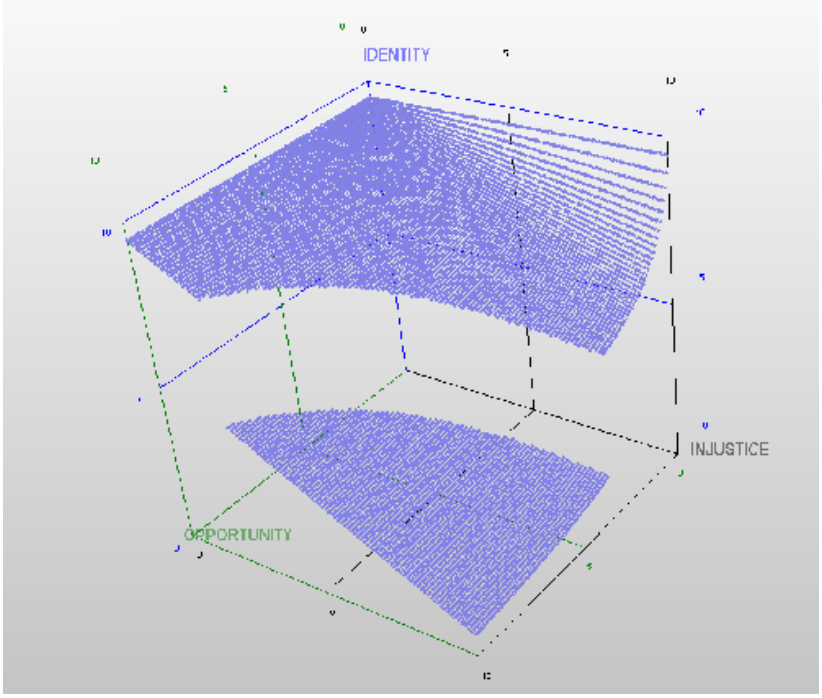



**Figure 7: A Representative Identity Response Surface**

Together, the surface's two parts represent a hypothesized relationship between the three variables of injustice, opportunity, and identity for a single individual. Any point on the surface is a possible psychological state for this individual. Because injustice and opportunity are the model's independent variables, while identity is its dependent variable, the value of identity at any point on the surface is the joint product of the values of injustice and opportunity at that point. Using the standard language of catastrophe theory, identity is the *behavior variable*, and injustice and opportunity are *control variables*. The two-dimensional plane at the bottom of the figure defined by the control variables is the *control surface*, and the surface itself is the *response surface*. Thus the model generates what I call an "identity response surface."

Figure 7's identity response surface shows discontinuous change in the behavior variable, identity (I), given certain changes in the values of injustice (InJ) and opportunity (Opp). Along the *bifurcation line* where this discontinuous change occurs—that is, where the response surface splits—a range of values of the behavior variable (I) becomes inaccessible, regardless of the values of injustice and opportunity.

I call the upper part of the identity response surface the *humanization plane*. As the value of perceived opportunity increases from back to front, this plane's width (in the injustice dimension, from left to right) starts to diminish, because the individual's rising estimate of the capacity of others in the population to inflict harm makes him or her more sensitive to perceived injustice. As the value of perceived injustice rises from left to right, the plane's depth (in the opportunity dimension) also starts to diminish, because rising injustice increases the individual's distrust of others in the population, which makes him or her more sensitive to perceived opportunity. The lower part of the response surface, which I call the *dehumanization plane*, declines in a slope from the bifurcation line towards an exclusive and antagonistic identity in the bottom right edge of the cube. The lowest point of this plane (the corner nearest to the viewer in Figure 7) represents a mental state of full dehumanization of others in the population.

### 5.3 Feedback and the Bifurcation of the Identity Response Surface

The bifurcation of the identity response surface arises from feedback between the mathematical functions describing the EH→F and F→EH relationships. Recall that the individual's assessment of expected harm (EH) is the product of the individual's distrust of other members of the population (DT) and his or her estimate of their capacity to harm him or her (CtH). Expected harm therefore consists of the contents of the lower "DTxCtH" oval in Figure 6. In the model, expected harm (the entire contents of the oval) influences fear (EH→F). This relationship, I propose, is highly nonlinear and best represented by a Hill function that produces a curve of the form shown in Figure 8. The reverse relationship (F→EH) is linear and is also shown in Figure 8. (For this latter function, the independent variable F is on the figure's y or vertical axis.)

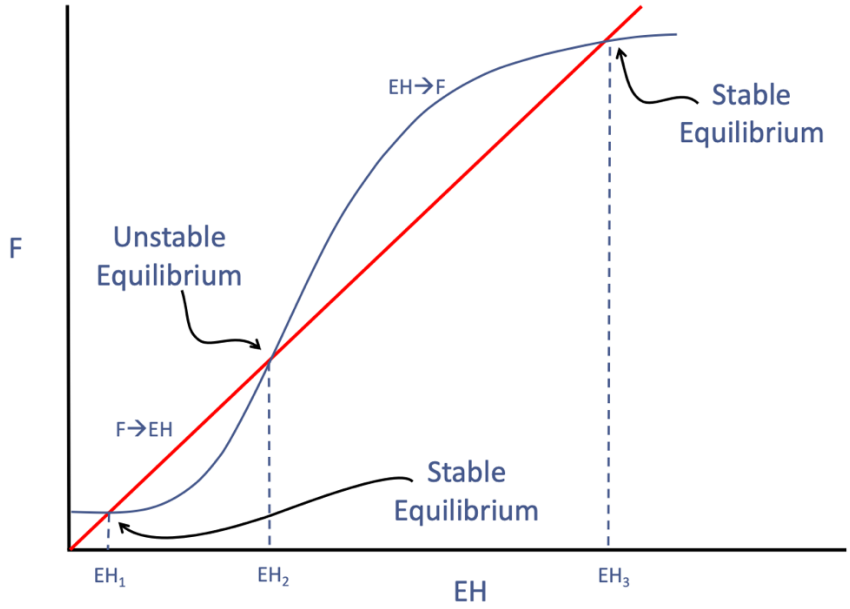



**Figure 8: Feedback Interaction between EH→F and F→EH Relationships**

The two curves intersect at three points creating two stable equilibria at $EH_1$ and $EH_3$ and one unstable equilibrium at $EH_2$. The iterative feedback between the two functions drives paired values of EH and F that are to the left of $EH_2$ to the equilibrium at $EH_1$, while it drives paired values of EH and F that are to the right of $EH_2$ to the equilibrium at $EH_3$, as shown in Figure 9. Through fear's effect on identity in the model, these two stable equilibria in turn bifurcate the identity response surface into two discrete surfaces—the humanization and dehumanization planes—with a catastrophic nonlinearity in the value of identity (I) for the particular values of InJ and Opp along the bifurcation line.

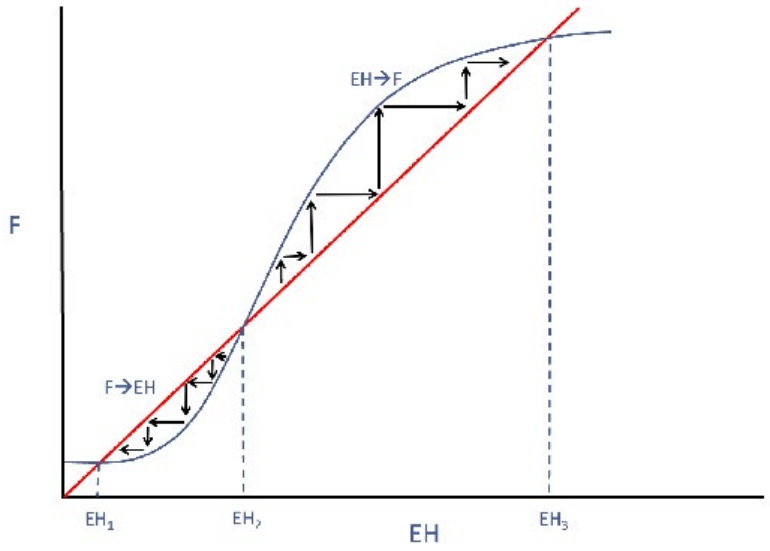



**Figure 9: Iterative Feedback Drives EH/F Values to Stable Equilibria**

## 5.4 Absence of Hysteresis in the Identity Response Surface

The mechanistic model used here does not generate a conventionally folded catastrophe response surface, so it does not produce hysteresis. In other words, the two parts of the identity response surface do not overlap for a subset of values of InJ and Opp such that each InJ/Opp pair in this subset is associated with two possible values of I. As a result, within the core model, a jump from one plane to another is reversible along the same route in the coordinate space.

I argue that this result reasonably represents the psychological dynamics of dehumanization at the *individual* level. It might be possible to reverse the discontinuous change in a single person's mental state from identification to non-identification—that is, to reverse the jump from the humanization to the dehumanization plane—by simply reversing the changes in perceived injustice and opportunity that produced the discontinuity in the first place.

But when dehumanization occurs across a large enough number of people who are interacting socially, it will likely produce hysteresis, or the impossibility of simple reversibility, at the group or *sociological* level. The interactive effect of many instances of individual-level dehumanization among communicating members of a group will likely harden an antagonistic "we-they" group identity. Then specific beliefs, norms, practices (including group defensive preparations), institutions and distributions of power will congeal around this antagonistic identity, justifying and reinforcing it.

If this process happens on each side of a conflict, group members on each side can then use the other side's dehumanization (and associated behaviors) to justify their own dehumanization (Kteily, Hodson, and Bruneau, 2016). Intergroup conflict thus becomes entrenched, sharply restricting reversibility and increasing the likelihood that both populations will be caught in a dehumanization trap.

In the full model shown in Figure 6, I capture this possibility and the resulting hysteresis by including a causal feedback from identity (I) to the individual's material and social circumstances (MSC), an input that lies outside the model's core. The development of entrenched, institutionalized, and interdependent antagonistic group identities—a situation in which each side's positive group self-esteem depends upon cultivation of a negative image of the other side—is a large change in the individual's broader material and social circumstances. Other feedbacks from changes in identity to model inputs are possible. For instance, entrenched and interdependent antagonistic group identities could shift perceptions of the structural constraints on, and the intrinsic power of, other groups in the surrounding population, affecting both the ExCon and EnCap inputs.

Such changes in dehumanization's broader social context—changes that involve, fundamentally, the breakdown of community sentiments—are *constitutive* of dehumanization, in the sense that they involve multiple, deep, and usually relatively slow alterations in the shared meanings or intensionality through which actors construct and understand their social environments. In contrast, the processes I represent in this paper's core model—occurring within a single individual and understandable, ultimately, as involving fast shifts in cognitive and brain states relating to interpersonal identification—are more strictly *causal*. The psychological outcome of de-identification that my core model explains might, indeed, be a necessary condition for broader constitutive changes that entrench intergroup antagonism and dehumanization at the sociological level.

## 5.5 Limit Hypotheses

In Figure 7, the proposed identity response surface intersects the cube's four sides. The points of intersection on a given side represent a hypothesized relationship between one of the control variables (injustice, InJ, or opportunity, Opp) and the behavior variable (identity, I) with the other control variable held constant at an extreme or "limit," value. Each of the four sides therefore depicts a "limit hypothesis" for the relationships between the control variables and the behavior variable.

Figure 10a shows the hypothesized relationship between injustice and identity when opportunity is maximally low—that is, when the individual perceives that others in the population have little or no capacity to act independently—represented by the intersection of the response surface with the back of the cube in Figure 7. In this case, change in the value of injustice does not produce change in the value of identity, which remains constant at the inclusive and tolerant end of the range of identity's values.

My reasoning here is straightforward: when an individual perceives that other members of the population have very little or no capacity to act independently, he or she will not attribute responsibility for circumstances perceived to be unjust to any member of that population. Instead, the individual will attribute the circumstances to fate or forces beyond any person's control. So rising injustice will not influence identity, which remains inclusive and tolerant.

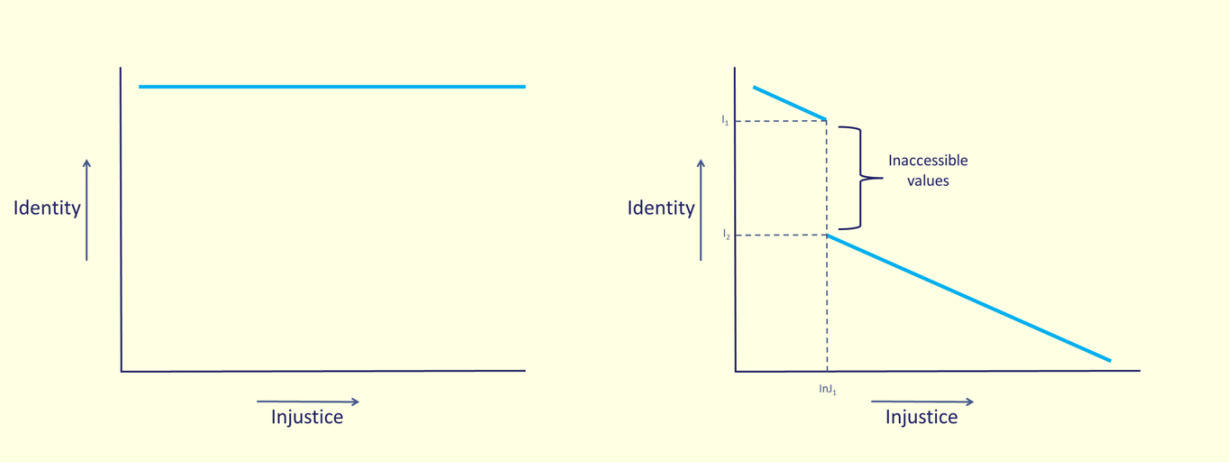
Figure 10b shows the contrasting hypothesized relationship between injustice and identity when opportunity is maximally high—that is, when an individual believes that others in the population have extremely wide latitude to act independently—represented by the intersection of the identity response surface with the front of the cube in Figure 7. In this case, a relatively small increase in injustice from the zero point on the left produces a discontinuous change from an inclusive and tolerant identity to a value in the lower-middle of the range of identity's values. The point on the response surface that represents the individual's psychological state falls off the humanization plane at $InJ_1$ and drops from $I_1$ to $I_2$, with the values of identity between $I_1$ and $I_2$ inaccessible.

Because opportunity is maximally high and therefore the perceived latitude of others to act independently is very large, responsibility for any amount of perceived injustice is immediately and fully attributed to others, which produces both anger towards those others and distrust of them. The rising anger causes the decline in identity up to $InJ_1$. High opportunity also means high perceived capacity to harm; this capacity to harm interacts with distrust to generate fear, in turn engaging the EH/F feedback loop. When injustice and therefore distrust become sufficiently high at $InJ_1$, this feedback loop causes discontinuous change in the value of identity from $I_1$ to $I_2$. As injustice continues to increase beyond $InJ_1$, anger likewise increases, and the value of identity declines towards the variable's exclusive and antagonistic extreme value.

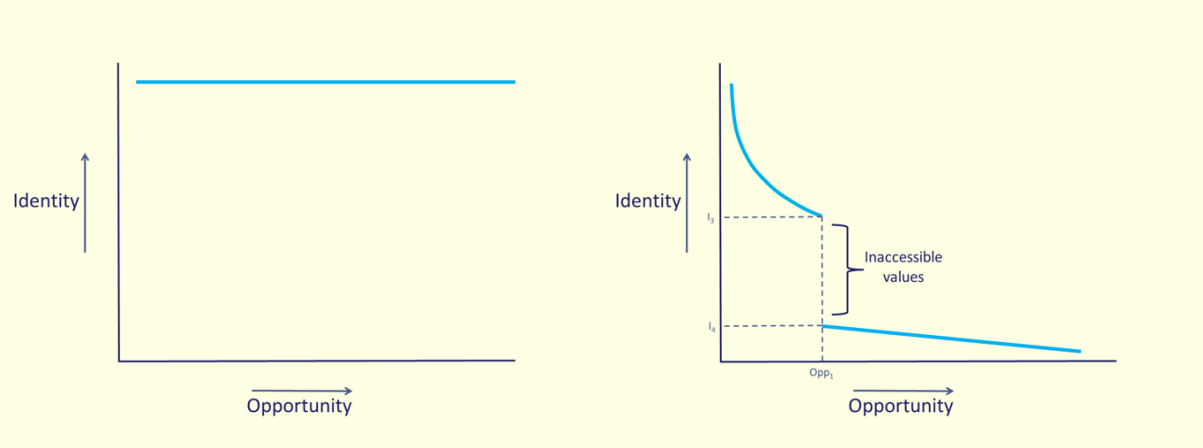
Figures 10c and 10d show the hypothesized relationships between opportunity and identity when injustice is maximally low and high respectively. When injustice is maximally low (represented in Figure 10c, the cube's left side in Figure 7, when viewed from outside the cube), changes in opportunity produce no anger or distrust, so identity stays inclusive and tolerant across all values of opportunity.

But when injustice is maximally high (in 10d, the cube's right side viewed from inside the cube itself), rising opportunity at first produces anger, which induces a rapid but not discontinuous decline in identity to a value in the middle of the variable's range. The curvilinear shape of this decline—at first very rapid and then slow—reflects an assumption, embedded in the parameter settings of this particular implementation of the model, that the individual is susceptible to the fundamental error of attribution; as opportunity rises, he or she therefore attributes responsibility for the high level of perceived injustice very rapidly, which in turn quickly increases anger. Rising opportunity also increases perceived capacity to harm, which interacts with already high distrust to generate fear, once again engaging the EH/F feedback loop. When opportunity and therefore capacity to harm become sufficiently high at $Opp_1$, this feedback loop causes discontinuous change in the value of identity from $I_3$ to $I_4$, a value near the variable's exclusive and antagonistic extreme.

The phenomenon of catastrophic dehumanization is most obvious at this limit, when opportunity is maximally high, because the collapse of identity from $I_3$ to $I_4$ takes the individual immediately from a moderate degree of de-identification with other members of the population to an extreme degree.




**Figures 10a and 10b: Relationship between Injustice and Identity (when opportunity is maximally low and high, respectively)**




**Figures 10c and 10d: Relationship between Opportunity and Identity (when injustice is maximally low and high, respectively)**

## 5.6 Sequencing of Anger and Fear

The contrast between Figures 10b and 10d is instructive. In the former, where opportunity is held maximally high but injustice starts out low and increases, the discontinuity in identity happens relatively early in the change in InJ's value. Then, as InJ's value continues to increase, identity's value steadily declines to an extreme low. This sequence of changes in identity's value arises because fear's impact largely precedes anger's. With opportunity at its maximum, the EH/F feedback causes the discontinuity in identity's value early, and then as injustice increases, rising anger continues to drive down identity's value.

In contrast, in Figure 10d, where injustice is held extremely high but opportunity starts out low and increases, anger's impact largely precedes fear's. Because the parameter settings in this particular implementation of the model reflect an assumption that attributed responsibility rises very rapidly with opportunity, increasing anger first causes identity's rapid decline to the middle of the variable's range. Only subsequently does the EH/F feedback engage to cause the discontinuity in identity's value.

## 6. DECISION LOGIC AT BIFURCATION

Given the model's underlying logic, what rational-choice processes might explain the shift from the stable equilibrium that creates the humanization plane to the stable equilibrium that creates the dehumanization plane?

Such a shift, I hypothesize, can be understood as a product of the individual's use of minimax expected regret or worst-case avoidance reasoning.[32] I can illustrate this reasoning and the resulting shift with a somewhat stylized description of the change in the individual's mental state along a diagonal trajectory, as shown in Figure 11, from low to progressively higher levels of perceived injustice and opportunity, first across the humanization plane and then, after crossing the bifurcation line, down to the dehumanization plane.

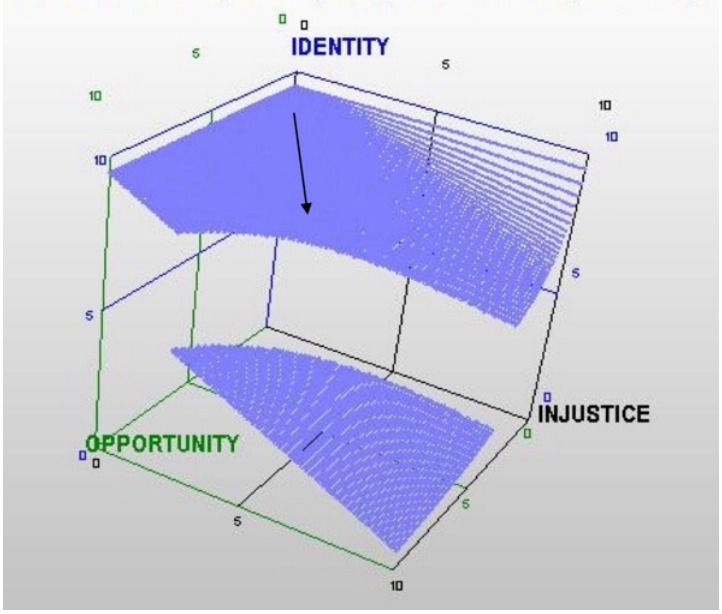



**Figure 11: Illustrative Trajectory across Humanization Plane**

Starting in the upper-rear corner of the coordinate space, as perceived injustice and opportunity rise, anger starts to rise. The individual also starts to scan the surrounding population to identify people or subgroups to blame for the perceived injustice, and he or she starts to divide the population into those perceived as potentially responsible for the injustice and those perceived not responsible. In other words, he or she begins to identify members of a potentially blameworthy "they" group or outgroup that is distinct and excluded from a blameless "we" group or ingroup.[33] This psychological change is represented by the continuous decline in identity as the trajectory proceeds towards the bifurcation line.[34]

As the individual's psychological state approaches the line, he or she faces an increasingly clear choice about how to act towards the newly distinguished outgroup: he or she must choose between identifying with or not identifying with this group. The decision process producing this choice is likely largely unconscious and heavily influenced by emotions like anger and fear, but it represents a real choice nonetheless.

Despite the fact that the individual's level of distrust of the newly defined outgroup—that is, his or her estimate of the probability that the outgroup is untrustworthy—is rising, a choice to identify with the outgroup is a choice to act towards the outgroup *as if* it were trustworthy. By identifying with the outgroup, the individual willfully leaves himself or herself potentially vulnerable to the outgroup's actions, which implies trust. Conversely, a choice to not identify with the outgroup is a choice to act towards the outgroup *as if* it were untrustworthy; it psychologically legitimizes defensive preparations to reduce vulnerability, implying distrust.

Whichever way the individual decides, he or she faces a risk of being wrong. Identifying with the outgroup, by leaving the individual vulnerable, might bring substantial costs in the form of harm, should the outgroup turn out to be untrustworthy. On the other hand, not identifying with the outgroup could also bring substantial costs in the form of unnecessary defensive preparations and lost community, should the outgroup turn out to be trustworthy.

I argue that the individual will choose the option that minimizes the expected cost of being wrong—in other words, his or her expected regret. The key issue is whether the expected cost of being wrong about identifying with the outgroup is less than or greater than the expected cost of being wrong about not identifying with the outgroup. The bifurcation line marks the point at which the ratio of these respective costs shifts from less than one to greater than one.

Figure 12 shows that the individual's decision to identify or not identify with the outgroup has four possible discrete outcomes: two outcomes should the individual decide to identify and two should he or she decide not to identify. For each of these four outcomes, the figure shows the individual's expected-value calculation. This calculation consists of the individual's estimate of the probability of that outcome multiplied by the sum of his or her estimate of the costs and benefits of that outcome should it come to pass.

The probabilities associated with each outcome are the individual's estimates of the likelihood of the outgroup's trustworthiness or untrustworthiness. Because the individual arrives at these estimates prior to making the decision about whether or not to identify with the outgroup, I assume the probabilities are the same across the two decision paths. In other words, the individual's estimated probability for outcome 1 is the same as that for outcome 3, and his or her estimated probability for outcome 2 is the same as that for outcome 4. I also assume, importantly, that trustworthiness is a binary variable—that the outgroup is either trustworthy or not—so the estimated probabilities of outcomes 1 and 2 must add up to 100 percent as must those of outcomes 3 and 4.

We can introduce two simplifications to Figure 12's rather complicated picture. First, we can reasonably assume that the estimated costs and benefits marked by asterisks will be near zero. For instance, the individual will probably think that he or she will bear little cost from identifying with the outgroup if it is actually trustworthy. Similarly, the individual will probably think that he or she will accrue little benefit from identifying with the outgroup if it is actually untrustworthy. Second, the estimated benefits labeled A and D are likely roughly equivalent in the individual's mind. Both sets of benefits arise from accurately judging the trustworthiness of the outgroup and thereby being better able to maintain something approximating the status quo.[35]

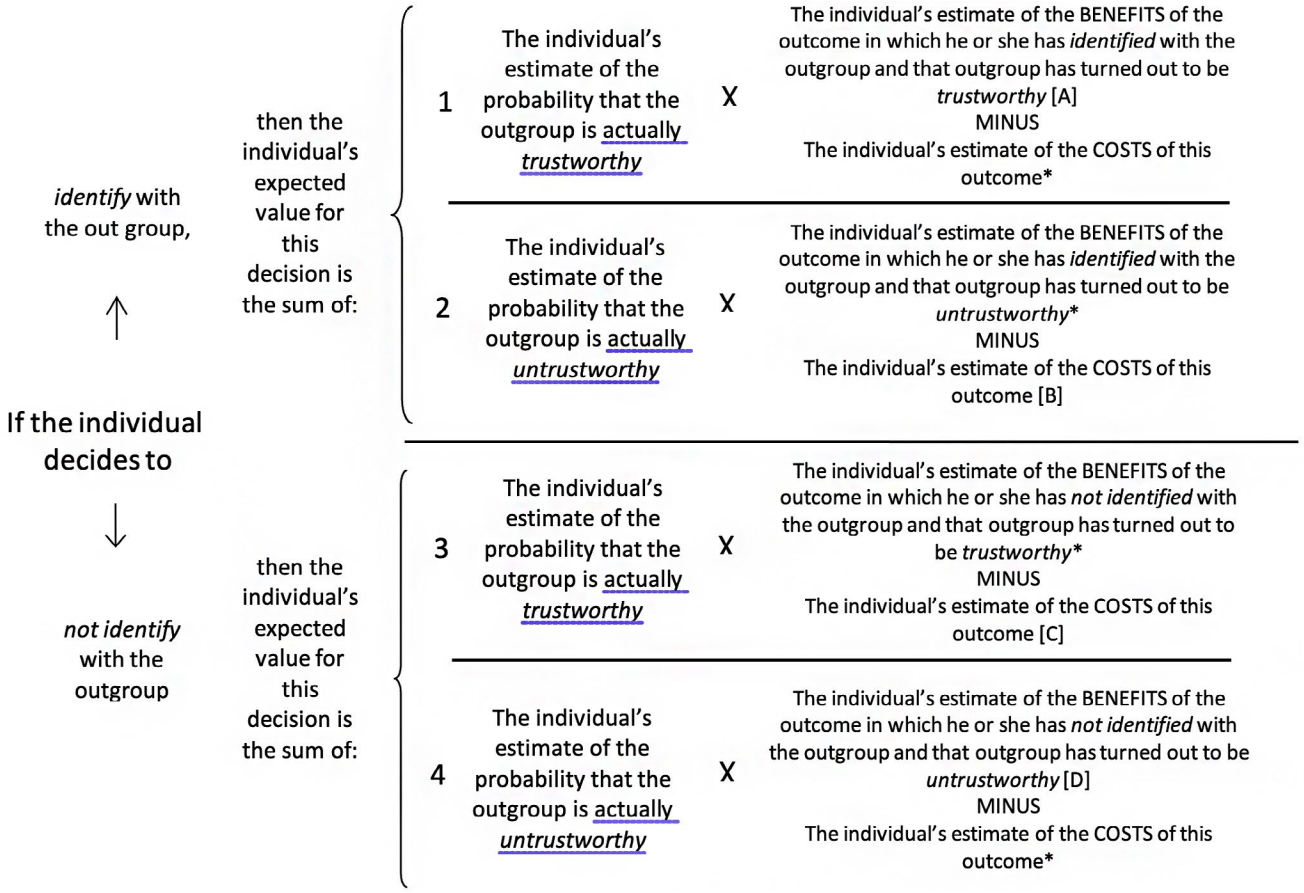



**Figure 12: The Individual's Expected Value Calculations at the Bifurcation Line**

Thus of the eight benefits and costs listed across the four outcomes, we can assume that four (marked by asterisks) have negligible effect on the individual's judgment and two (A and D) have differential effects on this judgment only through their interaction with the estimated probabilities in their respective outcomes. Costs B and C, I therefore argue, have a particularly important influence on the individual's expected value calculations.

Cost B is an element of the expected value calculation for outcome 2, and this outcome is an example of what statisticians call a Type II error. Cost C is an element of the expected value calculation for outcome 3, which is an example of a Type I error. Type I error arises when a true null hypothesis is rejected, and Type II error arises when a false null hypothesis is not rejected. On the humanization plane, the individual has the null or baseline hypothesis that the outgroup is trustworthy. In outcome 3 in Figure 12, the individual rejects a true null hypothesis, and thus makes a Type I error, when he or she concludes the outgroup is untrustworthy when the group is actually trustworthy. In outcome 2, the individual fails to reject a false null hypothesis, and makes a Type II error, when he or she concludes the outgroup is trustworthy when it's actually untrustworthy.

On the humanization plane approaching the bifurcation line in Figure 11, increasing injustice and opportunity combine to increase distrust and therefore to increase commensurately the individual's estimate of the probability that the outgroup is untrustworthy. This is the probability in the expected value calculations for outcomes 2 and 4. An equivalent *decrease* takes place in the individual's estimate of the probability that the outgroup is *trustworthy*, which is the probability in the expected value calculations for outcomes 1 and 3.

Simultaneously, increasing opportunity increases the outgroup's perceived capacity to harm and therefore the individual's estimate of the *costs* of the outcome in which he or she has identified with the outgroup but the outgroup has turned out to be untrustworthy (B in Figure 12). Put simply, higher opportunity makes the outgroup potentially more dangerous, which in turn makes identifying with it more costly should it turn out to be untrustworthy.

These changes in turn affect the individual's overall estimates of the expected values of identifying with the outgroup or not identifying with it. Prior to reaching the bifurcation line, the individual's estimate of the expected value of identifying with the outgroup exceeds his or her estimate of the expected value of not identifying with it. *The line represents the point at which the order reverses*. Now the expected value of not identifying exceeds that of identifying. More specifically, the individual's estimate of the expected cost of Type II error (the probability that the outgroup is untrustworthy multiplied by cost B in outcome 2) starts to exceed his or her estimate of the expected cost of Type I error (the probability that the outgroup is trustworthy multiplied by cost C in outcome 3). This change, I argue, triggers the sudden flip in the individual's psychological state and the discontinuity in identity.

## 6.1 Reconciling the Two Explanatory Accounts

The above account is simply a working hypothesis and remains largely speculative. It provides a rational-choice or cold-cognition explanation of the discontinuity in mental state that the mathematical model explains by reference to both hot and cold psychological mechanisms.

These two explanations are entirely compatible and in fact complement each other. The rational-choice account unpacks the model's expected-harm calculation, showing that it can be interpreted using the logic of minimax expected regret. This logic is grounded in a possible-worlds ontology, because it involves a comparison of the expected gains and losses arising under four potential outcomes. The discontinuity occurs only when the individual's expected value of not identifying exceeds that of identifying.

Within the model, this threshold corresponds to the particular value of expected harm, $EH_2$, at the unstable equilibrium in Figure 8. Below this point, the EH/F feedback dampens both fear and expected harm to produce a stable equilibrium at $EH_1$ where he or she feels secure and therefore identifies with others in the population. Above $EH_2$, the feedback accentuates both fear and expected harm to produce a stable equilibrium at $EH_3$ where he or she feels insecure and therefore de-identifies with others in the population.

Unlike the rational-choice explanation, the model thus allows for the causal role of emotions: within the EH/F feedback loop, the individual's assessment of expected harm affects his or her fear, and this fear affects his or her assessment of expected harm.[36] More generally, cold cognition affects hot emotion, and hot emotion provides data for cold cognition. The psychological manifestation of the change in value of the ratio of the expected costs of being wrong from less than one to greater than one is akin to a Gestalt shift—a flip from a cognitive-emotional state of identification to one of non-identification.

## 7. CONCLUDING REMARKS

Despite its manifold complexities, the formal model of catastrophic dehumanization I propose here remains rudimentary. Yet even at this early stage of development, it offers theoretical insight and suggests multiple avenues for further research and testing.

Three theoretical insights are of particular note. First and most generally, the model shows how hot and cold cognition can be integrated into one causal account of a particular psychological phenomenon, with neither subordinate to the other. Second and more specifically, it shows how a positive (self-reinforcing) feedback between a rational calculation and an emotional state can produce multiple discrete and stable psychological equilibria, and how shifts between these equilibria can then account for observed sharp discontinuities in psychological state. Third and finally, by integrating three key factors—injustice, opportunity, and identity—that theories of human conflict usually considered separately and within different disciplinary domains, the model provides a starting point for development of a unified and interdisciplinary theory of conflict psychology.

With regards to further research and testing, researchers could investigate, through close examination of historical cases or structured field research, the extent to which perceived injustice and opportunity are precursors to anger and fear and, in turn, the extent to which anger and fear are precursors, at least on some occasions, to dehumanization. In laboratory settings, researchers could conduct more fine-grained tests of the model's specific hypothesized causal relationships—for instance, between opportunity and attribution of injustice; between attribution of injustice and distrust; and, vitally, between expected harm (as defined within the model) and fear. (Of course, when undertaking such work, researchers should strictly observe ethics protocols, because the psychological manipulations that testing this model would likely require could, if performed without care, profoundly traumatize experimental subjects.)

The model's explanation of the dehumanization discontinuity critically hinges not only on the existence of the EH/F causal feedback, but also on the particular characteristics of that feedback. The discontinuity requires the existence of stable and unstable psychological equilibria, and the model's functions must have particular forms and parameters to produce such equilibria. Whether or not they do is an empirical question. It should be possible, for example, to investigate in the laboratory whether changes in expected harm produced the hypothesized nonlinear, sigmoid-curve change in fear, whether fear has a reciprocal and linear relationship on expected harm, and whether an identity discontinuity, should it occur, is the product of some form of minimax regret reasoning.

Even prior to such investigations, however, the proposed model should stimulate new thinking about the nature and causes of dehumanization, a psychological phenomenon of enormous consequence to human well-being this century and beyond.

# APPENDIX:

## IMPLEMENTATION ASSUMPTIONS: FUNCTIONAL FORMS AND PARAMETERS

Each of the core model's causal relations, shown as arrows in Figure 5 (reproduced below), can be represented by a mathematical function. Together, the functions' forms and parameters reflect certain assumptions about the relations between the model's variables. Here I identify the main assumptions, illustrate how to adjust them within the model, and discuss some consequences of these adjustments.

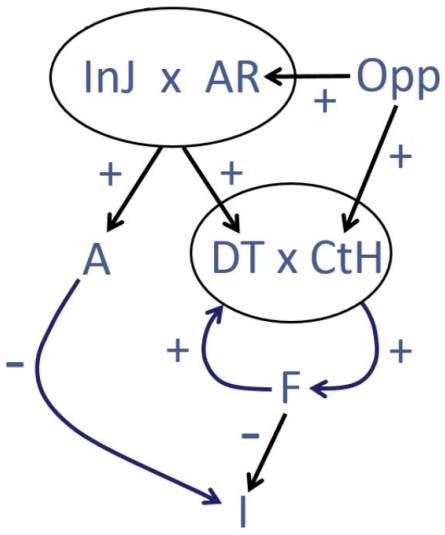



**Figure 5: The Core Model of Catastrophic Dehumanization** (reproduced from main text)

In all the implementations of the model discussed here, the values of the independent and dependent variables (InJ, Opp, and I) vary between a minimum of 0 and a maximum of 10. Starting with the causal relations at the top of figure 5 and working to the bottom, the following are the model's basic functions, reflecting some very simple assumptions about the psychological dynamics of dehumanization.

1. AR = Opp/10
2. A = (InJ∗AR)/2
3. DT = InJ∗AR/10
4. CtH = Opp
5. $EH_{T0}$ = DT∗CtH
6. $F_{T0}$ = $1+(8*(EH_{T0})^4/(K4+(EHT0)^4)$, with K=4.0
7. $EH_{T+1}$ (feedback) = $F_{T0}$
8. $F_{T+1}$ = $1+(8*(EH_{T+1})^4/(K^4+(EHT_{+1})^4)$, with K=4.0

(iterate functions 7 and 8, N times)

9. I = $10-(A+(F_{T+N}/2))$

In function 1 above, attributed responsibility (AR) is a coefficient between 0 and 1 that varies in direct proportion to change in opportunity (Opp). So a value of 5 for Opp produces a value of 0.5 for AR, which means, in the model's context, that the individual attributes to other members of the population responsibility for half of perceived injustice (InJ). The strict proportionality between Opp and AR reflects the assumption that the individual is not susceptible to the fundamental error of attribution.

In function 2, attributed responsibility (AR) then multiplies perceived injustice (InJ) to generate anger (A). Combining functions 2 and 9, anger can directly reduce the value of identity (I) up to 5 units, if InJ is at its maximum value of 10 and AR is at its maximum value of 1. In other words, within the model, change in the individual's level of anger can, by itself, reduce identity by half its maximum value.

The remaining functions, 3 through 8, concern the feedback relationship between expected harm (EH) and fear (F). Like AR, distrust (DT) is a coefficient that varies between 0 and 1; as discussed in the main body of this paper, the model interprets distrust as the individual's estimate of the probability that other members of the population are untrustworthy, which is assumed to be equivalent to his or her estimate of the probability that those other members will harm him or her if they have the capacity to do so. The simple linear relationships between injustice and distrust and between opportunity and distrust (via AR) reflect the assumptions that if the individual perceives no injustice, or if he or she perceives that other members of the population have no opportunity, then he or she will feel no distrust towards those other members.

The value of capacity to harm (CtH) is a simple linear function of opportunity (Opp), on the assumption that the individual's defensive capability is low and stays constant over all values of injustice and opportunity. Thus, as the perceived opportunity of other members of the population rises, so does, in direct proportion, their perceived capacity to harm the individual (CtH). Expected harm (EH), as discussed in the main body of the paper, is a function of the individual's estimate of the probability that other members will inflict harm (DT) multiplied by the perceived capacity of those members to inflict harm (CtH).

Expected harm (EH) then produces fear (F) in an amount determined by a Hill function of the form:

$$f(s) = \frac{V_{\max} s^n}{(K_m)^n + s^n}$$

Where expected harm (EH) is S in the above function; $V_{max}$ is a parameter that establishes the maximum possible range of values of the dependent variable fear (F); K is a parameter that determines the shape of the function's sigmoid curve; and the exponent n determines the range of values of EH over which the curve changes from its minimum to maximum values of F.

If indeed human beings exhibit a psychological relationship between expected harm and fear (EH→F), then the best mathematical representation of this relationship—including its functional form and parameter settings—is an empirical question. The functional form and parameter settings may in fact differ across individuals or across groups, depending on their beliefs, histories, or cultures. As this model is a hypothesis about the cause of discontinuous dehumanization, in all implementations of the model, I have used functional forms and parameters that create the equilibria the hypothesis requires.

Specifically, I set $V_{max}$ to 8 and the intercept in the Hill function to 1, so the value of fear (F) varies between 1 and 9. An intercept value of 1 reflects an assumption that the individual is never entirely free of fear. Even when expected harm is 0, he or she still experiences fear, albeit at a very low level.

I also set the values of the K parameter and the exponent *n* to 4. The result is the curve shown in the Figure 13 below. The shape of this curve ensures that the feedback function F→EH ($EH_{T+1} = F_{T0}$), which is linear at the origin, intersects the Hill function cleanly in three locations, corresponding to the two stable and one unstable equilibria described in the main paper (and illustrated in Figure 8).

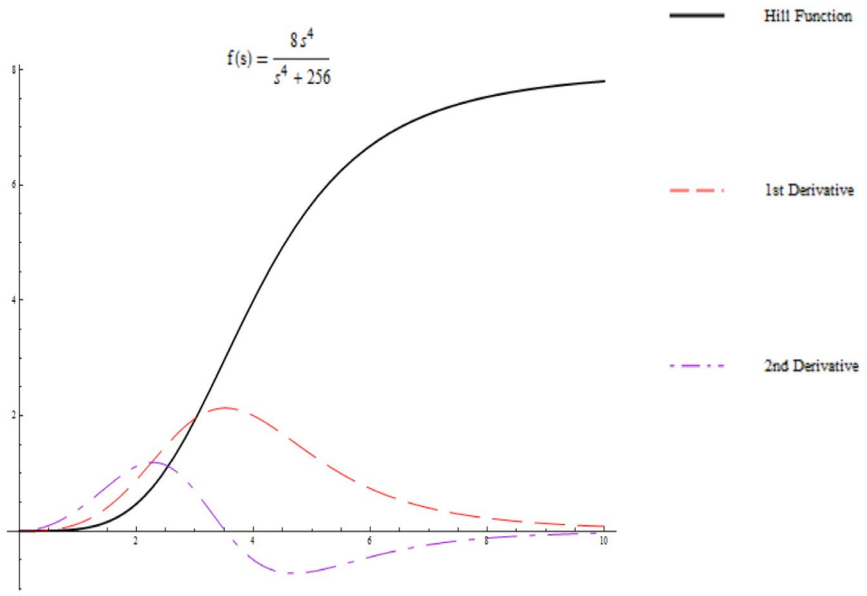



**Figure 13: The Hill function applied in the model**

The "T0" time subscripts that appear in functions 5 and 6 indicate that these two functions establish the initial or "time zero" values for EH and F. The other time subscripts in functions 7 and 8 indicate that these two functions are iterated N times in the feedback loop.

Implementing the above basic functions (in an Excel spreadsheet, for instance) for values of the model's independent variables between 0 and 10 at 0.1 increments produces 10,000 discrete data points. Running the feedback loop through 18 iterations then generates the following identity response surface composed of the 10,000 data points:

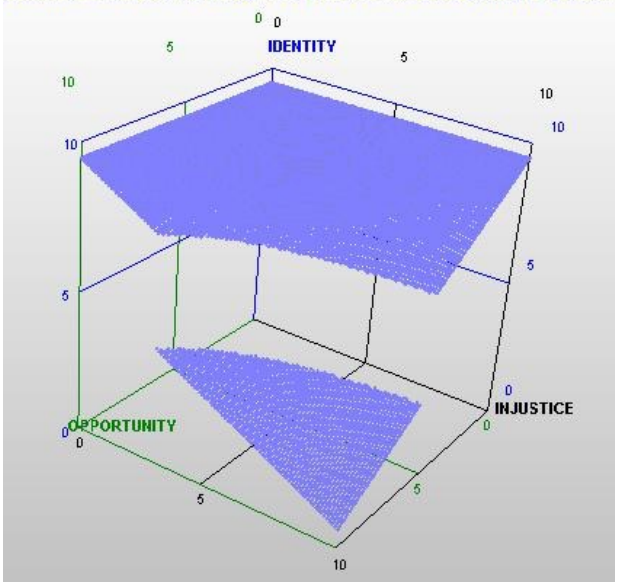



**Figure 14: Identity Response Surface Produced, Basic Functions**

This response surface is noticeably different from that shown in Figure 7, which I have reproduced again immediately below: most significantly, the humanization plane is larger, the bifurcation line is straighter, and the decline in identity at maximum injustice, as opportunity increases through its lower values, is neither curvilinear nor as rapid.

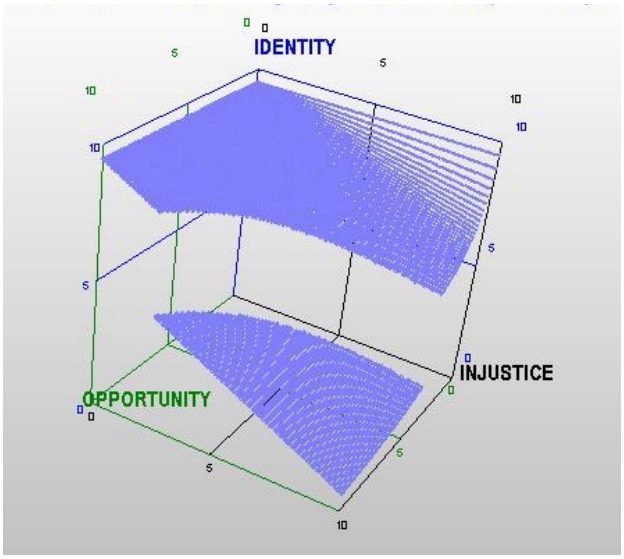



**Figure 7: Identity Response Surface, Adjusted Functions** (reproduced from main text)

Figure 7's identity response surface is generated by the following adjusted functions. The adjustments reflect, I argue, somewhat more realistic assumptions about the psychological dynamics of dehumanization.

1. AR = $(12-(10*(Opp + 0.76)^{-2/3}))/10$
2. A = $(InJ*AR)/2$
3. DT = $0.1+(InJ*AR*0.09)$
4. CtH = Opp
5. $EH_{T0}$ = $DT*CtH$
6. $F_{T0}$ = $1+(8*(EH_{T0})^4/(K^4+(EH_{T0})^4)$, with K=4.0
7. $EH_{T+1}$ (feedback) = $F_{T0}$
8. $F_{T+1}$ = $1+(8*(EH_{T+1})^4/(K^4+(EH_{T+1})^4)$, with K=4.0

(iterate functions 7 and 8, N times)

9. I = $10-(A+(F_{T+N}/2))$

Functions 1 and 3 above (highlighted) are different from those in the previous set. The new function 1, which produces the curve below, reflects an assumption that the individual is highly susceptible to the fundamental error of attribution. As perceived opportunity increases, the individual's attribution of responsibility rises rapidly. At a value for Opp of 1.0, for instance, AR is about 0.5, which means that the individual already attributes to other members of the population responsibility about half of perceived injustice.

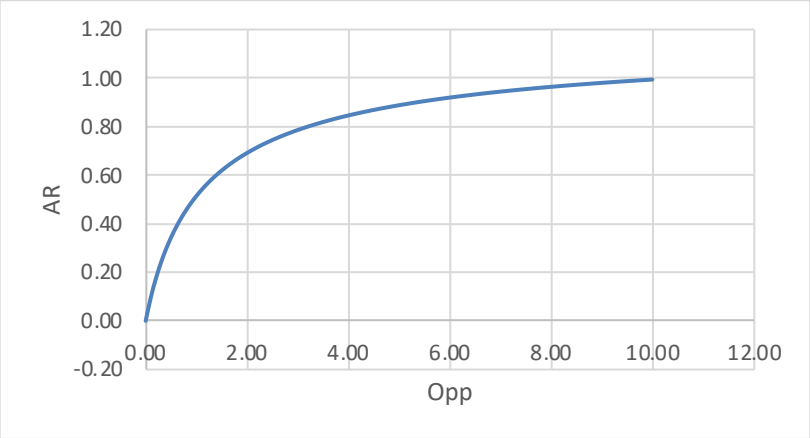



**Figure 15: Functional Relationship between Opportunity and Attributed Responsibility, Reflecting Susceptibility to the Fundamental Error of Attribution**

When combined with high levels of perceived injustice (InJ), rapidly rising AR generates rapidly rising anger, which in turn (by functions 2 and 9) rapidly decreases identity (I). The result is the swift curvilinear decline in identity prior to the discontinuity that can be seen at maximum injustice on the right side of Figure 7's identity response surface and also in the limit hypothesis shown in Figure 10d in the main body of the paper. Function 1 can be adjusted to reflect a range of alternative assumptions about the individual's attribution behavior.

The adjusted function 3 above reflects an assumption that, even in the absence of attributed injustice, the individual exhibits some distrust towards other members of the population. I have set this irreducible or "base" level of distrust (DT) at a low level, at ten percent (or 0.1) of distrust's potential range between 0 and 1.0. The InJ x AR product determines how much additional distrust the individual exhibits within the remaining 90 percent of DT's potential range.

By manipulating the base level of distrust (DT), the modeler can adjust the individual's sensitivity to rising opportunity at low levels of injustice (InJ)—that is, along the left side of the cube. For instance, if base-level distrust is set at one-third of distrust's potential range (or at 0.33) in function 3, as follows:

3. DT = 0.33+(InJ*AR*0.067)

and if all the other functions in the above set remain the same, then the bifurcation line intersects the left side of the cube, as seen below.

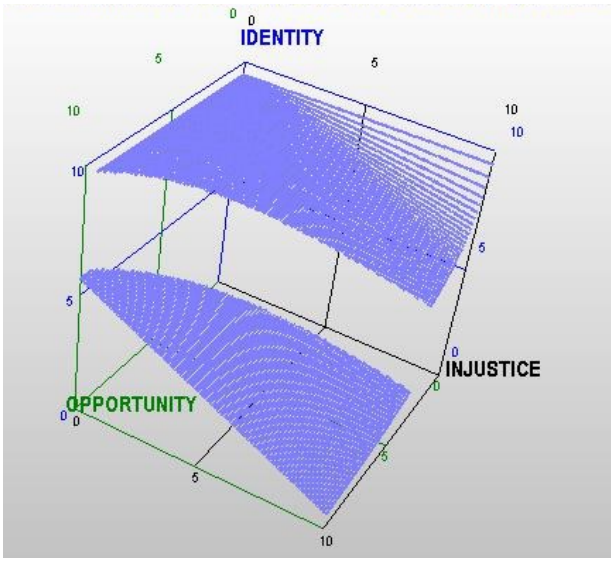



**Figure 16: Identity Response Surface, Adjusted Distrust**

Now rising opportunity can by itself cause a discontinuity in identity. In other words, if the base level of distrust is high, a phenomenon not uncommon in some cultures, perceived opportunity can generate sufficient fear to produce a collapse of identity; perceived injustice is unnecessary.

Finally, depending on modelers' assumptions about people's propensity to dehumanize others, they may wish to manipulate the relative dimensions of the entire humanization and dehumanization planes. If, for instance, the modeler thinks the identity discontinuity can occur more easily, at generally lower levels of injustice and opportunity, than I have proposed in Figure 7, then the bifurcation line can be moved backwards in the cube so that the humanization plane shrinks in size.

Changes in the relative sizes of the humanization and dehumanization planes can be accomplished by moving either or both of the EH→F and F→EH curves in Figure 8 (shown in the main body of the paper) to the right or left. Below I illustrate shifts to the left of both curves (1.0 units for the EH→F curve and 0.75 units for the F→EH curve).

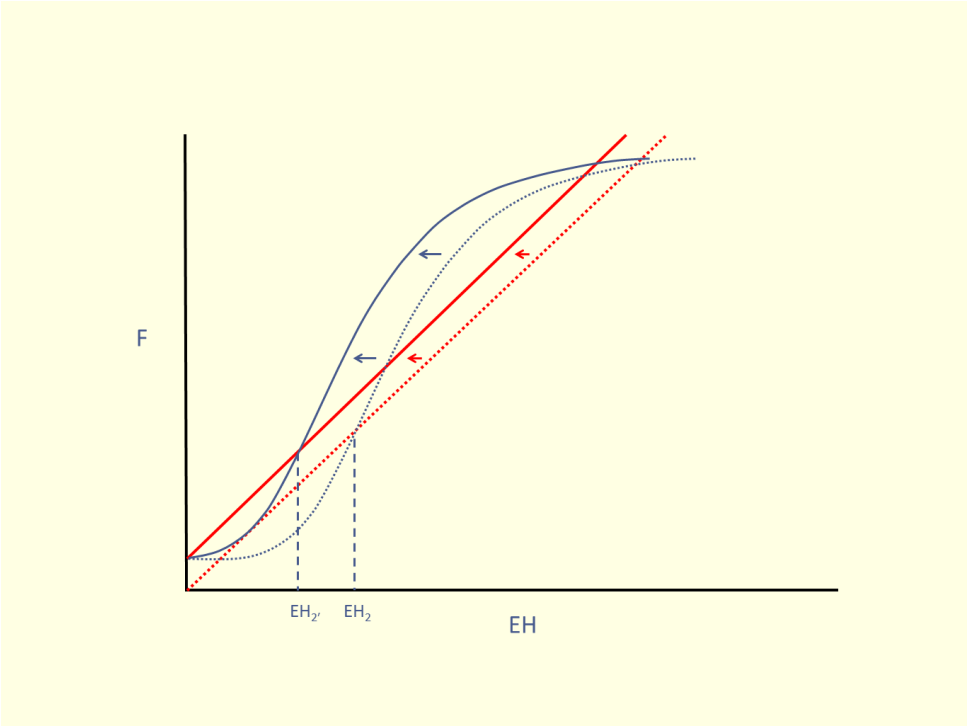



**Figure 17: Impact on Equilibria of Adjustments of Relative Positions of Feedback Functions**

These shifts move the location of the unstable equilibrium from $EH_2$ to a lower value at $EH_2'$, which indicates that the identity discontinuity happens at lower levels of distrust (DT) and capacity to harm (CtH), and thus at lower levels of injustice (InJ) and opportunity (Opp). Modifying functions 6 through 8 as shown below produces these shifts.

1. AR = $(12-(10*(Opp + 0.76)^{-2/3}))/10$
2. A = $(InJ*AR)/2$
3. DT = $0.1+(InJ*AR*0.09)$
4. CtH = Opp
5. $EH_{T0}$ = $DT*CtH$
6. $F_{T0}$ = $1+(8*(1+EH_{T0})^4/(K^4+(1+EH_{T0})^4)$, with K=4.0
7. $EH_{T+1}$ (feedback) = $F_{T0}-0.75$
8. $F_{T+1}$ = $1+(8*(1+EH_{T+1})^4/(K^4+(1+EH_{T+1})^4)$, with K=4.0
   (iterate functions 7 and 8, N times)
9. I = $10-(A+(F_{T+N}/2))$

The resulting response surface has a smaller humanization plane.

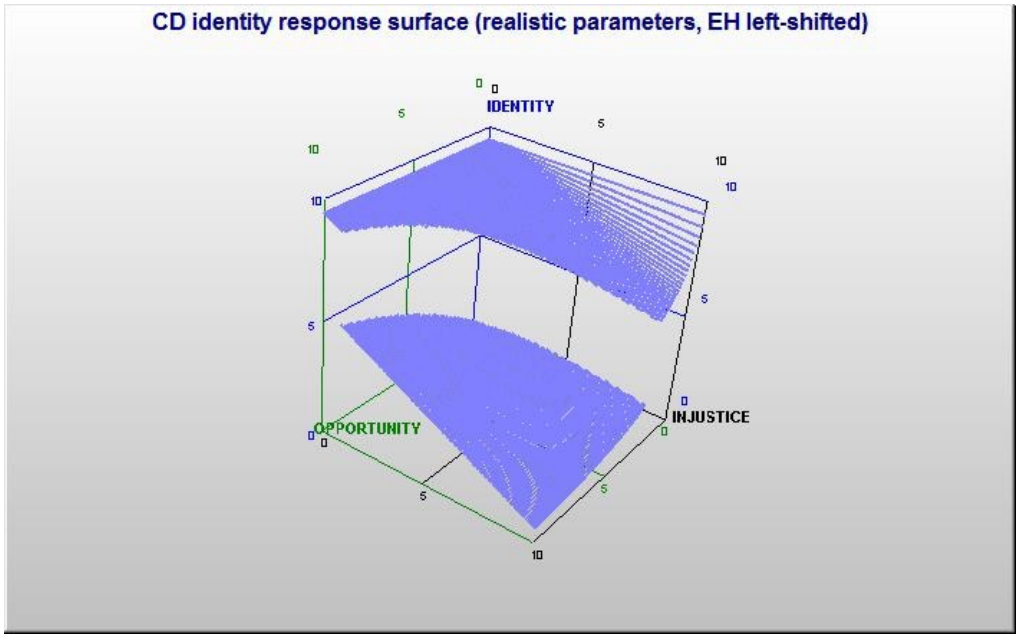



**Figure 18: Identity Response Surface, Adjusted Feedback Functions**

# REFERENCES


Armon-Jones, Claire, "The Social Function of Emotions," in Rom Harre, ed. *The Social Construction of Emotions* (New York: Basil Blackwell, 1986).

Baele, Stephane J., Oliver C. Sterck, and Elisabeth Meur, "Theorizing and Measuring Emotions in Conflict: The Case of the 2011 Palestinian Statehood Bid," *Journal of Conflict Resolution* 60, no. 4 (2016): 718-747.

Baron-Cohen, Simon. *The Science of Evil: On Empathy and the Origins of Human Cruelty* (New York: Basic, 2011).

Berreby, David. *Us and Them: The Science of Identity (Chicago: University of Chicago Press, 2005)*.

Betancourt, Hector, and Irene Blair, "A Cognition (Attribution)-Emotion Model of Violence in Conflict Situaitons," *Personality and Social Psychology Bulletin* 18, no. 3 (June 1992): 343-350.

Billig, M., and Henri Tajfel, "Social Categorization and Similarity in Intergroup Behavior," *European Journal of Social Psychology 3*, no. 1 (January/March 1973): 27-52.

Bodenhausen, Galen V., Lori A. Sheppard, and Geoffrey P. Kramer, "Negative Affect and Social Judgement: the Differential Impact of Anger and Sadness," *European Journal of Social Psychology 24* (1994): 45-62.

Brown, Roger. *Social Psychology*: The Second Edition (New York: Free Press, 1986).

Boix, Carles, "Economic Roots of Civil Wars and Revolutions in the Contemporary World," *World Politics 60* (April 2008): 390-437.

Buhaug, Halvard, Lars-Eric Cederman, and Kristian Skrede Gleditsch, "Square Pegs in Round Holes: Inequalities, Grievances, and Civil War, *International Studies Quarterly* 58 (2014): 418-431.

Collier, Paul, Anke Hoeffler, and Dominic Rohner, "Beyond Greed and Grievance: Feasibility and Civil War," *Oxford Economic Papers* 61 (2009): 1-27.

Crockett, Molly J., Luke Clark, Golnaz Tabibnia, Matthew D. Lieberman, and Trevor W. Robbins, "Serotonin Modulates Behavioral Reactions to Unfairness," *Science 320* (27 June 2008): 1739.

Costalli, Stefano, and Andrea Ruggeri, "Indignation, Ideologies, and Armed Mobilization: Civil War in Italy, 1943-45," *International Security 40*, no. 2 (Fall 2015): 119-157.

Damasio, Antonio. Descartes' *Error: Emotion, Reason, and the Human Brain* (New York: Putnam, 1994).

DeSteno, David, Nilanjana Dasgupta, Monica Y. Bartlett, and Aida Cajdric, "Prejudice From Thin Air: The Effect of Emotion on Automatic Intergroup Attitudes," *Psychology Science* 15, no 5 (2004): 319-324.

Elster, Jon. S*trong Feelings: Emotion, Addiction, and Human Behavior* (Cambridge, MA: MIT Press, 1999).

Emanuele, Enzo, Natascia Brondino, Marco Bertona, Simona Re, and Diego Geroldi, "Relationship between Platelet Serotonin Content and Rejections of Unfair Offers in the Ultimatum Game," *Neuroscience* Letters 437 (2008): 158-161.

Esses, Victoria, Scott Veenvliet, Gordon Hodson, and Ljiljana Mihic, "Justic, Morality, and the Dehumanization of Refugees," *Social Justice Research* 21 (2008): 4-25.

Fearon, James, and David Laitin, "Ethnicity, Insurgency, and Civil War," *American Political Science Review* 97, no. 1 (February 2003): 75-90. https://doi.org/10.1017/S0003055403000534.

Frijda, Nico H. *The Emotions* (Cambridge, UK: Cambridge University Press, 1987).

Frijda, Nico H. "The Laws of Emotion," *American Psychologist* 43, no. 5 (1988): 349-58.

Gilmore, Robert. *Catastrophe Theory for Scientists and Engineers* (New York: John Wiley 1981).

Giner-Sorolla, Roger, and Angela T. Maitner, "Angry at the Unjust, Scared of the Powerful: Emotional Responses to Terrorist Threat," *Personality and Social Psychology Bulletin* 39, no. 8 (2013): 1069-1082.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto, "Chapter Two, Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism," in Patricia Devine and Ashby Plant, eds., *Advances in Experimental Social Psychology*, 47 (2013, Academic Press): 55-130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4.

Greenberg, Jeff, Tom Pyszczynski, Sheldon Solomon, Abram Rosenblatt, Mitchell Veeder, Shari Kirkland, "Evidence for Terror Management Theory II: The Effects of Mortality Salience on Reactions to Those Who Threaten or Bolster the Cultural Worldview," *Journal of Personality and Social Psychology* 58, no. 2 (1990): 308-318.

Hagan, John, and Wenona Rymond-Richmond, "The Collective Dynamics of Racial Dehumanization and Genocidal Victimization in Darfur," *American Sociological Review* 73 (December 2008): 875-902.

Haidt, Jonathan, "The Moral Emotions," in R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, eds., Handbook of Affective Sciences (Oxford: Oxford University Press, 2003), pp. 852-870.
Hall, Todd H., and Andrew A.G. Ross, "Affective Politics after 9/11," *International Organization* 69 (2015): 847-879, doi: 10.1017/S0020818315000144.

Hall, Todd H., and Andrew A.G. Ross, "Affective Politics after 9/11," *International Organization* 69 (2015): 847-879, doi: 10.1017/S0020818315000144.

Hardin, Russell. One for All: *The Logic of Group Conflict* (Princeton, NJ: Princeton University Press, 1995).

Haslam, Nick, "Dehumanization: An Integrative Review," *Personality and Social Psychology Review* 10, no. 3 (2006): 252-64.

Hirschberger, Gilad, Tom Pyszczynski, and Tsachi Ein-Dor, "Why Does Existential Threat Promote Intergroup Violence? Examining the Role of Retributive Justice and Cost-Benefit Utility Motivations," *Frontiers in Psychology* 6: 1761 (November 2015), doi:10.3389/fpsyg.2015.01761.

Isenberg, Daniel J. "Group Polarization: A Critical Review and Meta-Analysis," *Journal of Personality and Social Psychology* 50, no. 6 (1986): 1141-1151.

Jervis, Robert, "Cooperation under the Security Dilemma," *World Politics* 2 (1978): 167-213 [check].

Kelman, Herbert, "Violence without Moral Restraint: Reflections on the Dehumanization of Victims and Victimizers," *Journal of Social Issues* 29, no. 4 (1973): 25-61.

Kteily, Nour, Gordon Hodson, and Emile Bruneau, "They See Us as Less than Human: Metadehumanization Predicts Intergroup Conflict via Reciprocal Dehumanization," *Journal of Personality and Social Psychology* 110, no. 3 (2016): 343-370.

Kurzban, Robert, John Tooby, and Leda Cosmides, "Can Race Be Erased? Coalition Computation and Social Categorization," *Proceedings of the National Academy of Sciences* 98, no. 26 (December 18, 2001): 15387-92.

Leader Maynard, Jonathan, "Combating Atrocity-justifying Ideologies," in Serena J. Sharma and Jennifer Welsh, eds., *The Responsibility to Prevent: Overcoming the Challenges to Atrocity Prevention* (Oxford: Oxford University Press, 2015).

Lerner, Jennifer, Julie Goldberg, and Philip Tetlock, "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility," *Personality and Social Psychology Bulletin* 24, no. 6 (June 1998): 563-574.

Lerner, Jennifer S., Roxana M. Gonzalez, Deborah A. Small, and Baruch Fischhoff, "Effects of Fear and Anger on Perceived Risks of Terrorism: A National Field Experiment," *Psychological Science* 14, no. 2 (March 2003): 144-150.

Lerner, Jennifer S., and Dacher Keltner, "Fear, Anger, and Risk," *Journal of Personality and Social Psychology* 81, no. 1 (2001): 146-159. Lichbach, Mark Irving, "An Evaluation of 'Does Economic Inequality Breed Political Conflict?' Studies," *World Politics* 41, no. 4 (July, 1989): 431-470.

Lichbach, Mark Irving, "An Evaluation of 'Does Economic Inequality Breed Political Conflict?' Studies," *World Politics* 41, no. 4 (July, 1989): 431-470.

Lindenfors Patrik, Andreas Wartel and Johan Lind, "'Dunbar's Number' Deconstructed," *Biology Letters* (2021), https://doi.org/10.1098/rsbl.2021.0158.

Mackie, J.L., *The Cement of the Universe: A Study of Causation* (Oxford: Oxford University Press, 1974).

Mackie, Diane M., Thierry Devos, and Eliot R. Smith, "Intergroup Emotions: Explaining Offensive Action Tendencies in an Intergroup Context," *Journal of Personality and Social Psychology* 79, no. 4 (2000): 602-616.

Mahoney, James, "Toward a Unified Theory of Causality," *Comparative Political Studies* 41 (2008): 412-36.

McDoom, Omar, "The Psychology of Threat in Intergroup Conflict: Emotions, Rationality, and Opportunity in the Rwandan Genocide," *International Security* 37, no. 2 (Fall 2012): 119-155.

Mercer, Jonathan, "Emotional Beliefs," *International Organization* 64 (Winter 2010): 1-31.

Midlarsky, Manus. *Origins of Political Extremism: Mass Violence in the Twentieth Century and Beyond* (Cambridge, UK: Cambridge University Press, 2011).

Panksepp, Jaak, and Lucy Biven. *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions* (New York: Norton, 2012.

Petersen, Roger. *Understanding Ethnic Violence: Fear, Hatred, and Resentment in Twentieth-Century Eastern Europe* (Cambridge, UK: Cambridge University Press, 2002).

Phelps, Elizabeth A., "Emotion and Cognition: Insights from Studies of the Human Amygdala," *Annual Review of Psychology* 57 (2006):27–53.

Posen, Barry R., "The Security Dilemma and Ethnic Conflict," *Survival* 35 (1993): 27-47.

Quigley, Brian, and James Tedeschi, "Mediating Effects of Blame Attribution on Feelings of Anger," *Personality and Social Psychology Bulletin* 22, no. 12 (December, 1996): 1280-88.

Scheffer, Marten, *Critical Transitions in Nature and Society* (Princeton: Princeton University Press, 2009).

Schimel, Jeff, and Linda Simon, Jeff Greenberg, Tom Pyszczynski, Sheldon Solomon, Jeannette Waxmonsky, "Stereotypes and Terror Management: Evidence that Mortality Salience Enhances Stereotypic Thinking and Preferences," *Journal of Personality and Social Psychology* 77, no. 5 (1999): 905-926.

Schwarz, Norbert, "Feelings as Information: Informational and Motivational Functions of Affective States," in E. Tory Higgins and Richard M. Sorrentino, eds., *Handbook of Motivation and Cognition: Foundations of Social Behavior* Vol. 2 (New York: Guilford, 1990), pp. 527-561.

Slovic, Paul, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor, "The Affect Heuristic," *European Journal of Operational Research* 177 (2007): 1333-1352.

Tajfel, Henri, and John Turner. “An Integrative Theory of Intergroup Conflict,” in W. G. Austin and S. Worchel, eds., *The Social Psychology of Intergroup Relations* (Monterey, CA: Brooks/Cole, 1979), pp. 33-47.

Thom, René, *Structural Stability and Morphogenesis* (Reading, MA: Benjamin, 1975).

van der Maas, Han L.J., Rogier Kolstein, and Joop van der Pligt, “Sudden Transitions in Attitudes,” *Sociological Methods & Research* 32, no. 2 (2003): 125-52.

Weiner, Bernard. Judgments of Responsibility: *A Foundation for a Theory of Social Conduct* (New York: Guilford, 1995).

Weingast, Barry R., “Constructing Trust: The Political and Economic Roots of Ethnic and Regional Conflict,” in Karol Soltan, Eric Uslaner, and Virginia Haufler eds., *Institutions and Social Order* (Ann Arbor: University of Michigan Press, 1998), pp. 161-200.

Zeeman, E.C., “Catastrophe Theory,” *Scientific American* 234 (1976): 65-83.

Zeeman, E.C., *Catastrophe Theory: Selected Papers*, 1972-1977 (New York: Addison-Wesley, 1977).

# ENDNOTES

[1]My thanks to Jonathan Leader Maynard, Michael Lawrence, Marten Scheffer, and Manjana Milkoreit for their detailed comments and advice and to Greg Baribeau, Frances Westley, Yonatan Strauch, and Paul Thagard for their additional contributions.

[2]"I would propose that to perceive another as human we must accord him identity and community . . . . To accord a person identity is to perceive him as an individual, independent and distinguishable from others, capable of making choices, and entitled to live his own life on the basis of his own goals and values" Herbert Kelman (1973, 48-49).

[3]Berreby (2005: 183-223) provides an extensive an insightful discussion of such moral exclusion of the "other."

[4]Empathy is the ability to understand the experiences, thoughts, and feelings of another. In my formulation, it requires neither sympathy with, nor moral approbation of, the other; and while being able to identify with someone likely entails being able to empathize with them, the converse is not true. In a major study, Baron-Cohen (2011: 18-19, 209-210) argues that empathy includes both a cognitive component, which involves recognition of another's thoughts and feelings, and an affective component, which involves responding to those recognized thoughts and feelings with an "appropriate emotion." His affective component roughly corresponds to what I call identification. He does not acknowledge, however, that when an observer assesses the appropriateness of a person's emotional response to another in a particular situation, he or she is making a normative judgment. I would argue that this normative judgment involves assessing whether the person in question is identifying with the other person—i.e., whether he or she is accepting or denying the moral legitimacy of the other person's way of life, interests, and actions. Baron-Cohen appears to believe that identifying with the other in this way—accepting their moral legitimacy—is the only "appropriate" emotional response.

[5]For instance, Adolf Hitler was a vegetarian and abhorred the mistreatment of animals. My thanks to Michael Lawrence for this observation.

[6]In contrast, for Baron-Cohen, low empathy is a necessary condition for intentional cruelty towards another (2011: 15). Because he includes an "appropriate" affective response in his definition of empathy, and because intentional cruelty in never appropriate, by definition empathy renders such cruelty impossible. By Cohen's formulation, psychopaths can be intentionally cruel, because they exhibit the cognitive (recognition) component of empathy but not its affective (response) component. Their condition results from neurological deficiencies and is therefore permanent, so it is pathological. In my view, in contrast, psychopaths are able to empathize with others, but they cannot identify with them. The mental state of dehumanization similarly combines empathy with non-identification, but because it does not involve neurological deficiencies and is therefore potentially temporary, it is not pathological.

[7]In discussing the functional role of hatred, Armon-Jones (1986: 73) notes that it facilitates dehumanization and that "dehumanization is necessary to the agent's justification of his otherwise immoral treatment of the object."

[8]I am alluding here to INUS causal logic. See Mackie (1974) and Mahoney (2008).

[9]Examples include Petersen (2002), Midlarsky (2011), Hirschberger et al. (2015), and Baele, Sterck, and Meur (2016). Buhaug, Cederman, and Gleditsch (2014) examine the "emotional power of grievances" arising from perceived intergroup injustice as a key cause of civil war.

[10]See, for instance, Fearon and Laitin (2003) and Collier, Hoeffler, and Rohner (2009). Boix (2008) and McDoom (2012) seek to reconcile and integrate the motive and opportunity perspectives.

[11]The division between emotion and rational cognition implicit in the distinction between hot and cold cognition is empirically untenable, because emotion and cost-benefit cognition are intimately entangled in almost all human cognition (Damasio, 1994; Phelps, 2006; Mercer, 2010; Hall and Ross, 2015). Nonetheless, the literature on the causes of human conflict remains surprisingly segregated into the categories described here. See Lichbach (1989) for an early attempt to reconcile the perspectives.

[12]Anger and fear figure prominently in other analyses of conflict psychology; see, for instance, Mackie, Devos, and Smith (2000), Petersen (2002), McDoom (2012), and Giner-Sorolla and Maitner (2013). Drawing in particular on Frijda (1987), Petersen elaborates a detailed functional theory of the emotional causes of ethnic conflict: a person's emotional response to changed environmental conditions in turn alters the relative salience of that person's desires and thereby prepares him or her to respond to those changed conditions. Petersen distinguishes between the effects of fear, hatred, resentment, and rage. I treat the last three emotions as types of anger, although, as Petersen notes, hatred is generally a product of long-standing and often culturally embedded grievances towards a specific individual or group. Costalli and Ruggeri (2015) distinguish between anger and indignation, where the former arises from a wrong done to oneself or one's next of kin, while the latter arises from a wrong done to a third party; again, I subsume indignation within anger.

[13]This assumption acknowledges that human cognitive limits by themselves can produce at least some degree of the de-individuation of the other group's members that I identified in the Introduction above as one of three cognitive shifts characteristic of dehumanization.

[14]Scholars have recently given these cross-scale interactions detailed attention. For instance, Hall and Ross (2015: 856-860) identify three pathways by which individual-level emotional phenomena can scale up to "collective affective experience"; and Baele, Sterck, and Meur (2016) present a theory of the "interplay between the individual and collective levels of emotion in conflict."

[15]"Group polarization is said to occur when an initial tendency of individual group members toward a given direction is enhanced following group discussion" (Isenberg, 1986).

[16]A discontinuity or catastrophe is distinct from the vernacular notion of a "tipping point," in that it involves more than just the system's rapid shift to another state, but also a jump across intermediate states that are inaccessible to the system.

[17]The general form of the equation for a Riemann-Hugoniot surface, which is the basis for the well-known cusp catastrophe model, is $x^3 - bx - a = 0$.

[18]In this model, therefore, the variable Identity (I) combines two distinct dimensions of group identity that social psychologists sometimes distinguish: inclusive/exclusive and tolerant/antagonistic. It assumes that a person with a highly inclusive group identity is tolerant towards members of the ingroup, while a person with a highly exclusive identity is antagonistic towards members of the outgroup.

[19]I use the somewhat dated gendered pronouns "he," "she," "him," "her," "himself," "herself," and "his" and "hers" in this and some subsequent passages to ensure no confusion arises about the intended referent, since I need plural pronouns such as "them" and "they" to refer to outgroup members.

[20]Once the outgroup is defined, movement towards the exclusive/antagonistic end of the identity scale is accompanied by a shift from perceiving the outgroup as consisting of people who share certain properties to perceiving it as an entity or thing in itself (Berreby, 2005: 210). Midlarsky (2011: 345-7) uses Haidt's moral- foundations theory (Haidt, 2003; see also Graham et al. 2013) to identify degrees of political extremism. He argues that as extremism worsens, people become less inclined to act on the moral impulses of care and reciprocity in their dealings with certain other people, who then come to be defined as members of the outgroup, while they become more inclined to act of the moral impulses of loyalty, respect, and purity with regards to members of the ingroup.

[21]My thanks to Jonathan Leader Maynard for bringing this distinction to my attention. See Leader Maynard (2015) for an explication of how ideology can encourage people to hold certain groups responsible for past injustice or to consider them a future threat.

[22]Weiner (1995) shows that attribution to an actor of causal responsibility for a negative outcome depends in part on an assessment of how much causal control the actor had over events that led to that outcome.

[23]An enormous range of scholarship bears on these causal relationships. On the link between perceived injustice and anger, for instance, Haidt (2003) labels anger a "moral emotion," because it almost always arises in response to some kind of moral judgment, especially about injustice. "Anger," he writes, "may be most frequently triggered by perceived injustices against the self." Emanuele et al. (2008) and Crockett et al. (2008) provide evidence that, at the neurological level, the serotonin system affects the strength of this relationship. On the link between attribution of responsibility and anger, Betancourt and Blair (1992) establish that attribution to an actor of causal responsibility for a negative outcome helps cause anger towards that actor, although Quigley and Tedeschi (1996) and Lerner, Goldberg, and Tetlock (1998) argue that the relationship is reciprocal. Some social psychologists also contend that the mental state of blame is an essential intermediate stage between attribution of responsibility and anger; but Weiner (1995: 15) argues persuasively that this relationship is unmediated. Finally, on the link between anger and identity, Bodenhausen, Sheppard, and Kramer (1994) show that anger makes people more likely to stereotype other people, while DeSteno et al. (2004) find that anger creates "automatic prejudice toward the outgroup."

[24]I thus take as equivalent the following three variables describing the individual's psychological state with respect to other members of the population: 1. the individual's level of distrust of those other members, 2. the individual's estimate of the probability that those other members are untrustworthy, and 3. the individual's estimate of the probability that those other members will inflict harm on him or her if they have the capacity to do so.

[25]I assume that the individual judges that other members of the population, should they decide to inflict harm, will use their capacity to the fullest or maximum extent possible. So the greater the perceived capacity of others to inflict harm, the higher the individual's estimated disutility of any harm those others might inflict.

[26]Analysts of the security dilemma in international relations and ethnic conflict—for instance, Jervis (1978) and Posen (1993)—have highlighted the implications of changes in opportunity. They refer to the situation described in this sentence as emerging anarchy.

[27]Petersen (2002: 72-73) identifies exactly this feedback. "Information regarding the collapse of previous protections and the build-up of the 'other side' produces a belief that one's life, family, and property are in danger. This belief produces an emotion, an action tendency, that affects not only the nervous system of individuals, but further sensitizes their cognitive capacities. The fearful individual is 'activated' to become exceptionally alert to any signals regarding the safety of the environment. The emotion then works to filter information, selecting out the evidence of danger. The result is a confirmation of previous beliefs and a further heightening of fear." Petersen acknowledges, however, that social scientists have not studied the feedback or its implications.

[28]My thanks to Paul Thagard for the insight that a person's emotional responses can serve as an information input into his or her rational calculation of costs and benefits. On how emotional responses influence cognition, see Schwarz (1990), Elster (1999: 165), and Slovic et al. (2007).

[29]I assume here that a person's emotional responses to his or her environment can have consequences for the person's meaningful (or intensional) understanding of his or her environment. Specifically, in the present model anger and fear produce changes in identity, which is a richly intensional mental state. On the link between anger and identity, see for instance Bodenhausen, Sheppard, and Kramer (1994) and DeSteno et al. (2004). On the link between fear and identity, McDoom (2012) shows that fear activates intergroup boundaries, encourages homogenization and denigration of outgroups, and stimulates ingroup solidarity. Greenberg et al. (1990) and Schimel et al. (1999) also show that fear, specifically mortality salience, causes stereotyping, ingroup favoritism, outgroup derogation, prejudice, and intolerance of deviance. These behaviors, the authors argue, help bolster cultural worldviews that buffer people against fear of vulnerability. This argument is grounded in the work of, among others, Henri and Tajfel and John Turner (see, for instance, Tajfel and Turner, 1979) on the relationship between self-esteem and group identity.

[30]On the role of elites, see Hardin (1995).

[31]My thanks to Jonathan Leader Maynard for this formulation.

[32]Barry Weingast (1998: 166-170) develops a similar argument about the motivational effects, in situations of weak institutions or emerging anarchy, of uncertainty about another group's trustworthiness and the perceived high cost of victimhood.

[33]Kurzban et al. (2001) propose an analogous psychological process: the individual scans the population to establish "coalitional affiliations" that he or she then uses to establish population categories.

[34]Underpinning my argument here are two assumptions concerning people's folk beliefs about social causation. First, I assume that, if perceived opportunity is sufficiently high, people prefer to attribute responsibility for injustice to other identifiable people (either individuals or groups) rather than to non-human forces or things, such as fate, nature, or God. People also blame others to avoid the possibility that they bear some responsibility for the situation. Second, I assume that people generally believe that injustice has a unitary cause. In other words, people prefer to blame a single agent—that is, a single individual or clearly identified group—for perceived injustice rather than multiple agents. These two folk beliefs combine, I suggest, to encourage an individual—in a situation of high perceived injustice and opportunity—to identify in the surrounding population the members of a single malign outgroup.

[35]Because the individual's decision not to identify with the outgroup legitimizes preparation of defenses against the outgroup, one way of interpreting Figure 12 is to substitute "not prepared defenses against" for each instance of "identified with" and "prepared defenses against" for each instance of "not identified with."

[36]On the relationship between fear and expected harm, Schwarz (1990: 527) notes that "negative affective states, which inform the organism that its current situation is problematic, foster the use of effortful, detailed-oriented, analytical processing strategies." Lerner and Keltner (2001) and Lerner et al. (2003) find that fear induces more pessimistic estimates of risk.