Affective Science
https://doi.org/10.1007/s42761-021-00089-z

RESEARCH ARTICLE

The 3d Mind Model Characterizes How People Understand Mental States Across Modern and Historical Cultures

Mark A. Thornton[1] · Sarah Wolf[2] · Brian J. Reilly[3] · Edward G. Slingerland[4] · Diana I. Tamir[5]

Received: 4 June 2021 / Accepted: 25 October 2021

Abstract
Humans rely on social interaction to achieve many important goals. These interactions rely in turn on people's capacity to understand others' mental states: their thoughts and feelings. Do different cultures understand minds in different ways, or do widely shared principles describe how different cultures understand mental states? Extensive data suggest that the mind organizes mental state concepts using the 3d Mind Model, composed of the following psychological dimensions: rationality (vs. emotionality), social impact (states which affect others more vs. less), and valence (positive vs. negative states). However, this evidence comes primarily from English-speaking individuals in the USA. Here, we investigated mental state representation in 57 contemporary countries, using 163 million English-language tweets; in 17 languages, using billions of words of text from internet webpages; and across more than 2,000 years of history, using curated texts from four historical societies. We quantified mental state meaning by analyzing the text produced by each culture using word embeddings. We then tested whether the 3d Mind Model could explain which mental states were similar in meaning within each culture. We found that the 3d Mind Model significantly explained mental state meaning in every country, language, and historical society that we examined. These results suggest that rationality, social impact, and valence form a generalizable conceptual backbone for mental state representation.

Keywords Emotion · Social cognition · Cross-cultural · Text analysis

Introduction

For millennia, humans have engaged in, and been defined by, complex social interactions (Bowles & Gintis, 2003; Boyd & Richerson, 2009; Tomasello & Vaish, 2013). Humans rely on cooperation and competition with conspecifics to achieve their most critical goals, from self-defense to food acquisition to childrearing. Successful social interactions are predicated, at least in part, upon the ability of individuals to understand the thoughts and feelings of others. Existing evidence suggests that cultures vary considerably in how they understand mental states (Adams et al., 2010; Jackson et al., 2019; Lillard, 1998; Liu et al., 2008). This raises an important question: are any principles of mental state representation shared widely across cultures, or do people in each culture understand others' minds in an idiosyncratic way?

Data from contemporary English-speakers in the USA suggests that the 3d Mind Model organizes mental state representation (Tamir & Thornton, 2018; Tamir et al., 2016; Thornton & Tamir, 2017, 2020; Thornton et al., 2019b). This model describes how the mind and brain represent others' mental states—both cognitive states, such as reasoning and decision-making, and affective states, including emotions and moods. It posits that when people think about these states, they attend to three key features: rationality, social impact, and valence. Attending to rationality means that one can distinguish cognitive states like

Handling editor: Kristen Lindquist

✉ Mark A. Thornton
Mark.A.Thornton@dartmouth.edu

[1] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA

[2] The Jewish Theological Seminary, New York City, NY 10027, USA

[3] Department of Modern Languages and Literatures, Fordham University, Bronx, NY 10458, USA

[4] Department of Philosophy, University of British Columbia, Vancouver, BC V6T 1Z2, Canada

[5] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

Published online: 22 January 2022

calculation and planning from affective states like ecstasy and grief; attending to social impact means that one can distinguish intense, social states like love and envy from low energy, asocial states like exhaustion and stupor; and attending to valence means that one can distinguish positive states like awe and gratitude from negative states like sadness and anger. The mind uses these dimensions to represent others' complex internal states parsimoniously (Tamir et al., 2016).

The 3d Mind Model was derived from other theories in the social and affective science literatures, including the Circumplex Model of Affect (Russell, 1980), the Stereotype Content Model (Cuddy et al., 2008; Fiske et al., 2002), the agency and experience model of mind perception (Gray et al., 2007), and eight other theoretically important dimensions (Tamir et al., 2016). The 3d Mind Model outperforms these (and nearly 40 other) candidate dimensions in capturing how people think about mental states. For example, the 3d Mind Model can predict how people's brains will respond to thinking about mental states (Thornton & Tamir, 2020). Indeed, this model captures over 80% of reliable variance in neural patterns that arise as people consider others' mental states. The 3d Mind Model also captures how people make explicit conceptual judgments about mental states—for instance, judging which of two mental states is more similar to a third. Together, these findings provide strong evidence for the validity and explanatory power of the 3d Mind Model, at least within contemporary U.S. American culture.

The 3d Mind Model is useful for capturing how people *think* about mental states because it also reflects a key element of how people *experience* mental states. Specifically, it encodes how people dynamically transition from state to state. People are more likely to experience transitions between states that are closer in this space than further away (Thornton & Tamir, 2017). For example, people are highly likely to feel pride one moment and then happiness the next, and indeed, happiness and pride are located close together in this three-dimensional space. By encoding transition probabilities in this way, people can easily and accurately predict others' future mental states (Koster-Hale & Saxe, 2013; Lin & Thornton, 2021; Tamir & Thornton, 2018; Thornton et al., 2019a). The ability to predict others' future emotions is highly adaptive. For example, if one sees a person experiencing pride, one could predict that they will soon experience happiness and engage with them on that basis. People learn this model by observing emotion dynamics in the world and, in doing so, acquire an accurate and efficient way to predict the social future (Thornton & Tamir, 2017; Thornton et al., 2019b, 2020).

This prior research answers the questions "*How* are mental states represented?" and "*Why* are they organized that way?" However, they do not tell us *where* in the world, or *when* in history, this model applies. To date, this model has been tested primarily in a single culture, the contemporary USA, that is far from representative of all cultures (Henrich et al., 2010; Schulz et al., 2018). Indeed, prior studies have demonstrated considerable cross-cultural variation in emotion expression, perception, and linguistic representation (Gendron et al., 2014a, 2014b, 2018; Jackson et al., 2019). Thus, cross-cultural validation is critical to determine how widely the 3d Mind Model can be generalized.

In the present investigation, we test the 3d Mind Model in a wide array of cultures, including 57 countries, 17 languages, and 4 historical societies. We evaluate people's concepts of emotions and other mental states by analyzing large text corpora, including social media posts from across the globe (Fig. 1), multilingual web content, and historical archives. Words reflect the ways that people understand concepts (Mikolov et al., 2013). Analyzing texts from diverse cultures offers a unique window into commonalities and variations in human psychology (Jackson et al., 2021). Text analyses can recover the explicit meaning of words as well as the implicit connotations associated with them (Caliskan et al., 2017). The modern data are highly naturalistic, capturing the language produced spontaneously by ordinary people in their everyday lives. The archival data allow us to query the mental state representations of people from historical societies (e.g., seventeenth-century France)—populations otherwise inaccessible to modern psychologists. Together, these texts offer novel insight into mental state representation across cultures.

Our computational text analyses offer a scalable, naturalistic test of whether each of these diverse cultures mentalizes using the 3d Mind Model. The results indicate how broadly and effectively the 3d Mind Model generalizes from contemporary America to a broad range of other peoples, places, and times.

Methods

Text Corpora

To test whether people in different cultures conceptualize mental states in line with the 3d Mind Model, we drew upon text from a variety of sources. We operationalized culture in three ways: spatial, linguistic, and temporal. Spatially, we examined English-language text produced in contemporary countries around the world. Linguistically, we examined texts from an array of different languages. Temporally, we investigated text from different historical societies and periods. None of these operationalizations on its own cuts culture cleanly at its joints. For example, people of different cultures live within the same country, and people in different countries can share the same culture. Likewise, people speaking the same language may have very similar or very

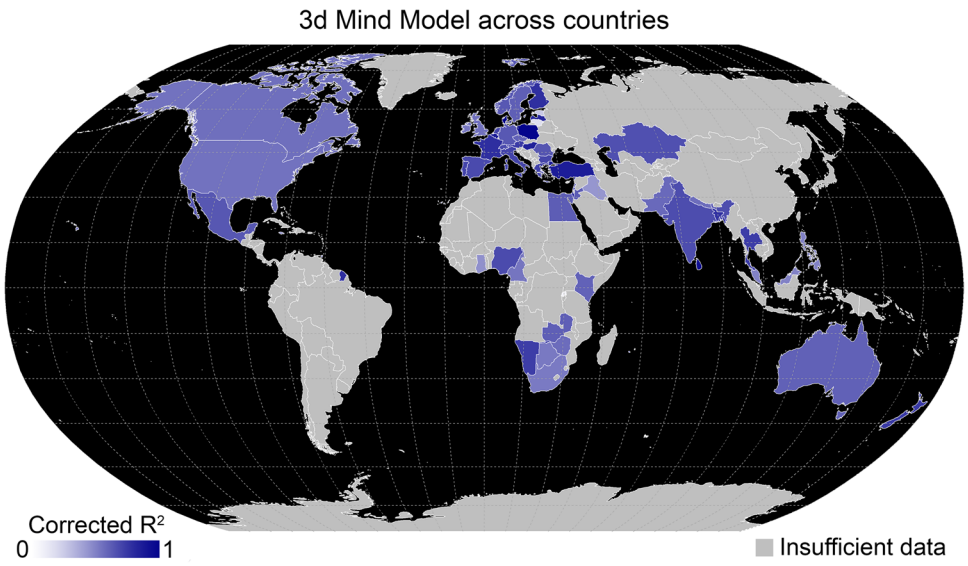

Fig. 1 The 3d Mind Model across countries. The 3d Mind Model was tested in each of the countries shown in blue. The model performed statistically significantly in all countries tested. Countries shown in gray were either not targeted for analysis due to small English-speaking populations or were excluded after data collection due to insufficient sample sizes (less than 10,000 English tweets)

different cultural practices in other respects. However, by triangulating culture from all three of these perspectives, we can conduct a thorough test of the generalizability of the 3d Mind Model. The following subsections describe the text corpora examined at the country, language, and historical levels of analysis.

Country-Level Text We selected an initial set of countries for analysis based on the sizes of their English-speaking populations (https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population). Specifically, we identified 67 countries with total populations of at least 1 million and at least 10% of that population speaking English as a first or additional language. For each country in this set, we identified bounding boxes (longitudes and latitudes) that encompassed the main geographic units of the country (https://github.com/graydon/country-bounding-boxes). Some countries were described by multiple bounding boxes due to their size and shape (e.g., three separate boxes were used for Alaska, Hawaii, and the lower 48 U.S. states). This resulted in a final set of 82 bounding boxes. Having identified regions of interest, we next used Twitter's streaming API to collect tweets from within the bounding boxes. Custom Java code was used to collect 167 million English-language tweets between March 16 and July 19, 2018 (with brief interruptions due to technical issues). These dates were selected based on convenience. Non-English-language tweets—as determined via Twitter's automatic identification system—were filtered out so that language could be held constant for country-level analyses.

We verified the country of origin of each tweet using the "geo" field information provided by the streaming API (see SI). We discarded tweets whose locations could not be verified. We also excluded 10 of the original 67 countries for having fewer than 10,000 total tweets: Croatia, Mauritius, Slovakia, Slovenia, Madagascar, Liberia, Papua New Guinea, Sierra Leone, Lesotho, and Morocco. This left a total of 163 million successfully located tweets across the remaining 57 countries. We preprocessed the Twitter text to maximize the natural language content and minimize artifacts and noise. We removed all Twitter and web operators, including @handles, #hastags, and URLs. We removed all punctuation, transformed all uppercase letters to lowercase letters, and collapsed whitespace down to single spaces between words.

This country-level data allows us to test whether the 3d Mind Model can explain the meaning of mental state words across a wide array of countries. By using geolocated English-language Twitter data, this analysis provided spatial specificity to our investigation while holding language constant.

Language-Level Text To complement the country-level data, we also evaluated text from 17 languages: Arabic, Bulgarian, Chinese, Dutch, English, French, German, Gujarati, Hebrew, Hungarian, Japanese, Korean, Marathi, Romanian, Serbian, Spanish, and Turkish. Together, these languages are spoken by over 3 billion people (Eberhard et al., 2019), including many people from countries not included in our country-level texts. Moreover, these languages represent many different linguistic families and writing systems, providing another test of cross-cultural generalizability.

For the language-level analysis, we drew upon large pre-existing corpora: each language's version of Wikipedia (https://www.wikipedia.org/) from 2017, and the Common Crawl (https://commoncrawl.org/). The Common Crawl is an ongoing archive of all publicly available webpages on

the internet—at the time consisting of over 600 billion word tokens.

Historical Text Cultures evolve, hybridize, and disappear over historical time. As a result, existing cultures may not be representative of all cultures that have ever existed. Indeed, recent historical trends like Western colonialism, globalization, and mass media have arguably made the world's cultures more interconnected than at any other point in history. To complement our study of contemporary cultures, we examined four historical societies: classical period China (c. 1,000 BCE–220 CE), Jewish texts from antiquity through the Middles Ages, and seventeenth–nineteenth-century English and French writers. If the 3d Mind Model offers a generalizable description of how people represent mental states, then it should also apply to these temporally distant societies.

To study classical China, we examined 96 texts from the Chinese Text Project, totaling 5.7 million characters (Slingerland et al., 2017; Sturgeon, 2006). Texts in the Chinese Text Project were stripped of punctuation.

To study historical Jewish culture, we retrieved texts from Sefaria.org, which maintains a large library of Jewish texts. Since metadata was not available in this corpus, we manually identified texts from three periods of interest: pre-Talmudic, the Talmud, and the High Middle Ages. The pre-Talmudic texts consisted of the Tanakh, Apocrypha, Mishnah, Tanaitic, and Halakhic Midrash. The Talmud group consisted of both the Bavli and Yerushalmi portions of the Talmud. The High Middle Ages texts included the Zohar portion of the Kabbalah, Maimonides' Guide for the Perplexed and commentary on the Mishnah, Rashi's commentaries on the Tanakh and Talmud, and additional commentaries on the Tanakh by Ibn Ezra and Nachmanides. These groups of texts were composed of mixtures of Hebrew and Aramaic scripts.

Finally, we used the Standardized Project Gutenberg Corpus (Gerlach & Font-Clos, 2018) to assemble corpora of 31,927 English-language texts and 2,158 French language texts written by authors born in the 1600s, 1700s, and 1800s (author birth years but not dates of publication were available in the metadata). Although we refer to these as English and French cultures, this should not necessarily be construed narrowly as applying only to Britain and France. Both cultures colonized substantial overseas empires during the eras in question, and some portion of the literature may have been written in these places. These cultures were also undergoing rapid transformation during this period due to factors like the industrial revolution and political revolutions. These changes contributed to making these cultures into the outliers which they are today. Generalizing across the period thus provides a valuable test of the 3d Mind Model's cross-cultural applicability.

Mental State Terms

To determine whether language use was consistent with the 3d Mind Model, we focused on the meaning of mental state terms, such as "happiness," "planning," and "pity." If the 3d Mind Model predicted the meanings of such words within a given culture, this would suggest that people in that culture think about mental states in a way described by the model. To this end, we examined a set of 166 English mental state terms (see Fig. 2A for a complete list). This set was generated and extensively validated as part of previous research (Tamir et al., 2016; Thornton & Tamir, 2017, 2020). Part of this validation consisted of using judgments from a large sample of online participants to locate the coordinates of each of the mental states on the dimensions of the 3d Mind Model (see SI). These coordinates indicated the rationality, social impact, and valence of each state. As described in the representational similarity analysis section below, we used these coordinates to derive the independent variables in our statistical analysis.

For the language-level analyses, the 166 English mental state terms were translated into each of the 16 other languages under investigation. Volunteer translators fluent in English and the target language provided the translated terms. Translators were allowed to provide multiple translations for the same English term, if they thought it appropriate. Likewise, they could translate different English terms into the same non-English term. This flexibility meant that even words lacking a direct translation could be approximated via a combination of state terms, reducing the problem of untranslatable words. Nonetheless, starting with English words places limitations on our findings, in that there might be mental states in other languages that were not even indirectly encompassed by the English terms.

For two of our four historical analyses, we relied on expert translations of our mental state terms. A scholar of Classical Chinese provided translations of each of the 166 English mental state words into Classical Chinese characters. A scholar of the Jewish liturgy provided Hebrew and Aramaic translations for our examination of the Jewish corpora. Of the 166 English mental state terms, 104 English terms could be translated into historically appropriate Hebrew terms, and 44 into Aramaic. For our analyses of historical English and French texts, we used the same English to French translations for modern French in the language-level analysis. The centuries in question featured relatively standard orthography compared to previous historical epochs, meaning that the spellings in these texts are generally consistent with modern spellings. Some semantic drift in mental state word meaning may have occurred over this period. However, such drift is likely modest and could only be expected to weaken our observed effects, not produce false positives.

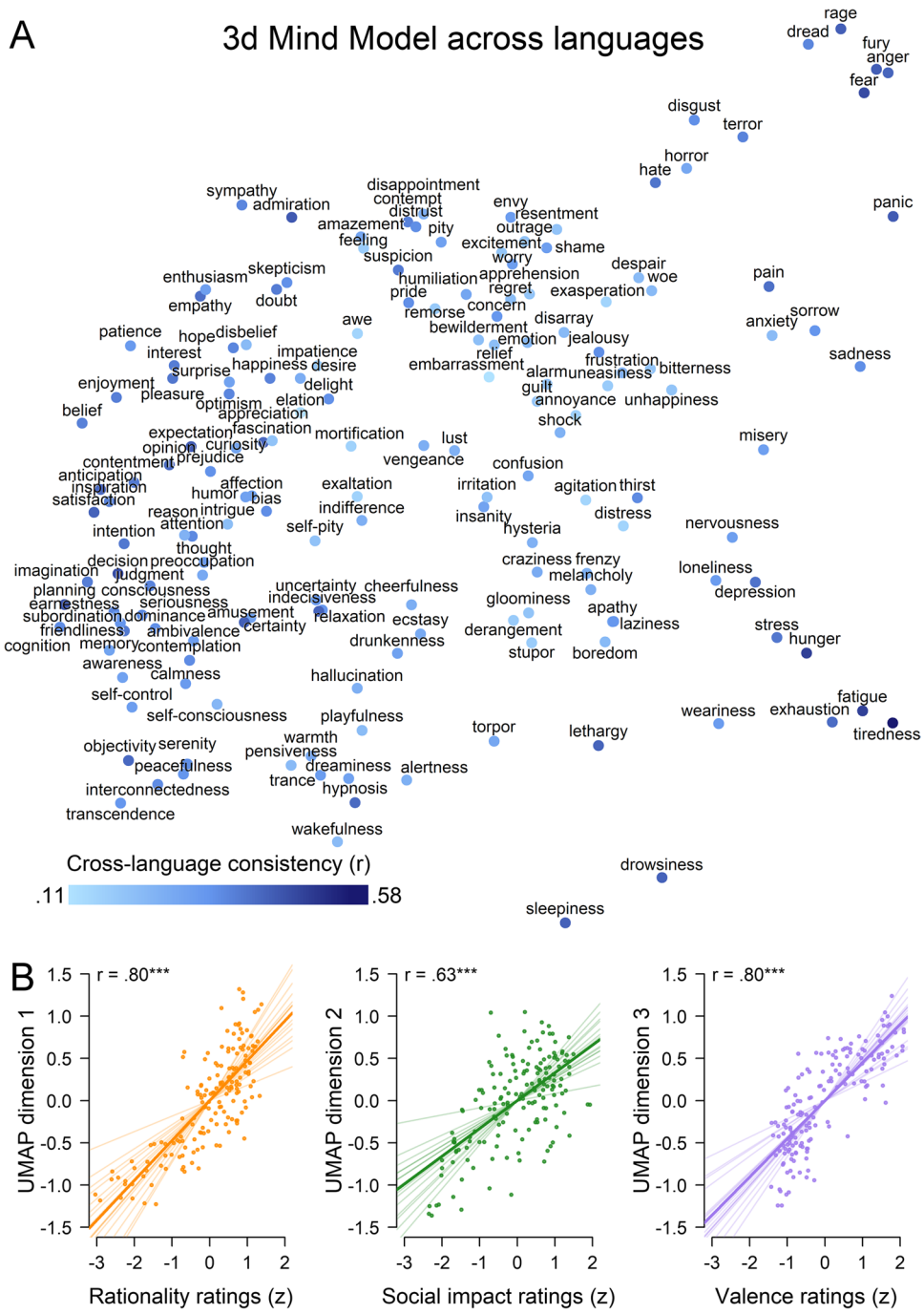

Fig. 2 The 3d Mind Model across languages. **A** A UMAP (McInnes et al., 2018) projection illustrates the semantic similarity between mental states words as proximity in a 2d space (see SI). The closer the two words are to each other, the more similar their vector representation in the 300d fastText embedding. The placement of words reflects the average across all 17 languages we examined. Darker blue points indicate states with more consistent meanings across languages. **B** Ratings of states on the dimensions of the 3d Mind Model (*x*-axes) were strongly correlated with scores from a 3d UMAP projection of the word embedding (*y*-axis). The solid line shows the linear best fit between the ratings and the cross-language average fastText embedding (***$=p<.001$). The lighter lines represent the fit of each of the 17 languages

Word Embeddings

We capitalized on recent developments in computational text analysis to automatically quantify the meanings of mental state words. Specifically, we used a word embedding algorithm known as fastText (Bojanowski et al., 2017). Word embedding algorithms estimate meaning from text using statistical regularities in natural language (Bojanowski et al., 2017; Grave et al., 2018; Mikolov et al., 2013). For example, the mental states happiness and excitement are similar in meaning and often occur in close temporal proximity to one another in real life (e.g., "I was excited to catch up with my friend and felt so happy spending time with them."). As a result, the words which denote the emotions (e.g., "happy" and "excited") tend to co-occur with each other—and with other related words (e.g., "friend" in the example above)—in text. These co-occurrences are precisely the sorts of associations that a human observer might want to learn to predict mental states in daily life (Tamir & Thornton, 2018). They

are also what allow word embeddings to estimate word meaning.

Word embeddings represent words as points (known as "vectors") in high-dimensional (e.g., 100–300d) spaces. As an embedding is trained on a text corpus, it learns to predict which words will appear in the text based on the other nearby words. For example, if the algorithm observed that the word "excited" and "happy" in the same sentence, it would place the corresponding word vectors for excited and happy closer together. By the end of the training process, vectors that are closer together in the embedding represent words with more similar meanings.

Conceptual meaning is not the only factor that leads to statistical regularities in natural language. Other factors, such as purely syntactic rules, also produce characteristic patterns of word co-occurrence. These co-occurrences may not always reflect meaning, but they do help the algorithm predict which words are likely to occur in the text, and as such, they are incorporated into the embedding. As a result, an embedding will typically represent not only the meanings of words, but also grammatical features unrelated to meaning. Our goal is to test the extent to which the 3d Mind Model captures the part of the embedding space that defines the meanings of words.

FastText is a state-of-the-art word embedding algorithm (Bojanowski et al., 2017). It differs from earlier word embedding algorithms in that it enriches its word vectors using subword information. For example, "joyful" and "joyous" are treated as entirely separate words by other algorithms. In contrast, fastText recognizes the shared stem "joy" and uses this to inform the vectors assigned to both of these words, and the word "joy" itself. This feature helps train the embedding more efficiently and minimizes the need for preprocessing steps such as lemmatization. Subword information also allows fastText to compose vectors for new words that never appeared verbatim in its training corpus. As described in the following sections, we applied fastText to the text from each of the cultures under investigation.

Country-Level Embeddings We applied fastText to the text of each country's tweets. Each tweet was treated as a separate document in this analysis, meaning that fastText only considered word co-occurrences within the same tweet. FastText produced a 100d word embedding for each country. This dimensionality is the default for the fastText algorithm, and prior research suggests that it is within the range of optimal dimensionality for word embeddings in general (Yin & Shen, 2018). From within each country's embedding, we extracted 100d vectors for each of the 166 mental state terms described above.

Language-Level Embeddings To obtain word vectors for each mental state in each language, we turned to a set of fastText embeddings pretrained on large web corpora (https://fasttext.cc/docs/en/crawl-vectors.html). These 300d embeddings were made available by the developers of the fastText (Grave et al., 2018). From each embedding, we extracted vectors corresponding to each of our 166 mental state terms. In cases where our translators had provided multiple translations of the same English term, we averaged the vectors for these different translations. Some translated terms consisted of two words separated by a space. fastText necessarily treats these as separate words, so in these cases, we extracted the vectors for each part of the term and averaged them. This may have introduced an additional source of noise into the non-English results, as the meanings of these terms may not be the sum of their parts.

Historical Embeddings We trained fastText separately on each of our historical corpora, including the Chinese Text Project, the three periods of Jewish liturgical texts, and each of the three centuries of English and French texts. The default settings in fastText assume that it is dealing with an alphabetic script, but Classical Chinese uses logograms. To address this, we changed the ngram setting of the algorithm to 1 so that it would only analyze individual logograms, rather than inadvertently treating multiword phrases as if they were words. From each embedding, we extracted 166 100d vectors corresponding to each of our 166 mental state terms.

Statistical analyses

Representational Similarity Analysis The 3d Mind Model makes specific predictions about how similar two states should be, based on their closeness on its dimensions. For example, two states that are similarly positive (e.g., happiness and awe) are closer together in the 3d Mind Model than a positive and a negative state (e.g., happiness and disgust). If the 3d Mind Model captures mental state meaning across countries, languages, and time, then the mental states that are closer together on the dimensions of rationality, social impact, and valence should be represented by word vectors that are closer to each other in the text embeddings. The presence of such associations would indicate that these dimensions provide a generalizable description of the representations that people use to understand the minds of others.

The dimensionality of the word embeddings (100d or 300d) and 3d Mind Model (3d) differed greatly. To compare how mental states are represented across these spaces, we turned to a technique known as representational similarity analysis (Kriegeskorte et al., 2008). This method functions by first computing the similarity (i.e., closeness) between words within each space. Specifically, we separately computed how close each mental state was to every other mental state on the 3d Mind Model dimensions and computed how

close each mental state vector was to every other mental state vector within the word embeddings. The resulting values capture the organization of mental states—reflected in patterns of similarities and dissimilarities—within each representational space. Next, we correlated these values with each other to determine how similarly the mental states were arranged within these two spaces. The stronger the correlation, the more similar the two spaces organize information about mental states, and thus the stronger the evidence that the 3d Mind Model describes mental state meaning within the culture.

For all three levels of analysis, we performed representational similarity analysis in the same way. First, we preprocessed the word vectors by mean-centering each of the 100 or 300 dimensions. We do this to remove the background signal shared across all of the word vectors, which merely reflects the fact that they are all mental state terms. Second, for each culture's trained embedding, we estimated the similarities between the mental state word vectors by Pearson correlating them with each other. More correlated word vectors indicate mental states with more similar meanings. Third, we computed the similarities between mental states on the dimensions of the 3d Mind Model. In a previous investigation (Tamir et al., 2016), we located the coordinates of each of the 166 mental state terms on each of the dimensions using human ratings (see SI). We used these data to compute distances between states as their absolute differences separately on each dimension of the 3d Mind Model. We reverse-coded (i.e., sign-flipped) each dimension's distances to make them into similarities rather than dissimilarities. This process yielded three variables, indicating how similar each mental state was to each other state on the dimensions of rationality, social impact, and valence. Finally, we regressed the word vector similarities onto the 3d Mind Model similarities. This analysis reveals how well the 3d Mind Model describes the meaning of mental state words in a given culture.

To supplement this analysis, in which we consider the 3d Mind Model as a whole, we computed the zero-order Pearson correlations between the word vector similarities and the similarities on rationality, social impact, and valence, respectively (see SI). We also performed parallel representational similarity analyses on the Circumplex Model of Affect (Russell, 1980) and compared the fits of the two models (see SI).

We evaluated the performance of the 3d Mind Model in three complementary ways:

Permutation Testing We assessed the statistical significance of each representational similarity analysis using the Mantel test—a permutation test suitable for comparing similarity matrices via correlation or regression (Smouse et al., 1986). This test allowed us to compute a p value for the 3d Mind Model based on the F statistic of the multiple regression. This procedure was conducted separately for each culture at each level of analysis (e.g., for each country in the country-level analysis).

Prevalence Estimates At the country and language levels, we estimated the prevalence of the 3d Mind Model's influence. Specifically, we examined the proportion of countries/languages within which the 3d Mind Model was statistically significant. To do so, we applied a binomial test to the observed numbers of significant and nonsignificant countries/languages, as determined by the Mantel tests. This provided a 95% confidence interval around the observed prevalence. This confidence interval reflects how widely spread the 3d Mind Model is in the population of countries/languages, though not necessarily the effect size within each culture. Since our α level for the permutation testing was 0.05, we would expect a prevalence of at least 5% statistical significance (false positives) even if the 3d Mind Model had no effect in any culture. Therefore, we also used these binomial tests to compute p values comparing the observed prevalence at each level of analysis to this chance expectation, thereby addressing the problem of multiple comparisons.

Noise Ceiling Analysis The raw R^2 values and correlation coefficients produced by the representational similarity analyses in this paper suffer from two serious biases. First, they show a strong sample size bias (see SI), such that effect sizes are strongly positively correlated with sample sizes ($rs > 0.85$). This occurs because larger samples of text lead to richer and more reliable word embeddings, thereby arbitrarily increasing the apparent performance of the 3d Mind Model within cultures that furnished more text. This bias is problematic because it misleadingly makes it appear as if the 3d Mind Model explains mental state representation far better in cultures that furnished us with more text.

Second, the raw R^2 and correlations are deflated by the presence of systematic but irrelevant variance in the word embeddings. Word embeddings reflect all of the statistical regularities in the text that help predict the presence of other words in the same context (e.g., predicting one word in a tweet based on all the other words in that tweet). Some of those statistical regularities reflect the conceptual meaning of words—hence, our ability to use word embeddings to study how people represent mental states. However, many language features are minimally related to conceptual meaning, such as parts of speech, frequency of occurrence, or some syntactic rules. The variance associated with these features cannot, and need not, be explained by a theory of mental state representation.

We conducted a noise ceiling analysis that instead compares the performance of our model to the proportion of variance that *does* need to be explained by a theory of mental

state representation (Kriegeskorte et al., 2008; Lage-Castellanos et al., 2019). This analysis estimated the best possible performance of a (U.S./English-based) theory of mental state representation in explaining the word embedding data from each culture (see SI). The resulting corrected effect sizes ($R^2_{corrected}$) indicate the variance in the word embeddings accounted for by the 3d Mind Model, as a proportion of the variance accounted for by the ideal model. In other words, the corrected values reflect how well the 3d Mind Model performs, relative to the best possible generalization from English-speaking American MTurk participants to the text generated by people of other cultures. The corrected values are conservative—equivalent to the lower bound of a 95% confidence interval. As a result, they are unsuitable for use in significance testing, and we did not use them for such.

Results

We tested the extent to which the 3d Mind Model could explain the meaning of mental state words across diverse countries, languages, and historical societies. We did so by estimating the meanings of 166 mental state words in each culture, using the fastText word embedding algorithm (Bojanowski et al., 2017). We calculated the similarities between mental state words within each culture and then tested how well similarities within the 3d Mind Model could predict these linguistic similarities. If the 3d Mind Model describes mental state meaning within a given culture, then state words with more similar meanings should also be closer together on the model's dimensions of rationality, social impact, and valence.

3d Mind Model Across Countries with English Speakers

Using three different analytic strategies, we found consistent evidence that the 3d Mind Model captures mental state meaning across 163 million Tweets from 57 contemporary nations. First, permutation tests indicated that the 3d Mind Model statistically significantly explained mental state meaning in all 57 countries we examined (Fig. 1; all $ps < 0.05$). This 100% prevalence is itself significantly above the 5% prevalence we would expect by chance ($p = 6.9 \times 10^{-75}$). Second, we inferred the likely global prevalence of the 3d Mind Model. We estimated that the 3d Mind Model would statistically significantly predict mental state meaning in 94.9% to 100.0% of all countries around the globe.

Third, we estimated the extent to which the 3d Mind Model could explain mental state meaning in each culture. On average, the 3d Mind Model explained a small proportion of the total variance in mental state meaning in each country: mean $R^2_{raw} = 0.012$ (95% CI = [0.0084, 0.017]). However, this effect size is not an accurate reflection of the variance explained by the model because it is both deflated and biased by sample size. We computed corrected effect sizes using a noise ceiling analysis. These $R^2_{corrected}$ values (Fig. 1) indicate that the 3d Mind Model achieved 65% of the best possible performance (mean $R^2_{corrected} = 0.65$, SD = 0.18). This shows that it accounts for the majority of mental state meaning that is shared between the USA and other countries. In sum, these results indicate that the 3d Mind Model predicts mental state meaning in a high proportion of contemporary countries with substantial English-speaking populations, at least among social media users.

3d Mind Model Across Languages

The country-level results provide initial evidence that the 3d Mind Model can explain mental state meaning across cultures. However, English-speaking Twitter users are not representative of the cultures from which they are sampled. For example, since most Twitter users are from the USA, users in other countries are likely to be relatively more exposed to American culture than their peers, merely by using the app. Moreover, tourists or migrants from predominantly English-speaking cultures may make up a disproportionate amount of the Twitter text from other countries. We addressed these concerns by operationalizing culture at the level of language rather than geographical regions.

We found consistent evidence that the 3d Mind Model captures mental state representations across 17 different languages. First, in the permutating test, we found that the 3d Mind Model significantly explained the meaning of mental states terms in all 17 languages we examined ($ps < 0.05$). This result of 100% prevalence is significantly above the 5% prevalence we would expect by chance ($p = 7.6 \times 10^{-23}$). Second, we inferred the prevalence of the 3d Mind Model across languages. We estimated that the 3d Mind Model would statistically significantly predict mental state meaning in 83.8% to 100.0% of all languages.

Third, we estimated the extent to which the 3d Mind Model predicted mental state meaning in each language. Across languages the mean R^2_{raw} of the 3d Mind Model was 0.052 (95% CI = [0.040, 0.065]). The mean $R^2_{corrected}$ of the 3d Mind Model was 0.50 (SD = 0.11), indicating that the 3d Mind Model explained on average half of the variance in mental state meaning shared between US raters and the languages we examined here. Together, these results indicate that the 3d Mind Model generalizes to explain mental state meaning across a wide variety of languages.

3d Mind Model Across History

Is the 3d Mind Model's ability to predict mental state meaning across contemporary countries and languages just an ephemeral effect of the modern zeitgeist? Widely applicable principles of cognition should generalize not just across space and language, but also across time. Historical societies make an excellent testbed for theories of cultural variability and generalizability (Muthukrishna et al., 2021). For instance, if the same psychological dimensions describe how temporally distant writers such as Colette, Jane Austen, Maimonides, and Confucius used mental state words, this would suggest that these dimensions have broad applicability across not just space and language, but also time. To test the 3d Mind Model across history, we analyzed corpora of Classical Chinese texts, the Jewish Liturgy from antiquity through the Middles Ages, and English and French language texts written by authors born in the seventeenth, eighteenth, and nineteenth centuries.

First, permutation testing results indicated that the 3d Mind Model significantly explained mental state word meaning in all the historical societies we examined (all $ps \leq 0.0001$). We did not perform binomial testing on the prevalence of the historical effects. Second, we examined the effect sizes of the 3d Mind Model in each historical society (Fig. 3). We observed the following R^2_{raw} values: 0.037 for Classical Chinese, 0.008 for pre-Talmudic Hebrew/Aramaic, 0.007 for the Talmud, 0.010 for High Middles Ages Hebrew/Aramaic, 0.051 for seventeenth-century English, 0.098 for eighteenth-century English, 0.12 for nineteenth-century English, 0.011 for seventeenth-century French, 0.036 for eighteenth-century French, and 0.064 for nineteenth-century French. Finally, in our noise ceiling analysis, we observed the following $R^2_{corrected}$ values: 0.35 for Classical Chinese, 0.45 for pre-Talmudic Hebrew/Aramaic, 0.56 for the Talmud, 0.52 for High Middles Ages Hebrew/Aramaic, 0.52 for seventeenth-century English, 0.55 for eighteenth-century English, 0.56 for nineteenth-century English, 0.59 for seventeenth-century French, 0.64 for eighteenth-century French, and 0.53 for nineteenth-century French. Together, these results indicate that the 3d Mind Model can explain mental state representation in societies spanning thousands of years of historical time.

Discussion

Complex social interactions have distinguished humanity for millennia and influenced the biological and cultural evolution of the species (Bowles & Gintis, 2003; Boyd & Richerson, 2009; Heyes & Frith, 2014; Tomasello & Vaish, 2013). To cooperate and compete effectively, individuals must understand each other's mental states: the thoughts, feelings, beliefs, intentions, and desires that shape behavior (Paal & Bereczkei, 2007). Although mentalizing appears ubiquitous across cultures, this does not mean that all cultures perform it in the same way (Slingerland et al., 2017). Here, we tested a modern theory of mental state representation—the 3d Mind Model (Tamir et al., 2016; Thornton & Tamir, 2020)—across a wide variety of modern and historical cultures. We found that this model explained mental state word meaning to a statistically significant degree in all 57 countries, 17 languages, and 4 historical societies we investigated. Together, these cultures span every permanently populated continent, most of the world's population, and nearly half of recorded history. The breadth of support for the 3d Mind Model across these cultures indicates that it provides a generalizable characterization of mental state representation.

These results still leave room for cultural diversity in mental state representation. Generalizability and variability are not mutually exclusive. A low-dimensional description like the 3d Mind Model may generalize comparatively well, but higher-dimensional descriptions may be necessary to

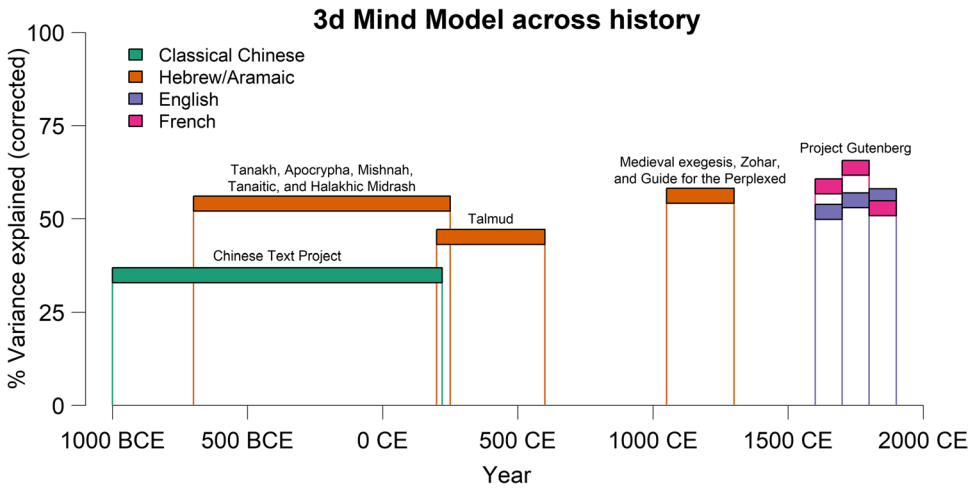

Fig. 3 The 3d Mind Model across historical societies. The 3d Mind Model significantly explained mental state meaning in all 10 historical cultures/periods we examined. The timeline indicates the years covered by each corpus and the variance explained by the model within each corpus ($R^2_{corrected}$). All dates are approximate

fully characterize mental state representation within individual cultures or people (Jolly & Chang, 2017). Moreover, the significance of the 3d Mind Model as a whole does not imply that every dimension was significant (see SI). Rather than undermining the importance of cultural variability in mental state representation, the present results offer a new starting point for describing how cultures differ.

The raw effect sizes we observed are relatively small. That said, these raw effects may be deflated due to the methods we used. In a noise ceiling analysis to correct this bias, we found that the 3d Mind Model achieves over half the performance of the best possible generalization from a US-based model to other cultures. This indicates that any shared meaning across cultures is captured well by the 3d Mind Model. However, this shared meaning appears to be modest compared to culturally idiosyncratic variance in mental state meaning.

The 3d Mind Model also outperforms another concrete theory (see SI), the Circumplex Model of Affect (Russell, 1980). The Circumplex Model is widely considered the gold-standard low-dimensional description of emotions. As a baseline comparison, it provides a better standard against which to test the 3d Mind Model than chance alone. In addition, the 3d Mind Model outperformed the Circumplex Model at the country-level and historical society level (see SI). These findings provide additional context for the 3d Mind Model performance in generalizing across cultures.

We could have compared the 3d Mind Model to other dimensional theories of word meaning, such as Osgood's factors of evaluation, potency, and activity (Osgood et al., 1957). However, we question the usefulness of comparison to domain-general theories. The semantic differential procedure used to generate the data behind Osgood's factors is problematic due to the polysemy of language. For example, it is possible to rate virtually anything on the "good–bad," but what makes a car "good" (e.g., affordable, good gas mileage) has virtually nothing in common with what makes a person "good" (e.g., moral, agreeable). Indeed, word embeddings reflecting the "goodness" mental states are *not* the same as those reflecting the "goodness" of objects (Thornton & Tamir, 2020). Moreover, the brain represents mental states in neural regions largely non-overlapping with those devoted to object representation (Konkle & Caramazza, 2013; Tamir et al., 2016). For these reasons, it is not clear that such domain-general models represent an appropriate alternative against which to compare the 3d Mind Model.

The present study features several limitations. First, it examines coarsely aggregated cultures, not individual people, subregions within countries, or specific years within historical periods. Second, it relies entirely on text analysis and starts from English mental state terms. As a result, mental state concepts that are not well captured by writing may have been poorly estimated. Our approach also implicitly excludes state words that originate in other languages and have no approximate equivalent in English. This could distort the apparent organization of mental states in such cultures. Third, our reliance on text inherently excludes cultures that do not produce text. Cultures that existed before the invention of writing, for example, might have conceptualized mental states in quite different ways from contemporary literate cultures. For these reasons, we cannot argue based on the present results that the 3d Mind Model is shared by all individuals or universal across cultures.

The current research joins a developing body of large-scale cross-cultural studies on the nature of mental states. Previous entries in this genre, such as Jackson et al. (2019), also drew upon language for insight into how people think about emotions. That work drew upon existing research by language experts to map out colexifications across a very large set of languages. Colexifications—in which two or more words in one language are represented by one word in another langue—provide a window into word meaning. Our approach complements this method by using a more data-driven approach to estimate the meaning of words from larger bodies of natural text. Additionally, instead of analyzing culture as coextensive with language, we triangulate culture from three different perspectives: spatial, linguistic, and temporal. Together, such studies also demonstrate how text provides a powerful new lens on the human mind emerging methods for understanding psychological content (Jackson et al., 2021).

The present investigation marks a substantial advance in our understanding of mental state representation. We find that a broad range of cultures may rely on a shared conceptual core to understand other people's thoughts, feelings, goals, and desires. These core concepts—characterized by the 3d Mind Model—offer scaffolding for understanding minds across cultures. This scaffolding may prove useful in constructing behavioral interventions to promote cross-cultural understanding or creating affective computing systems with greater cultural competence.

Acknowledgements The authors thank Abla Alaoui-Soce, Mihir Gandhi, Ece Hakim, Sagi Jaffe-Dax, Benedek Kurdi, Judith Mildner, Mira Nencheva, Ouafae Nobi, Danny Ochoa, DongWon Oh, Angela Radulescu, Milena Rmus, Karina Tachihara, Yinzong Wei, and Zidong Zhao for their assistance with translating mental state terms. We thank Dominic Burkart for their assistance with Twitter data collection and Abi Weber for her assistance with the Jewish liturgical texts. We thank Jason Mitchell for the use of pairwise similarity ratings drawn upon in the noise ceiling analysis. We are also grateful to Donald Sturgeon for the permission to download the CTP corpus in its entirety for analysis purposes and Ryan Nichols for cleaning and preparing this corpus for analysis.

Additional Information

Funding Information This work was supported by NIMH grant R01MH114904 to D.I.T.

Data Availability Data and materials are freely available on the Open Science Framework (https://osf.io/jqu3k/).

Ethical Approval The Committee on the Use of Human Subjects at Harvard provided ethical oversight and approval for the collection of the mental state dimension and similarity ratings.

Conflict of Interest The authors declare no competing interests.

Informed Consent Informed consent was obtained in a manner approved by Committee on the Use of Human Subjects at Harvard University from the participants who provided ratings of mental state dimensions and similarities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42761-021-00089-z.

Author Contribution Conceptualization: M.A.T., E.G.S, and D.I.T; methodology: M.A.T.; software: M.A.T.; validation: M.A.T. and E.G.S.; investigation: M.A.T.; resources: S.W., B.J.R. and E.G.S.; data curation: M.A.T. and E.G.S; writing—original draft: M.A.T.; writing—review and editing: M.A.T., S.W., B.J.R, E.G.S., and D.I.T; visualization: M.A.T. and D.I.T; supervision: D.I.T; project administration: M.A.T. and D.I.T.; funding acquisition: D.I.T.

References

Adams, R. B., Jr., Rule, N. O., Franklin, R. G., Jr., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., & Ambady, N. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience, 22*(1), 97–108.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Bowles, S., & Gintis, H. (2003). Origins of human cooperation. *Genetic and Cultural Evolution of Cooperation, 2003*, 429–443.

Boyd, R., & Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society b: Biological Sciences, 364*(1533), 3281–3288.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2019). *Ethnologue: Languages of the world*.

Fiske, S., Cuddy, A., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878–902.

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science, 27*(4), 211–219.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014a). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science, 25*(4), 911–920.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014b). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion, 14*(2), 251–262.

Gerlach, M., & Font-Clos, F. (2018). A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *ArXiv Preprint ArXiv:1812.08092*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *ArXiv Preprint ArXiv:1802.06893*.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD: To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners. *Nature, 466*(7302), 29–30.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*(6190), 1243091–1243091.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science, 366*(6472), 1517–1522.

Jackson, J. C., Watts, J., List, J.-M., Drabble, R., & Lindquist, K. (2021). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*.

Jolly, E., & Chang, L. J. (2017). *The Flatland fallacy: Moving beyond low dimensional thinking*.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *The Journal of Neuroscience, 33*(25), 10235–10242.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Lage-Castellanos, A., Valente, G., Formisano, E., & De Martino, F. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Computational Biology*, *15*(3), e1006397.

Lillard, A. (1998). Ethnopsychologies: Cultural Variations in Theories of Mind. *Psychological Bulletin, 123*(1), 3–32.

Lin, C., & Thornton, M. A. (2021). *Linking attributions of mental states and traits: Evidence for bidirectional causation*.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2008). Theory of mind development in chinese children: A meta-analysis of false-belief understanding across cultures and languages. *Developmental Psychology, 44*(2), 523–531.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a historical science. *Annual Review of Psychology, 72*, 717–749.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences, 43*(3), 541–551.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.

Schulz, J., Bahrami-Rad, D., Beauchamp, J., & Henrich, J. (2018). *The origins of weird psychology*.

Slingerland, E., Nichols, R., Neilbo, K., & Logan, C. (2017). The distant reading of religious texts: A "big data" approach to mind-body concepts in early China. *Journal of the American Academy of Religion, 85*(4), 985–1016.

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology, 35*(4), 627–632.

Sturgeon, D. (2006). *Chinese text project*. D. Sturgeon.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences, 22*(3), 201–212.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences, 113*(1), 194–199.

Thornton, M. A., Rmus, M., & Tamir, D. (2020). Mental state dynamics explain the origin of mental state concepts. *PsyArXiv*.

Thornton, M. A., & Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences, 114*(23), 5982–5987.

Thornton, M. A., & Tamir, D. I. (2020). People represent mental states in terms of rationality, social impact, and valence: Validating the 3d Mind Model. *Cortex, 125*, 44–59.

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019a). The brain represents people as the mental states they habitually experience. *Nature Communications, 10*(1), 1–10.

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019b). The social brain automatically predicts others' future mental states. *Journal of Neuroscience, 39*(1), 140–148.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64*, 231–255.

Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. *ArXiv Preprint ArXiv:1812.04224*.