EJIS

RESEARCH ARTICLE

Effort versus accuracy: How well do we understand why others perceive threats?

Marika Landau-Wells (iD)

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

Abstract

Threat perception provokes a range of behaviour, from cooperation to conflict. Correctly interpreting others' behaviour, and responding optimally, is thought to be aided by 'stepping into their shoes' (i.e. mentalising) to understand the threats they have perceived. But IR scholarship on the effects of attempting this exercise has yielded mixed findings. One missing component in this research is a clear understanding of the link between effort and accuracy. I use a US-based survey experiment (study N = 839; pilot N = 297) and a novel analytic approach to study mentalising accuracy in the domain of threat perception. I find that accurately estimating why someone feels threatened by either climate change or illegal immigration is conditional on sharing a belief in the issue's overall dangerousness. Similar beliefs about dangerousness are not proxies for shared political identities, and accuracy for those with dissimilar beliefs does not exceed chance. Focusing first on the emotional states of those who felt threatened did not significantly improve accuracy. These findings suggest that: (1) effort does not guarantee accuracy in estimating the threats others see; (2) emotion understanding may not be a solution to threat mis-estimation; and (3) misperception can arise from basic task difficulty, even without information constraints or deception.

Keywords: climate change; emotion; experiments; immigration; threat perception

Introduction

Threat perception can provoke a wide range of behaviour in individuals.[1] Theories of international relations (IR) have credited threat perception with causing aggressive, defensive, cooperative, and non-cooperative actions by leaders and their states.[2] Given these many possibilities, scholars have

---

[1]By 'threat perception', I mean the apprehension that something or someone could be dangerous and cause harm. I do not mean the detection of a communicative signal intended to coerce. For further discussion on this distinction between the two ways in which 'threat perception' is used in IR, see David A. Baldwin, 'Thinking about threats', *Journal of Conflict Resolution*, 15:1 (1971), pp. 71–8; Marika Landau-Wells, 'Building from the brain: Advancing the study of threat perception in International Relations', *International Organization* (forthcoming).

[2]Zoltán I. Búzás, 'The color of threat: Race, threat perception, and the demise of the Anglo-Japanese alliance (1902–1923)', *Security Studies*, 22:4 (2013), pp. 573–606; Thomas J. Christensen and Jack Snyder, 'Chain gangs and passed bucks: Predicting alliance patterns in multipolarity', *International Organization*, 44:2 (1990), pp. 137–68; Raymond Cohen, 'Threat perception in international crisis', *Political Science Quarterly*, 93:1 (1978), pp. 93–107; Barbara Farnham, 'The theory of democratic peace and threat perception', *International Studies Quarterly*, 47:3 (2003), pp. 395–415; Kai He, 'Polarity and threat perception in foreign policy: A dynamic balancing model', in Nina Græger, Bertel Heurlin, Ole Wæver, and Anders Wivel (eds), *Polarity in International Relations: Past, Present, Future* (Cham: Springer International Publishing, 2022), pp. 45–61; Robert Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976); Jack S. Levy, 'Misperception and the causes of war: Theoretical linkages and analytical problems', *World Politics*, 36:1 (1983), pp. 76–99; John J. Mearsheimer, *The Tragedy of Great Power Politics* (New York: W. W. Norton & Company, 2001); Glenn H. Snyder, 'Alliance

---

noted how difficult it is to correctly interpret others' actions in light of fundamental uncertainty[3] and imperfect information,[4] even when relationships are not inherently adversarial and there is no intention to deceive. Interpreting others' behaviour and responding optimally are thought to be aided by 'stepping into their shoes' and understanding the threats they have perceived.[5]

The terminology used to describe this mental exercise is highly varied in the IR literature and includes 'empathy'[6] and 'perspective-taking',[7] as well as variants, e.g. 'strategic empathy',[8] and metaphors, e.g. 'stepping into someone else's shoes'. To avoid picking through these competing conceptualisations, I follow recent social cognitive science literature in referring to the bundle of mental exercises that constitute thinking about the internal life of others as 'mentalising'.[9] Mentalising subsumes specific targets of inference (e.g. beliefs, traits, intentions) and techniques (e.g. physical perspective-taking, simulation).[10] As Schurz et al. demonstrate, thinking about another person's beliefs, their intentions, their traits, or their emotional states are each distinct mental exercises recruiting different combinations of the brain regions associated with social cognition and other functions (e.g. sensorimotor regions).[11] Since the inferential goal with regard to threat perception is quite specific, I refer to the task of trying to understand threats perceived by others as 'threat estimation'.

Regardless of the terminology used, however, the question investigated in much of the IR literature is one of *effort*: what is the effect of *trying* to estimate the threats that others see? This question arises because there is a strong suspicion that mentalising effort is not always applied: 'Awareness

theory: A neorealist first cut', *Journal of International Affairs*, 44:1 (1990), pp. 103–23; Stephen M. Walt, *Origins of Alliances* (Ithaca, NY: Cornell University Press, 1987); Kenneth N. Waltz, *Theory of International Politics* (New York: McGraw-Hill, 1979).

[3] Jervis, *Perception and Misperception in International Politics*; Mearsheimer, *The Tragedy of Great Power Politics*; Thomas C. Schelling, *Arms and Influence*, 1st ed. (New Haven, CT: Yale University Press, 1966).

[4] James D. Fearon, 'Rationalist explanations for war', *International Organization*, 49:3 (1995), pp. 379–414; Andrew H. Kydd, *Trust and Mistrust in International Relations* (Princeton, NJ: Princeton University Press, 2007).

[5] Joshua Baker, 'The empathic foundations of security dilemma de-escalation', *Political Psychology*, 40:6 (2019), pp. 1251–66; Neta C. Crawford, 'Institutionalizing passion in world politics: Fear and empathy', *International Theory*, 6:3 (2014), pp. 535–57; Naomi Head, 'Transforming conflict: Trust, empathy, and dialogue', *International Journal of Peace Studies*, 17:2 (2012), pp. 33–55; Jervis, *Perception and Misperception in International Politics*; Janice Gross Stein, 'Building politics into psychology: The misperception of threat', *Political Psychology*, 9:2 (1988), pp. 245–71; Claire Yorke, 'Is empathy a strategic imperative? A review essay', *Journal of Strategic Studies*, 46:5 (2023), pp. 1082–102.

[6] Baker, 'The empathic foundations of security dilemma de-escalation'; Robert Jervis, 'Cooperation under the security dilemma', *World Politics*, 30:2 (1978), pp. 167–214; Yorke, 'Is empathy a strategic imperative?'.

[7] Joshua D. Kertzer, Ryan Brutger, and Kai Quek, 'Perspective-taking and security dilemma thinking: Experimental evidence from China and the United States', *World Politics*, 76:2 (2024), pp. 334–78; Don Casler and Dylan Groves, 'Perspective taking through partisan eyes: Cross-national empathy, partisanship, and attitudes toward international cooperation', *The Journal of Politics*, 85:4 (2023), pp. 1471–486.

[8] Zachary Shore, *A Sense of the Enemy: The High Stakes History of Reading Your Rival's Mind* (Oxford: Oxford University Press, 2014).

[9] François Quesque, Ian Apperly, Renée Baillargeon, et al., 'Defining key concepts for mental state attribution', *Communications Psychology*, 2:29 (2024), pp. 1–5; Chris D. Frith and Uta Frith, 'Mechanisms of social cognition', *Annual Review of Psychology*, 63 (2012), pp. 287–313; Matthias Schurz et al., 'Fractionating theory of mind: A meta-analysis of functional brain imaging studies', *Neuroscience & Biobehavioral Reviews*, 42 (2014), pp. 9–34; Matthias Schurz, Joaquim Radua, Matthias G. Tholen, et al., 'Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind', *Psychological Bulletin*, 147:3 (2021), pp. 293–327.

[10] Some IR scholarship uses the word 'empathy' to cover a conceptual space similar to 'mentalising' (A. Burcu Bayram and Marcus Holmes, 'Feeling their pain: Affective empathy and public preferences for foreign development aid', *European Journal of International Relations*, 26:3 (2020), pp. 820–50 [p. 821]; Naomi Head, 'Costly encounters of the empathic kind: A typology', *International Theory*, 8:1 [2016], pp. 171–199 [pp. 174–175]; Yorke, 'Is empathy a strategic imperative?', p. 1082). However, in the study of mentalising, 'empathy' refers to the specific exercise of 'grasping and sharing others' emotional and sensory feelings' (Maria Arioli, Zaira Cattaneo, Emiliano Ricciardi, and Nicola Canessa, 'Overlapping and specific neural correlates for empathizing, affective mentalizing, and cognitive mentalizing: A coordinate-based meta-analytic study', *Human Brain Mapping*, 42:14 [2021], pp. 4777–4804 [p. 4779]) and I preserve that distinction here.

[11] Schurz et al., 'Fractionating theory of mind'.

of how and why an adversary feels threatened … is an important component of empathy but political leaders often display no sensitivity to their adversary's sense of vulnerability while they dwell heavily on their own perception of threat.'[12] A significant body of scholarship in this domain has concluded that the application of mentalising effort, deliberately or spontaneously, can improve decision-making by helping leaders better interpret one another's behaviours.[13]

Arguments in favor of mentalising effort imply that its value resides in providing a more *accurate* understanding of others' beliefs, perceptions, intentions, or emotions than would be possible without such effort.[14] Conversely, when suboptimal decisions are observed, there is a presumption that genuine effort was lacking, perhaps because stereotypes or heuristics were relied upon,[15] or confounding factors, such as deception, disrupted a mental exercise that would otherwise have yielded accurate results.[16] But IR scholars can rarely measure mentalising accuracy itself because, as Stein notes: 'Even with the advantage of hindsight, assessment of accuracy is often not obvious.'[17] Yet this lacuna means there little empirical evidence to support the presumed relationship between mentalising *effort* and mentalising *accuracy*.

Recent experimental studies that have directly manipulated mentalising effort have cast doubt on its straightforward effect on decision-making. Kertzer, Brutger, and Quek show that the consequences of mentalising effort are conditional on prior beliefs and can lead to both escalatory and de-escalatory responses after observing the same action.[18] Casler and Groves also find conditional effects of mentalising effort, but for cooperative choices.[19] In sum, knowing that someone has made the effort to 'step into someone else's shoes' is insufficient for predicting their next move.

One way to understand the varied effects of mentalising effort on decision-making is to more closely interrogate the relationship between *effort* and *accuracy*. If mentalising accuracy is not simply a function of effort, then effort alone does not produce a better understanding of others and, hence, more optimal decision-making.

In this article, I investigate mentalising accuracy in the domain of threat perception. Specifically, I ask: how accurate are people in estimating why other people feel threatened in a non-adversarial setting, and what are the drivers of accuracy? I also investigate a common assertion in the IR literature: is the capacity to understand others' emotions an asset when trying to estimate the threats that they perceive?

Evidence from cognitive science casts doubt on the idea that making the *effort* to understand the world from someone else's perspective generates an *accurate* rendering of that perspective. Humans have a mixed track record in accurately inferring others' thoughts, beliefs, and emotional states, even in controlled, non-adversarial conditions.[20] To study accuracy, cognitive scientists use

---

[12]Stein, 'Building politics into psychology', p. 250.

[13]Baker, 'The empathic foundations of security dilemma de-escalation'; Head, 'Transforming conflict'; Marcus Holmes, *Face-to-Face Diplomacy: Social Neuroscience and International Relations* (Cambridge: Cambridge University Press, 2018); Marcus Holmes and Keren Yarhi-Milo, 'The psychological logic of peace summits: How empathy shapes outcomes of diplomatic negotiations', *International Studies Quarterly*, 61:1 (2017), pp. 107–22; Jervis, 'Cooperation under the security dilemma'; Stein, 'Building politics into psychology'.

[14]Baker, 'The empathic foundations of security dilemma de-escalation'; Holmes, *Face-to-Face Diplomacy*; Ralph K. White, 'Empathy as an intelligence tool', *International Journal of Intelligence and CounterIntelligence*, 1:1 (1986), pp. 57–75.

[15]Jervis, *Perception and Misperception in International Politics*; Shore, *A Sense of the Enemy*.

[16]Charles A. Duelfer and Stephen Benedict Dyson, 'Chronic misperception and international conflict: The U.S.–Iraq experience', *International Security*, 36:1 (2011), pp. 73–100.

[17]Stein, 'Building politics into psychology', p. 247.

[18]Kertzer, Brutger, and Quek, 'Perspective-taking and security dilemma thinking'.

[19]Casler and Groves, 'Perspective taking through partisan eyes'.

[20]Daniel R. Ames and Malia F. Mason, 'Mind perception', in Susan T. Fiske and Neil Macrae (eds), *The SAGE Handbook of Social Cognition* (London: SAGE, 2012), pp. 115–37; Nicholas Epley and Eugene M. Caruso, 'Perspective-taking: Misstepping into others' shoes', in Keith D. Markman, William P. Klein, and Julie A. Suhr (eds), *Handbook of Imagination and Mental Simulation* (New York: Psychology Press, 2009), pp. 297–311; Tal Eyal, Mary Steffel, and Nicholas Epley, 'Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective', *Journal of Personality and Social Psychology*, 114:4 (2018), pp. 547–71; Céline Hinnekens, William Ickes, Liesbet Berlamont, and Lesley Verhofstadt,

relatively simple exercises with known 'ground truths', which differ in many ways from the high-stakes, adversarial, fast-moving events of interest to IR scholars. Nevertheless, how individuals perform in simplified conditions provides a useful baseline for theorising about more complex scenarios. Baselines for cognitive faculties during simplified tasks (e.g. reasoning in the domain of gains versus losses), has already been extended to IR contexts, such as McDermott's work on the role of Prospect Theory in foreign policy decision-making.[21]

To explore mentalising accuracy in the domain of threat perception, I borrow a research design from cognitive science. Within this literature, researchers study mentalising accuracy by deliberately establishing a 'ground truth' set of perceptions in one group of participants against which to compare estimates made by another group of participants.[22] I adapted these methods to create a survey experiment (main study N = 839; pilot N = 297). In the experiment, half of the sample was randomly assigned to provide their own reasons for perceiving either climate change (Issue 1) or illegal immigration (Issue 2) as dangerous and to describe the emotional responses they associated with the issue (the Self-Raters). Threat perception was measured along nine dimensions of potential harm (e.g. physical harm to themselves, financial losses). Emotional responses were measured with respect to ten emotions. All question wording is provided in the Supplementary Material (Table A2).

The other half of the sample engaged in two mentalising exercises. Mentalisers were instructed to think about 'people who are concerned about [climate change/illegal immigration]' while answering the same threat perception and emotional response question batteries. The ordering of the threat-estimating and emotion-understanding exercises was counterbalanced, producing two sub-conditions: Threats-First Mentalisers and Emotions-First Mentalisers. Figure 1 illustrates the assignment to conditions (Panel A) and the study design (Panel B).

In order to preserve the complexity of threat perception (nine dimensions of potential harm) and emotional responses (ten candidate emotions), as well as the heterogeneity of opinions offered by the Self-Raters, I adapted an analytic tool used to represent populations within high-dimensional ecological niches.[23] In this case, I represent the population of Self-Raters who are at least moderately concerned about their Issue, and the niche they occupy is defined by their collective responses to the threat perception (or emotional response) batteries for their Issue. The niche is constructed as a hypervolume, which is an enclosed n-dimensional space that includes all the Self-Ratings (barring extreme outliers) and many plausible nearby ratings (see Figure 2 for two examples). I treat any Mentaliser's guess about the perceptions of the Self-Raters that falls within a hypervolume as accurate because it is within the realm of plausible Self-Rater responses even if it does not correspond to a specific Self-Rater's response. Any guess outside the hypervolume is inaccurate, though some guesses are better (i.e. closer) than others. I benchmark the quality of both binary accuracy and the distance of inaccurate guesses by comparing Mentalisers' responses to the distribution of results generated by 500 samples of random responses to the survey questions. Benchmarking against random responses allows me to characterise the effects of effort by estimating how well one can perform on the task when applying no effort whatsoever.

I used this method of characterising ground truth perceptions and chance accuracy to derive empirical answers to my questions of interest. I found that Mentalisers, on average, were more accurate than chance when estimating the nature of the threats perceived by the Self-Raters, regardless

'Empathic accuracy: Empirical overview and clinical applications', in Michael Gilead and Kevin N. Ochsner (eds), *The Neural Basis of Mentalizing* (Cham: Springer International Publishing, 2021), pp. 149–70.

[21]Rose McDermott, *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy* (Ann Arbor: University of Michigan Press, 2001).

[22]William Ickes, 'Measuring empathic accuracy', in Judith A. Hall and Frank J. Bernieri (eds), *Interpersonal Sensitivity: Theory and Measurement* (Mahwah, NJ: Lawrence Erlbaum Associates, 2001), pp. 219–42; Rebecca Saxe, 'Why and how to study theory of mind with fMRI', *Brain Research*, 1079:1 (2006), pp. 57–65.

[23]Benjamin Blonder, Christine Lamanna, Cyrille Violle, and Brian J. Enquist, 'The n-dimensional hypervolume', *Global Ecology and Biogeography*, 23:5 (2014), pp. 595–609; Benjamin Blonder, Cecina Babich Morrow, Brian Maitner, et al., 'New approaches for delineating n-dimensional hypervolumes', *Methods in Ecology and Evolution*, 9:2 (2018), pp. 305–19.

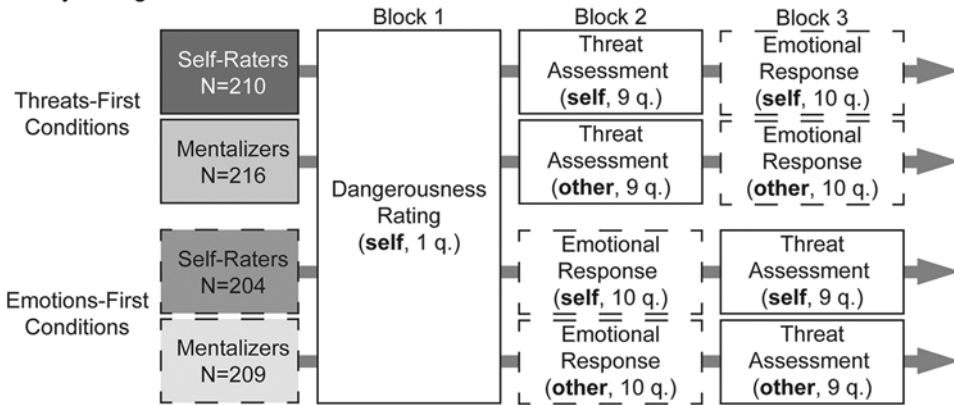## A. Survey Experiment Assignment to Condition
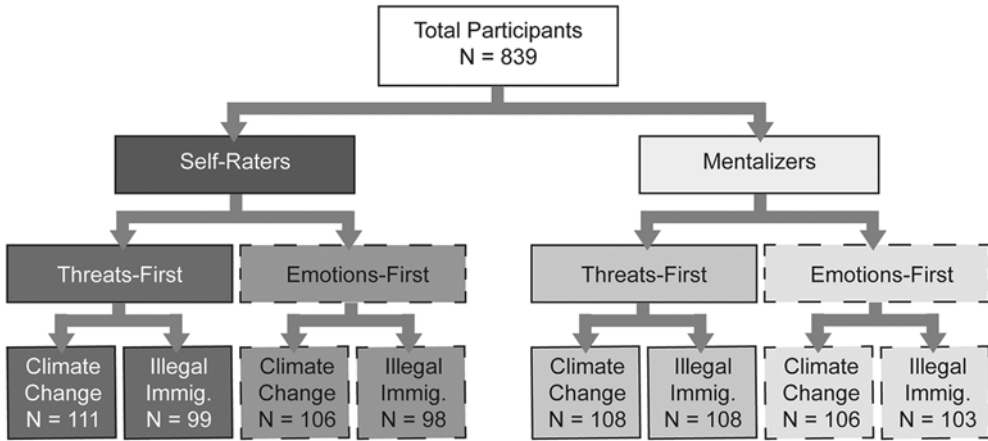
B. Survey Design

Figure 1. Survey experiment structure. (A) Survey experiment assignment to condition, (B) Survey design.

of Issue. However, the effect is driven entirely by Mentalisers who share the Self-Raters' belief in their Issue's dangerousness. That is, Mentalisers who believe climate change or illegal immigration is just as dangerous as the Self-Raters do are able to accurately estimate *why* the Self-Raters feel threatened. Mentalisers who did not associate their Issue with at least a moderate level of dangerousness had lower levels of binary accuracy in threat estimation than would be expected by completing the task with random guesses, but their inaccurate guesses were also better (i.e. closer to the ground truth) than would be expected if they had not exerted any effort. This pattern suggests these Mentalisers were trying to complete the task but did so with an incorrect mental model of those who felt threatened.

Using a series of regressions, I show that the similarity/dissimilarity distinction is not a proxy for either shared political partisanship or shared ideology. Similar beliefs about dangerousness are a stronger correlate of threat estimation accuracy than either variable, regardless of Issue. While cognitive science has identified social distance as a factor in mentalising accuracy,[24] prior focus has been on social groups (e.g. cultural ingroup versus cultural outgroup) and personal relationships

[24]Epley and Caruso, 'Perspective-taking'; Eyal, Steffel, and Epley, 'Perspective mistaking'; Shira Mor, Claudia Toma, Martin Schweinsberg, and Daniel Ames, 'Pathways to intercultural accuracy: Social projection processes and core cultural values', *European Journal of Social Psychology*, 49:1 (2019), pp. 47–62.
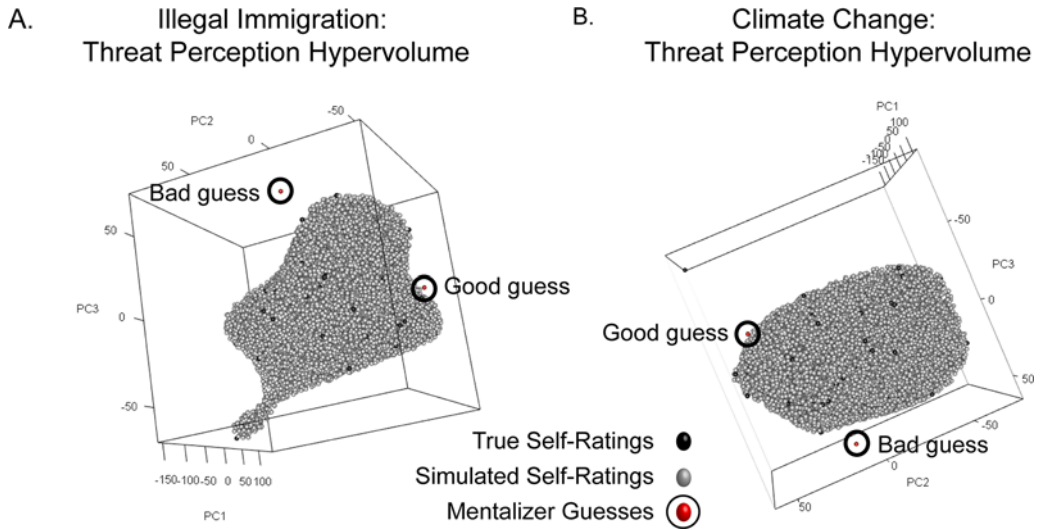
Figure 2. Ground truth threat perceptions. (A) Illegal Immigration: threat perception hypervolume, (B) Climate Change: Threat perception hypervolume.
Note: Axes for both hypervolumes consist of the first three Principal Components (PCs) derived from the nine-dimensional ratings data to avoid collinearity issues. Units are arbitrary.

(e.g. marital partners). This finding highlights the importance of mentalising proximity in a new dimension with relevance in IR: beliefs about danger.

I also test an idea derived from the IR literature that emotion understanding could enhance threat estimation accuracy, either by encouraging a process of simulating others' internal states that then acts as a 'gateway' for other inferences,[25] or by providing important contextual information for others' perceptions of threat.[26] I do not find support for a gateway effect on threat estimation accuracy. Mentalisers in the Emotions-First conditions were no more accurate in their threat perception estimates than those in the Threats-First conditions. I also find no evidence of an incremental context effect. The correlation between emotion understanding accuracy and threat estimation accuracy was only greater than chance for those Mentalisers who already held similar beliefs about their Issue's dangerousness to the Self-Raters.

These findings have several implications. First, an automatic link between mentalising effort and accuracy should not be assumed, at least in the domain of threat perception. Instead effects appear to be conditional on prior beliefs, broadly consistent with Kertzer, Brutger, and Quek and Casler and Groves.[27] Second, mentalising effort on its own is unlikely to aid in the correct interpretation of others' behaviours if differences of opinion about what does and does not constitute a danger are a point of contention (e.g. in the security dilemma). Third, despite the inherent entanglement between emotions and threat perception *for perceivers*, considering (and even accurately understanding) emotional responses does not produce more accurate estimates of threat perception *for mentalisers*. These two mental exercises are at least somewhat distinct. Finally, these findings suggest that a notion of baseline mentalising task difficulty should be integrated into the literature on the sources of misperception, which to date has emphasised the ability of confounders, such

[25] Marcus Holmes, 'The force of face-to-face diplomacy: Mirror neurons and the problem of intentions', *International Organization*, 67:4 (2013), pp. 829–61; Holmes, *Face-to-Face Diplomacy*.

[26] Baker, 'The empathic foundations of security dilemma de-escalation'; Ken Booth and Nicholas J. Wheeler, *The Security Dilemma: Cooperation and Trust in World Politics* (New York: Palgrave Macmillan, 2008); Head, 'Transforming conflict'; Yorke, 'Is empathy a strategic imperative?'.

[27] Kertzer, Brutger, and Quek, 'Perspective-taking and security dilemma thinking'; Casler and Groves, 'Perspective taking through partisan eyes'.

as imperfect information and deception, to undermine mentalising's beneficial effect on decision-making. This study suggests that suboptimal responses to others' behaviours may not be the result of failing to make an effort, but rather of failing to succeed at the task.

In addition to these substantive implications, this paper also makes a methodological contribution by demonstrating how mentalising accuracy can be explored empirically. I show that a combination of experimental design and analytic tools from other fields can capture complex subjective 'ground truths' and create the conditions for estimating accuracy, which could be extended to a variety of other topics. I also show how simulations, in combination with a mathematical representation of complex perceptions, can characterise accuracy relative to chance, which provides a way to validate the application of effort to mentalising tasks while not conflating effort and accuracy. The representation of threat perceptions in a multidimensional space simplifies the open-ended mentalising task one would find in the real world, but it still preserves the potential for a high degree of variability in subjective perception. Preserving a complicated 'ground truth' makes it easier to see the tremendous capacity for, and fundamental challenge of, grasping why others perceive danger.

The paper proceeds in several sections. First, I review the literature related to threat perception and mentalising in IR and cognitive science. Second, I introduce the survey experiment, including a discussion of its design and data collection procedures. Third, I discuss the main analytic methods. Fourth, I present the main study's results. The final section concludes. The Supplementary Material includes detail on the pilot study, the main study's participants and survey instrument, as well as technical aspects of the analyses. All data and code required to replicate the results within the paper are available on the Harvard Dataverse (doi:10.7910/DVN/VLDEQU).

Threat perception and mentalising in International Relations

How people interpret and respond to perceived danger in the world around them is a central question in International Relations.[28] But the actions that leaders and their groups or states take in response to perceived threats are theoretically quite variable and include: preemptive or preventive aggression,[29] alliance offers,[30] and both policy coordination and subversion, in the case of nuclear weapons for example.[31]

The task of interpreting the actions others take is rarely straightforward, even outside of adversarial or time-sensitive contexts, because humans are not mind-readers. There is always uncertainty about why other people do what they do (i.e. 'the problem of other minds'),[32] as well as the fundamental uncertainty and incomplete information that accompany most real-world interactions in the IR domain.[33]

---

[28]Baldwin, 'Thinking about threats'; Cohen, 'Threat perception in international crisis'; Jervis, *Perception and Misperception in International Politics*; Levy, 'Misperception and the causes of war'; Janice Gross Stein, 'Threat perception in international relations', in Leonie Huddy, David O. Sears, and Jack S. Levy (eds), *Oxford Handbook of Political Psychology*, 2nd ed. (New York: Oxford University Press, 2013), pp. 364–94; Walt, *Origins of Alliances*.

[29]Mearsheimer, *The Tragedy of Great Power Politics*; Jonathan Renshon, *Why Leaders Choose War: The Psychology of Prevention* (New York, NY: Praeger, 2006).

[30]Snyder, 'Alliance theory'; Walt, *Origins of Alliances*.

[31]Glenn Chafetz, 'The political psychology of the nuclear nonproliferation regime', *The Journal of Politics*, 57:3 (1995), pp. 743–75; Vipin Narang, *Seeking the Bomb: Strategies of Nuclear Proliferation* (Princeton, NJ: Princeton University Press, 2022).

[32]Anita Avramides, 'Other minds', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Stanford University, 2019), available at:{https://plato.stanford.edu/archives/sum2019/entries/other-minds/}.

[33]Brian C. Rathbun, 'Uncertain about uncertainty: Understanding the multiple meanings of a crucial concept in International Relations theory', *International Studies Quarterly*, 51:3 (2007), pp. 533–57; Janice Gross Stein, 'Radical uncertainty and pragmatism: Threat perception and response', this Special Issue.

Interpreting others' actions

The significance of understanding how people resolve 'the problem of other minds' is well understood in the IR literature. Some theoretical approaches propose that the problem is solved by assumption (e.g. assuming a particular form of rationality guides others' actions).[34] But a significant body of work has demonstrated that there is a great deal of variation in how people, including leaders, approach the task of interpreting others' threat-related behaviour. In cooperative contexts, such as the maintenance of security cooperation agreements, failure to understand threats as they are perceived by one's partners can undermine potential gains to cooperation.[35] In adversarial contexts, such as the dispute over a piece of territory[36] or an arms race,[37] scholars have associated escalation with the same failure.

One straightforward explanation for the failure to understand the threats perceived by others, and thus suboptimal responding, is a lack of genuine effort. That is, people do not bother to 'step into someone else's shoes' and see the world from their perspective before interpreting their actions[38] and so fail to engage meaningfully in mentalising.

In case studies, IR scholars have noted an apparent lack of genuine consideration of the world as seen by others, even in high-stakes situations. Many of the documented cases are adversarial and concern the failure to understand that one's own actions could be perceived as threats,[39] or that one's own actions are less significant than other potential threats.[40] But similar failings have also been documented within established cooperative relationships that face new, unevenly perceived threats (e.g. climate change,[41] refugee flows[42]). In both cases, failures to respond optimally to others' actions have been attributed to an unwillingness to see the world from a different point of view.

Recent work has highlighted the role of effort in mentalising.[43] In some settings, such as face-to-face diplomacy, scholars have argued that it happens relatively easily and spontaneously.[44] In this view, mentalising is aided by simulating others' internal states (e.g. 'feeling what they are feeling') in order to make inferences about their actions and intentions.[45] That is, simulation offers a 'gateway' into making inferences about how another person is thinking.[46] Holmes argues that this type of mentalising can operate even to understand culturally or physically different others.[47]

---

[34] Fearon, 'Rationalist explanations for war'; Bruce Bueno de Mesquita, 'An expected utility theory of international conflict', *American Political Science Review*, 74:4 (1980), pp. 917–31.

[35] Tal Dingott Alkopher and Emmanuelle Blanc, 'Schengen Area shaken: The impact of immigration-related threat perceptions on the European security community', *Journal of International Relations and Development*, 20:3 (2017), pp. 511–42; Jörg Monar, 'Common threat and common response? The European Union's counter-terrorism strategy and its problems', *Government and Opposition*, 42:3 (2007), pp. 292–313; Gary Winslett, 'Differential threat perceptions: How transnational groups influence bilateral security relations 1', *Foreign Policy Analysis*, 12:4 (2016), pp. 653–73.

[36] Levy, 'Misperception and the causes of war'.

[37] Jervis, *Perception and Misperception in International Politics*.

[38] Jervis, 'Cooperation under the security dilemma'; Stein, 'Building politics into psychology'.

[39] Booth and Wheeler, *The Security Dilemma*; Jervis, 'Cooperation under the security dilemma', p. 181; Richard Ned Lebow and Janice Gross Stein, *We All Lost the Cold War* (Princeton, NJ: Princeton University Press, 1995), p. 67; Levy, 'Misperception and the causes of war', p. 90; White, 'Empathy as an intelligence tool'.

[40] Duelfer and Dyson, 'Chronic misperception and international conflict'.

[41] Hannes Sonnsjö and Niklas Bemberg, 'Climate change in an EU security context: The role of the European External Action Service', Research Report (Stockholm International Peace Research Institute & The Swedish Institute of International Affairs, 2016).

[42] Alkopher and Blanc, 'Schengen Area shaken'.

[43] Booth and Wheeler, *The Security Dilemma*; Head, 'Costly encounters of the empathic kind'; Yorke, 'Is empathy a strategic imperative?'.

[44] Holmes, 'The force of face-to-face diplomacy'; Holmes, *Face-to-Face Diplomacy*.

[45] Holmes, 'The force of face-to-face diplomacy', p. 830.

[46] The brain-level foundations of this theory, which posits a critical role for mirror neurons, is contested (Cecilia Heyes and Caroline Catmur, 'What happened to mirror neurons?', *Perspectives on Psychological Science*, 17:1 [2022], pp. 153–68; Rebecca Saxe, 'Against simulation: The argument from error', *Trends in Cognitive Sciences*, 9:4 [2005], pp. 174–79).

[47] Holmes, 'The force of face-to-face diplomacy', pp. 846–7.

But the effortful version of mentalising is also seen as beneficial for decision-making.[48] Some argue that the success of this type of mentalising hinges only on sufficient information.[49] But another viewpoint argues that understanding emotions and the context that gives rise to them is critical for mentalising success, particularly if one is trying to infer the 'why' of a particular action,[50] and that this type of understanding is a skill that should be cultivated.[51]

Yet, in research that experimentally manipulates mentalising effort, IR scholars have shown that effort does not have consistent effects on the interpretation of others' behaviour and on response decisions. Instead, interpretation and response decisions are conditional on prior beliefs, even when mentalising is attempted. Kertzer, Brutger, and Quek show that mentalising effort can provoke both escalatory and de-escalatory behaviors in US and Chinese participants, depending on prior beliefs about the adversary.[52] Casler and Groves show that mentalising effort can spur cooperative behaviour, but that this effect is limited to those with specific partisan leanings in the US context.[53] This research has also shown that the dispositional tendency to mentalise[54] generally mimics the experimentally induced effects of effort. Thus, mentalising effort, applied after encouragement or due to a dispositional tendency, interacts with prior beliefs in a way that can produce decisions which do not always appear optimal. The relationship between mentalising effort and decision-making is therefore not straightforward.

Effort and accuracy

These mixed findings on the effects of mentalising effort on response decisions suggest that the logic behind arguments in favour of 'stepping into someone else's shoes' to improve decision-making outcomes is incomplete. One missing piece is an understanding of the relationship between mentalising *effort* (i.e. trying) and mentalising *accuracy* (i.e. succeeding). The link between effort and accuracy is rarely tested in IR because, as Stein notes: 'Even with the advantage of hindsight, assessment of accuracy is often not obvious … the dangers inherent in a situation are rarely unambiguous.'[55] Despite the challenge in assessing accuracy, its link to effort is not trivial. If this link does not hold, then one explanation for the varied results of mentalising effort documented above is that some people *try* but do not conjure up an accurate representation of the other person. Their response decisions might be optimal *for the person they imagined*, but their imagination failed. Vitally, this suboptimal behaviour does not arise from misperceptions induced by extenuating circumstances or deception, but rather from misperceptions attributable to the difficulty of the task itself.

Evidence from cognitive science provides reasons to be sceptical that mentalising effort will yield accurate models of other minds.[56] People perceive mentalising as difficult and effortful, rather than

---

[48]Baker, 'The empathic foundations of security dilemma de-escalation'; Crawford, 'Institutionalizing passion in world politics'; Holmes and Yarhi-Milo, 'The psychological logic of peace summits'; H. R. McMaster, *Battlegrounds: The Fight to Defend the Free World* (New York, NY: HarperCollins, 2021); Robert McNamara, *In Retrospect: The Tragedy and Lessons of Vietnam* (New York, NY: Knopf Doubleday Publishing Group, 2017); Shore, *A Sense of the Enemy*; White, 'Empathy as an intelligence tool'; Yorke, 'Is empathy a strategic imperative?'.

[49]Shore, *A Sense of the Enemy*; White, 'Empathy as an intelligence tool'.

[50]Baker, 'The empathic foundations of security dilemma de-escalation'; Booth and Wheeler, *The Security Dilemma*; Head, 'Transforming conflict'; Yorke, 'Is empathy a strategic imperative?'.

[51]Crawford, 'Institutionalizing passion in world politics', p. 544.

[52]Kertzer, Brutger, and Quek, 'Perspective-taking and security dilemma thinking'.

[53]Casler and Groves, 'Perspective taking through partisan eyes'.

[54]Mark H. Davis, 'Measuring individual differences in empathy: Evidence for a multidimensional approach', *Journal of Personality and Social Psychology*, 44:1 (1983), pp. 113–26.

[55]Stein, 'Building politics into psychology', p. 247.

[56]Ames and Mason, 'Mind perception'; Epley and Caruso, 'Perspective-taking'; Eyal, Steffel, and Epley, 'Perspective mistaking'.

easy.[57] Our success as a species suggests we cannot be completely incapable of making inferences about others; on the other hand, systematic errors have been demonstrated in tests of mentalising accuracy,[58] which suggests that we may be wrong quite often, but in ways that are not necessarily detrimental.[59] One type of systematic error arises because people simply struggle to imagine those who are quite different from themselves, e.g. demonstrating greater accuracy for one's own cultural ingroup than an outgroup.[60]

The mentalising exercises explored by cognitive scientists are relatively simple (e.g. inferring someone's emotions from their facial expressions) when set against the real-world cases explored by IR scholars. Nevertheless, simple baseline exercises of general cognitive faculties have provided insight into a range of scenarios in IR and foreign policy decision-making.[61] There is thus some value in establishing a baseline for how well people can perform the task of estimating the threats that others perceive under simplified conditions. A better understanding of this baseline can then inform theoretical expectations for more complex situations (e.g. adding an adversarial dimension, adding stress and time sensitivity to model crises). In the next section, I lay out a method for defining this baseline and for testing whether it can be improved upon with a technnique proposed in the IR literature: emotion understanding.

Survey experiment

In order to study mentalising accuracy, it is essential to establish 'ground truth' mental states. In the case of threat perception, one of the primary issues raised in the literature is the difficulty in understanding *why* one person would see a particular scenario, state, or phenomenon as dangerous when another person might not (i.e. what kind of harm is the source of concern?).[62] This mentalising challenge arises due to the fundamentally subjective nature of threat perception.[63]

To better understand baseline accuracy in estimating the why other people feel threatened, I borrow a paradigm used in cognitive science that measures the ability to 'accurately infer the specific content of another person's covert thoughts and feelings'.[64] In a common version of the paradigm, participants perform an exercise while being recorded and then watch their own recording to describe their thoughts and feelings at particular time-points.[65] A second set of participants

[57]C. Daryl Cameron, Cendri A. Hutcherson, Amanda M. Ferguson, et al., 'Empathy is hard work: People choose to avoid empathy because of its cognitive costs', *Journal of Experimental Psychology: General*, 148:6 (2019), pp. 962–76.

[58]Epley and Caruso, 'Perspective-taking'; Matthew R. Jordan, Theresa Gebert, and Christine E. Looser, 'Perspective taking failures in the valuation of mind and body', *Journal of Experimental Psychology: General*, 148:3 (2019), pp. 407–20; Saxe, 'Against simulation'; Andrew R. Todd and Diana I. Tamir, 'Factors that amplify and attenuate egocentric mentalizing', *Nature Reviews Psychology*, 3:3 (2024), pp. 164–80.

[59]William Ickes and Jeffry A. Simpson, 'Motivational aspects of empathic accuracy', in Garth J. O. Fletcher, Margaret S. Clark (eds), *Blackwell Handbook of Social Psychology: Interpersonal Processes* (Maldon, MA: Blackwell Publishers Ltd, 2003), pp. 229–49; Dominic D. P. Johnson, *Strategic Instincts: The Adaptive Advantages of Cognitive Biases in International Politics* (Princeton, NJ: Princeton University Press, 2020).

[60]Mor et al., 'Pathways to intercultural accuracy'.

[61]Daniel Kahneman and Jonathan Renshon, 'Hawkish biases', in A. Trevor Thrall and Jane K. Cramer (eds), *American Foreign Policy and the Politics of Fear: Threat Inflation since 9/11* (New York: Routledge, 2009), pp. 79–96; Jervis, *Perception and Misperception in International Politics*; Johnson, *Strategic Instincts*; McDermott, *Risk-Taking in International Politics*; Thomas C. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960).

[62]Baker, 'The empathic foundations of security dilemma de-escalation', p. 1257; Stein, 'Building politics into psychology', p. 247.

[63]Landau-Wells, 'Building from the brain'; Eitan Oren, 'How leaders perceive security dangers: The neglected dimension of unfolding experience', this Special Issue; Natalia Chaban and Ole Elgström, 'Expectations–(threat) perceptions gap', this Special Issue.

[64]Ickes, 'Measuring empathic accuracy', p. 220.

[65]Ickes, 'Measuring empathic accuracy'; Hinnekens et al., 'Empathic accuracy'; Nuria K. Mackes, Dennis Golm, Owen G. Daly, et al., 'Tracking emotions in the brain: Revisiting the empathic accuracy task', *NeuroImage*, 178 (2018), pp. 677–86; Jamil Zaki, Jochen Weber, Niall Bolger, and Kevin Ochsner, 'The neural bases of empathic accuracy', *Proceedings of the National Academy of Sciences*, 106:27 (2009), pp. 11382–7.

then watch the same video and are asked to describe the thoughts and feelings of the person in the video at those same time-points. The exercises in question are naturalistic and unrehearsed, but otherwise highly variable (e.g. therapy sessions, the discussion of autobiographical events). Because the data provided in the traditional version of this task are unstructured comments and thus arbitrarily complex, trained raters are needed to consistently assess the similarity between the ground truth self-descriptions and the assessments provided by mentalising participants. These hand-coded estimates of agreement then serve as the accuracy measure. However, in cases where experimenters radically simplify the task down to a single dimension (i.e. asking for a positive or negative affect score), the correlation between self-ratings and the mentalisers' guesses can be used directly as an accuracy measure.[66]

To balance between the open-ended design of the original accuracy task as described in Ickes[67] and the one-dimensional valence task described by Zaki et al.,[68] I use question batteries, which are close-ended but multidimensional. This method has the advantage of capturing some of the complexity inherent in subjective perceptions. It also allows for substantial individual-level variation in those subjective perceptions. I detail the analytic approach required to take advantage of the richness of these subjective assessments in the Analysis section.

Design

The conventional accuracy task uses pre-recorded, annotated videos as representations of the ground truth, often relying on the same stimuli across multiple studies.[69] But a temporal separation between collection of ground truth perceptions of threat and mentalising efforts is problematic. Events can drastically alter the estimation of particular threats, as shown by rapid shifts in the extent to which Russia is viewed as a threat by Europeans and Americans.[70] To avoid any risk that external events interfere with the ability to accurately mentalise, I designed a survey experiment to simultaneously collect ground truth perceptions and mentalising estimates.

I adapted the traditional accuracy task to the study of threat perception in two ways. First, while conventional accuracy tasks use self-reported reflections about *personal events* (e.g. a therapy session), I use self-reported reflections about two familiar international political phenomena: illegal immigration and climate change. In the American context, individuals vary in the extent to which they believe these phenomena are dangerous and in their reasons why.[71] These two issues have traditionally generated mirror-image patterns of concern across the partisan divide in the United States. Republicans are more likely to be concerned about illegal immigration,[72] and Democrats are more likely to be concerned about climate change.[73] Including both issues makes it possible to separate partisan identification (e.g. identifying as a Democrat) from other factors that might affect mentalising accuracy. By using issues which evoke different levels of concern across the population, I also avoid conflating the propensity for threat perception with conservatism.[74]

---

[66]Zaki et al., 'The neural bases of empathic accuracy'.

[67]Ickes, 'Measuring empathic accuracy'.

[68]Zaki et al., 'The neural bases of empathic accuracy'.

[69]Ickes, 'Measuring empathic accuracy'.

[70]'Large shares see Russia and Putin in negative light, while views of Zelenskyy more mixed', Pew Research Center (July 2023).

[71]Marika Landau-Wells, 'Dealing with danger: Threat perception and policy preferences', PhD diss., MIT (2018); Matthew T. Ballew, Anthony Leiserowitz, Connie Roser-Renouf, et al., 'Climate change in the American mind: Data, tools, and trends', *Environment: Science and Policy for Sustainable Development*, 61:3 (2019), pp. 4–18; Nicholas A. Valentino, Stuart N. Soroka, Shanto Iyengar, et al., 'Economic and cultural drivers of immigrant support worldwide', *British Journal of Political Science*, 49:4 (2019), pp. 1201–26.

[72]J. Baxter Oliphant and Andy Cerda, 'Republicans and Democrats have different top priorities for U.S. immigration policy', Pew Research Center (September 2022).

[73]Alec Tyson, Cary Funk, and Brian Kennedy, 'What the data says about Americans' views of climate change', Pew Research (August 9, 2023), available at: {https://www.pewresearch.org/short-reads/2023/08/09/what-the-data-says-about-americans-views-of-climate-change/}.

[74]Marika Landau-Wells and Rebecca Saxe, 'Political preferences and threat perception: Opportunities for neuroimaging and developmental research', *Current Opinion in Behavioral Sciences*, 34 (2020), pp. 58–63; Fade R. Eadeh and Katharine K. Chang,

The survey experiment used a fully factorial between-subjects design with eight conditions (two issues × two perspectives × two question orders) to which participants were randomly assigned. Figure 1A shows the assignment to conditions. In all conditions, participants answered three question blocks, visualised in Figure 1B. Participants in all conditions received the same initial question in Block 1, which asked them to rate the dangerousness of their Issue (Climate Change or Illegal Immigration) *from their own perspective* on a 0–100 scale (where 0 = 'Not at all dangerous' and 100 = 'Extremely dangerous'). This provided a common scale for the belief in dangerousness across issues and allowed me to identify the subset of participants who felt threatened by their Issue across conditions.

Participants in the four Self-Rater conditions were directed to answer all subsequent questions (Blocks 2 and 3) with reference to themselves (i.e. 'Please use the scales to indicate how relevant these specific concerns are **for you** when you think about [climate change/illegal immigration]'). Participants in the four Mentalising conditions were directed to answer all subsequent questions thinking about the views of others worried about their Issue ('Please use the scales to indicate how relevant you believe these specific concerns are **for other people who are worried** about [climate change/illegal immigration]').

For those in the Threats-First conditions, Block 2 consisted of a question battery about the relevance of nine 'specific concerns' associated with their Issue.[75] All nine concerns are listed in the Supplementary Material (Table A2). The list of concerns drew on prior research and captured physical threats (e.g. bodily harm), non-material threats (e.g. compromised spiritual purity), personal harms (e.g. loss of an economic asset), and collective harms (e.g. loss of group status). The rating scale for the relevance of each concern ranged from 0 ('Not at all relevant') to 100 ('Extremely relevant'). Block 3 consisted of a question battery asking how intensely ten emotions were evoked by each Issue. All ten emotions are listed in the Supplementary Material (Table A2). The list of emotions also drew on prior research and included both basic emotions (e.g. fear) and complex emotions (e.g. contempt). The rating scale for the each emotion ranged from 0 ('Do not feel at all') to 100 ('Feel strongly'). In the Emotions-First conditions, Block 2 consisted of the emotion battery, and Block 3 consisted of the concern/threat battery. All other question wording was the same across conditions.

Participant characteristics

839 research subjects participated in this experiment (53 per cent female; mean age 41 years old).[76] All subjects were recruited through Survey Sampling International (SSI, now Dynata), and the study was administered on the Qualtrics platform. The study was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology. Many aspects of this study, including the question design, parameters of the accuracy analysis, and sample size requirements were established based on a pilot study conducted on Amazon's Mechanical Turk platform (N = 297). Details of the pilot study are included in the Supplementary Material.

While the sample was not weighted to be nationally representative, it closely tracks a contemporaneous American National Election Survey (ANES) report of the electorate's composition at the national level on metrics of gender composition, political party identification, and race/ethnicity

---

'Can threat increase support for liberalism? New insights into the relationship between threat and political attitudes', *Social Psychological and Personality Science*, 11:1 (2020) pp. 88–96.

[75] I used 'concerns' in the question wording to avoid use of the word 'threat'.

[76] 1,060 participants completed some portion of the survey in which this study was included. Responses were excluded from the final analysis for three reasons. First, participants who did not complete the entire survey were dropped, regardless of whether they completed the primary questions of interest. Second, responses were dropped if they failed both of the attention checks embedded in the survey and completed the survey in the fastest quintile of all responses, suggesting they were 'speeders'. Third, responses were excluded if their answers to either the nine threat perception questions or the ten emotion questions had a variance of zero.

self-identification.[77] The share of women in the sample was slightly higher than in the electorate (53.5% versus 52%). The sample self-identified as slightly less White than the electorate (66% versus 69%).[78] There were also more political partisans in the sample than in the electorate (Democrats: 39% versus 35%; Republicans: 30% versus 28%).[79] See Table A1 in the Supplementary Material for additional demographic details of the full sample and the balance across experimental conditions.

Analysis

Measuring ground truths

To establish the ground truth of both the threat perceptions and the emotional responses for each Issue, I use only the Self-Ratings provided by subjects who found their Issue at least moderately dangerous (i.e. provided a Dangerousness rating in Block 1 of 50 or greater).[80] This restricts ground truth perceptions to 'those who are worried' about their Issue, which corresponds to the people Mentalisers were instructed to consider.

As noted above, both climate change and illegal immigration are issues where people hold a variety of beliefs about why they feel threatened. Any measure of ground truth perceptions of threat needs to account for these subjective differences. These differences also have implications for any judgement about mentalising accuracy. An accurate guess is one which could reasonably fall within this heterogeneous collection of perceptions, without necessarily corresponding to a specific response provided by one of the Self-Raters. Therefore, I treat the 'ground truth' not as a single point generated by a Self-Rater or as a single summary statistic of the Self-Rater group as a whole, but rather as the high-dimensional space occupied by the sample of Self-Rater responses.

To carry this out analytically, I borrow the concept of hypervolumes as used in the population ecology literature.[81] In that context, hypervolumes are a method for representing the ecological niche occupied by a population in a multidimensional space (e.g., terrain type, food sources). In this case, the populations of interest are the Self-Raters and the dimensions of interest are either (1) the nine 'concerns' for which each Self-Rater provided a relevance rating (i.e. threat perception ground truths) or (2) the ten emotions for which they provided an intensity rating (i.e. emotional response ground truths). The advantage of using hypervolumes instead of summary statistics as a way of characterising the ground truth views of a heterogeneous group is that the approach allows researchers to preserve the multidimensional nature of the underlying construct (i.e. threat perception) while making relatively few assumptions about the data's distributions. Pilot Study data showed that Self-Ratings were not univariate normal, multivariate normal, or uniformly related in any two dimensions and could potentially have discontinuities (e.g. clusters, holes).[82] Therefore, a hypervolume approach was appropriate.

The size and shape of hypervolumes are determined by several parameters. I used the Pilot Study data to set those parameters and then applied them unchanged to the main study. Based on the Pilot Study data, I constructed the hypervolumes with a one-class support vector machine

---

[77] American National Election Studies, University of Michigan, and Stanford University, 'The ANES Guide to Public Opinion and Electoral Behavior' (Ann Arbor, M.I.: Inter-University Consortium for Political and Social Research, 9 September 2017).

[78] Within the sample, 12 per cent self-identified as Black or African-American, and 10 per cent as Latino or Hispanic. For the 2016 ANES, 11 per cent self-identified as Black (non-Hispanic) and 12 per cent as Hispanic.

[79] The detailed breakdown of the sample is: Strong Democrat = 24%, Weak Democrat = 15%; Independent, lean Democrat = 10%; Independent = 15%; Independent, lean Republican = 6%; Weak Republican = 12%; Strong Republican = 18%. For the 2016 ANES: Strong Democrat = 21%; Weak Democrat = 14%; Independent, lean Democrat = 11%; Independent = 15%; Independent, lean Republican = 11%; Weak Republican = 12%; Strong Republican = 16%.

[80] I will continue to refer to the Self-Raters as a group, but properly they are a subset of the subjects in their experimental condition.

[81] Blonder et al., 'The n-dimensional hypervolume'; Blonder et al., 'New approaches for delineating n-dimensional hypervolumes'.

[82] Alternative symmetric methods for enclosing a set of points (e.g., spheres or ellipsoids) rely on these assumptions. Convex hulls, another possibility, assume all points should be continuously enclosed, which eliminates the possibility of clusters of response 'types'.

(SVM) instead of the Gaussian kernel method, which created volumes that substantially exceeded the scale boundaries even with small bandwidths and was sensitive to outliers. The SVM method is recommended by Blonder et al. for generating a volume that fits smoothly around the data without being overly sensitive to outliers. All SVM tuning parameters were kept at defaults.

Pilot data also revealed correlations within the nine-dimensional threat perception ratings and the ten-dimensional emotional response ratings. In such cases, Blonder et al. recommend using Principal Component Analysis (PCA) to define independent axes for the hypervolume. For each example in the Pilot Study data, the first three Principal Components (PCs) accounted for at least 75 per cent of the variance, so I chose three PCs as the compromise between complexity and analytic tractability. Data was centred before computing the PCs, but not scaled as all scales were identical to begin with.[83]

Based on the methods established with Pilot Study data, I constructed four hypervolumes from the Main Study data, one for each set of ground truth ratings provided by the Self-Raters across four conditions: Illegal Immigration Threat Perception (Threats-First Condition); Climate Change Threat Perception (Threats-First Condition); Illegal Immigration Emotional Response (Emotions-First Condition); and Climate Change Emotional Response (Emotions-First Condition). As with the Pilot Study data, the three PCs used to construct the hypervolumes always accounted for at least 75 per cent of the total variance. Tables A3 and A4 in the Supplementary Material provide the loadings for all items on the first three PCs. To summarise, while the first PC appears to capture mean differences, the second and third capture clusters of threats (or emotions) that hang together for participants. In the case of climate change, PC2 captured concerns about both individual and group status and moral purity, while PC3 captured environmental concerns as well. For illegal immigration, PC2 captured concerns about physical harm (to oneself and loved ones) as well as the loss of personal rights, while PC3 primarily captured economic loss concerns. Negative basic emotions (fear, anger, sadness) dominated PC2 for climate change, while negative complex emotions (resentment, contempt) dominated PC3. Negative basic emotions (anger, disgust) also dominated PC2 for illegal immigration, while sadness dominated PC3.

The ground truth threat perception hypervolumes for the Threats-First Illegal Immigration and Climate Change conditions are shown in Figure 2. Each axis in the figure captures one of the three PCs scaled in arbitrary units. Black points within the volume correspond to the true Self-Rater responses. Grey points represent 'nearby' simulated responses.[84] Conceptually, the collection of black and grey points is the set of responses that *could have been given* by the Self-Rater group and thus constitutes the ground truth. Any threat perception estimate provided by a Mentaliser that falls within this space is accurate in that it could have been given by a Self-Rater. And, as the circled points in Figure 2 indicate, estimates falling outside this volume can be either near (good guesses) or far (bad guesses) from the edge of the ground truth hypervolume.

Mentalising accuracy

In order to determine whether each Mentaliser's guess fell inside or outside their respective hypervolume, I first transformed each guess into the relevant rating space.[85] This transformation provided *binary accuracy*: a point inside the hypervolume was an accurate guess, and a point outside was inaccurate. I also measured the Euclidean distance of each inaccurate guess to the nearest edge of the hypervolume. This provided a measure of *miss distance*, which captures the quality of the Mentaliser's guess, even if it is inaccurate. Section A5 in the Supplementary Material contains additional detail on these procedures.

---

[83]Rescaling has no substantive effect on the first PC or the total variance explained, but the unscaled items yield results that are more interpretable.

[84]The number of simulated points varies by hypervolume and ranges from approximately 14,000 to 21,000.

[85]This required centring the data, rotating it using the eigenvector matrix of the PCA model for the corresponding ground truth Self-Ratings, and using the first three principal components as coordinates in the Self-Raters' space.
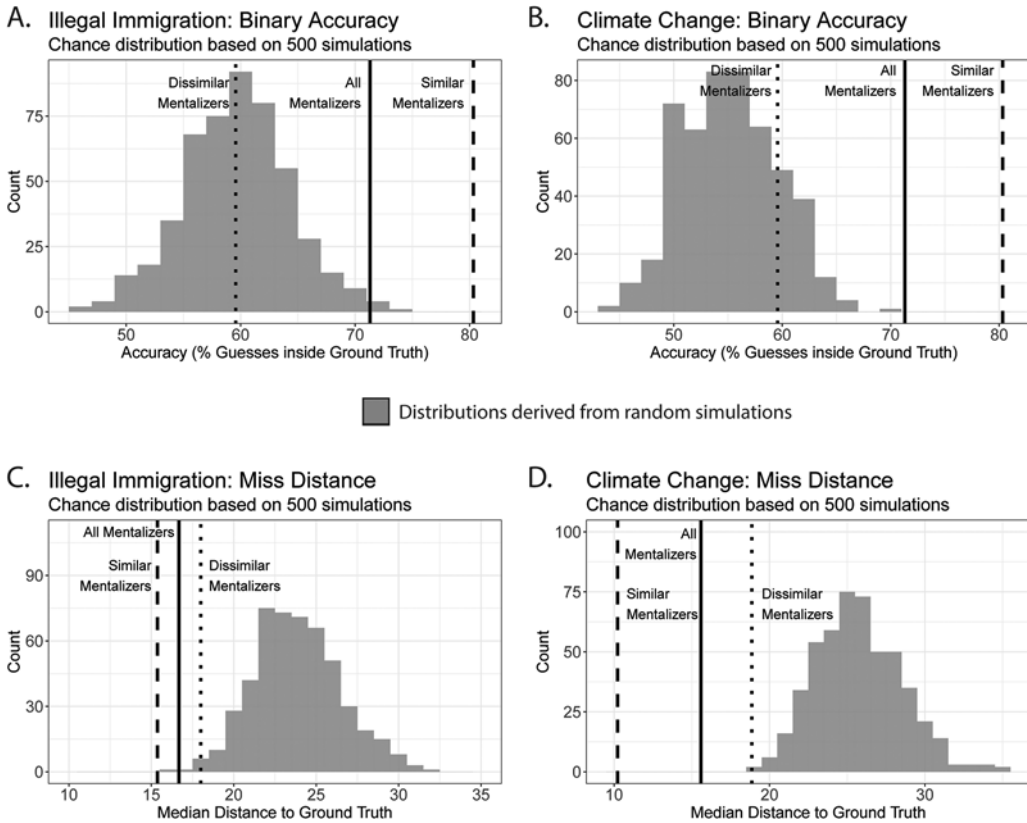
Figure 3. Threat estimation accuracy. (A) Illegal Immigration: binary accuracy. (B) Climate Change: binary accuracy. (C) Illegal Immigration: Miss distance. (D) Climate Change: Miss distance.
Note: Solid black lines represent results for all Mentalisers in the Threats-First Conditions. Dashed lines represent the results for the subset of Similar Mentalisers. Dotted lines represent the results for the subset of Dissimilar Mentalisers.

Chance accuracy as a reference point

I contextualise the *binary accuracy* and the *miss distances* generated in the experiment's Mentalising conditions by comparing them to values achieved by chance. To do this, I ask: how accurate would participants in the Mentalising conditions have been if they had completed the nine-dimensional threat estimation measure and the ten-dimensional emotion understanding measure by randomly selecting values for each item? Conceptually, this process captures the results of completing the task without applying any effort.

I first simulated 500 datasets of randomly completed responses with the same number of 'observations' as the true data. For each randomly constructed dataset, I transformed the each observation into the respective Self-Rating space using the true PCA eigenvectors. I then measured the *binary accuracy* of each observation and, if applicable, its *miss distance*. I then generated the distributions for chance binary accuracy (see Panels A and B of Figure 3 for examples) and for the median miss distance (see Panels C and D of Figure 3 for examples). These distributions formed the basis for judging how good subjects were at each Mentalising exercise and what a total lack of effort looked like.

Results

I focus first on threat estimation accuracy among those who completed the threat estimation task first (Threats-First conditions). As shown in Panels A and B in Figure 3, the guesses given by all

Mentalisers (solid lines) were more accurate than random guessing would generally have achieved for both Issues.[86] The difference in overall accuracy between Issues was not statistically significant, suggesting the tasks had similar levels of difficulty.[87]

However, estimation accuracy was substantially better for a particular subgroup: those Mentalisers who shared the Self-Raters' belief that their Issue represented a danger (i.e. Similar Mentalisers). The Similar Mentalisers were participants who had also rated their Issue as at least moderately dangerous on the survey's first question. For Similar Mentalisers (dashed lines), threat estimation binary accuracy[88] was significantly better than the Dissimilar Mentalisers,[89] i.e. those who had provided a Dangerousness score of less than 50 (dotted lines).[90] Since the Similar Mentalisers *would have been the Self-Raters* but for random assignment, their high level of estimation accuracy is unsurprising. It is notable, however, that the Dissimilar Mentalisers were not only less accurate than Similar Mentalisers, but also were also less accurate than could have been achieved by random guesses in many cases.[91]

However, as Panels C and D in Figure 3 show, there is no sign that participants were actually providing responses at random. On the contrary, while Dissimilar Mentalisers' guesses did not land inside the realm of plausible responses at a rate better than chance, their incorrect guesses were significantly *closer* (i.e. better) than would be achieved by truly random answers, within the conventional standard false positive allowance of 5 per cent.[92] This provides evidence that Dissimilar Mentalisers applied some *effort* to the mentalising task but did not succeed at it. That is, their mental models of those concerned by the Issue were wrong.

Does the significance of the Similar/Dissimilar distinction simply reflect the effect of shared partisanship or shared ideological orientation? Democrats (and liberals) might be more likely to self-report thinking of Climate Change as dangerous; Republicans (and conservatives) might be more likely to report thinking of Illegal Immigration as dangerous. Thus, a shared belief in dangerousness may be picking up on broader political alignment.

In a series of logistic regressions using binary accuracy as the dependent variable, I show that Similarity is not a proxy for either shared partisanship or shared ideological orientation. Tables 1 and 2 show that Similarity maintains its explanatory significance (Models 7 and 8) even after accounting for the partisan or ideological alignment that would lead to shared views on each Issue (Republican/conservative for Illegal Immigration; Democrat/liberal for Climate Change). Substantively, holding a similar belief about dangerousness increases the odds of threat estimation accuracy by two to three times, depending on the model.[93]

Model 1 in each table demonstrates the main effect of Similarity, accounting only for the experimental condition differences between Mentalisers (Threats-First versus Emotions-First). Model 2 adds three demographic covariates as controls (self-identifying as female, age, and self-identifying as White). These controls have no substantive effect on the importance of belief similarity. Models

---

[86]For the Illegal Immigration condition, 5 out of 500 random simulations have accuracy rates greater than the observed value for all Mentalisers, equivalent to $p = 0.01$. For the Climate Change condition, 0 out of 500 simulations have an accuracy rate greater than the observed value, equivalent to $p < 0.002$.

[87]Chi-squared test of differences in binary accuracy for all Mentalisers in the Illegal Immigration and Climate Change Threats-First conditions with p-values simulated based on 2,000 replicates: $\chi^2 = 2.03$, $p = 0.20$.

[88]80 per cent and 86 per cent for Illegal Immigration and Climate Change, respectively.

[89]60 per cent and 59 per cent for Illegal Immigration and Climate Change, respectively.

[90]Chi-squared test of differences in binary accuracy for Similar versus Dissimilar Mentalisers in the Illegal Immigration Threats-First condition with p-values simulated based on 2,000 replicates: $\chi^2 = 5.59$, $p = 0.02$. Similar versus Dissimilar Mentalisers in the Climate Change Threats-First condition: $\chi^2 = 9.21$, $p = 0.004$.

[91]For the Illegal Immigration condition, 237 out of 500 random simulations have accuracy rates greater than the observed value for Dissimilar Mentalisers, equivalent to $p = 0.47$. For the Climate Change condition, 105 out of 500 simulations have an accuracy rate greater than the observed value, equivalent to $p = 0.21$.

[92]For Dissimilar Mentalisers in the Illegal Immigration condition, 6 out of 500 random simulations have a median miss distance smaller than the observed value, equivalent to $p = 0.01$. For Dissimilar Mentalisers in the Climate Change condition, 1 out of 500 simulations had an median miss distance smaller than the observed value, equivalent to $p = 0.002$.

[93]I derive the range of odds by exponentiating the coefficients for Similarity across the full set of models.

Table 1. Correlates of binary threat estimation accuracy

	Issue: Illegal Immigration							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Similarity	0.73*	0.75*					0.73*	0.75*
	(0.32)	(0.33)					(0.34)	(0.34)
Republican			0.21	0.27			0.04	0.13
			(0.33)	(0.35)			(0.40)	(0.42)
Conservative					0.19	0.19	−0.06	−0.10
					(0.33)	(0.33)	(0.40)	(0.41)
Emotions-First	0.32	0.33	0.33	0.34	0.33	0.33	0.33	0.33
	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)	(0.32)
Female		−0.15		−0.16		−0.15		−0.16
		(0.33)		(0.32)		(0.33)		(0.33)
Age		0.0001		0.003		0.002		0.001
		(0.01)		(0.01)		(0.01)		(0.01)
White		−0.23		−0.21		−0.16		−0.25
		(0.37)		(0.37)		(0.36)		(0.38)
Constant	0.53*	0.75	0.84***	0.93	0.84***	0.94	0.53	0.75
	(0.27)	(0.57)	(0.24)	(0.57)	(0.24)	(0.57)	(0.28)	(0.57)
Observations	211	211	211	211	211	211	211	211
Log Likelihood	−116.86	−116.50	−119.25	−118.89	−119.27	−119.02	−116.84	−116.45
Akaike Inf. Crit.	239.71	245.00	244.49	249.78	244.54	250.04	243.69	248.89

Note: *p < 0.05; **p < 0.01; ***p < 0.001

3 and 4 perform the same comparison (main effect and the effect after controls) for the partisan identity that is most likely to be shared with the Self-Raters (Republican for Illegal Immigration; Democrat for Climate Change). Models 5 and 6 repeat this process for shared ideology (conservative for Illegal Immigration; liberal for Climate Change). These measures of political and ideological affinity do not appear to account for accuracy. As Models 7 and 8 demonstrate, adding these affinity measures does not reduce the substantive or statistical significance of Similarity, which indicates that there is a meaningful distinction between belief Similarity and these other constructs.

I next test the idea that the capacity to understand others' emotions is an asset when trying to understand why they feel threatened. One possibility is that the encouragement to understand emotions *before* attempting the threat estimation task makes that task easier. In this view, simulating (or at least focusing on) others' emotions provides a relatively intuitive and reliable way to access other aspects of their mind. However, as shown in Figure 4 and Tables 1 and 2, there was no statistically significant effect of considering others' emotional responses to either Issue before estimating threat perception in this study.[94] Nor did those in the Emotions-First condition become better guessers; there was no difference between the Threats-First and Emotions-First conditions

[94]Chi-squared test of differences in binary accuracy for Similar Mentalisers in Threats-First versus Emotions-First conditions, with p-values simulated based on 2,000 replicates: Illegal Immigration $\chi^2 = 0, p = 1$ (groups have identical accuracy rates); Climate Change: $\chi^2 = 0.27, p = 0.66$. For Dissimilar Mentalisers: Illegal Immigration $\chi^2 = 2.01, p = 0.18$; Climate Change: $\chi^2 = 0.76, p = 0.55$.

Table 2. Correlates of binary threat estimation accuracy

	Issue: Climate Change							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Similarity	1.13**	1.18**					0.96*	1.00*
	(0.37)	(0.39)					(0.39)	(0.40)
Democrat			0.82*	0.87*			0.58	0.56
			(0.36)	(0.38)			(0.42)	(0.44)
Liberal					0.58	0.71	0.10	0.29
					(0.39)	(0.40)	(0.45)	(0.46)
Emotions-First	0.03	0.001	0.08	0.09	0.15	0.15	0.05	0.04
	(0.35)	(0.36)	(0.35)	(0.36)	(0.35)	(0.36)	(0.36)	(0.37)
Female		0.84*		0.76*		0.86*		0.86*
		(0.37)		(0.37)		(0.37)		(0.38)
Age		−0.01		−0.01		−0.01		−0.01
		(0.01)		(0.01)		(0.01)		(0.01)
White		0.20		0.35		0.22		0.38
		(0.37)		(0.38)		(0.37)		(0.39)
Constant	0.59	0.41	1.00***	0.87	1.15***	1.01	0.40	−0.05
	(0.34)	(0.69)	(0.28)	(0.64)	(0.27)	(0.64)	(0.36)	(0.73)
Observations	214	214	214	214	214	214	214	214
Log Likelihood	−101.54	−98.28	−103.19	−99.97	−104.75	−101.14	−100.11	−96.57
Akaike Inf. Crit.	209.08	208.55	212.37	211.94	215.49	214.28	210.22	209.14

Note: $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

on miss distances.[95] While the treatment effect on Dissimilar Mentalisers appears to improve accuracy to better rates than chance for both Issues (dashed horizontal line), the size of the treatment effect itself is not statistically significant.

A second possibility is that understanding others' emotions provides useful context for understanding the reasons why they feel threatened. This effect would only hold for those whose emotion understanding was accurate, however. In this case, there should be a positive relationship between emotion understanding accuracy and threat estimation accuracy in the Emotions-First conditions. As Figure 5 shows, this positive relationship (measured as a Pearson correlation) is only greater than chance among Similar Mentalisers.[96] There is also no significant difference between the correlations for Similar Mentalisers in the Emotions-First and Threats-First conditions, suggesting the information added by considering emotions first is not substantial for that group.[97] As such, it does not seem as though access to others' emotions provides a consistent benefit in the threat estimation task, above and beyond the influence of shared beliefs about dangerousness.

[95]Mann-Whitney U test of differences in miss distance for Similar Mentalisers in Threats-First versus Emotions-First conditions with exact p-values: Illegal Immigration $W = 57, p = 0.41$; Climate Change: $W = 52, p = 0.18$. For Dissimilar Mentalisers: Illegal Immigration $W = 114, p = 0.70$; Climate Change: $W = 28, p = 0.66$.

[96]Correlations between binary accuracy for Emotion Understanding and for Threat Estimation, p-values derived from 500 simulations. For all Mentalisers in the Emotions-First conditions: Illegal Immigration, $\Phi = 0.19, p = 0.06$; Climate Change, $\Phi = 0.24, p = 0.07$. For Similar Mentalisers: Illegal Immigration, $\Phi = 0.22, p = 0.02$; Climate Change, $\Phi = 0.32, p = 0.01$. For Dissimilar Mentalisers: Illegal Immigration, $\Phi = 0.14, p = 0.15$; Climate Change, $\Phi = 0.04, p = 0.72$.

[97]Comparison of correlations between independent groups of Similar Mentalisers: Illegal Immigration, Fisher's $z = 1.28, p = 0.20$; Climate Change, Fisher's $z = -0.52, p = 0.60$.
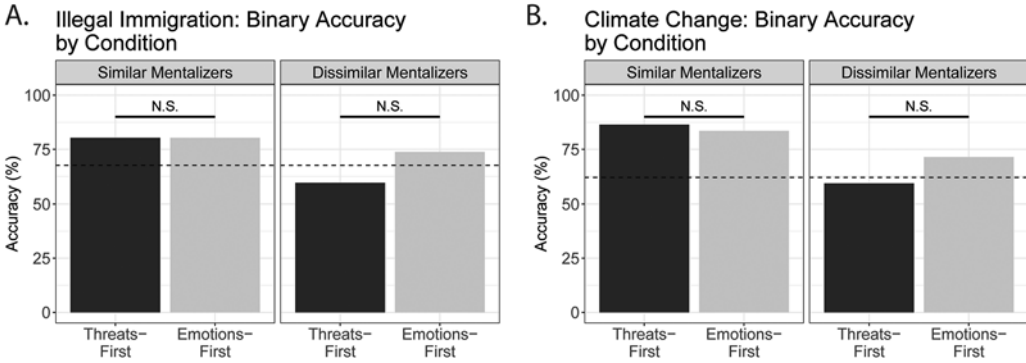
Figure 4. Effects of mentalising task order on threat estimation accuracy. (A) Illegal Immigration: Binary accuracy by condition, (B) Climate Change: Binary accuracy by condition.
Note: Dotted black line represents 95th percentile of Chance simulations. N.S. indicates the difference between columns is not statistically significant at the $p < 0.05$ level.
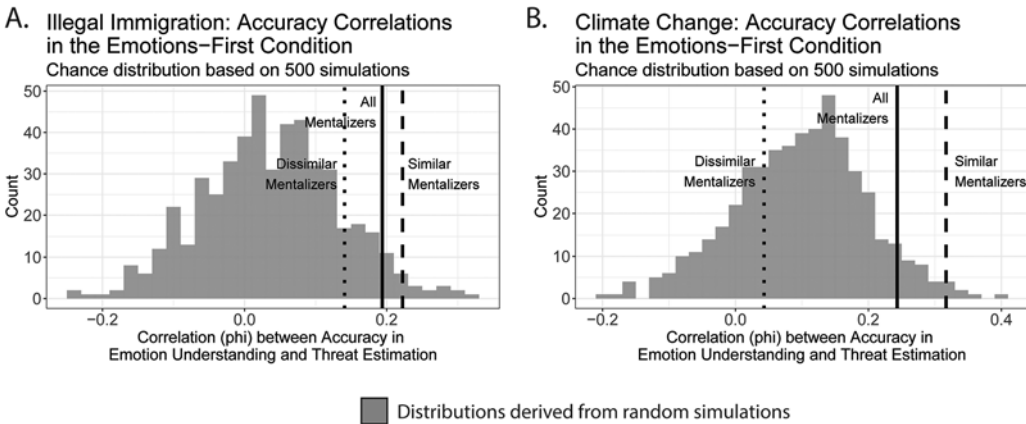

Figure 5. Correlation between emotion understanding and threat estimation accuracies. (A) Illegal Immigration: Accuracy correlations in the Emotions-First Condition. (B) Climate Change: Accuracy correlations in the Emotions-First Condition.
Note: Solid black lines represent results for all Mentalisers in the Emotions-First Conditions. Dashed lines represent the results for the subset of Similar Mentalisers. Dotted lines represent the results for the subset of Dissimilar Mentalisers.

Conclusion

The objective of this study was to provide a better understanding of the link between mentalising *effort* and mentalising *accuracy* in the domain of threat perception. The study's design was a deliberate compromise between the need to cleanly identify a challenging phenomenon like mentalising accuracy and tackling a question with relevance to IR. As such, the study's findings are best interpreted as a baseline against which to compare more complex mentalising tasks that take place in foreign policy decision-making. The threat perceptions Mentalisers were asked to estimate were multidimensional, but structured, which is not always a feature of real-world decisions. The 'others' whom they imagined were also not explicitly defined as adversaries, and recent work suggests that mentalising in adversarial settings may present unique challenges to inferential accuracy.[98] Future research is thus needed to identify ways in which these baseline findings are affected by such factors.

[98]Landau-Wells, 'Building from the brain'.

The baseline findings provided three insights into the relationship between mentalising effort and accuracy. First, mentalising effort was not sufficient to generate mentalising accuracy above rates that could be achieved by chance for subjects who did not already view their Issue as dangerous. Yet the failures of accuracy were not random. The pattern of missed guesses indicated that the Dissimilar Mentalisers applied effort but simply had the wrong mental model of their target. Second, this critical difference in beliefs about dangerousness was not simply politics in disguise. Neither self-reported party identification or ideological orientation provided as much explanatory power for binary accuracy as sharing beliefs about dangerousness. Thus, while finding that the effects of mentalising effort are conditional is consistent with prior work, the differentiator in this case is neither explicitly political (as in Casler and Groves)[99] nor adversarial (as in Kertzer, Brutger, and Quek).[100] Instead, the simplified structure of the study makes it possible to identify the basic significance of shared beliefs about dangerousness for threat estimation task performance.

The third insight provided by the study is the disambiguation between threat estimation and other forms of mentalising that are theoretically posited to assist that mental exercise. Specifically, I found that understanding another person's emotional responses to the Issue that concerned them (Climate Change or Illegal Immigration) provided no incremental benefit for threat estimation accuracy. While threat perception itself cannot occur without an emotional response, accurately understanding the latter did not correlate with accurately estimating the former. This finding is consistent with a literature on the multifaceted nature of social cognitive skills and their situation-specific utility.[101] But it suggests that – to the extent threat estimation is considered important for interpreting others' actions – focus should be on fostering a better understanding of their beliefs about danger.

While I have argued for this experiment's utility, these findings should be treated as provisional, as with any single study. While this study tested threat estimation of two prominent issues in international politics, new tasks and issues are necessary to determine whether the findings in this study hold elsewhere. Another limitation in applying this study's findings is its focus on recovering group-wide, rather than individual-level, threat perceptions. This set-up rendered the mentalising task easier because there were many possible 'right answers'. Indeed, one reason for the high absolute (and chance) binary accuracy levels in the study was the size of the ground-truth hypervolumes, since the Self-Raters disagreed amongst themselves about why either Climate Change or Illegal Immigration presented a danger. But in the case of a single individual, or a group with more internally consistent perceptions (e.g. advisors), the space of plausible answers shrinks significantly. The mentalising precision required might be offset by specific information about the targets of inference (e.g. knowing a person's history), but the difficulties faced by Dissimilar Mentalisers suggest that recovering a precise point, particularly from an adversary, is inherently challenging.

This study has several implications. First, future research into the effects of mentalising effort on inferences about and responses to others' behaviour should explicitly incorporate a theoretical position on accuracy. It is quite possible that *inaccurate* mentalising is the driver of optimal behaviour in certain contexts.[102] But if this is the case, it is not clear that greater information from intelligence sources[103] or personal exchanges[104] is what will foster better decisions. Instead, much of the work arguing for mentalising's beneficial effects does so on the presumption of accuracy. But

[99]Casler and Groves, 'Perspective taking through partisan eyes'.

[100]Kertzer, Brutger, and Quek, 'Perspective-taking and security dilemma thinking'.

[101]Adam D. Galinsky, William W. Maddux, Debra Gilin, and Judith B. White, 'Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations', *Psychological Science*, 19:4 (2008), pp. 378–84; Debra Gilin, William W. Maddux, Jordan Carpenter, and Adam D. Galinsky, 'When to use your head and when to use your heart: The differential value of perspective-taking versus empathy in competitive interactions', *Personality and Social Psychology Bulletin*, 39:1 (2013), pp. 3–16.

[102]Ickes and Simpson, 'Motivational aspects of empathic accuracy'; Johnson, *Strategic Instincts*.

[103]Shore, *A Sense of the Enemy*; White, 'Empathy as an intelligence tool'.

[104]Holmes, *Face-to-Face Diplomacy*; Holmes and Yarhi-Milo, 'The psychological logic of peace summits'.

this presumption should be clearly stated and explicitly theorised, given that there are alternative possibilities (e.g. useful inaccuracy).

The second implication is that the drivers of misperception deserve renewed scrutiny. Misperception is often attributed to a lack of mentalising effort, consistent with Stein's observation, or to confounders like deception.[105] The difficulty of mentalising, even in relatively optimal circumstances, suggests that trying and failing can have observably the same effects as a lack of effort. But there are consequences in theory and in practice for treating misperception as mostly correctible or treating perceptual accuracy as a rare event. The study of event forecasting, which takes the latter approach, offers some inspiration for IR scholars interested in better understanding the interaction between raw capabilities, individual differences, and systemic conditions.[106] Further, by studying threat perception accuracy as a rare event, scholars and practitioners might identify new interventions that go beyond encouraging the *effort* to understand others' perceptions and actually improve our *accuracy* when doing so.

Beyond these substantive implications, this paper contributes a multidisciplinary solution to the problem of studying mentalising and threat perception. Experiments are nothing new in IR research. But new tools and methods can expand the boundaries of the phenomena open to study. Using analytical approaches from other disciplines, I was able to separate mentalising effort from mentalising accuracy. The study of mentalising in IR has focused on situations where this type of disambiguation is extremely difficult, if not impossible. Yet this study suggests we should find new ways to investigate this distinction because the inherent difficulty of mentalising may be a much more fundamental challenge to optimal decision-making than previously assumed.

Marika Landau-Wells is an Assistant Professor in the Charles and Louise Travers Department of Political Science at the University of California, Berkeley. Her work focuses on international security, foreign policy decision-making, political behaviour, and the application of cognitive science to the study of politics.

[105]Stein, 'Building politics into psychology'.

[106]Welton Chang, Pavel Atanasov, Shefali Patil, Barbara A. Mellers, and Philip E. Tetlock, 'Accountability and adaptive performance under uncertainty: A long-term view', *Judgment and Decision Making*, 12:6 (2017), pp. 610–26; Christopher W. Karvetski, Carolyn Meinel, Daniel T. Maxwell, et al., 'What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters?' *International Journal of Forecasting*, 38:2 (2022), pp. 688–704.