Research Article

Psychological Science
1–19

DOI: 10.1177/09567976251362120
www.psychologicalscience.org/PS

What Is Rationality, Whom Is It Ascribed To, and Why Does It Matter? Evidence From Internet Text for 66 Social Groups and 101 Occupations

Charles A. Dorison[1] and Tessa E. S. Charlesworth[2]
[1]Management Area, McDonough School of Business, Georgetown University, and [2]Management & Organizations Department, Northwestern Kellogg School of Management, Northwestern University

Abstract

Scholars have extolled the virtues of rationality for centuries while also debating what rationality is and who is rational. Advancing these debates, we used word embeddings trained on 840 billion words of internet text—and validated with Prolific workers in the United States—to uncover the representation, group stereotypes, and occupational correlates of rationality at scale in naturalistic language. Four results emerged. First, rather than being synonymous with competence, representations of rationality included both an analytic/logic component and an interpersonal/trust component. Second, irrationality was not merely the opposite of rationality but contained its own unique subcomponents (volatility and unfairness). Third, rationality was consistently ascribed to high-power targets across 66 social groups. Last, rationality (especially its analytic component) was consistently associated with both earnings and wage gaps across 101 occupations. Associations with demographic representation were less consistent. Complementing normative approaches, these descriptive findings advance canonical debates about rationality, extending understanding of its components, stereotypes, and correlates.

Keywords

rationality, social groups, stereotyping, occupations, text analysis, word embeddings

Received 7/10/24; Revision accepted 7/8/25

Scholars have long extolled the virtue of rationality. Over a century ago, Aristotle proposed that the rational soul distinguishes humans from lower animals (Kraut, 2022). More recently, the virtues of rationality have captured the attention of scholars not only within psychology (Cusimano, 2025; Shafir & LeBoeuf, 2002) but also in related fields such as economics, sociology, philosophy, and political science (Knauff & Spohn, 2021; Pinker, 2021; Viale, 2021).

Despite widespread agreement that rationality is valuable, scholars have long disagreed on what rationality actually is. As Frank (1988) quipped, "there are almost as many definitions of rationality as there are people who have written on the subject" (p. 2). Definitions are sometimes complementary but are just as often conflicting. For example, within economics, rational choice theory centers the definition of rationality on analytic cost–benefit analysis as well as rigorous maximization of perceived self-interest (Becker, 1962). The rational decision maker is often portrayed as competent, cold, and calculating. However, other models of rationality within economics, psychology, and related disciplines define rationality as fundamentally about cooperation with others (Henrich et al., 2001; Kreps et al., 1982; Rand et al., 2012). Even when scholars begin with a common foundation (e.g., cognitive psychology), debates over rationality can turn

Corresponding Authors:
Charles A. Dorison, Management Area, Georgetown McDonough School of Business, Georgetown University
Email: charles.dorison@georgetown.edu

Tessa E. S. Charlesworth, Management & Organizations Department, Northwestern Kellogg School of Management, Northwestern University
Email: tessa.charlesworth@kellogg.northwestern.edu

caustic (Gigerenzer, 1996; Tversky & Kahneman, 1996), with some calling them "rationality wars" (e.g., Bermúdez, 2022). As Tetlock and Mellers (2002) summarized, "the debate over human rationality is a high-stakes controversy that mixes primordial political and psychological prejudices in combustible combinations" (p. 97).

The current research sidesteps such debates of how rationality *ought* to be represented. Instead, we used word embeddings trained on massive internet text to offer novel insights into how rationality *is* represented in naturalistic human language. Word embeddings, a natural language processing tool, represent all words in a text as numeric representations (vectors or strings of numbers) that can then be compared against each other to understand the relationships among words and concepts. Consequently, word embeddings have been used to uncover the existing cultural usage of several concepts, including group stereotypes (Caliskan et al., 2016). By taking a bottom-up, descriptive approach, word embeddings can cut across theoretical traditions (e.g., psychology vs. economics) and identify the concept and possible subcomponents of rationality as it is used at scale in human language.

Recognizing the benefits of a descriptive approach, Grossmann et al. (2020) examined humans' lay intuitions about rationality and reasonableness and whether these intuitions aligned with economic and legal scholarship. With a multimethod approach (including some related language-based analyses), the authors found that rationality could be differentiated from reasonableness and that rationality was relatively associated with an instrumental focus on preference maximization. But, like any good scholarship, this work raises as many questions as it answers. To what extent does rationality itself contain multiple underlying dimensions (vs. a single dimension)? For example, is rationality simply reducible to competence? Further, should irrationality be considered its own construct with unique subcomponents (rather than just the absence of rationality)?

Beyond assessing *what* rationality is, it is also critical to understand *whom* rationality is ascribed to. Given rationality's standing as a valued trait (Kraut, 2022), we can gain new insights into group disadvantages by considering whether and how social groups are stereotyped as rational (or irrational). Understanding stereotypes of intelligence/brilliance (Storage et al., 2020) and more general competence (Fiske et al., 2002) have helped shed light on, for example, disparities in representation across scientific fields (Leslie et al., 2015). Word embeddings can help expand such an understanding of group stereotypes into the domain of rationality. In particular, word embeddings provide unique insights by allowing us to first discover, from the bottom up, the rationality and irrationality constructs but also then to use those discovered constructs for understanding group stereotypes (see also Nicolas et al., 2021).

Last, there is an opportunity to link rationality and its group stereotypes to consequential societal outcomes. In an initial test, we explored whether an occupation that is stereotyped as more rational is also associated with greater earnings, larger gender gaps in earnings, and more White and male representation in the workforce. Moreover, we also considered how potential subcomponents of rationality may differentially be tied to these consequential outcomes—and even above and beyond previously well-studied domains of warmth, competence, or general valence. Altogether, such findings shed new light onto what rationality is, who it is ascribed to, and, ultimately, why it may matter for society.

Research Transparency Statement

General disclosures

Conflicts of interest: The authors declare no conflicts of interest. **Funding:** There was no funding for this research. **Artificial intelligence:** We used the generated outputs of ChatGPT as though it were a human rater or research assistant and compared its outputs with those obtained from our human participants and from our static embedding models. No other AI-assisted technologies were used in this research or in the creation of this article. **Ethics:** This research received approval from the Georgetown University ethics board (Study 00006750).

Study disclosures

Preregistration: The natural language processing (NLP) analyses of secondary data were not preregistered. However, the hypotheses, method, and analysis plan of the validation study in which we compared NLP results with human raters' results was preregistered (https://aspredicted.org/2pbc-f5bp.pdf). This preregistration was submitted prior to data collection, and there were no deviations from the preregistration for this study. **Materials:** There were no relevant materials for the secondary data analysis. For the validation study, we provide all materials (i.e., instructions and rating task given to participants). Materials are available at https://osf.io/cgjne under the folder "Additional study materials." **Data:** All primary data are publicly available (https://osf.io/cgjne). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/cgjne). **Open Science Framework (OSF):** To ensure long-term preservation, all OSF files were registered at https://doi.org/10.17605/osf.io/yktud. **Computational**

reproducibility: The computational reproducibility of the results in the main article (but not the Supplemental Material available online) has been independently confirmed by the journal's STAR team.

Method

We used word embeddings to explore the collective representations, group stereotypes, and occupational correlates of rationality in internet text. We begin by validating our approach and data setting (including with human data) before turning to the three central topics of interest. This research received approval from the Georgetown University ethics board (Study 00006750).

First, we mapped the semantic concepts of rationality and irrationality, including discovering latent subcomponents (Grossmann et al., 2020). Second, we examined whether these discovered concepts and subcomponents were systematically ascribed versus denied to 66 prominent social groups. Last, we assessed societal correlates, including earnings, gender gaps in earnings, and demographic representation, across 101 occupations. Although our analyses were not preregistered, we tested our theorizing across multiple data sets and analytic approaches (for full details, see the Supplemental Material). All files necessary to reproduce the analyses and figures are available on OSF (https://osf.io/cgjne).

Data source

The primary source of data was the largest existing set of pretrained static embeddings that was trained using the GloVe algorithm on 840 billion English words of internet text from Common Crawl (Pennington et al., 2014). The underlying Common Crawl text data are argued to represent a comprehensive "snapshot" of the entire internet (Caliskan et al., 2016). Although Common Crawl includes other languages in addition to English, we limited our investigation to English in the current study (but return to this point in the Discussion section). The GloVe algorithm is one of the earliest approaches to training static embeddings but has been extensively validated to show that it provides results analogous to human's representations of social concepts (e.g., Bhatia & Walasek, 2023; Caliskan et al., 2016). In fact, in recent work, this set of pretrained embeddings showed the strongest and most consistent correlations to human's attitudes compared with other commonly used static embeddings (Charlesworth et al., 2024). Thus, although we are not claiming that word embeddings have the same representations or learning processes as human minds (see Bhatia et al., 2019; Grossmann et al., 2023), it is indeed plausible to understand the outputs of word embedding associations as providing a window into the same associations held, on average, in society.

In essence, the GloVe algorithm is an unsupervised machine learning approach. First, the algorithm creates a large matrix of word co-occurrences within a specified context window (e.g., the co-occurrence of "bread" and "butter," "bread" and "jam," "bread" and "bird," and so on, within 10 words of one another), resulting in a matrix of size $N \times N$, where N specifies the number of words in the vocabulary. The algorithm then aims to reduce the dimensionality of that matrix to an $N \times 300$ matrix (where 300 is a dimensionality chosen by the user but is a standard dimension choice). Each word (i.e., a row in the matrix) then has an associated vector of 300 numbers that are used to encode its relative co-occurrence probabilities with all other words in the text. Formally, the objective of the algorithm is to minimize the difference between the dot product of two vectors (e.g., "bread" and "butter") and their co-occurrence probabilities so that words that appear often together (such as "bread" and "butter") will have similar vectors of numbers. In short, we aimed to have a set of word embeddings (an $N \times 300$ matrix) in which words that have similar meanings also had similar embeddings.

In addition to this primary source of GloVe (chosen because of its strong correlations to human attitudes), we also ensured the robustness of conclusions by looking at other static word embedding algorithms (e.g., fastText, word2vec) and other underlying text data sources (e.g., Wikipedia text alone), as well as combinations of these approaches (i.e., overlapping lists of words). Results from these replications generally converged with our primary conclusions and are fully reported in the Supplemental Material. Further, we conducted several analyses aimed at addressing common limitations of static embeddings, including the role of antonyms (e.g., Ali et al., 2019) and the interpretation of polysemous words. These additional analyses supported the results reported below and are also fully discussed in the Supplemental Material.

There are at least two reasons for using static embeddings rather than contextualized embedding algorithms (e.g., BERT) or generative language models (e.g., GPT). First, the relative simplicity and flexibility of static embeddings (such as GloVe) enable future investigations across new language settings (such as from historical text or other cultures; e.g., Wirsching et al., 2025). Second, the relative transparency and explainability of static embeddings enable greater researcher control than transformer-based models that are often proprietary (e.g., GPT) and have high complexity (with unexplainable parameters and layers; for a full discussion, see Bender et al., 2021). Perhaps most critically,

our exploratory analyses and robustness checks indicated that analyses with GPT (described below) would yield similar conclusions. Thus, we primarily relied on static embeddings, although we return to this choice in the Discussion section.

Validation of word embedding approach

Face validity of top-associated words. We began by discovering the bottom-up trait associates of the concepts "rational" and "irrational" using the seed words "rational" and "irrational." Specifically, for each trait (e.g., "able") in a long list of 408 traits (all traits available in the Common Crawl data; traits drawn from a longer list of approximately 600 traits—see Peabody, 1987), we computed the cosine similarity (essentially a correlation score) between the trait's vector (embedding) and the "rational" vector. We repeated this, separately, to compute the cosine similarity between the trait and the "irrational" vector. We then took the difference between the trait's similarity to "rational" versus "irrational." Additionally, to ensure conclusions generalized when we also considered nouns, verbs, and so on (i.e., not just limited to our selection of traits), we repeated this approach with 13,811 words (all words available in the Common Crawl data; words drawn from a longer list of approximately 14,000 words—see Warriner et al., 2013).

Table 1 reports the top 50 traits and words that are most relatively associated with rationality (and irrationality). These top associated traits are common synonyms of the concepts "rational" and "irrational." For instance, according to the Merriam-Webster online thesaurus, the top synonyms of "rational" include "logical," "analytic," "practical," and "intelligent," all of which were also included in the top 10 traits of "rational" in the bottom-up language-based analysis.

Baseline differences in average valence, warmth, and competence. In addition to face validity, there are also quantitative differences in the valence, warmth, and competence of these top "rational" and "irrational" concepts that provide convergent evidence for the validity of our approach. We selected these dimensions because valence is a core dimension of word meaning (Warriner et al., 2013), whereas warmth and competence are known to be central dimensions of social cognition (Fiske et al., 2007). Valence scores were taken from a prior study in which human raters rated the positivity/negativity of 408 traits (among a longer list of approximately 14,000 words; Warriner et al., 2013). Additionally, following the projection method discussed in Charlesworth et al. (2023) and Bolukbasi et al. (2016), each trait was given a score on its relative warmth (vs. coldness) and competence (vs. incompetence) by looking at the relative cosine similarities between the target trait and a set of seed words reflecting warmth/coldness and competence/incompetence (seed words taken from Nicolas et al., 2021).

As expected, relative to the rationality "neighborhood" of traits (i.e., the top 50 rational traits), the irrationality neighborhood of traits (i.e., the top 50 irrational traits) was much more negative (rational $M_{valence} = 1.38$; irrational $M_{valence} = -1.69$), $t(97.48) = 18.85$, $p < .001$, $d = 3.77$, less warm (rational $M_{warm} = 0.12$; irrational $M_{warm} = -0.14$), $t(91.66) = 32.13$ $p < .001$, $d = 6.43$, and less competent (rational $M_{competence} = 0.11$; irrational $M_{competence} = -0.11$), $t(94.83) = 25.90$, $p < .001$, $d = 5.18$. To put such effect sizes in perspective, a parallel analysis using the seed words "good" and "bad" (instead of "rational" and "irrational") found that the effects for rationality/irrationality were at least about 59% the magnitude of pure valence associations (which had effect sizes ranging from $d = 5.10$ to $d = 7.69$; see the Supplemental Material). Thus, the baseline differences in traits associated with rationality versus irrationality were almost as large and meaningful as the fundamental distinction of good versus bad, in line with the general classification of rationality as a widely held virtue.

Convergent validity with human data. Beyond assessing face validity and baseline differences, we also validated the rational and irrational traits against how laypeople classify these traits. To do so, we conducted a preregistered online study in which we recruited 225 participants from Prolific Academic; as preregistered, we excluded those participants who did not pass the attention check, resulting in a final sample of 187 participants ($M_{age} = 40.2$ years; 65% female, 34% male, < 2% other). Participants rated traits on a scale from 1 (*strongly irrational*) to 7 (*strongly rational*). This task matched the way we extracted trait projection scores from the language approach (i.e., how much each trait is associated with rationality vs. irrationality in language space). Each participant rated a random set of 60 total traits, yielding more than 1,000 total ratings across the 408 total traits (approximately 20 to 30 ratings per trait). We then took the average rating for each trait to extract a human-rated rationality/irrationality trait score and compared it with the association of that trait with rationality versus irrationality extracted from the embedding model. Human ratings were strongly correlated with language scores, $r = .90$, 95% confidence interval (CI) = [.88, .92], $t(406) = 41.10$, $p < .001$ (Fig. 1), lending confidence that the language representations of rationality/irrationality are meaningfully capturing humans' ideas of these concepts. A second study in which we asked lab participants to think of rational and irrational people in their lives and self-generate other traits to describe these targets also yielded moderate correlations with scores extracted from the embedding model (for full details, see the Supplemental Material).

Table 1. Words and Traits Most Relatively Associated With Rationality Versus Irrationality in Internet Text

Concept	Top 50 words	Top 50 traits
Rational	rational, concise, practical, sensible, efficient, analytical, pragmatic, analysis, logical, comprehensive, intelligent, competent, objective, reliable, thorough, framework, empirical, orderly, functional, thoughtful, adequate, mathematical, evaluation, basic, understanding, principles, reasonable, essential, impartial, provide, approach, refinement, comparative, coherent, deliberation, presentable, consideration, relevant, mathematics, robust, enable, dignified, straightforward, theoretical, methodical, satisfactory, modeling, necessary, accessible, proper	rational, concise, practical, efficient, analytical, logical, intelligent, competent, objective, reliable, thorough, orderly, thoughtful, understanding, reasonable, impartial, dignified, straightforward, methodical, trustworthy, ethical, formal, articulate, considerate, smart, dependable, constructive, prudent, diligent, independent, perceptive, conscientious, precise, honest, tactful, sober, knowledgeable, sophisticated, tidy, courteous, critical, refined, resourceful, insightful, flexible, simple, stable, respectable, accurate, careful
Irrational	uncontrollable, irrational, inexplicable, unfounded, unexplainable, illogical, nauseating, paranoia, infatuation, hysterical, hysteria, homophobic, manic, panicky, phobia, delusional, freakish, unbearable, hurtful, delirious, hateful, paranoid, superstitious, obsessive, shameful, sickening, unrealistic, frenzy, egomaniac, crazed, debilitating, uncalled, jealousy, insane, hypochondriac, obnoxious, maniacal, bizarre, insufferable, abusive, erratic, outrageous, spiteful, vicious, feverish, untrue, loathing, irresponsible, chauvinist, intolerable	irrational, illogical, superstitious, unrealistic, obnoxious, abusive, erratic, spiteful, irresponsible, vindictive, fanatical, inconsiderate, unkind, temperamental, resentful, egotistical, fearful, disrespectful, insecure, unpredictable, uncooperative, antisocial, reckless, compulsive, bullheaded, unruly, fickle, thoughtless, intolerant, jealous, deceitful, unjust, angry, unfair, insensitive, irritable, cranky, cruel, unreliable, impetuous, unstable, stubborn, rash, indiscreet, foolhardy, arrogant, cowardly, extravagant, belligerent, obstinate

Of note, in the supplemental analyses, we also explored whether more sophisticated language modeling approaches (namely the generative large language model GPT) would have yielded even stronger convergence with human data and thus even more compelling and valid results. In fact, as described in the Supplemental Material, we found that the results of the GPT-generated ratings were strongly correlated with the results from the static embedding models from GloVe Common Crawl, $r = .88$, 95% CI = [.85, .90], $t(406) = 36.59$, $p < .001$, and showed similarly strong correlations with human ratings, $r = .91$, 95% CI = [.89, .93], $t(406) = 44.17$, $p < .001$. Thus, it seems that an approach using GPT or other large language models would likely provide convergent and similar conclusions to those we report here using a lighter static embedding approach that has its own advantages in terms of transparency, lessened environmental/energy impacts, and flexibility for future applications across languages, media settings, or historical texts.

Results

Overview

Having introduced and validated our approach, we turn now to our three key analyses. The Supplemental Material reports numerous robustness checks to ensure that all conclusions are consistent across different word embedding algorithms (e.g., GloVe, fastText), data sources (e.g., Wikipedia, all internet text in Common Crawl), and analytic choices (e.g., principal components analysis approaches vs. exploratory factor analysis, aggregation decisions).

We began by considering the *what* of rationality (and irrationality). We moved beyond a simple description of rationality (provided in Table 1) to dig deeper into the latent meanings or subcomponents of this concept. For rationality and irrationality separately, we computed the latent principal components of their trait associates, and we offer both a qualitative and quantitative interpretation about the meaning of those subcomponents below. Critically, we investigated whether rationality yields meaningful subcomponents even beyond its association with competence and warmth (Fiske et al., 2007).

Second, we turned to the *who* of rationality. Prior work has considered how some social groups are stereotyped along dimensions that may be related to rationality, such as intelligence, brilliance, and competence (Storage et al., 2020). However, no research to our knowledge has examined group stereotypes of rationality (including its more complex subcomponents and comparison to irrationality) in large-scale naturalistic text. Answering this question is critical given that

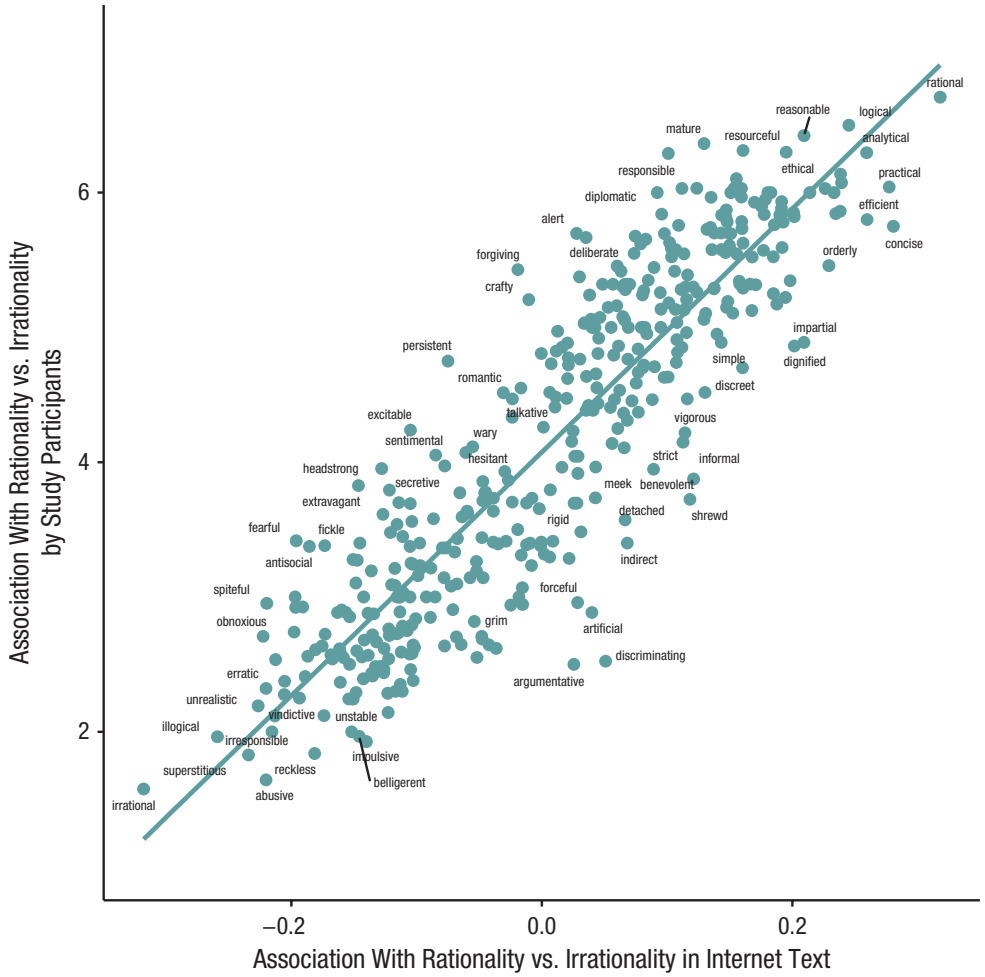

Fig. 1. Language associations and humans' ratings of a trait with rationality versus irrationality. The *y*-axis depicts the average rating from human participants for a trait's perceived rationality (vs. irrationality) on a scale from 1 (*strongly irrational*) to 7 (*strongly rational*). The *x*-axis depicts the relative association (average cosine similarity) between a trait and rationality (vs. irrationality) in contemporary internet text. The blue line and shaded gray area indicate the simple linear regression (and error) showing a significant and strong correlation between language associations of traits and humans' classifications of traits as rational versus irrational.

rationality is considered a valued trait (with some even calling it the essence of humanness; Kraut, 2022). Groups stereotyped as less rational may be less likely to be selected as social partners or leaders—and may even be more likely to be dehumanized (e.g., Kteily et al., 2015).

Relatedly, in our third and final analysis, we assessed why these representations of rationality matter for real-world, tangible outcomes. That is, we computed the stereotypes of 101 occupations not only on overall rationality but also its subcomponents and then linked those occupational stereotypes to earnings, gender gaps in earnings, and demographic representation (i.e., percentage women, Whites, and Blacks in each occupation). Prior related work has shown a correlation between the language stereotypes associating men with science and women with the arts and the demographic representation of women in science across countries (Lewis & Lupyan, 2020). Here, we extended this understanding to a new stereotype domain of rationality (rather than gender-science stereotypes), to variation across occupations (rather than across countries), and to more varied outcomes, including overall earnings and earnings inequalities.

Analysis 1: What is rationality and irrationality? To better understand the neighborhood of rationality and irrationality words, we performed a principal components analysis on the interrelationships among the top 50 traits associated with rationality (and, separately, the top

Table 2. Top 10 Traits Loading on Top Two Latent Principal Components of Rationality and Irrationality

	Top 10 traits	
	PC1	PC2
Rationality	"Interpersonal rationality": considerate, courteous, tactful, conscientious, diligent, resourceful, perceptive, knowledgeable, thoughtful, trustworthy	"Analytic rationality": understanding, insightful, objective, analytical, concise, critical, logical, perceptive, rational, practical
Irrationality	"Volatile irrationality": erratic, unpredictable, rash, unstable, irritable, compulsive, unreliable, temperamental, fickle, extravagant	"Unfair irrationality": unfair, unjust, unrealistic, illogical, irresponsible, extravagant, reckless, irrational, cruel, deceitful

Note: PC1 = Principal Component 1; PC2 = Principal Component 2.

50 traits associated with irrationality). Specifically, we first created a 50 × 300 matrix in which each of the top 50 traits had its corresponding 300-dimensional embedding. Then we computed the pairwise correlation matrix between all 50 traits and performed a principal components analysis on the corresponding correlation matrix using the prcomp() function in the R computing environment (Version 4.4.0; R Core Team, 2024), which, by default, uses singular value decomposition on the correlation matrix (this is mathematically equivalent to conducting a principal components analysis on the original 50 × 300 matrix).

The resulting scree plot and variance analysis suggested that a two-component solution both explained a substantial portion of variance and provided two interpretable subcomponents. Specifically, a two-component solution explained 41.61% of the variance in the rationality trait words and 43.90% of the variance in the irrationality trait words (for further descriptives of the principal components solution, see online R code). Although the third component explained approximately 14% of the additional variance, the traits that loaded highly on this component were not clearly interpretable with some united meaning (e.g., traits included "formal" and "dignified" but also "constructive" and "perceptive," suggesting more mixed content). We interpret the two-component solution as such below.

Decomposing rationality into subcomponents. The two principal components (PC1 and PC2) indicated diverging latent meanings (Table 2, Fig. 2). From qualitative inspection, rationality-PC1, which we term "interpersonal rationality," had the highest loadings on traits that reflected attributes of social reliability and trustworthiness (e.g., considerate, courteous, tactful, and conscientious). In contrast, rationality-PC2, which we term "analytic rationality," had the highest loadings on traits that appeared to reflect more typical attributes of cognitive intelligence and logic (e.g., insightful, objective, analytical, logical). Interestingly, we observed a marginal correlation of trait loadings on rationality-PC2 and valence: The more a trait loaded on rationality-PC2, the more negative that trait, perhaps reinforcing analytic rationality as cold and calculating.

It is worth pausing to consider how the two identified subdimensions of rationality are related to warmth and competence. To be clear, we are not claiming that rationality represents a third dimension of social cognition independent of warmth and competence (Abele et al., 2021). Instead, data indicate that the subdimensions of rationality are related to, but not reducible to, competence or warmth. Rationality-PC2 had high conceptual overlap with competence—and was indeed moderately correlated with it (r = .38; Table 3, Fig. 2f). However, we stress that the correlation and overlap were driven by a particular kind of cognitive competence (analytic and insightful) rather than a general competence construct that may also include attributes such as physical ability, strength, or agency (Fiske et al., 2002). Further, rationality-PC1 (interpersonal rationality) was, in fact, not meaningfully correlated with warmth (r = −.08; Table 3, Fig. 2c), although it was positively related to valence (r = .20; Table 3, Fig. 2a). Thus, the top loading traits of rationality-PC1 were specifically about how rational individuals are predictable and conscientious (more so than merely warm, friendly, or kind), underscoring that rationality-PC1 captured positive interpersonal attributes beyond general warmth.

This general pattern of two principal components—with one reflecting interpersonal rationality and the other reflecting analytical rationality—was generally consistent across our robustness checks, including using (a) exploratory factor analysis, (b) different data sources and embedding algorithms, and (c) different numbers of traits (e.g., 25 vs. 50, using only overlapping traits across embeddings; see the Supplemental Material). Although some approaches yielded less differentiated subcomponents (with each subcomponent containing both analytic and interpersonal traits), the key takeaway is that the concept of rationality in internet text consistently contained subcomponents of

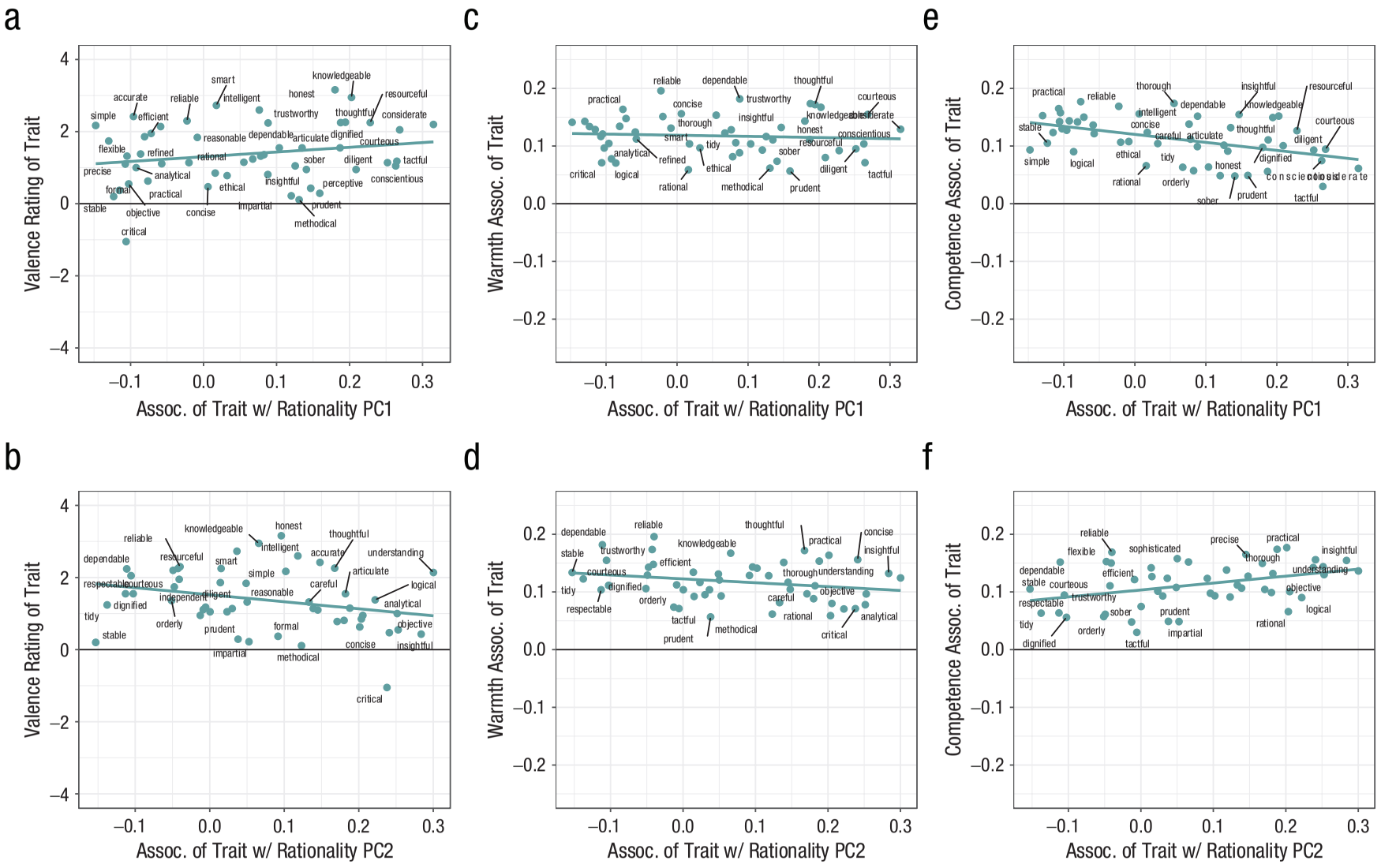

Fig. 2. Latent principal components of rationality and their relationship to valence, warmth, and competence. The *y*-axes depict the relative association between a trait and the latent principal components of rationality in contemporary internet text. The *x*-axes depict (a, b) the valence rating of the traits from human participants (Warriner et al., 2013) or the relative association of the traits with (c, d) warmth or (e, f) competence. The blue line and shaded gray area indicate the simple linear regression (and error) showing that PC1 reflects positive valence but negative competence and no association to warmth, whereas PC2 reflects negative valence and warmth but positive competence. PC1 = Principal Component 1; PC2 = Principal Component 2.

Table 3. Correlations of Trait Loadings on Principal Components With Trait Scores on Valence, Warmth, and Competence

	Correlations					
	PC 1 loadings			PC 2 loadings		
	Valence	Warmth	Competence	Valence	Warmth	Competence
Rationality	$r = .20$, 95% CI = [−.08, .46], $p = .16$	$r = -.08$, 95% CI = [−.35, .21], $p = .60$	$r = -.48$, 95% CI = [−.67, −.23], $p < .001$	$r = -.28$, 95% CI = [−.52, −.003], $p = .05$	$r = -.24$, 95% CI = [−.48, .05], $p = .10$	$r = .38$, 95% CI = [0.11, 0.60], $p = .006$
Irrationality	$r = .44$, 95% CI = [.19, .64], $p = .001$	$r = .82$, 95% CI = [.70, .89], $p < .001$	$r = .55$, 95% CI = [.32, .72], $p < .001$	$r = .10$, 95% CI = [−.18, .37], $p = .49$	$r = .08$, 95% CI = [−.20, .35], $p = .56$	$r = -.08$, 95% CI = [−.35, .21], $p = .60$

Note: A positive correlation indicates that, among the 50 traits that described rationality (or irrationality), the higher those 50 traits loaded on the principal component the more the traits were positive, warm, or competent (depending on the column). A negative correlation indicates that, among the 50 traits that described rationality (or irrationality), the higher those traits loaded on the principal component the more the traits were negative, cold, or incompetent. For example, although all rational traits were positive (see Fig. 2), the more they loaded on rationality-PC2 the more they were the relatively negative traits. Similarly, although all irrational traits were negative (see Fig. 2), the more they loaded on rationality-PC1 the more they were relatively positive and warm traits. PC1 = Principal Component 1; PC2 = Principal Component 2; CI = confidence interval.

interpersonal and analytic rationality. Moreover, each subcomponent contributed close to equal amounts of variance, suggesting that rationality in internet text is just as interpersonal as it is analytical. We discuss implications later, but for now note that this result aligns with a socially situated representation of rationality that goes beyond competence or the mere maximization of self-interest.

Decomposing irrationality into subcomponents. We reported above the large baseline differences in the positivity, warmth, and competence of irrationality (vs. rationality). However, the principal components analysis suggested that irrationality and its subcomponents were not simply the direct opposites of rationality and its subcomponents (see also a full discussion of antonym tests in the Supplemental Material). Rather, PC1-irrationality, which we term "volatile irrationality," appears to center on traits that reflect a volatile, erratic, or fickle personality that are not interpretable as the mere opposite of either the interpersonal or analytical dimensions of rationality.

Similarly, PC2-irrationality, which we term "unfair irrationality," appears to center on traits that refer more to immorality and cruelty, which, again, are not easily summarized as a mere opposite to rationality subcomponents. Interestingly, the correlations of these two PCs with word norms on valence, warmth, and competence (see Table 3) indicate that traits loading highly on PC1 (volatile irrationality) were relatively less negative and less incompetent than other irrationality traits. Thus, although all irrational traits were negative, the volatile irrationality traits were significantly less negative, especially compared with those that reflected more of the unfair or immoral content (e.g., "unkind" and "cruel," which loaded negatively on irrationality-PC1; Fig. 3). Like rationality, therefore, there are subcomponents of irrationality that can reflect important nuances in meaning: Irrationality is not always overwhelmingly negative, immoral, and unfair; instead, it can sometimes be more neutral in valence and simply capture unpredictable or erratic actions.

Analysis 2: Whom is rationality ascribed (or denied) to? The analyses so far provide two key conclusions. First, rationality contains two meaningful subcomponents, with approximately equal contributions of both interpersonal rationality and analytic rationality. Second, irrationality is not simply the opposite (or absence) of rationality but, rather, contains its own two subcomponents of volatile and unfair irrationality. Given the assumed importance of rationality for attributions of humanness (Kraut, 2022), we next considered how these more complex and nuanced representations of both rationality and irrationality are used for group stereotypes across advantaged and disadvantaged groups. In our second analysis we inspected the associations of rationality, irrationality, and their principal components across 66 social groups (33 stigmatized/low-power groups and 33 nonstigmatized/high-power groups). We use the terms "stigmatized" (or "low power") versus "nonstigmatized" (or "high power") following Link and Phelan (2001), who defined "stigma" as the co-occurrence of labeling, separation, status loss, and discrimination, but all critically in a context in which power is enacted. Thus, stigma/power are argued to be a broad construct that we can reasonably expect to also define how groups are ascribed valued characteristics such as rationality.

Group stereotypes of rational versus irrational overall. First, for the overall group rationality versus irrationality analyses, we used the Word Embeddings Association Test (WEAT; Caliskan et al., 2016) to compare the overall associations of 33 group contrasts (e.g., men/women, White/Black, rich/poor, young/old) with the top 10 rational versus irrational traits (the first 10 traits listed in Table 1). To represent group terms in text, we used word lists validated from previous research representing stigmatized/nonstigmatized social-group concepts in text (Caliskan et al., 2016; Charlesworth et al., 2023). Table S1 in the Supplemental Material lists all group terms.

Results showed that, overall, rationality (vs. irrationality) was more strongly associated with nonstigmatized/high-power groups (vs. stigmatized/low-power groups), with a mean WEAT effect size (d) of 1.25, 95% CI = [0.96, 1.53]. In fact, we found that this average effect was significantly different from zero across all 33 WEAT effect sizes, $t(32) = 9.01$, $p < .001$. For example, rationality was more strongly associated with rich (vs. poor; $d = 1.88$, $p < .001$), White (vs. Black; $d = 1.73$, $p < .001$), and men (vs. women; $d = 1.38$, $p < .001$). In fact, the nonstigmatized/high-power group (e.g., White, rich, abled) was associated with rationality for 31 of the 33 comparisons (except White vs. Aboriginal and tall vs. short, in which Aboriginal and short were more associated with rationality than White or tall, respectively; Fig. 4). To put these effect sizes in perspective, the overall magnitudes are similar to other studies of group-valence associations (e.g., Caliskan et al., 2016) as well as to our own tests (reported in the Supplemental Material) for group stereotypes on warmth ($d_{\text{warmth}} = 1.24$) and competence ($d_{\text{competence}} = 1.20$). Thus, rationality stereotypes provide meaningful and strong group associations, even when considered alongside well-studied stereotype dimensions.

Decomposing group stereotypes with rationality and irrationality separately. To further aid the interpretation of the overall stereotyping effects, we decomposed

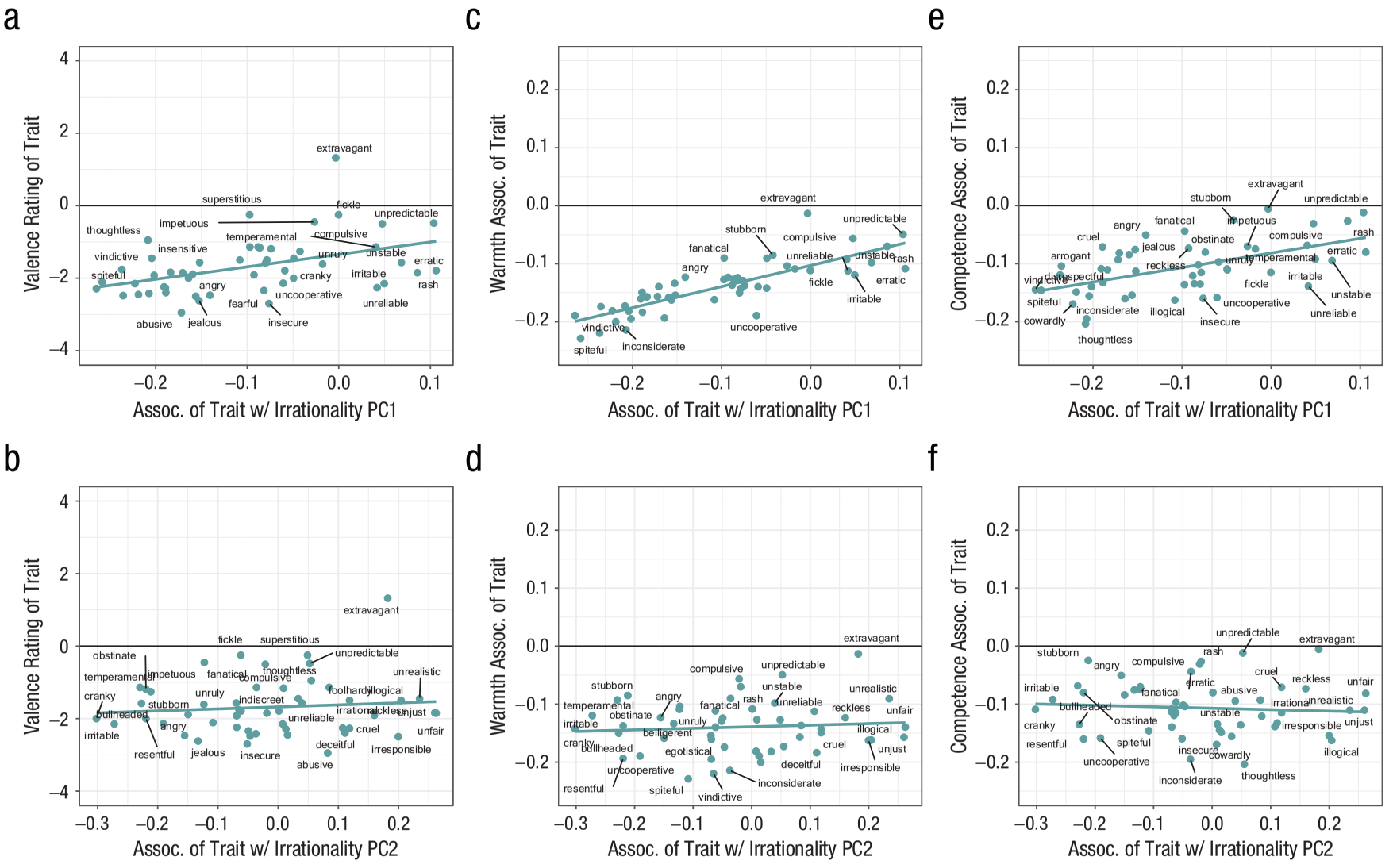

Fig. 3. Latent principal components of irrationality and their relationship to valence, warmth, and competence. The *y*-axes depict the relative association between a trait and the latent principal components of irrationality in contemporary internet text. The *x*-axes depict (a, c) the valence rating of the traits from human participants or (b, d) the relative association of the traits with competence. The blue line and shaded gray area indicate the simple linear regression (and error) showing that PC1 reflects less negative valence (noting that nearly all scores are below the zero line) and less incompetence, whereas PC2 does not seem to reflect latent dimensions of either valence or competence. PC1 = Principal Component 1; PC2 = Principal Component 2.

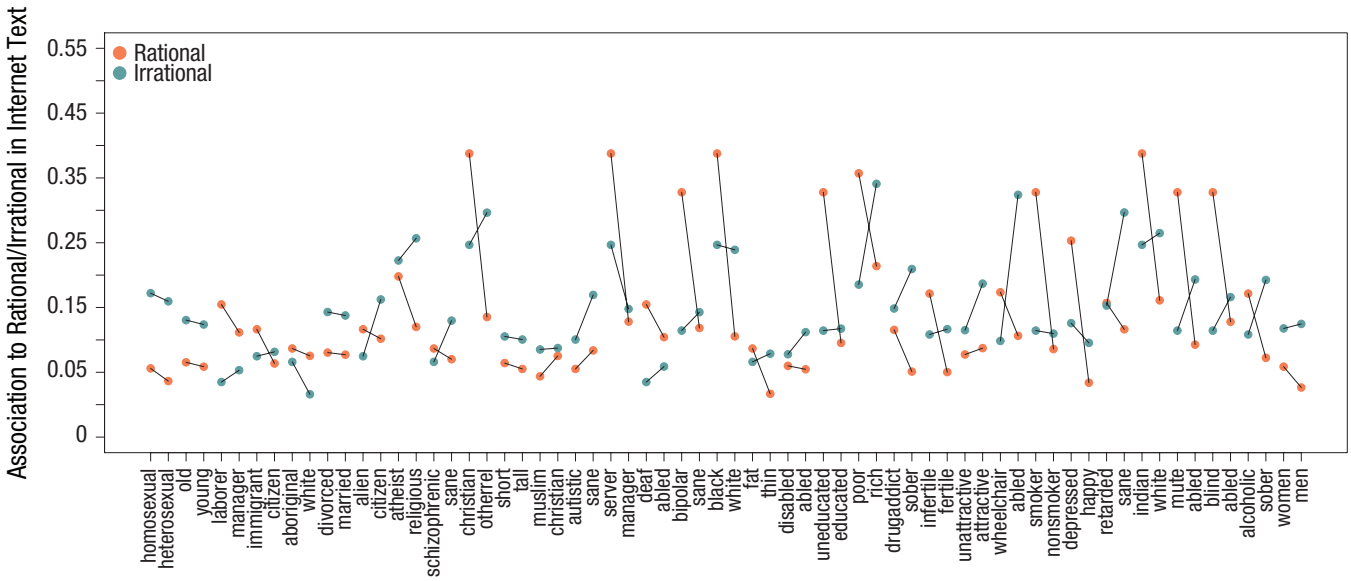

Fig. 4. Individual group-rationality and group-irrationality associations. The *y*-axis indicates the average cosine similarity (association) between the group terms and the rational (orange) or irrational (blue) words. The *x*-axis indicates the specific group under consideration. Black lines are used to highlight the contrasting groups (e.g., sober/alcoholic) reported in the overall WEAT effects. The larger black lines for rationality indicate that this dimension contributed more to group differences than the irrationality dimension. WEAT = Word Embeddings Association Test.

the overall associations to rationality/irrationality into the simplest associations of a single group (e.g., men) to a single concept (e.g., rationality). As depicted in Figure 4, we looked at the average cosine similarity between a set of group words (e.g., men) and the 10 rational traits and, separately, the average cosine similarity between the group words and the 10 irrational traits.

Across social groups, rationality (the orange dots in Fig. 4) was more strongly attributed to the dominant group (e.g., men, White, rich, young) than to the subordinate group (e.g., women, Black, poor, old), with generally large effect sizes (i.e., the slopes connecting the orange dots were often steep). In contrast, irrationality (the blue dots in Fig. 4) suggested less consistent group distinctions (i.e., the slopes connecting the blue dots were often flat). That is, across all 33 groups, the mean difference of cosine similarities on rationality, $|M_{\text{rational}}|$, was 0.10, whereas the mean difference of cosine similarities on irrationality, $|M_{\text{irrational}}|$, was 0.03. These mean difference effect sizes were significantly different from one another, $t(31) = 2.89$, $p = .007$, $d = 0.50$. In other words, rationality appears to carry more "weight" in group stereotypes than does irrationality. Although future research could dig deeper into the differences among social group pairs (e.g., irrationality was used more than rationality for stereotyping wheelchair use and intellectual disability), there was generally consistent evidence across most of the 33 group pairs.

Decomposing group stereotypes with rationality subcomponents. The analyses so far have shown that groups are (a) overall stereotyped but also (b) more stereotyped along rationality (vs. irrationality) subcomponents. Such results already lend new insights into rationality stereotypes but, as we know from our first analysis, rationality is not a unitary construct. Instead, rationality contains multiple subcomponents of meaning: analytic and interpersonal rationality. We thus considered how groups would be stereotyped along these subcomponents using a similar approach to the overall group-rationality analyses above but replacing the overall rationality words with the top 10 rational traits with the highest loadings on rationality-PC1 (and, in a separate analysis, the top 10 highest loading traits on rationality-PC2). Note that for all subsequent analyses, because we are no longer calculating double-difference scores (like the WEAT), we report the mean average cosine (MAC) similarities rather than the WEAT *d* scores.

Across all group contrasts, nonstigmatized/high-power groups were associated with both interpersonal ($M_{\text{interpersonal}} = 0.09$) and analytic ($M_{\text{analytic}} = 0.09$) rationality, with these two associations showing nearly identical magnitudes, $t(32) = 0.12$, $p = .90$. Indeed, group associations on the two components were highly correlated: Groups were systematically high (or low) on both interpersonal and analytic rationality, $r = .87$, 95% CI = [.74, .93], $t(31) = 9.65$, $p < .001$. We return to potential differences in the subcomponents of rationality in our final set of analyses below.

Analysis 3: Why do rationality stereotypes matter? Analysis 2 lends initial insight into the importance

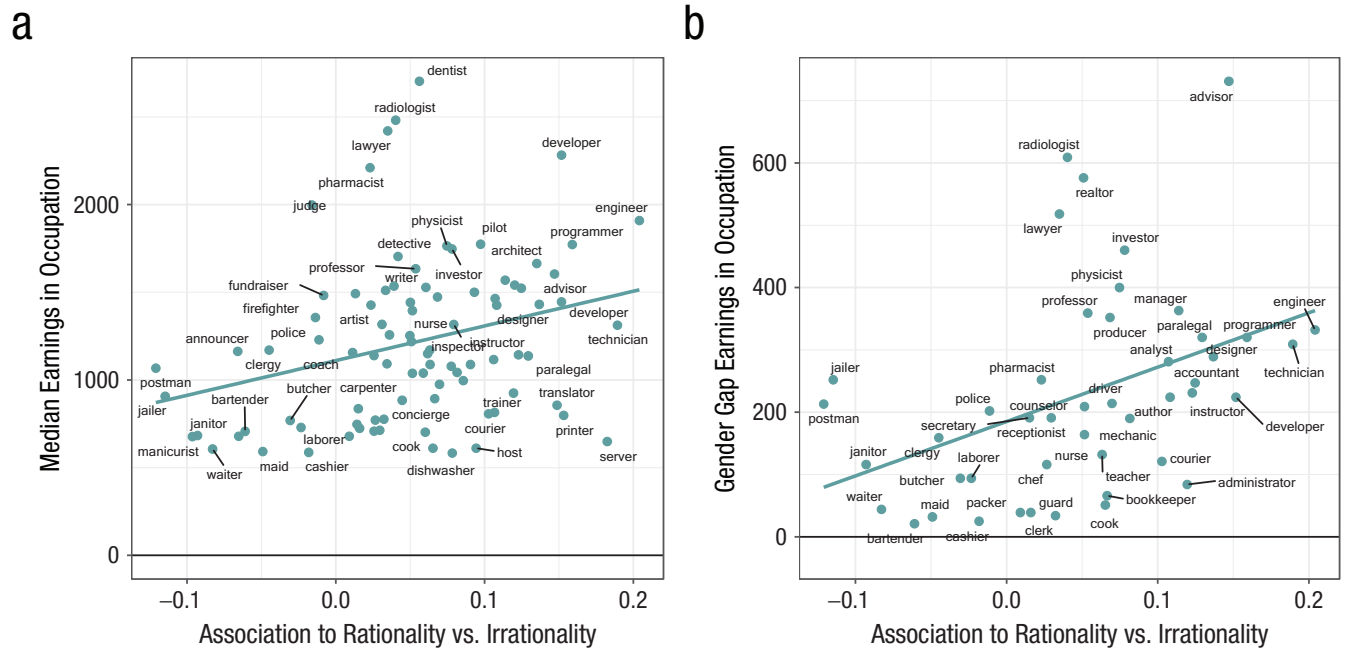

Fig. 5. Consequences for earnings across occupations. Occupations that are seen as more "rational" are those that have higher (a) earnings and larger (b) gender gaps in earnings (higher men advantages).

and stereotype consequences of rationality in language: It is a respected quality but is ascribed to high-power groups and denied to low-power groups. In Analysis 3, we extended our understanding of the consequences of rationality by also considering occupational stereotypes and the more tangible material outcomes of earnings, gender gaps in earnings, and demographic representations that are associated with an occupation being stereotyped as more (or less) rational. A list of 101 occupation labels (and the occupation workforce demographics median earnings overall and by gender) was created by taking the 2020 U.S. Bureau of Labor Statistics (BLS) report, translating all occupations into one-word labels (providing approximately 300 occupations), and then retaining all occupations that were available in the GloVe Common Crawl vocabulary (for similar approaches to modeling gender stereotypes of occupations, see also Charlesworth et al., 2021). Note that we had two cases with gendered occupation labels but only one BLS data entry: *waitress/waiter* and *actress/actor*. In these cases, we first calculated the language-based associations across the two individual gendered terms (i.e., "waitress" and "waiter" separately), but then, for the correlation analyses, we took the average across the individual labels and linked the average to the one BLS data entry. This provided a relatively conservative approach because, if anything, it should have led us to underestimate the correlations given that we were enforcing the gender–wage gap to be the same for the terms "waitress" and "waiter" despite the terms having possibly different representations in language.

Occupational stereotypes and correlates of rational versus irrational overall. To extract occupational stereotypes of rationality, we calculated the average cosine similarity between the occupation label (e.g., *carpenter*, *engineer*) and the top 10 rational words versus irrational words (from Analysis 1). Results of overall stereotypes showed clear face validity: The occupations stereotyped as most rational included *engineer*, *technician*, *programmer*, *developer*, and *advisor*, whereas the occupations stereotyped as least rational included *actress*, *bartender*, and *janitor* (see the Supplemental Material).

Next, we linked these occupational stereotypes of rationality with the key outcome data on earnings and representation. Results revealed that occupations stereotyped as more rational (vs. irrational) were associated with higher earnings in that occupation, $r = .29$, 95% CI = [.09, .47], $t(87) = 2.84$, $p = .006$, but also a larger gender gap in earnings (i.e., higher men advantages), $r = .41$, 95% CI = [.14, .63], $t(44) = 3.01$, $p = .004$ (Fig. 5). These results align with the understanding that rationality is a valued attribute and therefore that occupations stereotyped as more rational are compensated more highly; critically, however, those valued "rational" occupations also appear to be those for which masculine skills are particularly valued, giving rise to gender pay gaps. These correlations persisted at similar and

significant magnitudes even after controlling for occupational stereotypes along warmth, competence, and global valence (rs > .21, ps < .05; see the Supplemental Material). This robustness further reinforces that rationality is related to, but not redundant with, warmth/competence and can add explanatory value beyond those dimensions.

Correlations between occupational stereotypes of rationality and demographic representation within professions (e.g., percentage of employees who are women or non-White) were weaker and less consistent. Specifically, occupations stereotyped as more rational were those that have less Black people, $r = -.32$, 95% CI = [−.49, −.13], t (97) = −3.30, $p = .001$, as well as marginally more White people, $r = .19$, 95% CI = [−.006, .38], t (97) = 1.93, $p = .06$, and descriptively less women (although not significantly so), $r = -.13$, 95% CI = [−.32, .07], t (97) = −1.26, $p = .21$ (Fig. S7 in the Supplemental Material). At first glance, this may be surprising given that related work using language-based or survey-based stereotypes (i.e., of male-science or male-brilliant associations) has found significant relationships with gender representation (Leslie et al., 2015; Lewis & Lupyan, 2020). Perhaps these male-science stereotypes can be more concrete and observable (i.e., *describing* the observed absence of women in science), lending more robust correlations with representation. In contrast, rationality represents a more abstract and multifaceted stereotype that is thus more related to the relatively less observable outcomes (i.e., earnings). This interpretation encourages a deeper investigation of the nuances of rationality's subcomponents as correlates of both earnings and representation.

Decomposing occupation stereotypes and correlates with rationality and irrationality separately. As with the social-groups analysis above, we next assessed whether rationality or irrationality carried greater "weight" in defining the occupational stereotypes. Replicating the social-group results, we again found a greater magnitude of association to rationality, $|M_{\text{rational}}| = 0.14$ versus $|M_{\text{irrational}}| = 0.09$, that was significantly different across the occupations, $t(98) = 7.30$, $p < .001$, $d = 0.73$. Moreover, correlations to the key outcomes (earnings and representation) were entirely driven by occupation stereotypes on rationality, with null correlations to occupation stereotypes on irrationality. For instance, the correlation of median earnings to occupation-rationality stereotypes was $r = .39$, 95% CI = [.20, .56], $t(87) = 3.99$, $p < .001$, whereas the correlation of median earnings to occupation-irrationality stereotypes was $r = .07$, 95% CI = [−.14, .27], $t(87) = 0.61$, $p = .54$. Similarly, the correlation for gender gaps in earnings was moderate and significant for occupation-rationality associations, $r = .38$, 95% CI = [.10, .60], $t(44) = 2.71$, $p = .009$, but null (and even trending negative) for occupation-irrationality stereotypes, $r = -.21$, 95% CI = [−.48, .08], $t(44) = -1.46$, $p = .15$. As above, we found weaker and mostly null correlations for all representation outcomes for both rational and irrational stereotypes, $|r| < .24$, ps > .02. In sum, the ascription (or denial) of rationality to an occupation seems to be more meaningful than the ascription (or denial) of irrationality. For our final analysis, we therefore dug deeper into the related subcomponents of rationality (analytic and interpersonal) to understand whether the type of rationality stereotype mattered for occupational outcomes.

Decomposing occupation stereotypes and correlates with rationality subcomponents. In general, occupations were associated more strongly with interpersonal rationality, $|M_{interpersonal}| = 0.17$, than to analytic rationality, $|M_{analytic}| = 0.13$, $t(98) = 6.39$, $p < .001$, $d = 0.64$. Despite this overall difference, however, it was analytic (not interpersonal) rationality that was more strongly correlated with earnings. Indeed, there was a significant interaction between the strength of stereotype and the principal component of the stereotype, $b = -3{,}451$, $SE = 1{,}190$, $p = .0042$ (Fig. 6a). Breaking down this interaction we see that the correlation of analytic rationality and earnings was $r = .49$, 95% CI = [.31, .63], $t(87) = 5.18$, $p < .001$, whereas the correlation of interpersonal rationality and earnings was nonsignificant, $r = .04$, 95% CI = [−.17, .24], $t(87) = 0.36$, $p = .72$.

Similar results emerged for gender pay gaps, with a marginal interaction, $b = -998$, $SE = 579$, $p = .088$, reflecting a difference between the moderate, significant correlation for analytic rationality and pay gaps, $r = .45$, 95% CI = [.18, .65], $t(44) = 3.30$, $p = .002$, but nonsignificant correlation for interpersonal rationality and pay gaps, $r = .007$, 95% CI = [−.28, .30], $t(44) = 0.04$, $p = .97$. In sum, across both pay indicators, analytic rationality appears to be the subcomponent driving the earnings (especially the male earning advantage) of those "rational" occupations. Interpersonal rationality, despite being the more positive subcomponent of rationality, appears to give little advantage in occupational outcomes.

No significant interactions were found for the representation of women, Blacks, or Whites in the occupation (ps > .10). However, the descriptive pattern of the interaction for women representation was notable: If anything, the results for this outcome suggest that interpersonal rationality was a better predictor of the proportion of women representation (Fig. 6c), suggesting that either analytic or interpersonal rationality could be meaningful in real-world outcomes depending on which outcomes are being considered. For now, these results should be interpreted as preliminary.

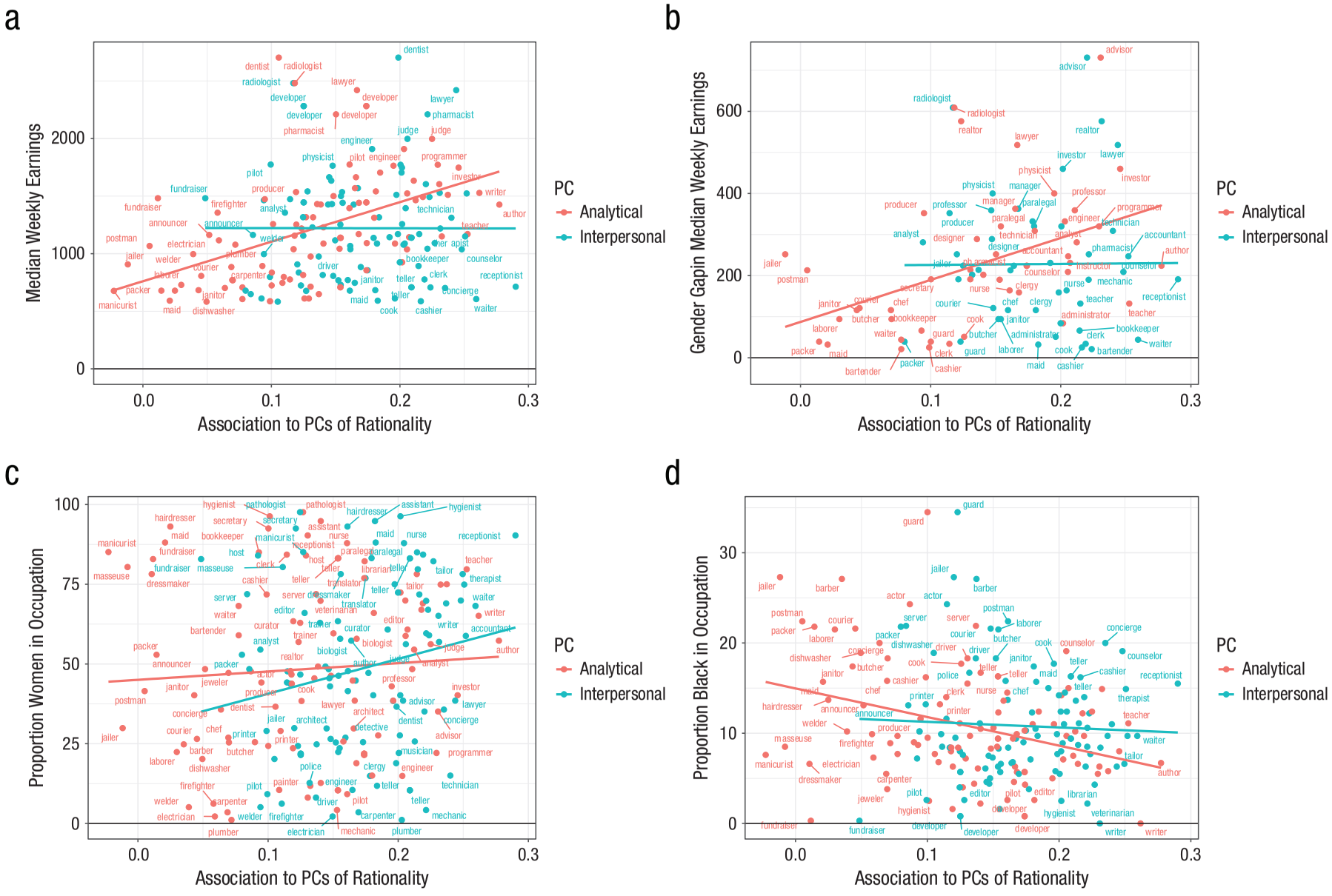

Fig. 6. Differences in consequences for earnings and representations across two rationality PCs. Although interpersonal rationality was generally more associated with most occupations, the analytical rationality (red lines) was a stronger predictor of (a) earnings and (b) gender gaps in earnings, both with significant interactions. It was also a descriptively better predictor of (d) the proportion of Black representations. Interestingly, however, interpersonal rationality was a descriptively better predictor of (c) the proportion of women representation across occupations, indicating that the relevance of the two PCs indeed depends on the outcomes under consideration. PC = principal component.

Summary of results

Altogether our analyses paint a nuanced picture regarding rationality, its subcomponents, group stereotypes, and occupational correlations. Two results underscore the importance of rationality overall: its associations with high- versus low-power groups and its associations with earnings and gender pay gaps in earnings. Intriguingly, when decomposing the differences between rationality and irrationality, rationality appears to be stereotyped more. But our results also underscore the importance of decomposing rationality into subcomponents. Although both analytic and interpersonal rationality were approximately equally associated with high- versus low-power groups, they were differentially associated with occupational outcomes. That is, analytic (rather than interpersonal) rationality was a stronger predictor of earnings and gender pay gaps in earnings, suggesting that, perhaps counterintuitively, it is the slightly more negative, cold, and calculating subdimension that may be particularly valued and rewarded in high-earning occupations.

Discussion

Although there is broad consensus that rationality is valuable, consensus over the construct of rationality has remained vexing (Cohen, 1981; Shafir & LeBoeuf, 2002; Stanovich & West, 2000; Tetlock & Mellers, 2002). Moreover, research has largely left untouched questions regarding which social groups rationality is ascribed (vs. denied) to and whether rationality has important societal correlates.

We used word embeddings trained on massive internet text to offer novel insights into the collective representations of rationality in society. By taking a bottom-up, exploratory methodology, this approach complements existing debates over how rationality *ought* to be represented, instead exploring how rationality *is* represented in society—along with which groups it is ascribed to and whether it is associated with occupational outcomes.

Nuances in the concept of rationality: analytic and interpersonal

Results revealed important complexities and nuances in representations of rationality. Rationality was associated with a subcomponent centered on analytic abilities, in line with models of rationality in neoclassical economics. This finding is consistent with that of Grossmann et al. (2020), who revealed that folk standards of rationality (vs. reasonableness) include associations with traits such as logical, smart, and intelligent. It is also consistent with the social cognition dimension of competence (Fiske et al., 2002).

More surprisingly, and moving beyond being synonymous with competence, rationality was associated—often in equal measure—with an *interpersonal* dimension centered on trust and conscientiousness. This subcomponent contradicts the portrait of a rational actor as predominantly a cold and calculating maximizer of self-interest (Becker, 1962). Instead, it is consistent with models of rationality that emphasize the inherent cooperation necessary for human success, sometimes at the expense of immediate self-interest (Henrich et al., 2001; Kreps et al., 1982; Miller & Ratner, 1998; Rand et al., 2012). Our work is most consistent with a nuanced and multidisciplinary perspective on rationality that paints humans as motivated by social pressures (Tetlock, 2002), accountability (Lerner & Tetlock, 1999), and a deep-seated need to belong (Baumeister & Leary, 1995). Our descriptive work converges with social-functionalist approaches to human judgment and choice that sometimes make the normative claim that decisions that appear irrational from a narrow cognitive lens can be interpreted as rational once the broader social goals of a decision maker are taken into account (Dorison & Heller, 2022; Lerner & Tetlock, 1999; Tetlock, 2002; see also Frank, 1987; Frank et al., 1993; Page, 2022).

One potential explanation for the existence of this interpersonal dimension of rationality is the fact that rationality is semantically linked with related social traits such as reasonable or moral (Grossmann et al., 2020). Recall that our goal here was to provide a comprehensive descriptive analysis and identify the broader semantic meaning of rationality and its subcomponents. Thus, rather than interpreting this result as rationality being *confounded* with semantically linked traits (such as reasonableness), we interpret this result as revealing the complex and multifaceted collective representation of rationality as it is used in everyday language.

Although most theorists focus on rationality, our work revealed distinct subcomponents of irrationality that were more than the mere absence or opposite of rationality (volatility and unfairness). Arguably, volatility can be seen as an opposite to the traditional rational attributes of consistency and coherence (but see Arkes et al., 2016). However, unfairness appears as a relatively surprising subcomponent of irrationality, connecting perhaps more closely to constructs such as immorality, cruelty, and deceit (Jackson et al., 2023). As with rationality, irrationality too may have an interpersonal undercurrent. Future work could explore intriguing cases of those who are "irrationally consistent" or "irrationally fair" (White et al., 2024).

Social-group stereotypes of rationality and irrationality

In addition to understanding what rationality is, our work also contributes to understanding who rationality

is ascribed to. There is a long history in social psychology of understanding which dimensions matter in group stereotypes, such as the relevance of warmth/competence (Fiske et al., 2002) or foreignness/inferiority (Zou & Cheryan, 2017). We add to this literature by showing the different relevance of subcomponents in shaping group stereotypes. The finding that rationality was generally used more in stereotyping compared with irrationality (i.e., showed greater group differences) aligns with work showing that greater group distinctions are often made in terms of the relative positivity of groups (e.g., Bergsieker et al., 2012). Thus, in contrast to conceptualizing group stereotyping as centered on antipathy (Allport, 1954), which may have been true in the past, studying group stereotyping today may fruitfully focus more on understanding which groups are relatively associated with positive or valued attributes.

Occupational correlates of rationality and irrationality

A third and final contribution of the current work is underscoring the importance of decomposing rationality into its subcomponents when considering societal correlates. Related work has linked language-based stereotypes to gender representation in occupations (Leslie et al., 2015; Lewis & Lupyan, 2020). We extend this framework to a new stereotype domain of rationality (rather than gender-science stereotypes), to variation across occupations (rather than across countries), and to more varied outcomes, including overall earnings and earnings inequalities. Although the current work revealed that rationality (especially its analytic subcomponent) consistently predicted earnings and gender pay gaps, we found less consistent associations with demographic representation. Our results reinforce the need for future work to examine how language-based stereotypes are linked to a wider set of value-laden outcomes beyond mere representation (e.g., workplace awards).

Limitations and conclusions

To be clear, the current research was not set up to address the critical question of how rationality should be defined or operationalized by scholars. Instead, we aimed to contribute to rationality debates by exploring how rationality (and irrationality) is collectively represented in human language. Inferences from our work are descriptive, not normative.

Other limitations also constrain our inferences. Perhaps most important, the current text corpora were limited to English and largely Western contemporary contexts. This limits the generalizability of the findings; future work is needed to generalize across time, cultures, and languages in which different conceptualizations of rationality may be uncovered. The monolingual analyses in the current work could (and should) be extended to multilingual word embeddings (e.g., Wirsching et al., 2025). In fact, one reason we chose to use relatively simple and flexible static embeddings (such as GloVe) was to enable future investigations across new language settings (e.g., historical texts). Similarly, our approach could be extended to explore how different academic disciplines (e.g., psychology vs. economics vs. philosophy) interpret rationality—and even whether research articles within a discipline may discuss rationality differently depending on the research focus (e.g., computer-science articles focusing on algorithmic fairness vs. algorithmic performance). Relatedly, whereas the present work included 66 social groups that were clearly divided on the basis of stigma and power, this meant omitting highly discussed and polarized groups such as liberals versus conservatives. Understanding rationality stereotypes across the political spectrum (and using text produced by different political groups) could help reveal who is stereotyping whom, including identifying asymmetries in whether some groups are particularly likely to use certain types of rationality stereotypes (e.g., liberals may value and use analytic rationality, whereas conservatives may value and use interpersonal rationality, or vice versa). Finally, the current work was not preregistered, a limitation that could be overcome in future confirmatory work.

Given the long-standing scholarly tradition of debating the normative meaning of rationality, it is perhaps surprising that relatively little attention has been paid to how laypeople use the word. By taking a descriptive approach—analyzing more than 840 billion words of internet text—we hope that the current work contributes productively to contemporary debates about what rationality is, who it is ascribed to, and why it matters.

Transparency

Action Editor: Vishnu Sreekumar
Editor: Simine Vazire
Author Contributions
Charles A. Dorison and Tessa E. S. Charlesworth contributed equally to this work.
Charles A. Dorison: Conceptualization; Data curation; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing – original draft; Writing – review & editing.
Tessa E. S. Charlesworth: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Validation; Visualization; Writing – original draft; Writing – review & editing.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
There was no funding for this research.

Artificial intelligence

We used the generated outputs of ChatGPT as though it were a human rater or research assistant and compared its outputs with those obtained from our human participants and from our static embedding models. No other AI-assisted technologies were used in this research or in the creation of this article.

Ethics

This research received approval from the Georgetown University ethics board (Study 00006750).

Open Practices

Preregistration: The natural language processing (NLP) analyses of secondary data were not preregistered. However, the hypotheses, method, and analysis plan of the validation study in which we compared NLP results with human raters' results was preregistered (https://aspredicted.org/2pbc-f5bp.pdf). This preregistration was submitted prior to data collection, and there were no deviations from the preregistration for this study. Materials: There were no relevant materials for the secondary data analysis. For the validation study, we provide all materials (i.e., instructions and rating task given to participants). Materials are available at https://osf.io/cgjne under the folder "Additional study materials." Data: All primary data are publicly available (https://osf.io/cgjne). Analysis scripts: All analysis scripts are publicly available (https://osf.io/cgjne). Open Science Framework (OSF): To ensure long-term preservation, all OSF files were registered at https://doi.org/10.17605/osf.io/yktud. Computational reproducibility: The computational reproducibility of the results in the main article (but not the Supplemental Material available online) has been independently confirmed by the journal's STAR team.

ORCID iD

Charles A. Dorison https://orcid.org/0000-0002-7072-2530

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976251362120

References

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*, *128*(2), 290–314.

Ali, M. A., Sun, Y., Zhou, X., Wang, W., & Zhao, X. (2019). Antonym-synonym classification based on new subspace embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 6204–6211.

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*(1), 20–39.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529.

Becker, G. S. (1962). Irrational behavior and economic theory. *Journal of Political Economy*, *70*(1), 1–13.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery.

Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, *102*(6), 1214–1238.

Bermúdez, J. L. (2022). Rational framing effects: A multidisciplinary case. *Behavioral and Brain Sciences*, *45*, Article e220. https://doi.org/10.1017/S0140525X2200005X

Bhatia, S., Richie, R, & Zou, W. (2019). Distributed semantic representations for modelling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36.

Bhatia, S., & Walasek, L. (2023). Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences, USA*, *120*(25), Article e2220726120. https://doi.org/10.1073/pnas.2220726120

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*. arXiv. https://doi.org/10.48550/arXiv.1607.06520

Caliskan, A., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Charlesworth, T. E. S., Morehouse, K., Rouduri, V., & Cunningham, W. A. (2024). Echoes of Culture: Relationships of implicit and explicit attitudes with contemporary English, historical English, and 53 non-English languages. *Social Psychological and Personality Science*, *15*(7), 812–823. https://doi.org/10.1177/19485506241256400

Charlesworth, T. E. S., Sanjeev, N., Hatzenbuehler, M. L., & Banaji, M. R. (2023). Identifying and predicting stereotype change in large language corpora: 72 groups, 115 years (1900–2015), and four text sources. *Journal of Personality and Social Psychology*, *125*(5), 969–990. https://doi.org/10.1037/pspa0000354

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*(3), 317–331.

Cusimano, C. (2025). The case for heterogeneity in metacognitive appraisals of biased beliefs. *Personality and Social Psychology Review*, *29*(2), 188–212. https://doi.org/10.1177/10888683241251520

Dorison, C. A., & Heller, B. H. (2022). Observers penalize decision makers whose risk preferences are unaffected by loss–gain framing. *Journal of Experimental Psychology: General*, *151*(9), 2043–2059.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Fiske, S. T., Cuddy, A. J., 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.

Frank, R. H. (1987). If homo economicus could choose his own utility function, would he want one with a conscience? *American Economic Review*, *77*(4), 593–604.

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. W. W. Norton & Company.

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives*, *7*(2), 159–171.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Grossmann, I., Eibach, R. P., Koyama, J., & Sahi, Q. B. (2020). Folk standards of sound judgment: Rationality versus reasonableness. *Science Advances*, *6*(2), Article eaaz0289. https://doi.org/10.1126/sciadv.aaz0289

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108–1109.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Jackson, J. C., Halberstadt, J., Takezawa, M., Liew, K., Smith, K., Apicella, C., & Gray, K. (2023). Generalized morality culturally evolves as an adaptive heuristic in large social networks. *Journal of Personality and Social Psychology*, *125*(6), 1207–1238. https://doi.org/10.1037/pspa0000358

Knauff, M., & Spohn, W. (Eds.) (2021). *The handbook of rationality*. MIT Press.

Kraut, R. (2022). Aristotle's ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Stanford University. https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, *27*(2), 245–252.

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *Journal of Personality and Social Psychology*, *109*(5), 901–931.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology*, *27*(1), 363–385.

Miller, D. T., & Ratner, R. K. (1998). The disparity between the actual and assumed power of self-interest. *Journal of Personality and Social Psychology*, *74*(1), 53–62.

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, *51*(1), 178–196.

Page, L. (2022). *Optimally irrational: The good reasons we behave the way we do*. Cambridge University Press.

Peabody, D. (1987). Selecting representative trait adjectives. *Journal of Personality and Social Psychology*, *52*(1), 59–71.

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Association for Computational Linguistics.

Pinker, S. (2021). *Rationality: What it is, why it seems scarce, why it matters*. Penguin.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*(1), 491–517.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(5), 701–717.

Storage, D., Charlesworth, T. E., Banaji, M. R., & Cimpian, A. (2020). Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology*, *90*, Article 104020. https://doi.org/10.1016/j.jesp.2020.104020

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471.

Tetlock, P. E., & Mellers, B. A. (2002). The great rationality debate. *Psychological Science*, *13*(1), 94–99.

Tversky, A., & Kahneman, D. (1996). On the reality of cognitive illusions. *Psychological Review*, *103*(3), 582–591.

U.S. Bureau of Labor Statistics. (2020). Employment and wages, annual averages. Retrieved from http://www.bls.gov

Viale, R. (2021). Why bounded rationality? In R. Viale (Ed.), *Routledge handbook of bounded rationality* (pp. 1–54). Routledge.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

White, M. W., Levine, E. E., & Kristal, A. C. (2024). Are rules meant to be broken? When and why consistent rule-following undermines versus enhances trust. *Journal of Experimental Social Psychology*, *111*, Article 104552. https://doi.org/10.1016/j.jesp.2023.104552

Wirsching, E. M., Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2025). Multilanguage word embeddings for social scientists: Estimation, inference, and validation resources for 157 Languages. *Political Analysis*, *33*(2), 156–163. doi:10.1017/pan.2024.17

Zou, L. X., & Cheryan, S. (2017). Two axes of subordination: A new model of racial position. *Journal of Personality and Social Psychology*, *112*(5), 696–717.