Building from the Brain: Advancing the Study of Threat Perception in International Relations

Marika Landau-Wells

Department of Political Science, University of California, Berkeley, USA
Email: mlw@berkeley.edu

Abstract "Threat perception" is frequently invoked as a causal variable in theories of international relations and foreign policy decision making. Yet haphazard conceptualization and untested psychological assumptions leave its effects poorly understood. In this article, I propose a unified solution to these two related problems: taking the brain into account. I first show that this approach solves the conceptualization problem by generating two distinct concepts that generalize across existing theories, align with plain language, and are associated with specific brain-level processes: *threat-as-danger* perception (subjectively apprehending danger from any source) and *threat-as-signal* perception (detecting a statement of the intention to harm). Because both types of perception occur in the brain, large-scale neuroimaging data capturing these processes offer a way to empirically test some of the psychological assumptions embedded in IR theories. I conduct two such tests using assumptions from the literatures on conflict decision making ("harms are costs") and on coercion ("intentions are inscrutable"). Based on an original analysis of fifteen coordinate-based meta-analyses comprising 500+ studies and 11,000+ subjects, I conclude that these assumptions are inconsistent with the cumulative evidence about how the brain responds to threats of either kind. Further, I show that brain-level data illuminate aspects of threat perception's impact on behavior that have not yet been integrated into IR theory. Advancing the study of threat perception thus requires a microfoundational approach that builds from what we know about the brain.

"Threat perception" is frequently deployed as a causal variable in theories of international relations (IR) and foreign policy decision making.[1] Both rationalist and behavioral explanations of war onset,[2] crisis behavior,[3] coercion,[4] and alliance

1. Stein 2013.
2. Fearon 1995; Levy 1983; Mearsheimer 2001.
3. Bueno de Mesquita, Morrow, and Zorick 1997; Cohen 1978; Davis 2000; Fordham 1998.
4. Jervis 1982; Powell 1987; Schelling 1966.

International Organization 78, Fall 2024, pp. 627–67

doi:10.1017/S0020818324000328

formation[5] assign "threat perception" a causal role. The term has appeared in approximately one thousand titles and abstracts of journal articles, books, and chapters[6] since Robert Jervis highlighted the concept in *Perception and Misperception in International Politics*,[7] a work that inspired a body of scholarship devoted to integrating the study of human cognition into IR.[8]

Yet the ubiquity of "threat perception" in the literature may owe more to ambiguity in its meaning than to consensus on its importance. The term suffers from two related problems. The first is inconsistent conceptualization of both "threat" and "perception." The definition of "threat" shifts across situations—such as crisis bargaining,[9] the security dilemma,[10] and alliance formation[11]—and across units of analysis, including leaders,[12] states,[13] and citizens.[14] Some scholarship constrains "threats" to the observable (for example, military capabilities or coercive communiqués), while other scholarship considers the ephemeral (for example, hostile intentions or environmental uncertainty). Some scholarship allows for intrinsic variation in how and whether threats are perceived, while other scholarship relies on a consistent correspondence between the existence of a threat and its apprehension. Ultimately, "threat perception" is a background concept with a "broad constellation of meanings and understandings."[15]

A second, related problem within IR theory is the reliance on untested psychological assumptions to link "threat perception" to outcomes of interest. Does the threat of existential harm consistently trump other considerations in decision making, as realist theory assumes?[16] Do people think about the potential for harmful outcomes of all kinds (for example, battle deaths or territorial losses) *as if* they were tallying potential monetary costs, as a number of rational-choice models require?[17] Do people assume that those who issue threats have inscrutable intentions that can only be inferred by estimating their strategic interests[18] or their personal characteristics?[19] Each of these assumptions constitutes a link in a theoretical chain

5. Christensen and Snyder 1990; Snyder 1990; Walt 1987.
6. References collected from Scopus on 4 February 2024 for 1976 through 2023. The query was restricted to "threat perception" in the title, abstract, or keywords of any publication in the social sciences. For context, "threat perception" occurred in 1,141 publications, "rising power" in 1,026, and "audience costs" in 177.
7. Jervis 1976.
8. For reviews, see Davis and McDermott 2021; Stein 2017.
9. Fearon 1994b; Schelling 1960.
10. Jervis 1976; Posen 1993.
11. Snyder 1990; Walt 1987.
12. Stein 2013; Yarhi-Milo 2013.
13. Cohen 1978; Mearsheimer 2001.
14. Myrick 2021; Rousseau and Garcia-Retamero 2007.
15. Adcock and Collier 2001, 531.
16. Mearsheimer 2001; Posen 1993.
17. Bueno de Mesquita 1983; Fearon 1994b.
18. Fearon 1997; Schelling 1966; Snyder 1971.
19. Hall and Yarhi-Milo 2012; Rathbun et al. 2016; Wong 2016.

between "threat perception" and an outcome of interest. Yet rarely, if ever, are these links tested.

The combined effect of conceptual ambiguity and untested assumptions is that little is known about what "threat perception" is or how it works to influence outcomes in IR. In this article, I argue that both problems can be addressed by integrating the brain into the conceptual and empirical microfoundations of the study of threat perception in IR.[20] I make the argument for taking the brain into account in two stages. First, I carry out a conceptual ground-clearing exercise to demonstrate that the proliferation of customized definitions of "threat" and "perception" in the IR literature is unnecessary. The plain-language meaning of "perception" entails a role for the brain. Combining this with the plain-language definitions of "threat" yields two generalizable, interpretable, systematized concepts:[21] the brain's apprehension of any potential danger, subjectively defined (*threat-as-danger* perception); and the brain's detection of a socially communicated statement indicating the conditional intention to harm (*threat-as-signal* perception).[22] Most existing customized definitions of "threat perception" can fit within one of these two concepts.

My proposed disambiguation is not entirely novel. David Baldwin called for a similar distinction in the service of developing a "science of threat systems."[23] But in the second stage of my argument, I show how explicitly acknowledging that perception is a brain-level phenomenon creates new opportunities to test assumptions within IR theories about how threat-as-danger perception and threat-as-signal perception work. These opportunities arise because both plain-language conceptualizations of "threat perception" are also topics of study in cognitive science.[24] This correspondence opens up new sources of data for IR scholars because the neuroscience literature contains a great deal of brain-level data on threat perception, in both senses of the term. Using two empirical examples, I demonstrate how these data can be used to test key assumptions of the IR literature. I also show how brain-level data can generate new insights and aid in developing new theories of threat perception's role in IR that rest on firmer microfoundations.

This article joins a small body of work linking neuroscientific evidence to IR theory.[25] Instead of relying on findings from single neuroimaging studies, as previous work has had to do, I introduce a new type of brain-level data: coordinate-based

20. I use the term "microfoundations" consistent with Kertzer 2017 to mean a lower level of analysis or the set of building blocks from which a phenomenon of interest is understood.

21. According to Adcock and Collier 2001, 531, a systematized concept is "a specific formulation of a concept used by a given scholar or group of scholars; [it] commonly involves an explicit definition."

22. These definitions follow from the two distinct meanings of "threat" in English, according to Stevenson 2010: (1) anything subjectively apprehended as dangerous or potentially damaging; and (2) a statement of a conditional intention to do harm. "Perception" means becoming aware of something through the senses and how that thing is understood (also from Stevenson 2010), processes which cannot occur without brain function.

23. Baldwin 1971, 77.

24. Blanchard 2017; Bulley, Henry, and Suddendorf 2017; Eilam, Izhar, and Mort 2011; Lantos and Molenberghs 2021; LeDoux 2022; Woody and Szechtman 2011.

25. Gammon 2020; Holmes 2013, 2018; McDermott 2004; Price and Sikkink 2021.

https://doi.org/10.1017/S0020818324000328

meta-analyses (CBMAs). CBMAs are one way to analyze large-scale neuroimaging data (that is, data from many individual neuroimaging studies). CBMAs did not exist when IR scholars initially explored the possibility of integrating neuroscience into the study of IR,[26] and I argue they are well suited to address some of the concerns raised about the viability of that project.

To demonstrate the utility of large-scale neuroimaging data, I analyze fifteen previously published, peer-reviewed CBMAs that represent the data collected by over 500 neuroimaging studies from more than 11,000 subjects. I use the data from these CBMAs to test two psychological assumptions found in theories in the IR literature that posit a causal role for "threat perception." I first test an assumption found within theories of conflict decision making about how people reason when confronted with threats-as-dangers (such as adversarial states, rebel groups, or hostile leaders). Specifically, I consider a proposition common in rational-choice models: that people think about the harms associated with prospective conflict (for example, battle deaths or territorial losses) *as if* they were economic costs. Using statistical comparisons of patterns of brain activity across relevant tasks from 126 unique studies, I find no support for this proposition. Behavioral evidence further suggests that relying on this simplifying assumption risks interpreting rational, but complex, choices as irrational or as mistakes. Instead of a cost- or value-based approach, findings in the cognitive science literature suggest that heuristic models of harm evaluation may better explain conflict-related decision making.

The second assumption, common to both rationalist and behavioral theories of coercion, is that people who receive threats-as-signals treat the intentions of the issuer as fundamentally inscrutable. From this perspective, people forgo the futile task of "mind-reading" to determine whether or not the issuer will act on their threat and instead reason about the issuer's *strategic interests* (such as domestic audience costs or reputational concerns) or their *characteristics* (such as a "madman" personality or a fearful emotional state). Using data from 392 unique studies, I find that neither of these theorized workarounds engages the same brain-level architecture as directly reasoning about those who intend to do harm. Further, activation of the brain-level architecture engaged by reasoning about those who intend to do harm (particularly the amygdalae) is associated with distorted perceptions of harm magnitude and a heightened desire for blame and punishment. These perceptions and preferences could help explain documented patterns of coercion failure,[27] since they are associated with resistance and retaliation, rather than capitulation. These same brain-level mechanisms also suggest why costly signals that make threats *more* credible may be associated with coercion failure. Finally, a brain-level account of threats-as-signals suggests that misperception of intended harm is a feature of the brain's social cognitive systems, not a bug to be corrected by better information. Thus, while the process of intention inference is challenging to observe directly,

26. McDermott 2004.
27. Dafoe, Hatz, and Zhang 2021; Powers and Altman 2023; Sechser 2011.

assuming it away may limit our ability to understand the consequences of coercion, and coercion failure in particular.

This article makes several contributions to the literature. As Joshua Kertzer has shown, many IR theories rest on psychological assumptions that are rarely tested.[28] A primary contribution of this article is to conduct two such tests within the literatures on conflict decision making and on coercion. I show how this testing can be conducted with brain-level data and how those data can provide evidence for theory building in addition to theory testing. As a second contribution, I introduce a new (low-cost) type of brain-level data to the study of IR: CBMAs. This type of data can help answer the call for better integration of neuroscientific evidence into IR,[29] which has been hindered, in part, by the prohibitive cost of original data collection. The utility of CBMAs is not limited to the domain of threat perception. Much neuroimaging data related to decision making under risk and uncertainty[30] and intergroup relations[31] has already been meta-analyzed and could be leveraged by IR scholars.[32] Finally, this article provides conceptual clarity. "Threat perception" appears in many IR theories, and there is a sense that it *matters*. Raymond Cohen went so far as to call threat perception "the decisive intervening variable between action and reaction in international crisis."[33] But advancing our understanding of conflict initiation, alliances, crises, coercion, and other contexts where "threat perception" might matter requires a clearer understanding of what it *is* and how it *works*.

The article proceeds in four sections. First, I conduct a conceptual ground-clearing exercise to show why taking the brain into account by reverting to plain-language definitions of "perception" and "threat" has value. I also highlight why—from a brain-level perspective—*threats-as-dangers* and *threats-as-signals* are distinct concepts. Second, I show how large-scale neuroimaging data represented by CBMAs can offer microfoundational evidence with which to test the psychological assumptions in the IR literature. In the third section, I analyze fifteen CBMAs to perform two such tests. The final section considers the implications of a closer connection between neuroscience and IR.

Threat Perception in International Relations

"Threat perception" means different things to different scholars of IR. David Baldwin identified the crux of the problem when he asked, "Precisely what is the phrase 'A threatens B' supposed to tell us?"[34] Baldwin showed that IR scholars use "threat"

28. Kertzer 2017.
29. Davis and McDermott 2021; Kertzer 2017; Kertzer and Tingley 2018.
30. Feng et al. 2022.
31. Saarinen et al. 2021.
32. In the online supplement, I provide guidance on conducting original meta-analyses as well.
33. Cohen 1978, 93.
34. Baldwin 1971, 71.

https://doi.org/10.1017/S0020818324000328

in at least two different ways. For one school, which Baldwin associated with game theorists, "threat … is an undertaking by A intended to change B's future behavior."[35] For another school, which he associated with social psychology, "threat refers to [B's] anticipation of harm," such that "B may be threatened by A regardless of what A is doing; B may even be threatened by A when A does not exist."[36] Baldwin argued that scholars of these two types of threat "are simply not referring to the same thing when they say 'A threatens B'."[37]

An Abundance of Conceptualizations

The problem of conceptual confusion Baldwin identified has only grown. IR scholars now variously define "threats" as a set of material properties, such as offensive military capabilities;[38] informational content, such as coercive communiqués;[39] intended bad outcomes, such as defection in the prisoner's dilemma;[40] unintentional vulnerabilities, such as environmental uncertainty;[41] and holistic impressions, such as enemy "images."[42] Some of these definitions (for example, coercive communiqués) follow plain-language use, but most are customized to the study of IR and to specific use cases. Only by comparing definitions is it possible to see whether two scholars are talking about the same thing when they talk about "threats."

How scholars link "threats" to "perception" adds a layer of ambiguity. From a mechanistic perspective, if a material danger exists (such as an enemy's tanks) then it will be perceived, though perhaps with some systemic uncertainty. This mechanistic perspective is most obvious in the survival-oriented logic of realism.[43] Other scholarship allows for individual-level variation in whether "threats" are "perceived." Some argue that features of the individual *perceiver* (B in Baldwin's formulation) matter most. These include the perceiver's disposition,[44] emotional state,[45] or personal experiences.[46] When B is a state or collective, its own values may structure what is (or is not) perceived as threatening.[47] For other scholarship, features of the *target* (A in Baldwin's formulation) explain perception. These include the content of holistic "images,"[48] an attachment to democratic norms,[49] or A's estimated

35. Baldwin 1971, 72.
36. Ibid.
37. Ibid.
38. Mearsheimer 2001.
39. Fearon 1994b.
40. Jervis 1978.
41. Stein 1988.
42. Herrmann 2013; Jervis 1989.
43. Mearsheimer 2001; Posen 1993.
44. Kahneman and Renshon 2009.
45. McDermott 2017.
46. Yarhi-Milo 2013.
47. Rousseau and Garcia-Retamero 2007.
48. Herrmann 2013; Jervis 1989.
49. Farnham 2003.

capabilities and intentions.[50] Finally, there is the surrounding context. For those emphasizing the perils of an anarchic world, for example, existential risk,[51] the distribution of power in the international system,[52] or the level of general uncertainty in the environment[53] determine whether threats are perceived. These varied notions of "perception" only make it more difficult to pin down what is known about "threat perception" in IR.

Hindering knowledge accumulation further are the many untested psychological assumptions that underpin theories of how "threat perception" (however defined) affects outcomes. Is one kind of danger more important than all others, serving as a reliable, driving force for human behavior? Realist theory posits that the avoidance of existential harm motivates people above all other considerations,[54] but, observably, considerations of nonmaterial threats (for example, spiritual concerns) can dictate behavior that runs counter to physical harm avoidance.[55] Can the decision to initiate war be reduced to the net expected value of war's gains set against its harmful consequences? Rational-choice theories assume people can tally the losses of both "blood" and "treasure" using a common scale,[56] but the value of abstract lives often defies a consistent calculus.[57] Assumptions like these reflect IR scholars' *models* of how the mind thinks about threats. The literature contains more models and assumptions than I can list. Testing them is challenging. But leaving assumptions about how threat perception works untested means that IR theories rest on fundamentally uncertain microfoundations.

Acknowledging the Brain and Returning to Plain Language

Without a clear understanding of what threat perception *is*, it is impossible to improve our theoretical models of how it *works*. What seems to have been lost in the conceptual proliferation of "threat perception" over time is that useful definitions of both "threat" and "perception" already exist in plain language. The plain-language meaning of "perception" is the becoming aware of something through the senses and how that thing is understood.[58] The integration and comprehension of sensory input occur in the brain. There is no escaping this at any level of analysis or aggregation. Institutions, groups, and systemic structures might affect the sensory input

50. Singer 1958; Walt 1987.
51. Tang 2008.
52. He 2022.
53. Jervis 1976.
54. Mearsheimer 2001; Posen 1993; Waltz 1954.
55. Benjamin and Simon 2003; Hassner 2016.
56. Bueno de Mesquita 1983.
57. Gold, Pulford, and Colman 2013; Tversky and Kahneman 1981. A well-known illustration of this inconsistency occurs in the Trolley Problem, where small variations in context can alter the value placed on a human life. For further discussion, see Lillehammer 2023.
58. Stevenson 2010.

https://doi.org/10.1017/S0020818324000328

(for example, information a person is exposed to), but they do not negate the brain's role as the site where integration and comprehension occur.

Of what does the brain become aware? Sticking with plain language, "threat" has two meanings (in English) aligned with Baldwin's distinction: anything subjectively apprehended as dangerous or potentially damaging (*threat-as-danger*); or a statement of a conditional intention to do harm (*threat-as-signal*).[59] Combining these terms partitions "threat perception" into two distinct concepts: "threat-as-danger perception" and "threat-as-signal perception." Both take place in the brain.

While it may seem that threat-as-danger perception is a macro-concept that simply subsumes the detection of threats-as-signals, a brain-level perspective highlights why the two concepts are distinct. As Baldwin noted, threats-as-dangers reside in the mind of the beholder and can include dangers with agency (for example, a rival leader), dangers without agency (for example, a deadly virus), and even dangers that do not exist (for example, a fire-breathing dragon). The mental exercises involved in thinking about threats-as-dangers do not inherently include thinking about other people. This can be because the danger in question is not a person (the deadly virus) or because the exercise does not require doing so (as with estimating the effect on trade of war with a rival state). But threats-as-signals are by definition *social communications*, even when they are misperceived. Reasoning about social signals requires reasoning about the content of another person's mind, which engages the brain's architecture for social cognition.[60]

Because reasoning about other people is enabled by specific brain-level architecture, threat-as-signal perception constitutes a distinct collection of mental exercises.[61] Preserving this distinction is particularly important in the study of coercion, where communication (implicit or explicit) between two or more actors is a focal point of inquiry.

Constraining "threat perception" to mean either "threat-as-danger perception" or "threat-as-signal perception" in lieu of customized definitions has several advantages. First, these two concepts can accommodate most definitions used in the literature, though some scholarly usage spans both concepts.[62] Second, these two definitions of "threat perception" are interpretable to other social scientists and to policymakers, because they align with the intuitions of nonspecialists. Third, the terminology is independent of any particular theory or use case, which enables consistency across scholarship. Finally, both systematized concepts of threat perception are compatible with other disciplines, including cognitive science.[63] This conceptual

59. Stevenson 2010.

60. Frith and Blakemore 2006; Saxe and Kanwisher 2003. This holds even when relying on stereotypes (Kobayashi et al. 2022) or when reasoning about strangers (Defendini and Jenkins 2023). I discuss the brain areas involved in social cognition in the test of Assumption 2.

61. Fehlbaum et al. 2022; Schurz et al. 2021. One way to discern the difference between mental exercises is based on patterns of brain activity, a point to which I turn in the next section.

62. Stein 2013.

63. Bertram et al. 2023; Blanchard 2017; LeDoux 2022.

https://doi.org/10.1017/S0020818324000328

compatibility opens up new sources of brain-level data for the study of threat perception in IR. These data can be used to test theoretical models and to illuminate aspects of threat perception that other theories miss. Before I illustrate these points through two empirical examples, I introduce the analysis of large-scale neuroimaging data as a new source of microfoundational evidence for theory development and testing in IR.

Brain-Level Microfoundations and Large-Scale Neuroimaging Data

By specifying that perception is a process that occurs in the brain, I have made an explicit connection between two constructs relevant to IR theory (threat-as-danger perception and threat-as-signal perception) and brain activity. Scholars have long seen the value in linking brain data to the study of political science generally[64] and IR in particular.[65] To date, work seeking to bridge neuroscience and IR has relied on insights from a handful of neuroscientific studies to inform theory building.[66] Here, I introduce a relatively new type of neuroscientific data that can be used for theory testing as well as theory building: large-scale neuroimaging data summarized by CBMAs. In this section, I connect the dots between functional neuroimaging, CBMAs, and how assumptions in IR theory can be tested.

Spatial Patterns of Brain Function as Data

Neuroscience encompasses the study of brain structure, function, and connectivity at various levels of granularity. For the purposes of this article, I focus on brain function, which refers to the connection between neuronal activity and mental or physical outputs (for example, cognition or behavior).[67] Functional neuroimaging data result from recording the brain's response to particular stimuli or during performance of a specific task.[68] When these functional data are collected over time using magnetic resonance imaging (MRI), the data are often summarized into a three-dimensional representation of average brain activity and stored as a single *image* (for a deeper discussion, see the online supplement).

The 3D images generated by functional MRI (fMRI) contain useful information because the brain is spatially organized. This means that collections of neurons in specific locations play consistent (replicable) roles in brain function, including responses to stimuli and other forms of cognition.[69] Most complex stimuli (such as

64. Haas 2016; Jost et al. 2014; Landau-Wells and Saxe 2020; Theodoridis and Nelson 2012.
65. Davis and McDermott 2021; Kertzer and Tingley 2018; Landau-Wells 2018; McDermott 2004; Price and Sikkink 2021.
66. Gammon 2020; Holmes 2013, 2018; Price and Sikkink 2021.
67. Park and Friston 2013; Shine and Poldrack 2018.
68. Poldrack, Mumford, and Nichols 2011.
69. Eickhoff, Yeo, and Genon 2018.

https://doi.org/10.1017/S0020818324000328

watching people interact) and tasks (such as playing a strategic game) require the involvement of multiple brain regions that are spatially distributed.[70] When this spatial distribution of neural activity is consistent across people, it can be treated as a *pattern*.[71] Because brains are always "on," these patterns are often expressed as a *contrast*, which is a comparison between the brain's activity while dealing with the stimuli of interest and its activity during a control condition or at rest (for example, feeling pain, contrasted with not feeling pain). This method cancels out much of the brain's background activity and isolates the effect of interest (that is, the brain's response to pain). The 3D images of contrasts are the basis of the popular notion that the brain "lights up" in response to certain stimuli. Figure A1 in the online supplement provides an example of this kind of canonical brain image, where brightly colored clusters indicate locations where there is a positive and significant difference in the brain's response between two experimental conditions (expressed as pain > no pain).

Contrast maps are a kind of basic data frequently produced (and published) as part of traditional neuroimaging studies. In the online supplement, I provide more detail on how contrast maps are calculated. Here, I will just note two aspects of these data that are important for what follows. First, each contrast image is made up of three-dimensional pixels, known as voxels. As with a pixel in a 2D image, each voxel has a location within the image (expressed in x, y, and z coordinates) and a value, which is often the result of a statistical test,[72] such as the t-statistics captured in Figure A1. Thus, the information shown in a contrast image resides in the values and spatial arrangement of statistically significant voxels. Second, because large collections of neurons are required for many cognitive functions, the activity associated with a particular contrast is captured by *clusters* of statistically significant voxels. The co-activation of these clusters constitutes the brain activity *pattern* associated with a given task.

A standard fMRI contrast map is simply a 3D image capturing a common pattern of activation across the brain for participants in a study. From that pattern, it is possible to identify the brain areas involved in the cognitive process of interest (for example, experiencing pain) using the image's coordinate system. Two images within the same coordinate system can be statistically compared for similarity in their *cluster-level* patterns and their *voxel-level* patterns. Weak (cluster-level) similarity implies involvement of the same brain areas, and thus possibly similar brain functions.[73] Strong (voxel-level) similarity implies that the same populations of neurons are involved, suggesting fundamentally similar brain-level representations.[74] I discuss how both types of similarity can be used for testing assumptions about how threat perception works, but first I introduce the concept of large-scale neuroimaging data analysis.

70. Fedorenko, Duncan, and Kanwisher 2013; Kanwisher 2010; Shine and Poldrack 2018.
71. Haxby, Connolly, and Guntupalli 2014.
72. Poldrack, Mumford, and Nichols 2011.
73. Eickhoff, Yeo, and Genon 2018.
74. Haxby, Connolly, and Guntupalli 2014; Shinkareva, Wang, and Wedell 2013.

Coordinate-Based Meta-Analyses

A single neuroimaging study produces a single contrast map for a given mental state of interest, representing average effects for the study's participants (for example, pain > no pain, as in supplementary Figure A1). Interpreting a single neuroimaging study requires the same caveats as interpreting results produced by other stand-alone experiments. A single neuroimaging study usually has between fifteen and fifty subjects, and these small sample sizes can raise concerns about both statistical robustness and generalizability.[75] While social scientists have turned to replication as a way to address similar concerns,[76] neuroimaging is an expensive method of data collection, which makes replication for its own sake unlikely.[77]

Neuroscientists have instead turned to data pooling as a means of establishing reliable and general patterns of activity across studies. Peer-reviewed, statistically driven CBMAs have become an increasingly common way to pool and report these results.[78] Where social scientists use meta-analyses to validate effect *sizes*, neuroscientists use CBMAs to validate effect sizes and *locations* across studies.[79] These effect- and location-based tests result in *concordance maps*, which are contrast maps indicating clusters where research shows some statistical consensus. CBMA concordance maps (hereafter, CBMA maps) are thus a form of aggregated large-scale neuroimaging data that summarizes the field's statistical consensus on the patterns of activity associated with a particular task (such as playing strategic games) or stimulus (such as experiencing pain) while masking one-off results.[80] This type of data is thus better suited to support general claims about the brain's responses than any single study with conventional sample sizes.

It is important to acknowledge that CBMAs cannot fix fundamental flaws in research design or analysis. "Garbage in, garbage out" still applies.[81] Researchers conducting meta-analyses also make several choices, beyond which studies to include, that affect a CBMA's stringency. I provide greater detail on CBMA construction and evaluation in the online supplement.

Conditional on reasonable research practices, then, CBMAs can address several concerns about generalizability that have limited the lessons IR scholars can draw from neuroimaging research. In Naoki Egami and Erin Hartman's terms, CBMAs

75. The within-subject design necessitated by most fMRI studies means the number of subjects is not equal to the number of (nonindependent) observations. One subject may provide dozens of observations per experimental condition. I discuss the aggregation of these observations in the online supplement. On sample sizes in fMRI, see Szucs and Ioannidis 2020; Turner et al. 2018.

76. Camerer et al. 2018; Coppock 2019.

77. While the cost of using MRI facilities varies from site to site, it is not uncommon for study costs to exceed USD 800 per subject in the United States.

78. Gilmore et al. 2017.

79. Samartsidis et al. 2017.

80. CBMAs require between seven and twenty studies, depending on the method. Depending on weighting criteria, clusters reported by a single study could contribute to the concordance map if that one study were relatively well-powered. See the online supplement for a discussion of CBMA construction methods.

81. Babbage 1864.

improve the ability to generalize from the sample (sample validity), from the treatment (treatment validity), and sometimes from the outcome measure (outcome validity).[82] Sample validity improves because CBMAs derive results from data collected at different research sites, sometimes in multiple languages, increasing the geographic and cultural diversity of the subject pool. Age and socioeconomic status diversity are likely to be lower in neuroimaging studies than in survey research, however.[83] Treatment validity improves when the same effects are found using multiple implementations.[84] Most CBMAs aggregate studies with a variety of stimuli presented in different modalities (such as video, audio, and text), which means that concordance maps reflect a variety of treatment implementations. Outcome validity can improve for the same reason, though some CBMAs consider only a single outcome by design.

Limitations

The main limitations of CBMAs for IR scholarship are limitations shared by single-study neuroscience. Neuroimaging studies rarely have access to leaders or elites, who are often the targets of IR theories. Kertzer demonstrated via a meta-analysis that elites and members of the public respond similarly to the vast majority of manipulations in the IR experiments sampled, which included 162 paired treatments.[85] Nevertheless, a concern with elite/non-elite comparisons from a neuroscience perspective is that elites would "think differently" about relevant tasks than non-elites, even if observable outcomes are similar. That is, the neural architecture used by elites might be different and so would not be represented in the neuroscientific literature.

To my knowledge, no studies have investigated directly whether elite/non-elite or leader/non-leader differences exist in any functional neuroimaging task. A few studies have considered the differences between experts and non-experts, however. In certain contexts, such as chess and clinical diagnoses, years of training does appear to alter the areas of the brain that process domain-relevant information for challenging problems ("thinking differently").[86] Yet, in the case of financial decision making, expertise was correlated with differences in the magnitude of neuronal activation and decision speed ("thinking faster"), but *not* with different patterns of activation or different choice behavior ("thinking differently").[87]

In the domain of threat perception, there is no evidence of chess-like "thinking differently" among those who are more sensitive to threats—such as highly anxious individuals or those who have been in combat—but there is evidence of "thinking faster."[88] This consistency is not surprising because the brain-level architecture for

82. Egami and Hartman 2022.
83. Dotson and Duarte 2020.
84. Shadish, Cook, and Campbell 2002.
85. Kertzer 2022.
86. Hruska et al. 2016; Krawczyk et al. 2011.
87. Laureiro-Martínez et al. 2014.
88. Chavanne and Robinson 2021; McCurry et al. 2020.

https://doi.org/10.1017/S0020818324000328

dealing with dangers has been largely conserved in evolutionary terms between humans and other mammals.[89] So it seems reasonable to provisionally extend the findings of threat-related neuroimaging studies to elite decision makers, but further research would be helpful for validation.

A second limitation of CBMA (and single-study) neuroimaging data is the context in which they are collected. Context validity is a challenge for experimentation in general,[90] but neuroimaging research takes place in a setting that is very removed from day-to-day life. Yet neuroscientists have started to demonstrate that how subjects think during fMRI tasks reflects how broader populations think when confronted with the same stimuli in the real world. For example, the "neural focus group" method pioneered by Emily Falk and colleagues demonstrates that it is possible to use patterns of neural response from small samples in neuroimaging studies to predict population-level behavioral responses to media stimuli (such as television ads and newspaper articles).[91] Thus, neuroimaging's unique data-collection context need not automatically invalidate the real-world applicability of its findings.

Testing Psychological Assumptions

CBMA maps can reflect the state of the neuroimaging literature's statistical consensus on how the brain responds during a particular mental exercise in the form of a single 3D image (see Panel A of Figure 1 for two illustrations). When a theoretical assumption about how threat perception works in IR theory can be reframed as positing the equivalence of two mental exercises—for example, thinking about physical harm is *like* thinking about an economic cost; or, reasoning about others' intentions is *like* reasoning about their strategic interests—then the assumption can be translated into a testable hypothesis. Specifically, the hypothesis is that the patterns of brain response associated with each mental exercise should be similar. The reason to use CBMAs instead of single studies for this comparison is that IR theories posit *general* models of how threat perception works, and only CBMAs offer sufficiently *generalizable* brain-level findings.

Similarity Analyses

To carry out tests of similarity, I use both cluster-level and voxel-level information from CBMA images. As a weak test, the *functional similarity* of two mental exercises compares the distribution of clusters of activity across major areas of the brain. Functional similarity does not require that two mental exercises activate the exact same voxels. Rather, it measures the extent to which two mental exercises activate a similar pattern of brain areas. I use Alejandro de la Vega and colleagues' definition

89. Blanchard 2017; LeDoux 2022.
90. Egami and Hartman 2022.
91. Berkman and Falk 2013; Falk et al. 2016; Scholz et al. 2017.

https://doi.org/10.1017/S0020818324000328

of these major brain areas as the basis for assessing functional similarity.[92] Following Huixin Tan and colleagues, I calculate the distributions of active voxels across the major brain areas and estimate functional similarity between two distributional vectors using Spearman's rho (ρ).[93]

As a stronger test, I also calculate voxel-level similarity within certain brain areas. This *representational similarity* quantifies the extent to which two mental exercises rely on the same voxels and thus potentially share some neuronal architecture.[94] While two mental exercises might engage the same brain *area*, representing thousands of voxels and millions of neurons, they may not actually activate the same voxels/neurons. Functional similarity may thus overstate the extent to which two mental exercises have the same brain-level microfoundations. To calculate *representational similarity*, I compare the binary patterns of activation for all the voxels in a given brain area using Pearson's phi (ϕ) as the measure of similarity. I account for the number of tests run on each pair of CBMAs (one test for each shared area of activity) using a simple Bonferroni correction.[95]

Data Sets

In this article, I rely on the output of fifteen previously published, peer-reviewed CMBAs from the neuroimaging literature. Table 1 gives the full list of publications and each meta-analysis used as data. I also reference some single-study findings where no meta-analytic equivalent yet exists.

As the table indicates, many published meta-analytic papers contain more than one CBMA. Comparison of CBMAs is an increasingly common means of summarizing meta-analytic knowledge within neuroscience.[96] Most comparisons conducted by the original authors are not directly related to my questions of interest, however, which is why I conduct my own analyses of their data. Additional information about the meta-analyses is provided in Table S1 in the online supplement. Throughout the text, I provide links to the raw CBMA data stored in the freely available archive, Neurovault.org, at <https://neurovault.org/>.

While each meta-analysis captures findings from a variety of tasks, many are analogous to the mental exercises described in the experimental and theoretical IR literature. The most direct analogues are the strategic gameplay CBMA, which is

92. de la Vega et al. 2016 divide the brain's anatomical regions into fifty functional areas based on the associations between voxel-level activation and a large number of labeled cognitive functions derived from Yarkoni et al. 2011's Neurosynth database. I visualize and list these parcels in supplemental Figure A2. Their map is available in Neurovault collection 2099 <https://identifiers.org/neurovault.collection:2099>.

93. Tan et al. 2022. I drop all parcels with no coverage, as well as those with less than a small "cluster" (that is, less than 1 percent of total active voxels), following similar comparative analyses by Briggs et al. 2022 and Tozzi et al. 2021.

94. Kriegeskorte, Mur, and Bandettini 2008.

95. I use the standard α of 0.05 divided by the number of brain areas in which the similarity test is run. For example, if activation similarity is tested in five areas, the Bonferroni-corrected α is 0.05/5 = 0.01.

96. Yeung et al. 2023.

https://doi.org/10.1017/S0020818324000328

TABLE 1. *Coordinate-based meta-analysis data sets*

Original publication	*Meta-analysis*	*Studies*	*Subjects*
Bartra et al. (2013)	Negative subjective value	77	1,371
Boccia et al. (2017)	Third-person harm	31	799
Bzdok et al. (2012)	Intentional harm	20	345
Eres et al. (2018)	Harm evaluation	84	1,963
Fede and Kiehl (2020)	Harm evaluation	58	1,410
Fehlbaum et al. (2022)	Mentalizing	204	4,786
Merritt et al. (2021)*	Out-group mentalizing	50	1,117
Schurz et al. (2021)	Infer others' beliefs	25	567
Schurz et al. (2021)	Infer others' emotions	12	346
Schurz et al. (2021)	Infer others' intentions	11	238
Schurz et al. (2021)	Infer others' traits	19	330
Schurz et al. (2021)	Simulate others' internal state	12	288
Schurz et al. (2021)	Strategic gameplay	13	236
Tan et al. (2022)	Physical harm	30	615
Tan et al. (2022)	Monetary loss	20	469
Total		666	14,880
Total, adjusted for duplicates		**518**	**11,814**

* Number of subjects is imputed.

neuroimaging data collected while subjects play the prisoner's dilemma, the ultimatum game, chicken, or other zero-sum, adversarial games, and the visceral simulation CBMA, which asks participants to engage in the processes posited by some IR scholars to play a role in intention inference.[97] The least comparable set of tasks, relative to the IR experimental literature, is found in the physical harm CBMA. These tasks capture the response to genuine pain, which represents a more direct test of IR's long-standing interest in both the threat and the experience of pain[98] than IR scholars have generally induced.[99] The other meta-analyses have tasks that are analogous to processes theorized in IR (for example, experiencing financial losses or reasoning about out-group members) but largely without explicit political overtones. For a deeper discussion of tasks included in the CBMAs, see the online supplement.

Applications

I next consider two assumptions from the IR literature about how threat-as-danger perception and threat-as-signal perception work. I first review the literature that leverages each assumption—conflict decision making in the first case and coercion in

97. Holmes 2013, 2018; Markwica 2018.
98. Hobbes 1651; Schelling 1966.
99. The absence of pain studies from the IR literature may be due to the additional training and facilities requirements rather than a lack of interest. Biomedical certification is a prerequisite for conducting this type of human-subjects research.

https://doi.org/10.1017/S0020818324000328

the second. I then translate each assumption into an expectation for patterns of similarity in brain-level responses. I next test these expectations empirically, using measures of functional (weak) and representational (strong) similarity. Finally, I discuss how these brain-level analyses affect our understanding of behavior and the implications for the study of threat-as-danger perception and threat-as-signal perception in IR.

Assumption 1: Thinking About Harms as Costs

Decisions surrounding violent conflict are some of the most scrutinized in the study of IR.[100] The perception of threats-as-dangers is considered an integral component of decisions to initiate conflict[101] and to form alliances.[102] Threats-as-dangers in this context could be adversarial states, rebel groups, hostile ideologies, or particular leaders, but in all cases, decision makers must evaluate the harms they could inflict. At minimum, two kinds of harm are relevant: losses of "blood" (lives) and losses of "treasure" (money). In the context of alliances, deciding whether to balance against, or bandwagon with, a powerful neighbor requires comparing losses of "blood" and "treasure" during a potential conflict to losses of political autonomy as well. Given the nature of these conflict-related decisions, it is unsurprising that scholars have assumed people have a way of evaluating a variety of harmful outcomes in a common space.

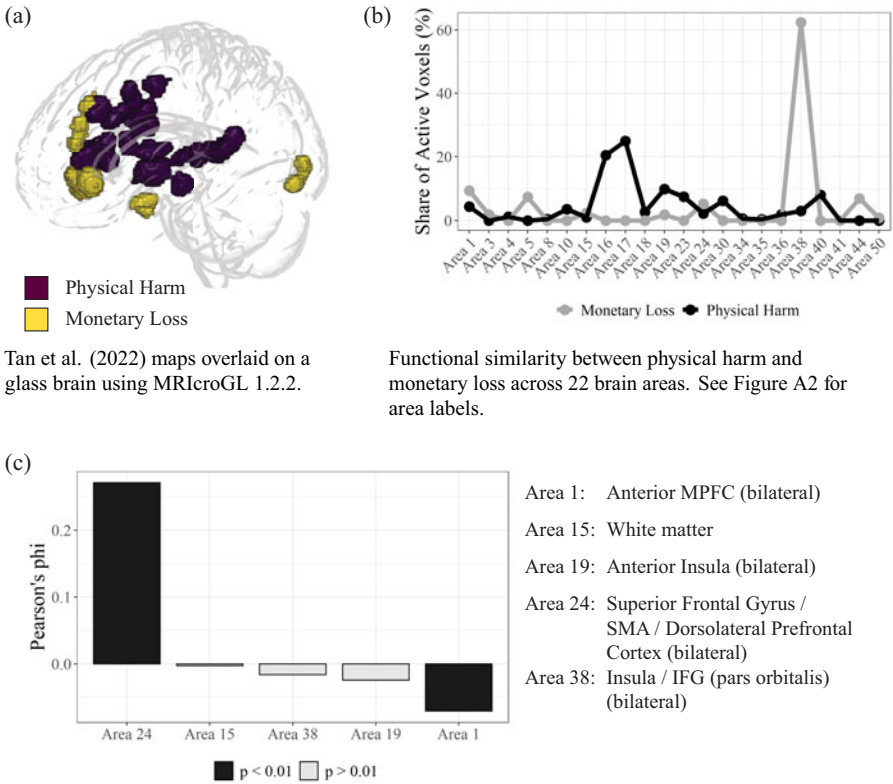
The language of "costs" is often used to characterize this comparison space.[103] This linguistic choice is reflected in the representation of harmful decision outcomes by either a quasi-numeric value (such as –10 in the double-defection cell in the prisoner's dilemma), an actual purported quantity (such as the monetary expenditure of equipping troops), or a variable that has ordinal properties, if not an exact value (such as c in a bargaining model of war). The core psychological assumption behind these implementations is that people reason about the *nonmonetary* harms associated with conflict, such as battle deaths or lost territory, *as if* they could be summarized in a single numeric (or quasi-numeric) value.

Assuming an equivalence between thinking about harms and thinking about costs is useful from a theoretical standpoint. Rational-choice models minimally require preferences for outcomes to be connected (that is, comparable) and transitive (that is, if A is preferred to B, and B to C, then A is preferred to C). When harmful outcomes can be summarized as a simple, one-dimensional quantity, like a cost, it is straightforward to maintain these two requirements.[104] Yet, constructs that are inherently subjective, such as privacy or well-being, are complex in that they are evaluated along multiple, sometimes incommensurable dimensions.[105] In these cases, researchers have found

100. For reviews, see Dafoe, Renshon, and Huth 2014; Levy 1998; Powell 2002; Stein 2002.
101. Fearon 1995; Jervis 1976; Levy 1983; Posen 1993; Powell 2006.
102. Christensen and Snyder 1990; Snyder 1990; Walt 1987.
103. Bueno de Mesquita 2010; Fearon 1995; Kydd 2010.
104. Baumann 2022; Quackenbush 2004; Zagare 1990.
105. Tomaino, Wertenbroch, and Walters 2023; Walasek and Brown 2023.

that preferences do not always maintain transitivity, though they may be comparable.[106] That is, the complexity of a concept has implications for how we *think* about it during decision making and for how we *behave* when we make choices.

If the brain simplifies physical and other nonmonetary harms such that they are processed similar to monetary losses, we could lean on the one-dimensional perspective favored by rational-choice frameworks. If, however, other types of harm are processed substantially differently and with greater brain-level complexity, we need to ask how decision making involving harm should be modeled.



Tan et al. (2022) maps overlaid on a glass brain using MRIcroGL 1.2.2.

Functional similarity between physical harm and monetary loss across 22 brain areas. See Figure A2 for area labels.

Representational similarity between physical harm and monetary loss in brain areas with overlapping activation. Bonferroni corrected p-value = 0.01.

FIGURE 1. *Comparison of CBMA maps for physical harm and monetary loss*

106. Hayden and Niv 2021; Juechems and Summerfield 2019; Kalenscher et al. 2010; Ranyard et al. 2020; Shenhav et al. 2014; Tsetsos et al. 2016; Tversky 1969; Walasek and Brown 2023.

https://doi.org/10.1017/S0020818324000328

Comparing the mental exercises of thinking about physical harm ("blood") and thinking about monetary losses ("treasure") offers the most direct test of harm–cost equivalence. Tan and colleagues conduct two meta-analyses to directly compare how the brain represents physical harm (30 studies, 615 subjects) and economic losses (20 studies, 469 subjects).[107] I use the corresponding CBMA maps to calculate the functional (weak) and representational (strong) similarity of these two mental exercises.[108] Figure 1 illustrates that these two exercises are not especially similar. Panel A visualizes the two CBMA maps and the distribution of significant clusters throughout the brain. The response to physical pain is more widely distributed than the response to economic losses, comprises almost three times as many active voxels (3,052 versus 1,118), and engages twice as many functional brain areas (18 versus 9). Panel B illustrates the functional activation profiles across brain areas and quantifies their lack of similarity (Spearman's $\rho = -0.28$, $p = 0.21$). While the two exercises engage five of the same functional brain areas, they each engage a number of regions uniquely (physical harm, 13; monetary loss, 4), suggesting that the brain functions supporting each exercise are quite different. Panel C shows that, within their five shared brain areas, representations are significantly similar in only one (dorsal midcingulate cortex, dMCC) and significantly dissimilar in another (anterior medial prefrontal cortex, aMPFC).[109]

Dorsal MCC is associated with integrating negative emotions and managing subsequent actions.[110] Similar patterns in this area may reflect the fact that both monetary losses and physical harm generate negative emotional experiences. Anterior MPFC, which shows significant *dissimilar* activation, is broadly associated with the encoding of subjective value, in addition to many other functions.[111] The significant *negative* correlation in activation across this region suggests that while both physical harm and monetary loss carry negative subjective value, the collections of neurons that represent this value are not the same. Thus, whether considering functional (weak) or representational (strong) similarity, the brain's responses to physical harm and economic losses are not much alike.

Physical pain may be too literal a translation of the harms associated with conflict, however. Excluding the experience of pain, Bartra and colleagues meta-analyzed 77 studies (1,371 subjects) of "negative subjective value" in which participants experienced a range of bad outcomes.[112] Harmful experiences used as stimuli included nonmaterial losses (such as losing a competition to a rival), material losses (such as losing money), and visceral unpleasantness for oneself

107. Tan et al. 2022.

108. I use the maps underlying Figures 1 and 2 in the original article, which are available at Neurovault, <https://identifiers.org/neurovault.collection:11982>.

109. Pearson correlation within dMCC:$\phi = 0.27$, $p < 0.01$. Pearson correlation within aMPFC:$\phi = -0.07$, $p < 0.01$. Bonferroni corrected α: 0.01.

110. de la Vega et al. 2016; Shackman et al. 2011; Tolomeo et al. 2016.

111. Clairis and Lopez-Persem 2023; Lieberman et al. 2019.

112. Bartra, McGuire, and Kable 2013.

https://doi.org/10.1017/S0020818324000328

(such as watching violent imagery) and for others (such as watching others experience fear).[113] I calculated the functional profile of negative subjective value and its similarity to Tan and colleagues' profile for economic losses. The functional profile of negative subjective value is represented across more brain areas (27) than monetary losses alone (9). As with physical pain, negative subjective value is not functionally similar to monetary losses (Spearman's $\rho = 0.11$, $p = 0.57$). Representational similarity analysis in aMPFC replicated the negative correlation found for physical harm ($\phi = -0.11$, $p < 0.001$). Negative subjective value and monetary losses were *positively* correlated in two areas associated with emotion experience (anterior insula: $\phi = 0.09$, $p = 0.002$) and emotion regulation (inferior frontal gyrus: $\phi = 0.11$, $p < 0.001$), respectively.[114] This suggests that while the *feeling* of losing money is processed in a manner similar to the feeling of other negative outcomes, they have little else in common as far as brain-level architecture is concerned. This meta-analytic finding is consistent with a single neuroimaging study showing that a classifier trained to detect monetary losses on the basis of neural responses could not detect other types of negative outcomes (such as visceral unpleasantness) for oneself or for others.[115]

Evidence from CBMAs thus indicates that nonmonetary harms engage a broader network of brain areas (18 for pain, 27 for other negative experiences) than the experience of monetary loss (9 areas), demonstrating a more complex brain-level representation.[116] The relative complexity of what is being valued in the brain has been linked to particular decision-making behavior, and to the intransitivity of preferences specifically. In such cases, intransitivity is not a matter of error; "instead, these data suggest that the irrationalities we observe in behavior reflect a fundamental irrationality in the neural representation of subjective value."[117] As Benjamin Hayden and Yael Niv summarize, "The fact that the brain *can* compute values to compare apples and oranges does not mean that it routinely does so, or that valuation is the primary process underlying choice."[118] Put differently, the brain may not assign one-dimensional values to "blood" and "treasure" to add them up, nor does it necessarily ever compute their cumulative "cost."

While the complex neural representation of the harms associated with conflict-related decision making may violate the requirements of rational-choice models in some cases, this does not mean that people make irrational decisions where harm is concerned. Frameworks for decision making that do not require transitivity or even value-based computation (such as heuristic decision-making models) often

113. Twenty-three of the seventy-seven studies used monetary losses as the negative experience, though only one study overlaps with Tan and colleagues' data set. The inclusion of monetary losses creates a bias *against* finding dissimilarities between negative subjective value and economic loss. The map for negative subjective value is available on Neurovault, <https://identifiers.org/neurovault.collection:917>.

114. Du et al. 2020; Gu et al. 2013; Morawetz et al. 2017; Touroutoglou et al. 2012.

115. Speer et al. 2023.

116. Shine and Poldrack 2018.

117. Glimcher 2022, 675.

118. Hayden and Niv 2021, 192, emphasis in the original.

https://doi.org/10.1017/S0020818324000328

perform better at predicting multi-attribute choices than rational-choice models and are themselves still rational.[119] For example, a single neuroimaging study evaluating threat-related decision making found that a heuristic policy best explained participants' choices for their survival strategy.[120] An optimally rational, utility-maximizing policy was a second-best explanation, and each decision policy was supported by different brain areas. The optimal policy also maximized a monetary incentive, but many adopted a heuristic approach anyway.

In sum, the brain does not appear to reason about all bad outcomes using the brain-level architecture for thinking about economic losses. While the language of costs has a simplifying appeal, the microfoundational evidence suggests that harm evaluation is a more complex process, engaging a broader range of brain areas. And behavioral evidence suggests that this type of complex brain-level representation yields preference patterns that do not obey the assumptions required by rational-choice models. Other logics of preference structure (such as heuristic models) might better capture how people evaluate the threats they perceive and the choices they make. Research into such models thus offers an alternative method for theory building in the domain of conflict-related decision making.[121]

Assumption 2: Treating Intentions as Inscrutable

The determinants of coercion's success or failure are a perennial topic of interest in IR.[122] Issuing threats-as-signals is one of the primary tools actors use to *attempt* to coerce others into doing something they would rather not.[123] Recent scholarship has demonstrated that coercion often fails in the real world[124] and that states with greater material power are not necessarily more successful,[125] suggesting that coercion's outcomes are not merely determined by resources. This scholarship has sought to explain variation in coercive outcomes by focusing on how threats are processed by those on the *receiving* end using experimental[126] and qualitative[127] data. These works provide an important step forward for the study of coercion, and

119. Brandstätter, Gigerenzer, and Hertwig 2006; Gigerenzer and Gaissmaier 2011; Hayden and Niv 2021; Juechems and Summerfield 2019.

120. Korn and Bach 2019.

121. Kahneman and Renshon 2009; Landau-Wells 2018.

122. Byman and Waxman 2002; Carnegie 2015; Davis 2000; Fearon 1997; George and Simons 1971; Greenhill 2011; Greenhill and Krause 2017; Jervis 1979; Markwica 2018; Pape 1996; Press 2005; Schelling 1960.

123. Other tools include the limited use of force, assurances, and positive inducements. But "coercion always involves some cost or pain to the target, or explicit threats thereof" (Art and Greenhill 2017, 4). Because explicit threats-as-signals play a role in both the compellent and deterrent forms of coercion, I do not distinguish between those concepts here, but for a recent discussion see Art and Greenhill 2017.

124. Dafoe, Hatz, and Zhang 2021; Morgan, Bapat, and Kobayashi 2021; Sechser 2011.

125. Chamberlain 2016; Sechser 2011.

126. Dafoe, Hatz, and Zhang 2021; Powers and Altman 2023.

127. Markwica 2018.

coercion failure in particular. But, by design, they can only speculate on the cognitive mechanisms involved.

How does the brain process threats-as-signals—that is, socially communicated statements of the conditional intention to harm? And do the cognitive processes involved affect how the recipients of threats *respond* to coercive attempts? Assuming that actors correctly identify that a threat-as-signal is meant for them,[128] scholarship on coercive threats often parses the first question into two problems: how do people assess *capabilities* (that is, can the issuer make good on their threat?) and how do people assess *intentions* (that is, will they?).[129] Scholarship often treats capabilities as somewhat uncertain but ultimately knowable, given sufficient information.[130] But intentions are generally treated as inscrutable: that is, they cannot be understood or investigated directly. This theoretical position is justified by a basic truth: we cannot know the content of another person's mind.[131]

One consequence of assuming that intentions are inscrutable is that both rationalist and behavioral approaches to the study of coercion posit that people do not try to puzzle out intentions directly. Rather, scholars have proposed that people use *work-arounds* to estimate others' intentions, which then feeds into the assessment of the *threat-as-signal* they have received.[132] A theoretical advantage of this stance is that it allows scholars to substitute a process that has no visible inputs (intention inference) with processes that do (such as emotion inference using facial expressions or audience cost inference using speech acts).

One commonly proposed workaround is that people reason about the *strategic interests* of the threat's issuer. Understanding the issuer's interests would then reveal whether they "should" carry out their threat. These interests can include the domestic costs and benefits of making good on the threat versus backing down,[133] as well as their potential gains and losses in the international arena.[134] Some scholars use well-established strategic games (such as chicken or the prisoner's dilemma) to capture the mental exercise of reasoning through another's interests in a coercive situation.[135] Others use situation-specific game-theoretic frameworks.[136] In both cases, reasoning about interests and payoff structures provides a shortcut to inferring intentions because payoff structures dictate what the threat's issuer *should* do and,

128. As Stein 2013 and others have noted, this is not always the case.

129. Bueno de Mesquita, Morrow, and Zorick 1997; Huth 1999; Jervis, Yarhi-Milo, and Casler 2021; Press 2005; Sartori 2013; Stein 2013. The product of these answers (and perhaps the added component of "resolve") is sometimes referred to as credibility.

130. Bueno de Mesquita, Morrow, and Zorick 1997; Fearon 1994a; Press 2005.

131. Holmes 2018; Jervis 1976; Levy 1983; Mearsheimer 2001; Rosato 2015.

132. An alternative approach is to treat intentions as having constant value, such as "aggressive," though Kertzer and McGraw 2012 note that this approach is not widely adopted among the public. For a discussion of this type of assumption, see Tang 2008.

133. Fearon 1997; Schultz 1998.

134. Press 2005; Slantchev 2010.

135. Kilgour and Zagare 1991; Schelling 1966; Snyder 1971.

136. Fearon 1994a, 1997; Powell 1987; Slantchev 2011.

https://doi.org/10.1017/S0020818324000328

therefore, what they likely *intend* to do as long as they are rational actors.[137] As Wilson, Stevenson, and Potts put it, "Game theory relies very little on assessing the motives of others."[138]

A second category of workarounds assumes that the threat issuer's personal *characteristics* provide clues about whether they intend to carry out their threat. Characteristics hypothesized as being informative here include relevant personality types or traits,[139] beliefs,[140] and emotional states.[141] From this perspective, assigning a value to one of these characteristics (such as *untrustworthy* for a trait or *afraid* for an emotional state) allows the recipient of a threat to infer the issuer's intentions using observable information, either from the current situation or from prior interactions. Taking the consideration of characteristics even further, Marcus Holmes has argued that when people have the opportunity to interact face to face, they viscerally simulate others' internal states and so gain insight into their intentions.[142] Seanon Wong makes a similar argument about the insights afforded by face-to-face interaction, but focuses on reading others' emotional states.[143]

Theories incorporating these indirect methods of intention inference generate the same testable implication: does the mental exercise of thinking about others' intentions to do harm look like something else (that is, one of the hypothesized workaround exercises)? I answer this question using two comparisons. First, I evaluate the similarity between direct intention inference (that is, the brain's response when specifically tasked with inferring intentions) and five possible workaround exercises derived from the coercion literature. Second, I compare patterns of neural response for inferences about those who intend harm and about those who do not, to more precisely characterize how people might respond to receiving threats-as-signals.

In social cognitive neuroscience, the study of how we think about the thoughts and mental states of other people is often captured by the umbrella term "mentalizing," which includes empathy, perspective taking, and reasoning about others' minds in a variety of contexts.[144] Mentalizing activities are supported by an identifiable brain network required for social cognition, which Lynn Fehlbaum and colleagues summarize in a CBMA (204 studies, 4,786 subjects). The mentalizing network as

137. Fearon 1997; Snyder 1971. As Stein 2013 notes, not all of the scholarship in this field explicitly articulates these assumptions from the point of view of the threat's recipient, focusing instead on the calculations performed by the issuer (Fearon 1997). Nevertheless, for these models to convey insights into the recipient's behavior (that is, coercive success or failure), recipients must use the same logic to interpret a threat as the issuer used to formulate it, or a separate theory of the recipient's calculus would be needed.

138. Wilson, Stevenson, and Potts 2006, 460.

139. Hall and Yarhi-Milo 2012; Sagan and Suri 2003.

140. Stein 1988.

141. Markwica 2018; Wong 2016.

142. Holmes 2013, 2018. Holmes's argument for visceral simulation is grounded in the idea of a mirror neuron system. While the nature of mirror neurons is debated (Heyes and Catmur 2022; Saxe 2005), the premise that visceral simulation provides insight into intentions does not require a mirror neuron system per se.

143. Wong 2016.

144. Frith and Blakemore 2006; Quesque et al. 2024.

https://doi.org/10.1017/S0020818324000328

represented in their CBMA spans many brain regions, including the bilateral temporoparietal junction, bilateral superior temporal sulci, precuneus, and portions of the medial and dorsolateral prefrontal cortex. As discussed, a crucial aspect of threats-as-signals is that receiving them *requires* mentalizing, even when considering strangers or relying on stereotypes.[145]

Matthias Schurz and colleagues conducted meta-analyses to compare the patterns of brain activation associated with *specific* mentalizing tasks, including inferring others' intentions (11 studies, 238 subjects), their traits (19 studies, 330 subjects), their factual beliefs (25 studies, 567 subjects), and their emotions (12 studies, 346 subjects), as well as simulating others' emotional or physical states (12 studies, 288 subjects) and reasoning about their actions during strategic gameplay (13 studies, 236 subjects).[146] I use the corresponding CBMA maps to test the empirical similarity of directly reasoning about others' intentions against the five workaround mental exercises derived from the coercion literature.[147]

As illustrated in panels A and B of Figure 2, *direct* intention inference is functionally (weakly) similar to three of the five workaround mental exercises advanced in theories of coercion and signaling. Panel A depicts the functional profiles of directly inferring someone's intentions and the functional profiles generated by each of the five workaround exercises. Panel B shows the functional similarities between the activation profile of direct intention inference depicted in panel A and each of the workarounds. In functional-similarity terms, the closest workaround activity is reasoning about others' emotional responses (Spearman's $\rho = 0.61$, $p < 0.001$), but belief inference ($\rho = 0.58$, $p < 0.001$) and trait inference ($\rho = 0.56$, $p < 0.001$) are also substantively similar. Neither strategic gameplay ($\rho = 0.24$, $p = 0.19$) nor visceral simulation ($\rho = 0.15$, $p = 0.38$) are even weakly similar to direct intention inference.

Much of the overlap in patterns of neural activation shown in panel A occurs within areas associated with the mentalizing network.[148] Panel C illustrates the representational (strong) similarity between direct intention inference and the three functionally similar mental exercises. Within mentalizing areas, the average representational similarity between direct intention inference and the three functionally similar workarounds is 0.40; within other brain areas, it is 0.18.[149] This suggests that inferring others' intentions directly is functionally similar to inferring their emotional states,

145. Defendini and Jenkins 2023; Kobayashi et al. 2022.

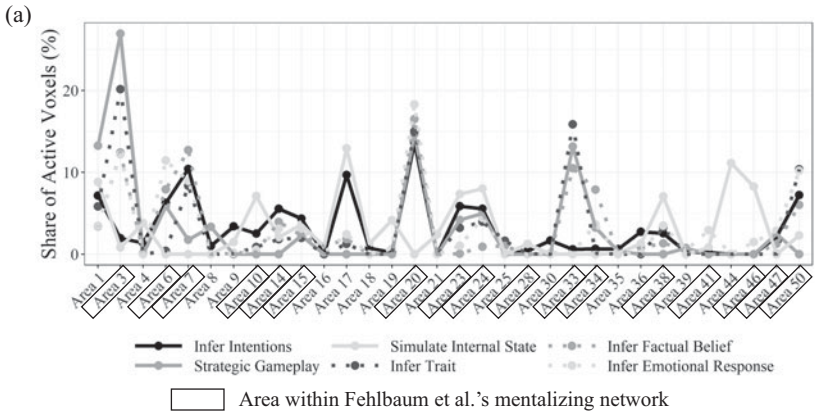
146. Schurz et al. 2021.

147. Schurz et al. provide exact task descriptions for all studies included in each meta-analysis referenced here in Tables S3.1, S3.2, S3.3, S3.5, and S4.4 of the original paper. Maps are available at Neurovault, <https://identifiers.org/neurovault.collection:9936>. I also discuss the composition of these CBMAs in the online supplement.

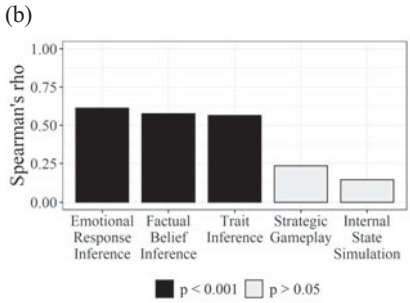
148. Of the voxels activated by more than one of Schurz et al.'s mentalizing activities, 88 percent fall within the mentalizing network.

149. I restrict this analysis to the nineteen regions containing at least 1 percent of the active voxels in Schurz et al.'s direct intention inference brain map to limit the comparison to areas with meaningful levels of activation. If a comparison task generates no activation in a region associated with intention inference, the correlation is set to zero.

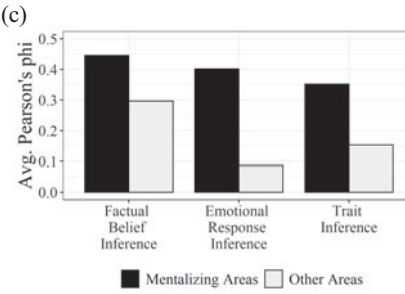
https://doi.org/10.1017/S0020818324000328



Functional profiles for mentalizing activities across 32 brain areas.



Functional similarity between intention inference and workaround mentalizing tasks.



Average representational similarity between intention inference and functionally similar workaround tasks.

FIGURE 2. *Comparison of brain activity during intention inference and workaround mentalizing tasks*

traits, or beliefs, in large part because all these tasks leverage the same basic brain-level architecture used for social cognition.

But the differences between direct intention inference and the theorized workarounds are also notable. The justification for proposing workaround mental exercises in theories of coercion is that people *know* they are not mind readers and therefore try some other method. Yet, if this substitution occurs because people know that intention inference is fundamentally futile, then we should see similar substitutions in other contexts, including neuroimaging studies. This would generate a high degree of representational similarity both *inside* and *outside* the mentalizing network, which is not what the comparison of CBMAs in panel C indicates. The implication for theories of coercion is that, while *accurate* intention inference may indeed be extremely challenging, people still *try*. Intentions may be uncertain, but

they are not inscrutable. The presumption of workarounds is convenient for scholarship because it is difficult to study cognitive processes that have no visible inputs or indicators. Nevertheless, this approach may be misleading.

The direct-intention-inference exercises reflected in Schurz and colleagues' CBMA are still one step removed from the mental exercise at the heart of coercion theories: thinking about those who might intend to *harm* us. None of the CBMAs analyzed here focus on inferences about those who intend harm, with the exception of the CBMA for strategic games.[150] Crucially, the CBMA for direct intention inference is derived from studies where subjects reasoned about others' actions when they did *not* result in serious harm.[151]

But evidence from cognitive science suggests that thinking about harm has consequences for reasoning. For example, across many cultures, people do not reason about actions that could cause harm to themselves or to others purely on the basis of the action's outcome.[152] Ryan Carlson and colleagues note that "some actions, especially those involving direct physical harm, are judged to be worse than others, even when the outcome is the same."[153] In particular, judgments about harm are influenced by *intent*. Intentional harms are often evaluated differently from accidents that cause identical bad outcomes, though not in all cultures.[154] When intent plays a role, it increases *subjective* perceptions of harm severity (sometimes called "harm magnification"),[155] even in damage-quantification tasks with objective answers and incentives for accuracy.[156] Attribution of intent to harm also increases the willingness to blame and punish[157] and the perception of moral wrongness.[158] Sandra Baez and colleagues observe a magnification effect of intent on perceived harm severity (but not punishment choices) in a study of judges and attorneys, suggesting motivation and expertise do not alter harm magnification as a perceptual phenomenon, but can attenuate its downstream behavioral consequences.[159]

Neuroimaging meta-analyses provide insight into the brain-level architecture supporting this perceptual bias. In meta-analyses that consider others' intentions, the mentalizing network is engaged when thinking about others' intentions in both

150. One of the eleven studies in the strategic games CBMA is quasi-cooperative (a version of the trust game). All other games are variants of the prisoner's dilemma, the ultimatum game, chicken, or other adversarial competitions. All outcomes rely on others' play, which Wilson, Stevenson, and Potts 2006 show requires more complex cognitive processing than games with deterministic outcomes.

151. One of the eleven studies in the meta-analysis included actions taken to deceive or surprise others, but not to harm them significantly.

152. Decety and Cowell 2018.

153. Carlson et al. 2022, 3.

154. Barrett et al. 2016; McNamara et al. 2019.

155. Elshout, Nelissen, and van Beest 2017; Freeman et al. 2015.

156. Ames and Fiske 2013.

157. Ames and Fiske 2015; Cushman 2008; Freeman et al. 2015; Monroe and Malle 2019.

158. Young and Tsoi 2013.

159. Baez et al. 2020.

https://doi.org/10.1017/S0020818324000328

harm-related and neutral scenarios.[160] But harm-related scenarios are associated with *additional* areas of brain activation. Specifically, multiple CBMAs of reasoning about others in harm-related versus non-harm-related scenarios find that harm-related scenarios engage either the left amygdala[161] or both the left and the right amygdalae.[162] A meta-analytic study of first- and third-person intentional harm showed amygdala activation only in the third-person case of observing *others* engaging in harmful behavior and not when engaging in harmful behavior oneself.[163] A meta-analysis of mentalizing directed at out-group members (50 studies, approximately 1,120 subjects) did not find amygdala activation,[164] suggesting that it is sensitive to *harm-related context* rather than adversarial social relationships per se.[165] Dorottya Lantos and colleagues found similar effects in a single neuroimaging study that compared videos of an out-group member issuing threats and the same out-group member making neutral statements, where the left amygdala was more active in the threat-as-signal condition.[166] Similarly, Ronald Sladky and colleagues found amygdala activation during the trust game, but only when subjects interacted with people they knew to be untrustworthy—that is, those who might intentionally harm them.[167]

Figure 3 visualizes the spatial relationship between the areas engaged in direct intention inference and the left and right amygdalae. Notably, the amygdalae are not active in any of the meta-analytic maps generated by Schurz and colleagues. The absence of amygdala activation in the workaround exercises is significant for two reasons. First, many of these CBMAs include subjects reasoning about negative outcomes, negative emotions, and others' pain, so valence alone cannot explain the lack of amygdala activation. Second, several studies included in the CBMAs for strategic games[168] and visceral simulation[169] report looking specifically for amygdala activation but not finding it, which suggests the null result is not a statistical fluke of the meta-analytic procedure. That is, the theorized workarounds are missing a key feature supporting how the brain reasons about those who intend harm.

Single-study research has also positively demonstrated that the amygdalae's engagement in the evaluation of harm-related scenarios is related to *intentionality*

160. Boccia et al. 2017; Bzdok et al. 2012; Eres, Louis, and Molenberghs 2018; Fede and Kiehl 2020 These meta-analyses contain many of the same studies, though each uses its own mix, as they were designed to study different aspects of social cognition. Across all four meta-analyses, the authors analyzed 97 unique studies with 2,407 subjects.

161. Bzdok et al. 2012 analyze 20 studies, 345 subjects; Eres, Louis, and Molenberghs 2018 analyze 84 studies, 1,963 subjects.

162. Boccia et al. 2017 analyze 31 studies, 799 subjects; Fede and Kiehl 2020 analyze 58 studies, 1,410 subjects.

163. Boccia et al. 2017.

164. Merritt et al. 2021.

165. This null result also held in a sub-analysis restricted to racial out-groups.

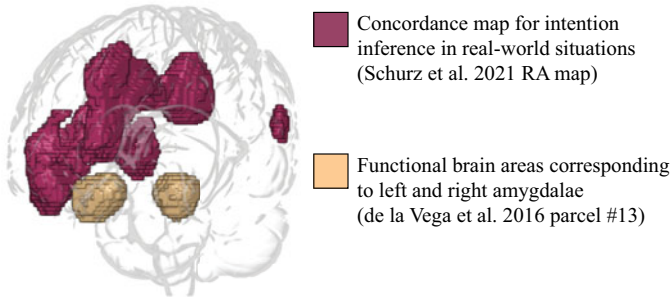
166. Lantos et al. 2020.

167. Sladky et al. 2021.

168. Takahashi et al. 2015; Zhou et al. 2014.

169. de Greck et al. 2012.

https://doi.org/10.1017/S0020818324000328



Note: Intention inference map provided by Schurz et al. and bilateral amygdala mask provided by de la Vega et al., overlaid on a glass brain using MRIcroGL 1.2.2.

FIGURE 3. *CBMA map of non-harm intention inference and the amygdalae*

and not purely to harm.[170] Eugenia Hesse and colleagues show in a small study of patients with electrodes in their left amygdala that the region responded more to intentional harm than to unintentional harm and did so faster than other brain regions.[171] In studies where people judge the wrongness of harms, the strength of the amygdala's response positively correlates with the severity of preferred punishment responses[172] and estimates of blameworthiness.[173]

Matthew Ginther and colleagues argue that the amygdala acts as a "gate," controlling whether emotional response information is passed along to brain regions engaged in decision making and complex social cognition (such as medial prefrontal cortex), and that the "gate" is open when harm by others is interpreted as intentional.[174] A behavioral meta-analysis argued for a similar process based on twenty-five studies of perceptions of intentional and unintentional harms, where perceptions of intent directly influenced preferences for aggression but also triggered negative emotional responses that additively contributed to preferences for aggression.[175] Consistent with these behavioral and brain-level findings, Hesse and colleagues posit that amygdala activity is related to harm magnification.[176] That is, perceptions of harm severity are distorted *because of the amygdala's role* in the cognitive process of evaluating intentional harms.

These findings regarding how the intention-inference piece of threat-as-signal perception is processed in the brain have several implications for the coercion literature.

170. Decety and Cacioppo 2012; Ginther et al. 2016; Hesse et al. 2016; Treadway et al. 2014.
171. Hesse et al. 2016.
172. Ginther et al. 2016.
173. Treadway et al. 2014.
174. Ginther et al. 2016.
175. Rudolph et al. 2004.
176. See the Supplementary Discussion in Hesse et al. 2016.

https://doi.org/10.1017/S0020818324000328

First, the findings regarding (fast) amygdala involvement in evaluating intentional harms suggests a brain-level bridge between theories of coercion and theories of emotion in IR. Recently, scholars have argued that receiving coercive threats-as-signals can prompt anger or even hatred, leading the recipient to resist the issuer's demands.[177] Todd Hall has proposed a similar dynamic for outrage.[178] To the extent that amygdala activity occurs in response to thinking about others' *intentions* on receipt of a threat-as-signal, Ginther and co-authors' "gate" model provides an explanation for the behaviors that these scholars have associated with hatred and anger (retaliation and resistance, respectively) and with outrage (aggression). This brain-level mechanism is also consistent with emotional accounts of coercion failures,[179] as well as accounts that are agnostic regarding the precise mental exercises involved in the choice to retaliate rather than capitulate.[180]

Second, these findings imply that efforts to make threats-as-signals more credible via costly signals may not work as the issuer intends. Costly signals provide additional information to the recipient of a threat about the issuer's intention to carry it out,[181] though the signal itself may be misinterpreted.[182] Greater credibility of a threat translates into greater certainty for the recipient that the issuer intends to harm them. In single neuroimaging studies, amygdala activation was greater during the anticipation of certain rather than uncertain harm[183] and when punishing identifiable rather than anonymous norm violators.[184] These findings suggest a positive relationship between (on the one hand) the confidence a threat's recipient has in the likelihood of harm at the hands of the issuer specifically and (on the other hand) amygdala-linked distortions and preferences (such as harm magnification or a desire to blame and punish). Given these distortions and preferences, resistance rather than capitulation is a plausible response to credible attempts at coercion.

Coercion failures may thus result from the same steps taken to improve coercion's chances of success (that is, costly signals). The potential for a null relationship between threat credibility established via costly signals and the probability of compliance has been identified experimentally.[185] In observational data, Todd Sechser has shown that compellant threats backed by demonstrations of military force (a costly signal) improve chances of success from a low baseline of 12.5 percent, but still fail half the time.[186] A brain-level account provides insight into the mechanisms behind these empirical findings.

177. McDermott, Lopez, and Hatemi 2017; Powers and Altman 2023.
178. Hall 2017.
179. Markwica 2018.
180. Dafoe, Hatz, and Zhang 2021.
181. Fearon 1997; Schelling 1966; Schultz 2001.
182. Jervis 1976; Kertzer, Rathbun, and Rathbun 2020.
183. Hur et al. 2020.
184. Feng, Tian, and Luo 2023.
185. Quek 2016.
186. Sechser 2011.

Finally, prior theoretical treatments of overblown responses to threats-as-signals argue that these responses are the product of misperceptions which can be corrected with information[187] or empathy,[188] or by addressing motivations for bias.[189] But this "correctable" account risks mistaking a feature of the brain's architecture for a bug in human reasoning. Baez and colleagues' study on judges and lawyers demonstrates that the behavioral effects of perceptual distortion can be mitigated, but not the misperception itself. A brain-level account thus suggests that to avoid escalation, short-circuiting the *consequences* of perceptual distortion may be a more fruitful intervention than correcting the misperception.

Conclusion

Within IR, the study of "threat perception" has suffered from a proliferation of conceptualizations and a dependence on untested psychological assumptions. The result is that we know less about what "threat perception" is and how it matters for IR than the volume of literature would suggest.

I have argued that the solution to both of these problems is the same: take the brain into account. Conceptually, this is straightforward. Returning to the plain-language meaning of "perception" renders the brain an essential aspect of any understanding of "threat perception." Pairing "perception" as a brain-based process with the two plain-language meanings of "threat" yields two systematized concepts grounded in the brain: threat-as-danger perception and threat-as-signal perception. These two definitions cover most custom formulations, generalize across situations, and are directly interpretable to non-experts. This conceptual ground-clearing also makes it easier to see why brain-level data offer a way to test assumptions about how threat-as-danger perception and threat-as-signal perception work in IR.

I argued that large-scale neuroimaging data are a particularly valuable resource for this kind of testing and introduced CBMAs as one way to leverage this type of data. CBMAs exist as peer-reviewed publications in their own right and thus are accessible to IR scholars. I argued that CBMAs are especially useful because they summarize the statistical concordance of many neuroimaging studies, offering generalizable insights and improving external validity in terms of samples, treatments, and outcomes. I then introduced pattern similarity analysis as a way of using CBMA data to test assumptions about psychological processes using the image-based outputs of fMRI.

In an original analysis of fifteen previously published CBMAs, I demonstrated this type of brain-level assumption testing. I first considered whether people think about

187. Fearon 1995.
188. Jervis 1978.
189. Stein 1988.

all types of harm associated with conflict as if they were economic costs, which is an assumption common in rational-choice models of conflict-related decision making. Analyzing the results of three CBMAs representing 126 unique studies, I found that there is minimal similarity between the brain's processing of monetary and nonmonetary harms. Nonmonetary harms are represented more widely across the brain, indicating greater complexity. Single neuroimaging studies and behavioral data demonstrate that complex constructs like harm are not always evaluated in ways that satisfy a rational-choice framework's requirements. That is, "blood" and "treasure" may not be so easily compared, much less added together. For this reason, alternative approaches to decision making that grant this complexity while preserving rationality (such as heuristic models) may provide better models of choices about violent conflict.

In the second example, I considered whether people treat the intentions of those who might harm them as inscrutable. Theories of coercion, both rationalist and behavioralist, often make this assumption based on the intuition that humans are not mind readers and yet must still assess the intentions of those who issue threats. Thus coercion theories often posit that people use mental workarounds, such as reasoning about the issuer's strategic interests or their characteristics, to indirectly estimate their intentions. Analyzing the results of twelve CBMAs representing 392 unique studies, I find that reasoning about those who intend harm does not look like the workarounds proposed in the literature in several important respects. Thinking about those who intend harm is characterized by amygdala activation and associated with magnified subjective perceptions of harm, heightened assessments of wrongness, and stronger preferences for blame and punishment. These perceptions and preferences shed light on several findings in the coercion literature, including high rates of coercion failure, the role of emotions, and the nature of misperceptions by those who receive threats-as-signals. There are also implications for the study of costly signaling, since brain-level processes suggest that rendering a threat-as-signal more credible may make a recipient less likely to capitulate.

Taken together, these tests illustrate the value of building an understanding of the brain into the study of "threat perception" in IR. The brain's architecture and functions are the ultimate arbiter of whether the psychological assumptions used in IR theories are valid, especially if these assumptions claim to characterize general cognitive processes underlying how most people think. Brain-level data also provide insights into processes that are unobservable and might otherwise remain subject to speculation (for example, intention inference).

While many scholars have discussed the potential for brain-level data to contribute to the study of IR, few have offered practical demonstrations of how this can be done. This article provides one such demonstration by introducing CBMAs as a source of accessible data for those looking to leverage neuroscientific evidence. This type of large-scale, cumulative data provides a more robust view of the field's findings than any single neuroimaging study. Moreover, with some training, any researcher with access to the neuroscientific literature can conduct

a new meta-analysis without the cost of original data collection. Even the existing set of CBMAs represent a trove of data on a variety of IR topics beyond threat perception, including risk and uncertainty processing[190] and intergroup relations.[191]

Skepticism is appropriate with neuroscientific data, however. CBMAs address only some of the concerns about statistical robustness and validity associated with brain-level data. Neuroscience as a field is also relatively young. Yet, relative to the state of psychology when Jervis published *Perception and Misperception in International Politics*, neuroscience is far more advanced and self-aware.[192] Jervis's book relied on single studies (some observational) for much of its theory building, while also acknowledging ongoing debates within psychology on many important fronts.[193] Even so, many of psychology's deepest issues were unknown or not acknowledged when it was published.[194] From the perspective of the field's "readiness" for outside use, integrating neuroscientific evidence into theory building and theory testing in IR today, when done carefully, represents a safer bet than Jervis made with cognitive and social psychology in the early 1970s.

Fundamentally, this article makes an argument for both conceptual and empirical consilience.[195] Aligning the study of threat perception across fields, including IR, is essential for the accumulation of knowledge. Not only does closer integration with cognitive science at a microfoundational level promise to advance the study of threat perception in IR, but it also enables IR scholarship to contribute to a broader understanding of human cognition and behavior when it comes to dealing with danger.[196]

Data Availability Statement

Replication files for this article may be found at <https://doi.org/10.7910/DVN/TBAFE2>.

Supplementary Material

Supplementary material for this article is available at <https://doi.org/10.1017/S0020818324000328>.

190. Feng et al. 2022.
191. Saarinen et al. 2021.
192. Gilmore et al. 2017.
193. See Chapter 10 of Jervis 1976 for examples.
194. Muthukrishna and Henrich 2019; Shrout and Rodgers 2018.
195. Wilson 1999.
196. Bendor 2020; Krosnick and McGraw 2002.

References

Adcock, Robert, and David Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95 (3):529–46.

Ames, Daniel L., and Susan T. Fiske. 2013. Intentional Harms Are Worse, Even When They're Not. *Psychological Science* 24 (9):1755–62.

Ames, Daniel L., and Susan T. Fiske. 2015. Perceived Intent Motivates People to Magnify Observed Harms. *Proceedings of the National Academy of Sciences* 112 (12):3599–3605.

Art, Robert J., and Kelly M. Greenhill. 2017. Coercion: An Analytical Overview. In *Coercion: The Power to Hurt in International Politics*, edited by Kelly M. Greenhill and Peter Krause, 3–32. Oxford University Press.

Babbage, Charles. 1864. *Passages from the Life of a Philosopher*. Longman.

Baez, Sandra, et al. 2020. The Impact of Legal Expertise on Moral Decision-Making Biases. *Humanities and Social Sciences Communications* 7 (1):1–12.

Baldwin, David A. 1971. Thinking About Threats. *Journal of Conflict Resolution* 15 (1):71–78.

Barrett, H. Clark, et al. 2016. Small-Scale Societies Exhibit Fundamental Variation in the Role of Intentions in Moral Judgment. *Proceedings of the National Academy of Sciences* 113 (17):4688–93.

Bartra, Oscar, Joseph T. McGuire, and Joseph W. Kable. 2013. The Valuation System: A Coordinate-Based Meta-analysis of BOLD fMRI Experiments Examining Neural Correlates of Subjective Value. *NeuroImage* 76:412–27.

Baumann, Peter. 2022. Rational Intransitive Preferences. *Politics, Philosophy and Economics* 21 (1):3–28.

Bendor, Jonathan. 2020. Bounded Rationality in Political Science and Politics. In *Oxford Handbook of Behavioral Political Science*, edited by Alex Mintz and Lesley Terris, 37–68. Oxford University Press.

Benjamin, Daniel, and Steven Simon. 2003. *The Age of Sacred Terror: Radical Islam's War Against America*. Random House.

Berkman, Elliot T., and Emily B. Falk. 2013. Beyond Brain Mapping: Using Neural Measures to Predict Real-World Outcomes. *Current Directions in Psychological Science* 22 (1):45–50.

Bertram, Teresa, Daniel Hoffmann Ayala, Maria Huber, Felix Brandl, Georg Starke, Christian Sorg, and Satja Mulej Bratec. 2023. Human Threat Circuits: Threats of Pain, Aggressive Conspecific, and Predator Elicit Distinct BOLD Activations in the Amygdala and Hypothalamus. *Frontiers in Psychiatry* 13. <https://doi.org/10.3389/fpsyt.2022.1063238>.

Blanchard, D. Caroline. 2017. Translating Dynamic Defense Patterns from Rodents to People. *Neuroscience and Biobehavioral Reviews* 76:22–28.

Boccia, Maddalena, C. Dacquino, Laura Piccardi, Pierluigi Cordellieri, Cecilia Guariglia, Fabio Ferlazzo, Stefano Ferracuti, and Anna Maria Giannini. 2017. Neural Foundation of Human Moral Reasoning: An ALE Meta-analysis About the Role of Personal Perspective. *Brain Imaging and Behavior* 11 (1):278–92.

Brandstätter, Eduard, Gerd Gigerenzer, and Ralph Hertwig. 2006. The Priority Heuristic: Making Choices Without Trade-Offs. *Psychological Review* (US) 113 (2):409–432.

Briggs, Robert G., et al. 2022. Parcellation-Based Tractographic Modeling of the Salience Network Through Meta-analysis. *Brain and Behavior* 12 (7):e2646.

Bueno de Mesquita, Bruce. 1983. The Costs of War: A Rational Expectations Approach. *American Political Science Review* 77 (2):347–57.

Bueno de Mesquita, Bruce. 2010. Foreign Policy Analysis and Rational Choice Models. In *Oxford Research Encyclopedia of International Studies*.

Bueno de Mesquita, Bruce, James D. Morrow, and Ethan R. Zorick. 1997. Capabilities, Perception, and Escalation. *American Political Science Review* 91 (1):15–27.

Bulley, Adam, Julie D. Henry, and Thomas Suddendorf. 2017. Thinking About Threats: Memory and Prospection in Human Threat Management. *Consciousness and Cognition* 49:53–69.

Byman, Daniel, and Matthew Waxman. 2002. *The Dynamics of Coercion: American Foreign Policy and the Limits of Military Might*. Cambridge University Press.

Bzdok, Danilo, Leonhard Schilbach, Kai Vogeley, Karla Schneider, Angela R. Laird, Robert Langner, and Simon B. Eickhoff. 2012. Parsing the Neural Correlates of Moral Cognition: ALE Meta-analysis on Morality, Theory of Mind, and Empathy. *Brain Structure and Function* 217 (4):783–96.

Camerer, Colin F., et al. 2018. Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour* 2 (9):637–44.

Carlson, Ryan W., Yochanan E. Bigman, Kurt Gray, Melissa J. Ferguson, and M.J. Crockett. 2022. How Inferred Motives Shape Moral Judgements. *Nature Reviews Psychology* 1 (8):468–78.

Carnegie, Allison. 2015. *Power Plays: How International Institutions Reshape Coercive Diplomacy*. Cambridge University Press.

Chamberlain, Dianne Pfundstein. 2016. *Cheap Threats: Why the United States Struggles to Coerce Weak States*. Georgetown University Press.

Chavanne, Alice V., and Oliver J. Robinson. 2021. The Overlapping Neurobiology of Induced and Pathological Anxiety: A Meta-analysis of Functional Neural Activation. *American Journal of Psychiatry* 178 (2):156–64.

Christensen, Thomas J., and Jack Snyder. 1990. Chain Gangs and Passed Bucks: Predicting Alliance Patterns in Multipolarity. *International Organization* 44 (2):137–68.

Clairis, Nicolas, and Alizée Lopez-Persem. 2023. Debates on the Dorsomedial Prefrontal/Dorsal Anterior Cingulate Cortex: Insights for Future Research. *Brain* 146 (12):4826–44.

Cohen, Raymond. 1978. Threat Perception in International Crisis. *Political Science Quarterly* 93 (1): 93–107.

Coppock, Alexander. 2019. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods* 7 (3):613–28.

Cushman, Fiery. 2008. Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment. *Cognition* 108 (2):353–80.

Dafoe, Allan, Sophia Hatz, and Baobao Zhang. 2021. Coercion and Provocation. *Journal of Conflict Resolution* 65 (2–3):372–402.

Dafoe, Allan, Jonathan Renshon, and Paul Huth. 2014. Reputation and Status as Motives for War. *Annual Review of Political Science* 17 (1):371–93.

Davis, James W. 2000. *Threats and Promises: The Pursuit of International Influence*. Johns Hopkins University Press.

Davis, James W., and Rose McDermott. 2021. The Past, Present, and Future of Behavioral IR. *International Organization* 75 (1):147–77.

de Greck, Moritz, Zhenhao Shi, Gang Wang, Xiangyu Zuo, Xuedong Yang, Xiaoying Wang, Georg Northoff, and Shihui Han. 2012. Culture Modulates Brain Activity During Empathy with Anger. *NeuroImage* 59 (3):2871–82.

de la Vega, Alejandro, Luke J. Chang, Marie T. Banich, Tor D. Wager, and Tal Yarkoni. 2016. Large-Scale Meta-analysis of Human Medial Frontal Cortex Reveals Tripartite Functional Organization. *Journal of Neuroscience* 36 (24):6553–62.

Decety, Jean, and Stephanie Cacioppo. 2012. The Speed of Morality: A High-Density Electrical Neuroimaging Study. *Journal of Neurophysiology* 108 (11):3068–3072.

Decety, Jean, and Jason M. Cowell. 2018. Interpersonal Harm Aversion As a Necessary Foundation for Morality: A Developmental Neuroscience Perspective. *Development and Psychopathology* 30 (1): 153–64.

Defendini, Ana, and Adrianna C. Jenkins. 2023. Dissociating Neural Sensitivity to Target Identity and Mental State Content Type During Inferences About Other Minds. *Social Neuroscience* 18 (2):103–121.

Dotson, Vonetta M., and Audrey Duarte. 2020. The Importance of Diversity in Cognitive Neuroscience. *Annals of the New York Academy of Sciences* 1464 (1):181–91.

Du, Jingnan, Edmund T. Rolls, Wei Cheng, Yu Li, Weikang Gong, Jiang Qiu, and Jianfeng Feng. 2020. Functional Connectivity of the Orbitofrontal Cortex, Anterior Cingulate Cortex, and Inferior Frontal Gyrus in Humans. *Cortex* 123:185–99.

Egami, Naoki, and Erin Hartman. 2022. Elements of External Validity: Framework, Design, and Analysis. *American Political Science Review* 117 (3):1070–1088.

Eickhoff, Simon B., B.T. Thomas Yeo, and Sarah Genon. 2018. Imaging-Based Parcellations of the Human Brain. *Nature Reviews Neuroscience* 19 (11):672–86.
Eilam, David, Rony Izhar, and Joel Mort. 2011. Threat Detection: Behavioral Practices in Animals and Humans. *Neuroscience and Biobehavioral Reviews* 35 (4):999–1006.
Elshout, Maartje, Rob M.A. Nelissen, and Ilja van Beest. 2017. Your Act Is Worse Than Mine: Perception Bias in Revenge Situations. *Aggressive Behavior* 43 (6):553–57.
Eres, Robert, Winnifred R. Louis, and Pascal Molenberghs. 2018. Common and Distinct Neural Networks Involved in fMRI Studies Investigating Morality: An ALE Meta-analysis. *Social Neuroscience* 13 (4): 384–98.
Falk, Emily B., Matthew Brook O'Donnell, Steven Tompson, Richard Gonzalez, Sonya Dal Cin, Victor Strecher, Kenneth Michael Cummings, and Lawrence An. 2016. Functional Brain Imaging Predicts Public Health Campaign Success. *Social Cognitive and Affective Neuroscience* 11 (2):204–214.
Farnham, Barbara. 2003. The Theory of Democratic Peace and Threat Perception. *International Studies Quarterly* 47 (3):395–415.
Fearon, James D. 1994a. Domestic Political Audiences and the Escalation of International Disputes. *American Political Science Review* 88 (3):577–92.
Fearon, James D. 1994b. Signaling Versus the Balance of Power and Interests: An Empirical Test of a Crisis Bargaining Model. *Journal of Conflict Resolution* 38 (2):236–69.
Fearon, James D. 1995. Rationalist Explanations for War. *International Organization* 49 (3):379–414.
Fearon, James D. 1997. Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs. *Journal of Conflict Resolution* 41 (1):68–90.
Fede, Samantha J., and Kent A. Kiehl. 2020. Meta-analysis of the Moral Brain: Patterns of Neural Engagement Assessed Using Multilevel Kernel Density Analysis. *Brain Imaging and Behavior* 14 (2):534–47.
Fedorenko, Evelina, John Duncan, and Nancy Kanwisher. 2013. Broad Domain Generality in Focal Regions of Frontal and Parietal Cortex. *Proceedings of the National Academy of Sciences* 110 (41): 16616–21.
Fehlbaum, Lynn V., Réka Borbás, Katharina Paul, Simon B. Eickhoff, and Nora M. Raschle. 2022. Early and Late Neural Correlates of Mentalizing: ALE Meta-analyses in Adults, Children and Adolescents. *Social Cognitive and Affective Neuroscience* 17 (4):351–66.
Feng, Chunliang, Xia Tian, and Yue-Jia Luo. 2023. Neurocomputational Substrates Underlying the Effect of Identifiability on Third-Party Punishment. *Journal of Neuroscience* 43 (47):8018–31.
Feng, Shuqing, Meng Zhang, Yunwen Peng, Shiyan Yang, Yufeng Wang, Xin Wu, and Feng Zou. 2022. Brain Networks Under Uncertainty: A Coordinate-Based Meta-analysis of Brain Imaging Studies. *Journal of Affective Disorders* 319:627–37.
Fordham, Benjamin. 1998. The Politics of Threat Perception and the Use of Force: A Political Economy Model of US Uses of Force, 1949–1994. *International Studies Quarterly* 42 (3):567–90.
Freeman, Daniel, Nicole Evans, Emma Černis, Rachel Lister, and Graham Dunn. 2015. The Effect of Paranoia on the Judging of Harmful Events. *Cognitive Neuropsychiatry* 20 (2):122–27.
Frith, Uta, and Sarah-Jayne Blakemore. 2006. Social Cognition. In *Cognitive Systems: Information Processing Meets Brain Science*, edited by Richard Morris, Lionel Tarassenko, and Michael Kenward, 138–62. London: Academic Press.
Gammon, Earl. 2020. Affective Neuroscience, Emotional Regulation, and International Relations. *International Theory* 12 (2):189–219.
George, Alexander L., and William F. Simons, eds. 1971. *Limits of Coercive Diplomacy*. Little, Brown.
Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. Heuristic Decision Making. *Annual Review of Psychology* 62:451–82.
Gilmore, Rick O., Michele T. Diaz, Brad A. Wyble, and Tal Yarkoni. 2017. Progress Toward Openness, Transparency, and Reproducibility in Cognitive Neuroscience. *Annals of the New York Academy of Sciences* 1396 (1):5–18.
Ginther, Matthew R., Richard J. Bonnie, Morris B. Hoffman, Francis X. Shen, Kenneth W. Simons, Owen D. Jones, and René Marois. 2016. Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. *Journal of Neuroscience* 36 (36):9420–34.

Glimcher, Paul W. 2022. Efficiently Irrational: Deciphering the Riddle of Human Choice. *Trends in Cognitive Sciences* 26 (8):669–87.
Gold, Natalie, Briony D. Pulford, and Andrew M. Colman. 2013. Your Money or Your Life: Comparing Judgements in Trolley Problems Involving Economic and Emotional Harms, Injury and Death. *Economics and Philosophy* 29 (2):213–33.
Greenhill, Kelly M. 2011. *Weapons of Mass Migration: Forced Displacement, Coercion, and Foreign Policy*. Cornell University Press.
Greenhill, Kelly M., and Peter Krause. 2017. *Coercion: The Power to Hurt in International Politics*. Oxford University Press.
Gu, Xiaosi, Patrick R. Hof, Karl J. Friston, and Jin Fan. 2013. Anterior Insular Cortex and Emotional Awareness. *Journal of Comparative Neurology* 521 (15):3371–88.
Haas, Ingrid J. 2016. Political Neuroscience. In *Neuroimaging Personality, Social Cognition, and Character*, edited by John R. Absher and Jasmin Cloutier, 355–70. Academic Press.
Hall, Todd H. 2017. On Provocation: Outrage, International Relations, and the Franco–Prussian War. *Security Studies* 26 (1):1–29.
Hall, Todd H., and Keren Yarhi-Milo. 2012. The Personal Touch: Leaders' Impressions, Costly Signaling, and Assessments of Sincerity in International Affairs. *International Studies Quarterly* 56 (3):560–73.
Hassner, Ron E. 2016. *Religion on the Battlefield*. Cornell University Press.
Haxby, James V., Andrew C. Connolly, and J. Swaroop Guntupalli. 2014. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience* 37 (1): 435–56.
Hayden, Benjamin Y., and Yael Niv. 2021. The Case Against Economic Values in the Orbitofrontal Cortex (or Anywhere Else in the Brain). *Behavioral Neuroscience* 135 (2):192–201.
He, Kai. 2022. Polarity and Threat Perception in Foreign Policy: A Dynamic Balancing Model. In *Polarity in International Relations: Past, Present, Future*, edited by Nina Græger, Bertel Heurlin, Ole Wæver, and Anders Wivel, 45–61. Springer International.
Herrmann, Richard K. 2013. Perceptions and Image Theory in International Relations. In *The Oxford Handbook of Political Psychology*, 2nd ed., 334–63. Oxford University Press.
Hesse, Eugenia, et al. 2016. Early Detection of Intentional Harm in the Human Amygdala. *Brain* 139 (1): 54–61.
Heyes, Cecilia, and Caroline Catmur. 2022. What Happened to Mirror Neurons? *Perspectives on Psychological Science* 17 (1):153–68.
Hobbes, Thomas. 1651. *Leviathan*. Edited by Richard Tuck. Cambridge University Press (1996).
Holmes, Marcus. 2013. The Force of Face-to-Face Diplomacy: Mirror Neurons and the Problem of Intentions. *International Organization* 67 (4):829–61.
Holmes, Marcus. 2018. *Face-to-Face Diplomacy: Social Neuroscience and International Relations*. Cambridge University Press.
Hruska, Pam, Kent G. Hecker, Sylvain Coderre, Kevin McLaughlin, Filomeno Cortese, Christopher Doig, Tanya Beran, Bruce Wright, and Olav Krigolson. 2016. Hemispheric Activation Differences in Novice and Expert Clinicians During Clinical Decision Making. *Advances in Health Sciences Education* 21 (5): 921–33.
Hur, Juyoen, Jason F. Smith, Kathryn A. DeYoung, Allegra S. Anderson, Jinyi Kuang, Hyung Cho Kim, Rachael M. Tillman, Manuel Kuhn, Andrew S. Fox, and Alexander J. Shackman. 2020. Anxiety and the Neurobiology of Temporally Uncertain Threat Anticipation. *Journal of Neuroscience* 40 (41):7949–64.
Huth, Paul K. 1999. Deterrence and International Conflict: Empirical Findings and Theoretical Debates. *Annual Review of Political Science* 2:25–48.
Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton University Press.
Jervis, Robert. 1978. Cooperation Under the Security Dilemma. *World Politics* 30 (2):167–214.
Jervis, Robert. 1979. Deterrence Theory Revisited. *World Politics* 31 (2):289–324.
Jervis, Robert. 1982. Deterrence and Perception. *International Security* 7 (3):3–30.
Jervis, Robert. 1989. *The Logic of Images in International Relations*. Columbia University Press.

Jervis, Robert, Keren Yarhi-Milo, and Don Casler. 2021. Redefining the Debate over Reputation and Credibility in International Security: Promises and Limits of New Scholarship. *World Politics* 73 (1): 167–203.

Jost, John T., H. Hannah Nam, David M. Amodio, and Jay J. Van Bavel. 2014. Political Neuroscience: The Beginning of a Beautiful Friendship. *Political Psychology* 35 (S1):3–42.

Juechems, Keno, and Christopher Summerfield. 2019. Where Does Value Come From? *Trends in Cognitive Sciences* 23 (10):836–50.

Kahneman, Daniel, and Jonathan Renshon. 2009. Hawkish Biases. In *American Foreign Policy and The Politics of Fear: Threat Inflation Since 9/11*, edited by A. Trevor Thrall and Jane K. Cramer, 79–96. Routledge.

Kalenscher, Tobias, Philippe Tobler, Willem Huijbers, Sander Daselaar, and Cyriel Pennartz. 2010. Neural Signatures of Intransitive Preferences. *Frontiers in Human Neuroscience* 4. <https://doi.org/10.3389/fnhum.2010.00049>.

Kanwisher, Nancy. 2010. Functional Specificity in the Human Brain: A Window into the Functional Architecture of the Mind. *Proceedings of the National Academy of Sciences* 107 (25):11163–70.

Kertzer, Joshua D. 2017. Microfoundations in International Relations. *Conflict Management and Peace Science* 34 (1):81–97.

Kertzer, Joshua D. 2022. Re-assessing Elite–Public Gaps in Political Behavior. *American Journal of Political Science* 66 (3):539–53.

Kertzer, Joshua D., and Kathleen M. McGraw. 2012. Folk Realism: Testing the Microfoundations of Realism in Ordinary Citizens. *International Studies Quarterly* 56 (2):245–58.

Kertzer, Joshua D., Brian C. Rathbun, and Nina Srinivasan Rathbun. 2020. The Price of Peace: Motivated Reasoning and Costly Signaling in International Relations. *International Organization* 74 (1):95–118.

Kertzer, Joshua D., and Dustin Tingley. 2018. Political Psychology in International Relations: Beyond the Paradigms. *Annual Review of Political Science* 21:1–23.

Kilgour, D. Marc, and Frank C. Zagare. 1991. Credibility, Uncertainty, and Deterrence. *American Journal of Political Science* 35 (2):305–334.

Kobayashi, Kenji, Joseph W. Kable, Ming Hsu, and Adrianna C. Jenkins. 2022. Neural Representations of Others' Traits Predict Social Decisions. *Proceedings of the National Academy of Sciences* 119 (22): e2116944119.

Korn, Christoph W., and Dominik R. Bach. 2019. Minimizing Threat via Heuristic and Optimal Policies Recruits Hippocampus and Medial Prefrontal Cortex. *Nature Human Behaviour* 3 (7):733–45.

Krawczyk, Daniel C., Amy L. Boggan, M. Michelle McClelland, and James C. Bartlett. 2011. The Neural Organization of Perception in Chess Experts. *Neuroscience Letters* 499 (2):64–69.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini. 2008. Representational Similarity Analysis: Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience* 2 (November).

Krosnick, Jon A., and Kathleen M. McGraw. 2002. Psychological Political Science Versus Political Psychology True to Its Name: A Plea for Balance. In *Political Psychology*, edited by Kristen Renwick Monroe, 79–94. Lawrence Erlbaum Associates.

Kydd, Andrew. 2010. Rationalist Approaches to Conflict Prevention and Resolution. *Annual Review of Political Science* 13 (1):101–121.

Landau-Wells, Marika. 2018. Dealing with Danger: Threat Perception and Policy Preferences. PhD diss., Massachusetts Institute of Technology. Available at <http://hdl.handle.net/1721.1/118222>.

Landau-Wells, Marika, and Rebecca Saxe. 2020. Political Preferences and Threat Perception: Opportunities forNeuroimaging and Developmental Research. *Current Opinion in Behavioral Sciences* 34:58–63.

Lantos, Dorottya, Yong Hui Lau, Winnifred Louis, and Pascal Molenberghs. 2020. The Neural Mechanisms of Threat and Reconciliation Efforts Between Muslims and Non-Muslims. *Social Neuroscience* 15 (4):420–34.

Lantos, Dorottya, and Pascal Molenberghs. 2021. The Neuroscience of Intergroup Threat and Violence. *Neuroscience and Biobehavioral Reviews* 131:77–87.

Laureiro-Martínez, Daniella, Nicola Canessa, Stefano Brusoni, Maurizio Zollo, Todd Hare, Federica Alemanno, and Stefano Cappa. 2014. Frontopolar Cortex and Decision-Making Efficiency: Comparing Brain Activity of Experts with Different Professional Background During an Exploration-Exploitation Task. *Frontiers in Human Neuroscience* 7. <https://doi.org/10.3389/fnhum.2013.00927>.

LeDoux, Joseph E. 2022. As Soon As There Was Life, There Was Danger: The Deep History of Survival Behaviours and the Shallower History of Consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377 (1844):20210292.

Levy, Jack S. 1983. Misperception and the Causes of War: Theoretical Linkages and Analytical Problems. *World Politics* 36 (1):76–99.

Levy, Jack S. 1998. The Causes of War and the Conditions of Peace. *Annual Review of Political Science* 1: 139–65.

Lieberman, Matthew D., Mark A. Straccia, Meghan L. Meyer, Meng Du, and Kevin M. Tan. 2019. Social, Self, (Situational), and Affective Processes in Medial Prefrontal Cortex (MPFC): Causal, Multivariate, and Reverse Inference Evidence. *Neuroscience & Biobehavioral Reviews* 99:311–28.

Lillehammer, Hallvard. 2023. *The Trolley Problem*. Cambridge University Press.

Markwica, Robin. 2018. *Emotional Choices: Howthe Logic of Affect Shapes Coercive Diplomacy*. Oxford University Press.

McCurry, Katherine L., B. Christopher Frueh, Pearl H. Chiu, and Brooks King-Casas. 2020. Opponent Effects of Hyperarousal and Re-experiencing on Affective Habituation in Posttraumatic Stress Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 5 (2):203–212.

McDermott, Rose. 2004. The Feeling of Rationality: The Meaning of Neuroscientific Advances for Political Science. *Perspectives on Politics* 2 (4):691–706.

McDermott, Rose. 2017. Emotions in Foreign Policy Decision Making. In *Oxford Research Encyclopedia of Politics*. <https://doi.org/10.1093/acrefore/9780190228637.013.418>.

McDermott, Rose, Anthony C. Lopez, and Peter K. Hatemi. 2017. "Blunt Not the Heart, Enrage It": The Psychology of Revenge and Deterrence. *Texas National Security Review* 1 (1). Available at <http://hdl.handle.net/2152/63934>.

McNamara, Rita Anne, Aiyana K. Willard, Ara Norenzayan, and Joseph Henrich. 2019. Weighing Outcome vs. Intent Across Societies: How Cultural Models of Mind Shape Moral Reasoning. *Cognition* 182:95–108.

Mearsheimer, John J. 2001. *The Tragedy of Great Power Politics*. Norton.

Merritt, Carrington C., Jennifer K. MacCormack, Andrea G. Stein, Kristen A. Lindquist, and Keely A. Muscatell. 2021. The Neural Underpinnings of Intergroup Social Cognition: An fMRI Meta-analysis. *Social Cognitive and Affective Neuroscience* 16 (9):903–914.

Monroe, Andrew E., and Bertram F. Malle. 2019. People Systematically Update Moral Judgments of Blame. *Journal of Personality and Social Psychology* (US) 116 (2):215–36.

Morawetz, Carmen, Stefan Bode, Birgit Derntl, and Hauke R. Heekeren. 2017. The Effect of Strategies, Goals and Stimulus Material on the Neural Mechanisms of Emotion Regulation: A Meta-analysis of fMRI Studies. *Neuroscience and Biobehavioral Reviews* 72:111–28.

Morgan, T. Clifton, Navin A. Bapat, and Yoshiharu Kobayashi. 2021. The Threat and Imposition of Economic Sanctions Data Project: A Retrospective. In *Research Handbook on Economic Sanctions*, edited by Peter A.G. van Bergeijk, 44–61. Elgar.

Muthukrishna, Michael, and Joseph Henrich. 2019. A Problem in Theory. *Nature Human Behaviour* 3 (3): 221–29.

Myrick, Rachel. 2021. Do External Threats Unite or Divide? Security Crises, Rivalries, and Polarization in American Foreign Policy. *International Organization* 75 (4):921–58.

Pape, Robert A. 1996. *Bombing to Win: Air Power and Coercion in War*. Cornell University Press.

Park, Hae-Jeong, and Karl Friston. 2013. Structural and Functional Brain Networks: From Connections to Cognition. *Science* 342 (6158):1238411.

Poldrack, Russell A., Jeanette A. Mumford, and Thomas E. Nichols. 2011. *Handbook of Functional MRI Data Analysis*. Cambridge University Press.

Posen, Barry R. 1993. The Security Dilemma and Ethnic Conflict. *Survival* 35 (1):27–47.

Powell, Robert. 1987. Crisis Bargaining, Escalation, and MAD. *American Political Science Review* 81 (3): 717–35.
Powell, Robert. 2002. Bargaining Theory and International Conflict. *Annual Review of Political Science* 5 (1):1–30.
Powell, Robert. 2006. War As a Commitment Problem. *International Organization* 60 (1):169–203.
Powers, Kathleen E., and Dan Altman. 2023. The Psychology of Coercion Failure: How Reactance Explains Resistance to Threats. *American Journal of Political Science* 67 (1):221–38.
Press, Daryl Grayson. 2005. *Calculating Credibility: How Leaders Assess Military Threats*. Cornell University Press.
Price, Richard, and Kathryn Sikkink. 2021. *International Norms, Moral Psychology, and Neuroscience*. Cambridge University Press.
Quackenbush, Stephen. 2004. The Rationality of Rational Choice Theory. *International Interactions* 30 (2):87–107.
Quek, Kai. 2016. Are Costly Signals More Credible? Evidence of Sender–Receiver Gaps. *Journal of Politics* 78 (3):925–40.
Quesque, François, Ian Apperly, Renée Baillargeon, Simon Baron-Cohen, Cristina Becchio, Harold Bekkering, Daniel Bernstein, et al. 2024. Defining Key Concepts for Mental State Attribution. *Communications Psychology* 2 (1):1–5.
Ranyard, Rob, Henry Montgomery, Emmanouil Konstantinidis, and Andrea Louise Taylor. 2020. Intransitivity and Transitivity of Preferences: Dimensional Processing in Decision Making. *Decision* 7 (4):287–313.
Rathbun, Brian C., Joshua D. Kertzer, Jason Reifler, Paul Goren, and Thomas J. Scotto. 2016. Taking Foreign Policy Personally: Personal Values and Foreign Policy Attitudes. *International Studies Quarterly* 60 (1):124–37.
Rosato, Sebastian. 2015. The Inscrutable Intentions of Great Powers. *International Security* 39 (3):48–88.
Rousseau, David L., and Rocio Garcia-Retamero. 2007. Identity, Power, and Threat Perception: A Cross-National Experimental Study. *Journal of Conflict Resolution* 51 (5):744–71.
Rudolph, Udo, Scott Roesch, Tobias Greitemeyer, and Bernard Weiner. 2004. A Meta-analytic Review of Help Giving and Aggression from an Attributional Perspective: Contributions to a General Theory of Motivation. *Cognition and Emotion* 18 (6):815–48.
Saarinen, Aino, Iiro P. Jääskeläinen, Ville Harjunen, Liisa Keltikangas-Järvinen, Inga Jasinskaja-Lahti, and Niklas Ravaja. 2021. Neural Basis of In-Group Bias and Prejudices: A Systematic Meta-analysis. *Neuroscience & Biobehavioral Reviews* 131:1214–27.
Sagan, Scott D., and Jeremi Suri. 2003. The Madman Nuclear Alert: Secrecy, Signaling, and Safety in October 1969. *International Security* 27 (4):150–83.
Samartsidis, Pantelis, Silvia Montagna, Thomas E. Nichols, and Timothy D. Johnson. 2017. The Coordinate-Based Meta-analysis of Neuroimaging Data. *Statistical Science* 32 (4):580–99.
Sartori, Anne E. 2013. *Deterrence by Diplomacy*. Princeton University Press.
Saxe, Rebecca. 2005. Against Simulation: The Argument from Error. *Trends in Cognitive Sciences* 9 (4): 174–79.
Saxe, Rebecca, and Nancy Kanwisher. 2003. People Thinking About Thinking People: The Role of the Temporo-Parietal Junction in "Theory of Mind." *NeuroImage* 19 (4):1835–42.
Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard University Press.
Schelling, Thomas C. 1966. *Arms and Influence*. 1st ed. Yale University Press.
Scholz, Christin, Elisa C. Baek, Matthew Brook O'Donnell, Hyun Suk Kim, Joseph N. Cappella, and Emily B. Falk. 2017. A Neural Model of Valuation and Information Virality. *Proceedings of the National Academy of Sciences* 114 (11):2881–86.
Schultz, Kenneth A. 1998. Domestic Opposition and Signaling in International Crises. *American Political Science Review* 92 (4):829–44.
Schultz, Kenneth A. 2001. *Democracy and Coercive Diplomacy*. Cambridge University Press.
Schurz, Matthias, Joaquim Radua, Matthias G. Tholen, Lara Maliske, Daniel S. Margulies, Rogier B. Mars, Jerome Sallet, and Philipp Kanske. 2021. Toward a Hierarchical Model of Social Cognition: A

Neuroimaging Meta-analysis and Integrative Review of Empathy and Theory of Mind. *Psychological Bulletin* 147 (3):293–327.
Sechser, Todd S. 2011. Militarized Compellent Threats, 1918–2001. *Conflict Management and Peace Science* 28 (4):377–401.
Shackman, Alexander J., Tim V. Salomons, Heleen A. Slagter, Andrew S. Fox, Jameel J. Winter, and Richard J. Davidson. 2011. The Integration of Negative Affect, Pain and Cognitive Control in the Cingulate Cortex. *Nature Reviews Neuroscience* 12 (3):154–67.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton, Mifflin.
Shenhav, Amitai, Mark A. Straccia, Jonathan D. Cohen, and Matthew M. Botvinick. 2014. Anterior Cingulate Engagement in a Foraging Context Reflects Choice Difficulty, Not Foraging Value. *Nature Neuroscience* 17 (9):1249–54.
Shine, James M., and Russell A. Poldrack. 2018. Principles of Dynamic Network Reconfiguration Across Diverse Brain States. *NeuroImage* 180:396–405.
Shinkareva, Svetlana V., Jing Wang, and Douglas H. Wedell. 2013. Examining Similarity Structure: Multidimensional Scaling and Related Approaches in Neuroimaging. *Computational and Mathematical Methods in Medicine* 2013. <https://doi.org/10.1155/2013/796183>.
Shrout, Patrick E., and Joseph L. Rodgers. 2018. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology* 69 (1):487–510.
Singer, David J. 1958. Threat-Perception and the Armament-Tension Dilemma. *Journal of Conflict Resolution* 2 (1):90–105.
Sladky, Ronald, Federica Riva, Lisa Anna Rosenberger, Jack van Honk, and Claus Lamm. 2021. Basolateral and Central Amygdala Orchestrate How We Learn Whom to Trust. *Communications Biology* 4 (1):1–9.
Slantchev, Branislav L. 2010. Feigning Weakness. *International Organization* 64 (3):357–88.
Slantchev, Branislav L. 2011. *Military Threats: The Costs of Coercion and the Price of Peace*. Cambridge University Press.
Snyder, Glenn H. 1971. "Prisoner's Dilemma" and "Chicken" Models in International Politics. *International Studies Quarterly* 15 (1):66–103.
Snyder, Glenn H. 1990. Alliance Theory: A Neorealist First Cut. *Journal of International Affairs* 44 (1): 103–123.
Speer, Sebastian P.H., Christian Keysers, Judit Campdepadrós Barrios, Cas J.S. Teurlings, Ale Smidts, Maarten A.S. Boksem, Tor D. Wager, and Valeria Gazzola. 2023. A Multivariate Brain Signature for Reward. *NeuroImage* 271:119990.
Stein, Janice Gross. 1988. Building Politics into Psychology: The Misperception of Threat. *Political Psychology* 9 (2):245–71.
Stein, Janice Gross. 2002. Psychological Expalanations of International Conflict. In *Handbook of International Relations*, edited by Walter Carlsnaes, Thomas Risse, and Beth A. Simmons, 292–308. Sage.
Stein, Janice Gross. 2013. Threat Perception in International Relations. In *Oxford Handbook of Political Psychology*, 2nd ed., edited by Leonie Huddy, David O. Sears, and Jack S. Levy, 364–94. Oxford University Press.
Stein, Janice Gross. 2017. The Micro-foundations of International Relations Theory: Psychology and Behavioral Economics. *International Organization* 71 (S1):S249–S263.
Stevenson, Angus, ed. 2010. *Oxford Dictionary of English*. Electronic resource. Oxford University Press.
Szucs, Denes, and John P.A. Ioannidis. 2020. Sample Size Evolution in Neuroimaging Research: An Evaluation of Highly-Cited Studies (1990–2012) and of Latest Practices (2017–2018) in High-Impact Journals. *NeuroImage* 221:117164.
Takahashi, Hideyuki, Keise Izuma, Madoka Matsumoto, Kenji Matsumoto, and Takashi Omori. 2015. The Anterior Insula Tracks Behavioral Entropy During an Interpersonal Competitive Game. *PLOS One* 10 (6):e0123329.
Tan, Huixin, Qin Duan, Yihan Liu, Xinyu Qiao, and Siyang Luo. 2022. Does Losing Money Truly Hurt? The Shared Neural Bases of Monetary Loss and Pain. *Human Brain Mapping* 43 (10):3153–63.

Tang, Shiping. 2008. Fear in International Politics: Two Positions. *International Studies Review* 10 (3): 451–71.
Theodoridis, Alexander G., and Amy J. Nelson. 2012. Of BOLD Claims and Excessive Fears: A Call for Caution and Patience Regarding Political Neuroscience. *Political Psychology* 33 (1):27–43.
Tolomeo, Serenella, David Christmas, Ines Jentzsch, Blair Johnston, Reiner Sprengelmeyer, Keith Matthews, and J. Douglas Steele. 2016. A Causal Role for the Anterior Mid-Cingulate Cortex in Negative Affect and Cognitive Control. *Brain* 139 (6):1844–54.
Tomaino, Geoff, Klaus Wertenbroch, and Daniel J. Walters. 2023. Intransitivity of Consumer Preferences for Privacy. *Journal of Marketing Research* 60 (3):489–507.
Touroutoglou, Alexandra, Mark Hollenbeck, Bradford C. Dickerson, and Lisa Feldman Barrett. 2012. Dissociable Large-Scale Networks Anchored in the Right Anterior Insula Subserve Affective Experience and Attention. *NeuroImage* 60 (4):1947–58.
Tozzi, Leonardo, Xue Zhang, Megan Chesnut, Bailey Holt-Gosselin, Carolina A. Ramirez, and Leanne M. Williams. 2021. Reduced Functional Connectivity of Default Mode Network Subsystems in Depression: Meta-analytic Evidence and Relationship with Trait Rumination. *NeuroImage: Clinical* 30:102570.
Treadway, Michael T., Joshua W. Buckholtz, Justin W. Martin, Katharine Jan, Christopher L. Asplund, Matthew R. Ginther, Owen D. Jones, and René Marois. 2014. Corticolimbic Gating of Emotion-Driven Punishment. *Nature Neuroscience* 17 (9):1270–75.
Tsetsos, Konstantinos, Rani Moran, James Moreland, Nick Chater, Marius Usher, and Christopher Summerfield. 2016. Economic Irrationality Is Optimal During Noisy Decision Making. *Proceedings of the National Academy of Sciences* 113 (11):3102–3107.
Turner, Benjamin O., Erick J. Paul, Michael B. Miller, and Aron K. Barbey. 2018. Small Sample Sizes Reduce the Replicability of Task-Based fMRI Studies. *Communications Biology* 1 (62). <https://doi.org/10.1038/s42003-018-0073-z>.
Tversky, Amos. 1969. Intransitivity of Preferences. *Psychological Review* 76 (1):31–48.
Tversky, Amos, and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211 (4481):453–58.
Walasek, Lukasz, and Gordon D.A. Brown. 2023. Incomparability and Incommensurability in Choice: No Common Currency of Value? *Perspectives on Psychological Science* 19 (6):1011–1030.
Walt, Stephen M. 1987. *Origins of Alliances*. Cornell University Press.
Waltz, Kenneth N. 1954. *Man, the State, and War: A Theoretical Analysis*. Columbia University Press.
Wilson, Edward O. 1999. *Consilience: The Unity of Knowledge*. Vintage Books.
Wilson, Rick K., Randolph Stevenson, and Geoffrey Potts. 2006. Brain Activity in the Play of Dominant Strategy and Mixed Strategy Games. *Political Psychology* 27 (3):459–78.
Wong, Seanon S. 2016. Emotions and the Communication of Intentions in Face-to-Face Diplomacy. *European Journal of International Relations* 22 (1):144–67.
Woody, Erik Z., and Henry Szechtman. 2011. Adaptation to Potential Threat: The Evolution, Neurobiology, and Psychopathology of the Security Motivation System. *Neuroscience and Biobehavioral Reviews* 35 (4):1019–1033.
Yarhi-Milo, Keren. 2013. In the Eye of the Beholder: How Leaders and Intelligence Communities Assess the Intentions of Adversaries. *International Security* 38 (1):7–51.
Yarkoni, Tal, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. 2011. Large-Scale Automated Synthesis of Human Functional Neuroimaging Data. *Nature Methods* 8 (8): 665–70.
Yeung, Andy Wai Kan, Michaela Robertson, Angela Uecker, Peter T. Fox, and Simon B. Eickhoff. 2023. Trends in the Sample Size, Statistics, and Contributions to the BrainMap Database of Activation Likelihood Estimation Meta-analyses: An Empirical Study of Ten-Year Data. *Human Brain Mapping* 44 (5):1876–87.
Young, Liane, and Lily Tsoi. 2013. When Mental States Matter, When They Don't, and What That Means for Morality. *Social and Personality Psychology Compass* 7 (8):585–604.
Zagare, Frank C. 1990. Rationality and Deterrence. *World Politics* 42 (2):238–260.

Zhou, Yuan, Yun Wang, Li-Lin Rao, Liu-Qing Yang, and Shu Li. 2014. Money Talks: Neural Substrate of Modulation of Fairness by Monetary Incentives. *Frontiers in Behavioral Neuroscience* 8. <https://doi.org/10.3389/fnbeh.2014.00150>.

Author

Marika Landau-Wells is an Assistant Professor in the Charles and Louise Travers Department of Political Science at the University of California, Berkeley. She can be reached at mlw@berkeley.edu.

Acknowledgments

I thank Ryan Brutger, Ron Hassner, Tyler Jost, Michaela Mattes, Spring Park, Rebecca Perlman, attendees at UC Berkeley's IR workshop and MIT's SSP seminar, and two anonymous reviewers for comments on this project.

Key Words

Threat perception; conflict; coercion; neuroscience

Date received: February 13, 2024; Date accepted: September 5, 2024