First impressions: The relative importance of what we say, how we say it, and how we act

Louis Hickman[1,2], Angela Duckworth[2,3], Cade Massey[2], Lyle Ungar[2,4], & Mandi Nerenberg[3]

[1]Department of Psychology, Virginia Tech

[2]The Wharton School, University of Pennsylvania

[3]Department of Psychology, University of Pennsylvania

[4]Computer and Information Science, University of Pennsylvania

First impressions: The relative importance of what we say, how we say it, and how we act

Abstract: Within moments, we form judgments about the people we meet. Yet surprisingly little is known about how dynamic verbal (what we say), paraverbal (how we say it), and nonverbal behaviors (how we act) shape first impressions. Observers (N = 1,413) rated the warmth, competence, and morality of strangers (N = 528) introducing themselves in 90-second videos. Separately, trained human raters and artificial intelligence (AI) algorithms scored various verbal, paraverbal, and nonverbal behaviors to examine their relative importance in impression formation. Observers agreed with each other most about first impressions of warmth and least about morality. With remarkably similar estimates across human and AI methods, warmth was conveyed primarily by nonverbal (and some verbal and paraverbal) behaviors, whereas both competence and moral character were conveyed by verbal and, to a lesser degree, paraverbal behaviors. What you say, how you say it, and how you act all matter—but in different ways—for conveying positive impressions.

Keywords: thin slice; social cognition; communication; voice; machine learning; person perception; open materials

Statement of relevance: First impressions form quickly and without conscious deliberation, and they bias subsequent judgments and behavior. Past research on first impressions has primarily relied on static images (i.e., photographs). To date, the nascent literature on dynamic cues and first impressions has examined verbal, paraverbal, and nonverbal behaviors in isolation, prohibiting a direct comparison of their relative contributions. Our study simultaneously investigated the relative contributions of all three types of behavior to impressions of warmth, competence, and moral character—the fundamental dimensions of social cognition. Both human and artificial intelligence approaches revealed that warmth is conveyed primarily by how we act (i.e., nonverbal behaviors) whereas competence and moral character are conveyed by what we say (i.e., verbal behaviors) and, to a lesser extent, how we say it (i.e., paraverbal behaviors).

First impressions: The relative importance of what we say, how we say it, and how we act

Within moments, we form judgments about the people we meet. First impressions emerge spontaneously and without conscious deliberation (Todorov et al., 2015). Regardless of accuracy, first impressions bias our subsequent judgments and behavior (Kahneman, 2011; Rule & Ambady, 2010). For example, hiring managers' first impressions of applicants in employment interviews predict final ratings and the likelihood of providing a job offer—even when controlling for the candidates' qualifications (Barrick et al., 2010; Barrick et al., 2012).

Research on first impressions has primarily explored the influence of static cues like attractiveness (Fiske et al., 2007). The frequent omission of dynamic behavioral cues from these studies is unfortunate because the influence of static cues decreases rapidly when behavioral information is available (Borkenau & Liebler, 1995; Kenny, 2004), limiting ecological validity. In particular, when we observe another person in real-time, we have the opportunity to observe their verbal (what they say), paraverbal (how they say it), and nonverbal behavior (what they do) (Ambady & Rosenthal, 1992; Brunswik, 1956). Here we investigate how dynamic verbal, paraverbal, and nonverbal behaviors predict first impressions of warmth, competence, and morality.

The nascent literature on dynamic cues and impression formation has not yet compared the relative contributions of verbal (e.g., analytical vocabulary, having a logical flow), paraverbal (e.g., pitch, speech rate), and nonverbal behaviors (i.e., facial expressions, body language). Instead, a handful of studies have focused on one type of behavior in isolation from the other two (Hall et al., 2019; Weisbuch et al., 2010). For instance, the most recent volume on first impressions (Ambady & Skowronski, 2008) included four chapters on facial cues (e.g., facial expressions, attractiveness) but none on verbal or paraverbal behavior, which necessitate

dynamic (rather than static) information, and a recent review of warmth and competence impressions in organizations emphasized nonverbal behaviors without mentioning verbal or paraverbal behaviors (Cuddy et al., 2011). One notable exception is an oft-quoted, highly cited claim that variation in liking and dominance perceptions are explained mostly by nonverbal behaviors (55%), followed by paraverbal behaviors (38%), and only minimally by verbal behaviors (7%) (Mehrabian, 1971). However, in the studies supporting this claim (e.g., Mehrabian & Wiener, 1967), speakers uttered just one word—a rare occurrence in real-world interactions. The current investigation heeds the call for investigating how multiple behaviors simultaneously influence impression formation (Hall et al., 2019).

Likewise, little is known about how verbal, paraverbal, and nonverbal behaviors relate to well-established dimensions of social cognition—warmth (i.e., friendliness, sociability) and competence (i.e., intelligence, skill; Fiske et al., 2007)—as well as a third, more recently identified dimension, moral character (i.e., trustworthy, principled; Goodwin, 2015). Some research suggests that dynamic nonverbal behaviors—including smiles and nods—relate to perceived warmth (Biancardi et al., 2017; Breil et al., 2021), whereas verbal and paraverbal behaviors—including speaking faster, speaking more, and using longer words—relate to perceived competence (Brown, 1980; Hickman et al., 2022; Schroeder & Epley, 2015). Given its more recent identification, moral character is almost entirely unexplored in research on dynamic behavior and first impressions. Claims that morality is distinct would be bolstered if its behavioral correlates differ from warmth.

One difficulty in this area of study: historically, research assistants coding videos for behaviors necessarily focused on a small set of features expected by researchers to be useful. This represents a logistical constraint (the time, effort, and cost of coding videos limits feasible

sample sizes), a conceptual limitation (*a priori* determination of important behaviors may miss key behaviors), and an external validity concern (findings may not generalize beyond the operationalizations of behavior employed in any single study; Bracht & Glass, 1968; Yarkoni, 2022). Therefore, in addition to this human approach (i.e., human raters and hierarchical regression), we also examined the relative importance of the three types of behavior for first impression formation using an artificial intelligence (AI) approach (i.e., computer-measured behavior and machine learning).

In this investigation, we compared how verbal, paraverbal, and nonverbal behaviors predict first impressions of warmth, competence, and moral character from 90-second, video self-introductions. First, 1,413 raters watched 528 individuals introduce themselves and provided 4,841 ratings of speaker warmth, competence, and moral character. Next, we used both human coders and AI to operationalize verbal, paraverbal, and nonverbal behaviors. We then used both explanatory and predictive (Yarkoni & Westfall, 2017) approaches to estimate the relative importance of these three types of behavior for first impressions. Because our goal was to understand the variance explained by these three qualitatively distinct modalities of behavior, rather than use data reduction techniques (e.g., factor analysis) to combine variables in each modality, we included specific behaviors as sets of verbal, paraverbal, and nonverbal predictors.

Open Practices Statement

The study reported in this article was not preregistered. The data have not been made available on a permanent third-party archive due to the identifiable, video-nature of the data—requests for the data extracted from the videos can be sent to the corresponding author. Our complete analysis scripts and study materials have been posted at https://osf.io/ezv4u/?view_only=a39a9e621a174455b809333e84112a13.

Method

Participants: Speakers

We recruited 528 participants through three channels: MBA students who participated for course credit ($N = 197$), undergraduate psychology students who participated for course credit ($N = 234$), and members of the Prolific platform who participated for a payment of $2 ($N = 97$). This sample size is nearly five times larger than the average sample size of studies in a recent meta-analysis on impression formation (Breil et al., 2021). In all cases, participants were U.S.-based and completed the same procedure. Specifically, they were informed that the best way to become a good public speaker is through practice and instructed to take time to prepare their response to an introductory prompt before recording it. The prompt was, "Please tell us about yourself (as you might be asked during a first introduction in a professional setting)." After recording their video, they could review their video and rerecord it as many times as they liked. They then self-reported their demographics.

Participants: Raters of Warmth, Competence, and Morality

Both samples of student speakers and additional Prolific participants watched other participants' videos and rated their warmth, competence, and morality on three-item adjective scales. The Online Supplement provides additional details about these raters. The final sample consisted of 1,413 raters' 4,641 ratings on 528 participants' videos. The ratings were averaged together within each speaker.

Measures

Here, we describe both the traditional survey measures and the AI approaches used to measure verbal, paraverbal, and nonverbal behaviors.

Warmth, Competence, and Morality

The warmth, competence, and morality survey items were preceded by the stem, "The speaker in this video is …" The three warmth items were "approachable," "warm," and "sociable." The three competence items were "talented," "skilled," and "capable." The three morality items were "trustworthy," "sincere," and "principled." These items were drawn from Landy et al. (2016). The Online Supplement reports additional details about our process of developing and validating these three scales. Responses were made on a 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree."

Public Speaking Behaviors

We developed a set of 14 public speaking behaviors by consulting with communications professionals at a large business school, reviewing several public speaking rubrics, and iteratively developing and revising the scale items and anchors. We recruited, conducted frame-of-reference training for, and paid nine people (who did *not* overlap with those who rated warmth, competence, and morality) such that at least two raters scored each of the 14 public speaking behaviors in each video (additional details provided in the Online Supplement).

The 14 behaviors were rated on a 4-point scale, with differing response options depending on the focal behavior. The OSF repository includes the rubric and response options. In all cases, higher scores indicate behavior expected *a priori* to indicate effective public speaking. We grouped these public speaking behaviors into verbal, paraverbal, and nonverbal behaviors. Any behaviors with ICC(1, k) < .50 were dropped, following Breil et al. (2022). Results were consistent with or without those behaviors—see the Online Supplement for additional details.

Verbal Behaviors. The verbal behaviors evaluated by trained raters included logical flow ("ideas flow logically with clear transitions between main ideas"), conclusion ("conclusion

reinforces key message"), relevance ("presentation is interesting and relevant to the audience"), and avoids fillers ("uses few non-words/fillers").

Paraverbal Behaviors. The paraverbal behaviors evaluated by trained raters included prosody ("uses vocal pace, volume, and pitch for emphasis and engagement") and pauses ("incorporates pauses to enhance impact and clarity").

Nonverbal Behaviors. The nonverbal behaviors evaluated by trained raters included eye contact ("establishes eye contact"), avoiding notes ("uses notes appropriately (minimal reliance on notes)"), facial expressions ("uses animated facial expressions"), gestures ("uses hand gestures"), and energy ("appears energetic").

Computer-Measured Behaviors

In addition to the human ratings of public speaking behaviors, we used computer software and AI to measure verbal, paraverbal, and nonverbal behaviors. The OSF repository reports the full set of computer-measured behaviors used in the machine learning (ML) models.

Verbal Behavior. To extract verbal behavior from the videos, we first transcribed them with Amazon's Transcribe. Amazon's Transcribe includes nonfluencies (e.g., "uh") and adds punctuation where it determines new sentences have occurred. Koenecke et al. (2020) found it to be one of the most accurate automatic speech recognition-based transcription services.

After transcribing each participant's speech, we applied Sentence-BERT's RoBERTa models from the *sentence_transformers* Python package (Reimers & Iryna, 2019) to convert each transcript into 768-dimension embeddings that reflect the style and meaning of speech. Due to the difficulty of interpreting the RoBERTa embeddings, we applied Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015) to the transcripts and correlated LIWC variables

with the warmth, competence, and morality impressions to elucidate specific verbal behaviors associated with first impressions.

Paraverbal Behavior. We applied openSMILE (Eyben, 2014) to measure paraverbal behaviors. We extracted a commonly used set of features, the Geneva Minimalistic Acoustic Parameter Set (Eyben et al., 2016), which includes pitch, measures of voice quality, and loudness. We extracted these features from overlapping 30-second windows, sliding the windows in 1-second steps, and then aggregated the results using their means. Additionally, we measured loudness peaks per second and voiced segments per second, which capture speech rate, for a final set of 53 paraverbal behaviors.

Nonverbal Behavior. We applied OpenFace (Baltrusaitis et al., 2018) to measure nonverbal behaviors. Instead of making *a priori* assumptions about the emotions reflected in facial expressions (Barrett et al., 2019), we extracted the presence and activation intensity of 18 facial action units (FAUs). For each FAU's frame-by-frame presence and activation intensity scores, we measured the mean, median, standard deviation, kurtosis, skewness, and (for activation intensity), minimum, 25th percentile, 75th percentile, and maximum. Additionally, OpenFace measured information about head pose along pitch, yaw, and roll axes yielding a final set of 352 nonverbal behaviors.

Analytic Approach

Human-Rated Verbal, Paraverbal, and Nonverbal Behaviors

To examine the relationship between human-rated speaker behaviors and warmth, competence, and morality impressions, we used hierarchical regression. Specifically, we regressed warmth, competence, and morality (separately) onto verbal, paraverbal, or nonverbal behaviors. Then we regressed each impression onto each pair of the types of behavior (i.e.,

verbal and paraverbal; verbal and nonverbal; paraverbal and nonverbal) and, finally, onto all human-rated public speaking behaviors. As an exploratory analysis, we also examined two-way interactions among the 11 behaviors, but when using a Bonferroni correction to limit Type I (false positive) errors, none of the interaction terms were significant in any of the three models that included all 11 behaviors and 55 interactions (see Table S11).

Computer-Measured Verbal, Paraverbal, and Nonverbal Behaviors

To provide triangulating evidence regarding the relationship between speaker behaviors and warmth, competence, and morality impressions, we trained ML elastic net regression (Zou & Hastie, 2005) models to use AI-measured verbal, paraverbal, and/or nonverbal behaviors to predict warmth, competence, and morality ratings in unseen data. This serves as a parallel analysis to hierarchical regression with three key differences: AI-measured predictors (vs. human-measured); elastic net regression (vs. ordinary least squares [OLS] regression); and cross-validating out-of-sample predictions (vs. in-sample). For both the human and AI approaches, we compared the multiple R of different models using Steiger's (1980) equations 3 and 10 to examine the incremental validity of each type of behavior, beyond the others, in predicting these impressions. An important distinction is that in OLS regression, without cross-validation, adding additional variables always increases R^2, whereas additional variables are not guaranteed to increase the R^2 obtained upon cross-validation.

Before training the models, we standardized all predictors to have mean = 0 and standard deviation = 1. We then trained ML models and tested the accuracy of their predictions using a 10 times repeated nested 10-fold cross-validation, in line with best-practice recommendations (Krstajic et al., 2014). The Online Supplement provides additional details about this procedure. We report the average correlation between y (i.e., observed warmth, competence, or morality

ratings) and $\hat{y}$ (i.e., predicted warmth, competence, or morality scores) across the 10 repetitions. The standard deviation of these correlations across the 10 repetitions tended to be quite small—for example, when predicting warmth, all seven models' R^2 across the 10 repetitions SD = .01. As a supplementary analysis, we applied the same procedure to human-rated behaviors, finding consistent results regarding which type of behavior was most important for each impression and which type(s) of behavior provides incremental validity (see Table S7).

Additionally, Online Supplement Tables S8-10 report the 10 highest weighted predictors, on average across the 10 repetitions, in each predictor set-dependent variable pair (*ML_model_coefficients.xlsx* in the OSF repository reports the average regression coefficients for all predictors in each predictor set-dependent variable pair).

Results

Observers formed clearer impressions of warmth and competence than of moral character. Interrater reliability was highest for warmth (ICC(1, k) = .80) and lowest for first impressions of moral character (ICC(1, k) = .61). Similarly, first impressions of warmth exhibited the highest variance (M = 3.87, SD = 0.58), and moral character impressions exhibited the lowest variance (M = 4.04, SD = 0.36) (see Online Supplement Table S1).

Warmth Is Best Predicted by Nonverbal Behaviors

Nonverbal behaviors (what speakers *did*) best predicted warmth impressions, accounting for 90.1%[1] and 82.3% of the association between behavior and warmth, respectively, in the human (r = .602)[2] and AI (r = .531) approaches, as Figure 1 illustrates. That said, verbal and paraverbal behaviors both contributed additional variance beyond nonverbal behaviors in the

[1] Calculated by dividing the r for the nonverbal behavior model by the r for the model that included all three types of behaviors.

[2] Although this correlation between fitted and observed values in multiple regression is generally referred to as multiple R, for ease of presentation we refer to it as r.

human (Δrs = .037 & .030, or 6.1% and 5.0% increases, respectively, for verbal and paraverbal behaviors, ps < .001; see Online Supplement Tables 2 & 4) and AI (Δrs = .080 & .076, or 15.1% and 14.3% increases, respectively, for verbal and paraverbal behaviors, ps < .01; see Online Supplement Table 3) approaches. When considering all three types of behavior together, the relationship between fitted/predicted and observed values increased beyond any pair of two types of behaviors (Δrs = .029 & .034, or 4.5% and 5.6% increases, respectively, in the human and AI approaches, compared to using verbal and nonverbal behaviors, ps < .01). Warmth impressions related more strongly to behavior in both methodological approaches than did competence or morality. Additionally, the results across the two methods were highly consistent: the r results for warmth in Tables S2-3 correlate r = .925.

What specific nonverbal behaviors lead observers to rate speakers as warm? Speakers were perceived as warmer for exhibiting animated, energetic facial expressions (human approach), and more specifically, when they smiled more intensely and frequently using both the mouth (FAU 12; zygomaticus major) and eye (FAU 6; orbicularis oculi, pars orbitalis) parts of the genuine, or Duchenne, smile (AI approach). Detailed results from both human (Figure 2 & Table S4) and AI (Table S8) approaches provide complementary information about the specific behaviors related to warmth.

Regarding paraverbal behaviors, observers rated speakers as warmer for more effectively using prosody (i.e., pace, volume, and pitch) and pauses (Figure 2, human approach), as well as for speaking faster and having a less nasal-sounding voice (Table S8, AI approach). In terms of verbal behaviors, observers rated speakers as warmer for using easy-to-understand speech, talking less about work and money, and being more cautious about what they said. Specifically, as Table 1 reports, observers rated speakers as warmer for using a greater proportion of function

words (e.g., it, no, very), tentative language (e.g., maybe, perhaps), and differentiation (e.g., hasn't, else) words, while observers rated speakers as *less* warm for spending more time talking about money-related concerns (e.g., cash, audit) and work (e.g., job, majors), using words longer than six letters, more analytical speech,[3] and a greater proportion of causation (e.g., because, effect) words.

Competence Is Best Predicted by Verbal (and Paraverbal) Behaviors

Verbal behaviors (what speakers *said*) best predicted competence impressions, accounting for 94.7% and 91.1% of the association between behavior and competence, respectively, in the human (r = .531) and AI (r = .438) approaches. However, the strength of the verbal behavior-competence relationship was not significantly different from the paraverbal behavior-competence relationship in either the human or AI approach (ps = .07 & .26, respectively). Adding paraverbal behaviors contributed additional variance beyond verbal behaviors in both the human (Δr = .026, a 4.9% increase, p < .01) and AI (Δr = .043, a 9.8% increase, p = .04) approaches, but adding nonverbal behaviors did not (Δrs = .006 & -.049, respectively, in the human and AI approaches). When considering all three types of behavior together, the relationship did not increase beyond using verbal and paraverbal behaviors (Δrs = .004 & -.031, respectively, in the human and AI approaches). Additionally, the results across the two methods were highly consistent: the r results for competence in Tables S2-3 correlate r = .936.

What specific verbal behaviors lead observers to rate speakers as competent? Observers rated speakers as more competent for having a logical flow to their speech and keeping it relevant and interesting to their audience (Figure 2 & Table S5, human approach). Specifically,

[3] The Analytical LIWC category is scored by summing the articles and prepositions categories and subtracting the adverbs, auxiliary verbs, impersonal pronouns, personal pronouns, conjunctions, and negation categories.

observers rated speakers as more competent for using more analytical speech and spending more time talking about achievement-related drives (e.g., success, better), work, and physical space (e.g., forward, surround). Observers rated speakers as less competent for using a higher proportion of informal language (including assent [e.g., agree, OK], and nonfluencies [e.g., er, um]), pronouns (e.g., I, itself), affective content (e.g., happy, cried), and present-focused speech (e.g., today, now) (Table 1, AI approach).

Regarding paraverbal behaviors, much as with warmth, observers rated speakers as more competent for more effectively using prosody (i.e., pace, volume, and pitch) and pauses (Figure 2, human approach) as well as talking faster (and saying more words overall) and having a less nasal-sounding voice (Table S9, AI approach).

Morality Is Best Predicted by Verbal Behaviors

Verbal behaviors (what speakers *said*) best predicted morality impressions, accounting for 85.5% and 92.7% of the association between behavior and morality, respectively, in the human (r = .353) and AI (r = .381) approaches. The strength of the verbal behavior-morality relationship was significantly different from the paraverbal behavior-morality relationship in the AI approach (p = .02) but not in the human approach (p = .08). Paraverbal behaviors contributed additional variance beyond verbal behaviors in both the human (Δr = .033, a 9.3% increase, p = .02) and AI (Δr = .024, a 6.3% increase, p = .048) approaches, but nonverbal behaviors only did so in the human approach (Δrs = .038 & .008, or 10.8% & 2.1% increases, respectively, in the human and AI approaches, ps = .02 & .35). When considering all three types of behavior together, the relationship between fitted/predicted and observed values increased in the human approach (Δr = .022, p = .04) but not the AI approach (Δr = .006, p = .35) compared to using only two types of behaviors. Notably, the behavior-morality relationship is much weaker than the

behavior-warmth relationship and somewhat weaker than the behavior-competence relationship. Further, the results across the two methods were highly consistent: the *r* results for morality in Tables S2-3 correlate $r = .857$.

What specific verbal behaviors lead observers to rate speakers as moral? Observers rated speakers as more moral when the speech was relevant and interesting to the audience (Figure 2 & Table S6, human approach). More specifically, observers rated speakers as less moral for using a higher proportion of words relevant to extrinsic motivations (i.e., money-related concerns, risk-related drives [e.g., danger, doubt]), assent, and netspeak (e.g., Twitter, Reddit), while observers rated speakers as more moral for using a higher proportion of conjunctions (e.g., and, but) (Table 1, AI approach).

Regarding paraverbal behaviors, observers rated speakers as more moral for more effectively using prosody and pauses (Figure 2, human approach), and for talking faster (and saying more words overall) and having a less nasal-sounding voice (Table S10, AI approach). Finally, regarding nonverbal behaviors, observers rated speakers as slightly more moral for having energetic facial expressions and, specifically, displaying mouth smiles more intensely and frequently (FAU 12).

Discussion

This investigation is the first to use both human and AI approaches to compare the relative importance of verbal, paraverbal, and nonverbal behavior in first impression formation. After viewing speakers introduce themselves in 90-second videos, human raters agreed about how warm speakers seemed more than they agreed about their competence; raters agreed least about speakers' moral character. Likewise, whether assessed by trained human raters or AI tools, dynamic behaviors collectively explained the most variance in impressions of warmth, less for

competence, and least for moral character. Regarding specific types of behaviors, human and AI approaches produced remarkably consistent findings: warmth was explained primarily by nonverbal behaviors (e.g., smiles), whereas both competence and moral character were explained primarily by verbal (e.g., logical flow of ideas) and, to a lesser degree, paraverbal behaviors (e.g., pace of speech).

While our findings support prior research linking nonverbal behaviors to warmth (Cuddy et al., 2011), they highlight the need for additional research on verbal and paraverbal behaviors, particularly as they relate to competence and moral character. Such research requires dynamic (e.g., videos), rather than static (e.g., photographs), information. If, like many prior studies, we used a human approach and only measured nonverbal behaviors, our findings would have misleadingly suggested that nonverbal behaviors were important for competence and moral character impressions. However, nonverbal behaviors did not contribute additional variance in the human approach beyond verbal and paraverbal behaviors for competence and only contributed minimal variance for moral character. The AI approach suggested that nonverbal behaviors hold little to no validity for understanding competence and morality impressions.

Methodologically, the AI approach provided results remarkably consistent with those from the human approach, opening the door to using AI in future studies of impression formation. We provide analytic codes on OSF (https://osf.io/ezv4u/?view_only=a39a9e621a174455b809333e84112a13) to enable researchers to capitalize on recent advances in AI for investigating dynamic behaviors and impression formation *at scale*. Studies included in a recent meta-analysis on how paraverbal and nonverbal behaviors relate to personality judgments had, on average, $\overline{N}$ = 111 speakers and measured 10.55 behaviors (Breil et al., 2021). In comparison, our study included N = 528 speakers and 1,173 AI-

measured behaviors. We expect that such new methods can spur future theoretical developments in this space (Greenwald, 2012).

Our findings also contribute to the literature on social cognition, suggesting that first impressions of warmth, competence, and morality differ in how strongly they relate to thin slices of behavior. Interrater reliability, variance, and variance explained were highest for warmth impressions and lowest for moral character impressions. If warmth and morality were indistinguishable (Fiske et al., 2007), we would expect them to relate similarly to behavior. Yet even though warmth and morality impressions correlated highly (r = .62), warmth was much more predictable from dynamic behaviors enacted in the videos. Further, the specific behaviors important for understanding impression formation differed for warmth and morality: Nonverbal behaviors best predicted warmth, while verbal behaviors best predicted morality impressions. Warmth and morality related similarly to certain verbal behaviors, such as being negatively related to discussing extrinsic motivations related to money and risk. However, they also exhibited differences; only warmth related negatively to analytical speech, using longer words, and spending more time talking about work. Together, this suggests that impressions of warmth and moral character, while strongly related, are not only conceptually distinct (Goodwin, 2015) but can be distinguished empirically as well. It is worth noting that the low interrater reliability of and variance explained in morality impressions suggests that morality impressions may not form as readily as warmth impressions from thin slices of behavior.

This investigation had two significant limitations that suggest important directions for future research. First, our study was observational rather than experimental. The possibility of unmeasured, third-variable confounds means that verbal, paraverbal, and nonverbal behaviors of speakers can not be assumed to be causally related to first impressions. In addition, for human

ratings, there is the possibility of reverse causation (i.e., that impressions of warmth, competence, and morality influenced ratings of verbal, paraverbal, and nonverbal behaviors). Importantly, there is no possibility of reverse causation in the AI approach, which partially alleviates this concern. Regardless, experimental paradigms are needed to confirm the causal role of the specific behaviors uncovered here.

A second limitation concerns the external validity of our findings (DePaulo & Friedman, 1998). We cannot assume that the pattern of results will hold in other contexts (e.g., in-person job interviews, cocktail party conversations) or with other populations (e.g., non-WEIRD individuals). However, some prior results align with ours in similar, digitally-mediated settings. Berry et al. (1997) found that warmth perceptions are better explained by nonverbal than verbal behaviors while competence perceptions are better explained by verbal than nonverbal behaviors in similar, introductory videos. Hickman et al. (2022) investigated ML models of interviewer-rated Big Five traits in three samples of mock employment interviews, where extraversion and conscientiousness can be considered analogues to warmth and competence. They found consistent results regarding the importance of nonverbal behaviors for extraversion/warmth and, specifically, the standard deviation of the activation intensity of the mouth part of the smile (FAU 12). They also found that nonverbal behaviors were relatively unimportant for conscientiousness/competence, and their conscientiousness models included several predictors (i.e., word count, proportion of words longer than six letters, and assent) that align with our findings in Table 1. Future work should seek to replicate these findings in a broader array of settings, including dyadic interactions and with demographically diverse raters and target samples.

In this investigation, speakers' dynamic verbal, paraverbal, and nonverbal behaviors in 90-second video self-introductions independently predicted strangers' impressions of warmth, competence, and moral character. In contrast, much prior research has relied on static facial cues (i.e., photographs). When available, dynamic information rapidly displaces static cues as causes of impressions (Kenny, 2004). By more closely simulating what it is like to meet a new acquaintance in real life, our focus on dynamic behaviors enhances ecological validity. Further, we found strikingly similar results when using either human raters or AI algorithms to operationalize verbal, paraverbal, and nonverbal behaviors. Therefore, our hope is that the current investigation sets a precedent for even more ambitious, AI-facilitated studies on the dynamics of social cognition.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256-274. https://doi.org/10.1037/0033-2909.111.2.256

Ambady, N., & Skowronski, J. J. (Eds.). (2008). *First impressions*. Guilford Press.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018, May). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 59-66). IEEE.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*(1), 1-68. https://doi.org/10.1177/1529100619832930

Barrick, M. R., Dustin, S. L., Giluk, T. L., Stewart, G. L., Shaffer, J. A., & Swider, B. W. (2012). Candidate characteristics driving initial impressions during rapport building: Implications for employment interview validity. *Journal of Occupational and Organizational Psychology, 85*(2), 330-352. https://doi.org/10.1111/j.2044-8325.2011.02036.x

Barrick, M. R., Swider, B. W., & Stewart, G. L. (2010). Initial evaluations in the interview: Relationships with subsequent interviewer evaluations and employment offers. *Journal of Applied Psychology, 95*(6), 1163-1172.

Biancardi, B., Cafaro, A., & Pelachaud, C. (2017, November). Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI)* (pp. 341-349).

Borkenau, P., & Liebler, A. (1995). Observable Attributes as Manifestations and Cues of Personality and Intelligence. *Journal of Personality, 63*(1), 1-25. https://doi.org/10.1111/j.1467-6494.1995.tb00799.x

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5*(4), 437-474. https://doi.org/10.3102/00028312005004437

Breil, S. M., Lievens, F., Forthmann, B., & Back, M. D. (2022). Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness. *Personnel Psychology*, advance online publication. https://doi.org/10.1111/peps.12507

Breil, S. M., Osterholz, S., Nestler, S., & Back, M. D. (2021). Contributions of nonverbal cues to the accurate judgment of personality Traits. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment* (pp. 195-218). Oxford University Press.

Brown, B. L. (1980). Effects of speech rate on personality attributions and competency evaluations. In H. Giles, W. P. Robinson, & P. M. Smith (Eds.), *Language* (pp. 293-300). Pergamon. https://doi.org/10.1016/B978-0-08-024696-3.50050-8

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Cuddy, A. J., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior, 31*, 73-98. https://doi.org/10.1016/j.riob.2011.10.004

DePaulo, B. M., & Friedman, H. S. (1998). Nonverbal communication. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (4th ed.) (pp. 3-40). McGraw-Hill.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S. & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., & Schuller, B. (2014). openSMILE:) The Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records*, *6*(4), 4-13. https://doi.org/10.1145/2502081.2502224

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. https://doi.org/10.1016/j.tics.2006.11.005

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38-44. https://doi.org/10.1177/0963721414550709

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*(2), 99-108. https://doi.org/10.1177/1745691611434210

Hall, J. A., Horgan, T. G., & Murphy, N. A. (2019). Nonverbal communication. *Annual Review of Psychology*, *70*(1), 271-294. https://doi.org/10.1146/annurev-psych-010418-103145

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, *8*(3), 265-280. https://doi.org/10.1207/s15327957pspr0803_3

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition.

Proceedings of the National Academy of Sciences, *117*(14), 7684-7689. https://doi.org/10.1073/pnas.1915768117

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(1), 1-15. https://doi.org/10.1186/1758-2946-6-10

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272-1290. https://doi.org/10.1177/0146167216655984

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software].

Mehrabian, A. (1971). *Silent messages* (Vol. 8, No. 152, p. 30). Wadsworth.

Mehrabian, A. & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology, 6*, 109-114. https://doi.org/10.1037/h0024532

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992, Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Rule, N., & Ambady, N. (2010). First impressions of the face: Predicting success. *Social and Personality Psychology Compass*, *4*(8), 506-516. https://doi.org/10.1111/j.1751-9004.2010.00282.x

Schroeder, J., & Epley, N. (2015). The sound of intellect: Speech reveals a thoughtful mind, increasing a job candidate's appeal. *Psychological Science, 26*(6), 877-891. https://doi.org/10.1177/0956797615572906

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251. https://doi.org/10.1037/0033-2909.87.2.245

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519-545. https://doi.org/10.1146/annurev-psych-113011-143831

Weisbuch, M., Ambady, N., Clarke, A. L., Achor, S., & Weele, J. V. V. (2010). On being consistent: The role of verbal–nonverbal consistency in first impressions. *Basic and Applied Social Psychology, 32*(3), 261-268. https://doi.org/10.1080/01973533.2010.495659

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45*, e1. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Tables

Table 1

Correlations Between Linguistic Inquiry and Word Count (LIWC) Variables and Warmth, Competence, and Moral Character

LIWC Category	Warmth	Competence	Morality
Descriptive Categories			
Word Count	.28	.35	.17
Words > 6 letters	-.13	.30	.03
Summary Language Variable			
Analytical thinking*	-.13	.26	-.03
Prepositions (e.g., to, with)	.00	.25	.10
Personal pronouns (e.g., I, her)	.05	-.24	-.05
Auxiliary verbs (e.g., am, will)	.04	-.29	.00
Negations (e.g., not, never)	-.04	-.19	-.07
Impersonal pronouns (e.g., it, those)	.10	-.15	.04
Conjunctions (e.g., and, but)	.19	.08	.17
Common adverbs (e.g., very, really)	.18	-.05	.12
Linguistic Dimensions			
Function words (e.g., it, no, very)	.18	-.12	.11
Total pronouns (e.g., I, itself)	.10	-.26	-.01
Common verbs (e.g., come, carry)	.06	-.27	-.09
Psychological Processes			
Personal concerns			
Work (e.g., job, majors)	-.19	.17	-.04
Money (e.g., cash, audit)	-.25	.09	-.25
Time orientation			
Present focus (e.g., today, now)	.06	-.32	-.02
Informal language (e.g., OK, um)	.01	-.30	.01
Assent (e.g., agree, OK)	-.13	-.32	-.15
Netspeak (e.g., twitter, reddit)	-.18	-.14	-.15
Nonfluencies (e.g., er, um)	.06	-.24	.08
Relativity (e.g., area, exit)	-.07	.17	-.07
Space (e.g., down, thin)	-.06	.20	.01
Affective processes (e.g., happy, cried)	.07	-.19	-.01
Positive emotion (e.g., love, sweet)	.08	-.18	.00
Cognitive processes (e.g., know, ought)	.10	-.10	.04
Differentiation (e.g., hasn't, else)	.17	-.16	.06
Tentative (e.g., maybe, perhaps)	.17	-.11	.10
Causation (e.g., because, effect)	-.14	.06	-.05
Drives (e.g., win, prize)	.00	.20	.04
Achievement (e.g., success, better)	-.10	.19	-.07
Power (e.g., superior, bully)	-.02	.14	.09
Risk (e.g., danger, doubt)	-.11	-.01	-.12

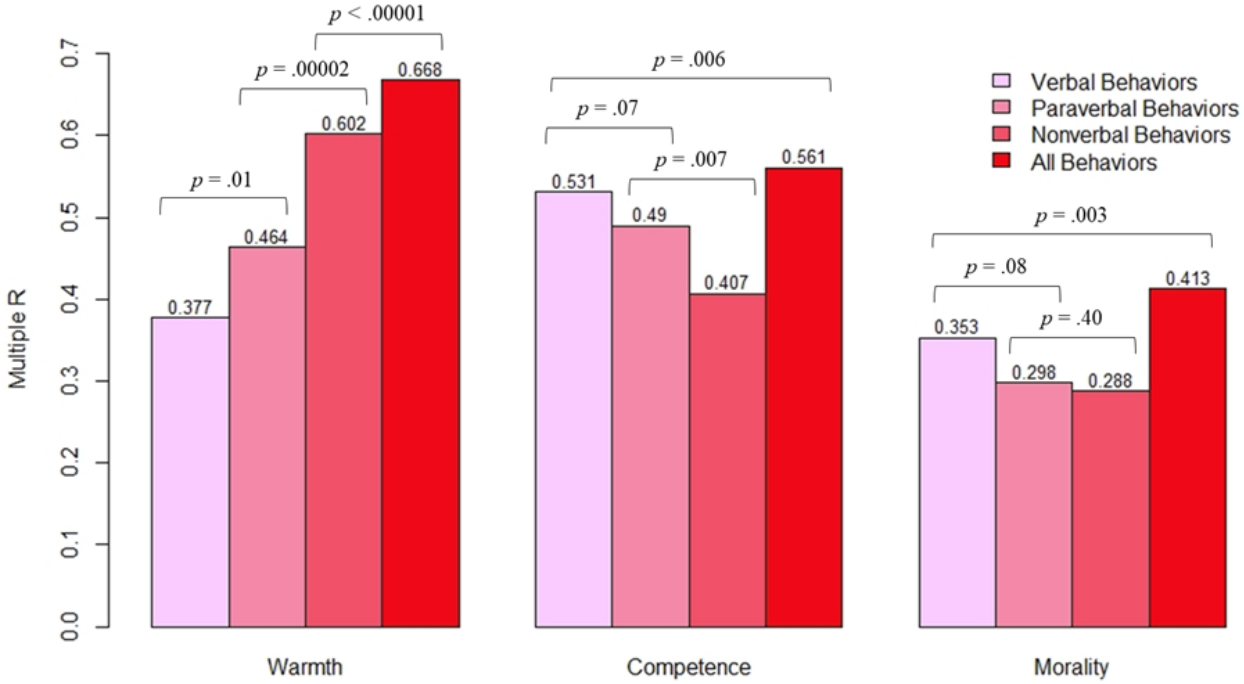
Note. $p < .05$ when $|r| > .08$; $p < .01$ when $|r| > .11$. For readability, variables are suppressed from the table when $|r| < .12$ for all three impressions or when their parent category correlated similarly to the impressions. *The Analytical LIWC category is the sum of the articles and prepositions categories minus adverbs, auxiliary verbs, impersonal pronouns, personal pronouns, conjunctions, and negations.

Figures

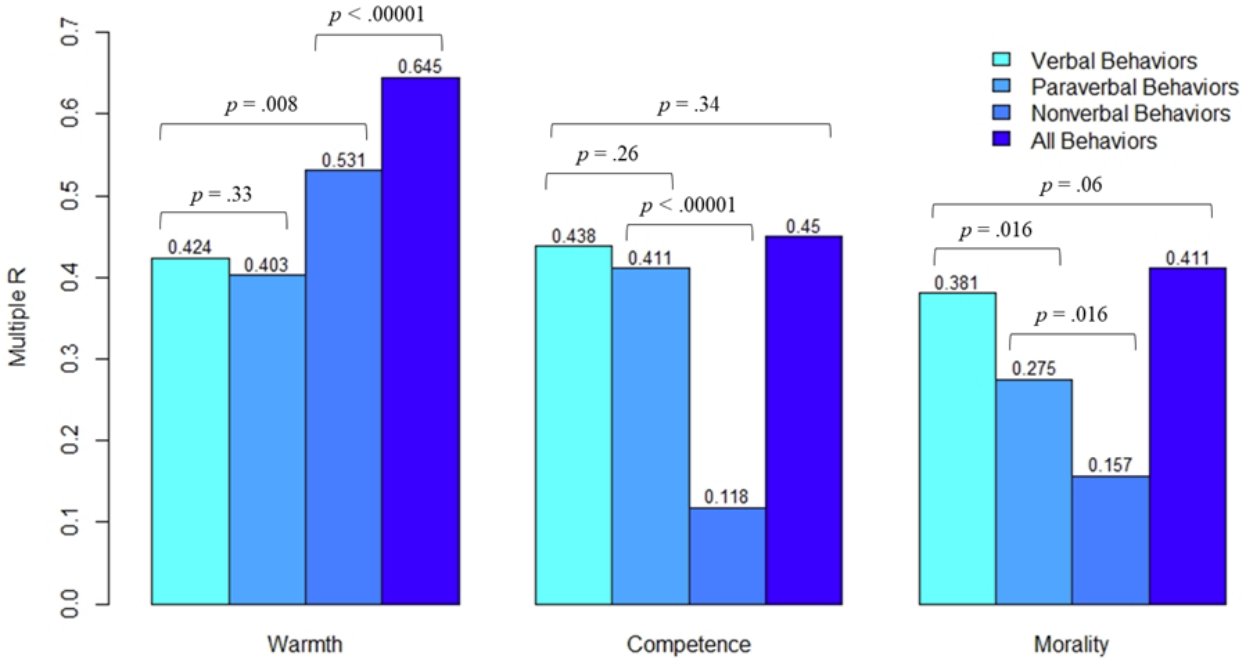
Figure 1

Multiple R from Verbal, Paraverbal, and Nonverbal Behaviors Across Human and AI Approaches

Panel A: Human-rated behaviors and multiple regression



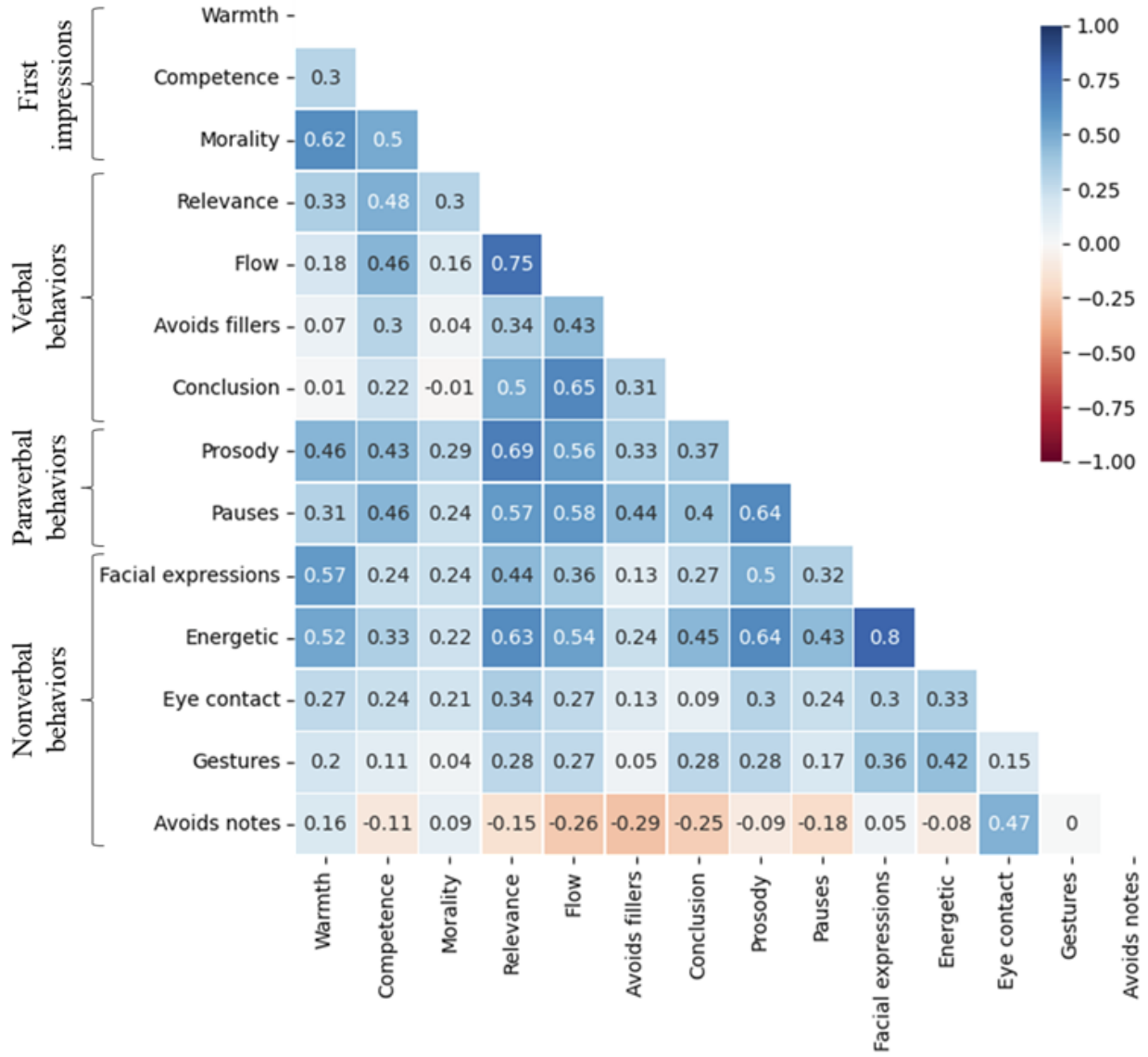
Panel B: Computer-measured behaviors and cross-validated elastic net regression



Note. *p*-values derived from *z*-score calculated using Steiger's (1980) equations 3 and 10 for two dependent correlations with Lee and Preacher's (2013) calculator.

Figure 2

Correlation Heatmap of Warmth, Competence, Morality, and Human-Rated Public Speaking Behaviors



Note. $N = 528$.

ONLINE SUPPLEMENT

Additional Method Details

Participants: Raters of Warmth, Competence, and Morality

The same MBA students who participated as speakers watched and rated six videos from other MBA students, while the undergraduate students who participated as speakers watched and rated three videos from other undergraduates. Two groups of Prolific participants (total N = 1,263 prior to removing those who failed attention checks) were paid $2 to watch and rate three videos, with one group rating other Prolific participants' videos and one group rating the undergraduate students' videos.

Because of the remote nature of the research and possibility of inattentiveness, raters also completed attention check items (e.g., "This speaker was born on earth."). For MBA students, three attention checks were included across the six videos, and for undergraduate students, two attention checks were included across the three videos. For the Prolific participants, one attention check item was included for every video rated. We discarded raters if they failed one or more of the attention checks. Before removing raters who failed one or more attention checks, there were N = 1,696 raters. After removing raters who failed one or more attention checks, the final sample consisted of 1,413 raters' 4,641 ratings on 528 participants' videos, as reported in the main text.

Measures

Warmth, Competence, and Morality

To derive the final set of adjectives used to measure warmth, competence, and moral character, we went through an iterative process of administering a larger set of items before winnowing them down to the final three. We created an initial list of adjectives by aggregating the results of the 10 exploratory factor analyses conducted by Landy et al. (2016) and

brainstorming additional, relevant items. Our initial list included eight items per construct. Specifically, for warmth the items were: sociable, extraverted, warm, friendly, enthusiastic, relatable, empathetic, and approachable. For competence the items were: skilled, competent, talented, effective, capable, clear, discerning, and prudent. And for moral character, the items were: moral, trustworthy, honest, respectful, fair, principled, authentic, and sincere. We aimed to winnow the scales down to three items per construct.

We then investigated the properties of these adjectival items in several ways, including through multilevel confirmatory factor analysis in pilot data, examining the semantic similarity of items in latent embedding space, and judging the extent to which the items narrowly or broadly reflected the intended constructs. First, we used multilevel confirmatory factor analysis in an exploratory manner to investigate the relationships among the items both between and within people. Additionally, we explored both the within-person, or rater-level correlations among the items, as well as the between-person, or ratee-level correlations among the items. According to factor analysis, the three highest loading items were: warm, friendly, and approachable for warmth; skilled, competent, and capable for competence; and trustworthy, honest, and sincere for moral character. When examining correlations, we subtracted heterotrait correlations for each item from their monotrait correlations. At the within-person level, this suggested: sociable, warm, and enthusiastic for warmth; skilled, competent, and capable for competence; and moral, honest, and fair for moral character. At the between-person level, this suggested: warm, friendly, and sociable for warmth; skilled, competent, and discerning for competence; and moral, honest, and fair for moral character.

Second, we calculated the semantic similarity of the 24 items using this website http://forwardflow.org/query_data from Gray et al. (2019). This provides data similar to a

correlation matrix, but instead of the correlations among the items, it represents the (dis)similarity of word meanings (i.e., semantics). Much like a correlation matrix, monotrait adjectives (e.g., only those that are meant to measure warmth) should be more strongly related semantically than are heterotrait adjectives (e.g., one adjective meant to measure warmth and one meant to measure competence). We again subtracted the heterotrait relationships from the monotrait relationships. This suggested: warm, approachable, and empathetic for warmth; competent, effective, and capable for competence; and honest, respectful, and principled for moral character.

Pairing these results with our judgment to ensure that the constructs were broadly captured (as opposed to having redundant, internally consistent items), we used the following items: approachable, warm, and sociable for warmth; talented, skilled, and competent for competence; and trustworthy, sincere, and principled for moral character. Although *talented* did not emerge in the top three competence items for the analyses mentioned above, it tended to fall fourth or fifth among the competence items and represents an innate aspect of competence, whereas skilled represents an acquired aspect of competence. Additional details about any of these specific analyses are available upon request.

In all cases, during the study, the nine warmth, competence, and morality items were presented in random order to the raters, with any attention checks included in the random order of presentation. We averaged the items from each scale within each rater for each speaker, and then calculated each speaker's final warmth, competence, and morality scores by averaging all raters' warmth, competence, and morality scores for each speaker.

Public Speaking Behaviors

The pool of nine raters included two sets of raters. Five of the raters were communications professionals at the same business school involved in helping develop the rubrics, and four of the raters were undergraduate and graduate research assistants. In both cases, raters were paid for their time and underwent frame-of-reference training that consisted of defining the dimensions rated, reviewing the scale items and scale anchors, conducting practice ratings, discussing sources of (dis)agreement, and receiving feedback. The communications professionals completed the ratings as part of their regular work tasks, and the research assistants were paid $13 per hour.

Verbal Behaviors with Low Interrater Reliability. The ratings of jargon use and word choice exhibited ICC(1, k) < .50, so we excluded them from our main analyses, following Breil et al. (2022). Jargon ("uses audience-appropriate language and avoids jargon") and word choice ("uses vivid and compelling word choices"). Notably, when these behaviors are included in the regression models (see Tables S12, S13, and S14), the results are consistent with those presented in the main text.

Paraverbal Behavior with Low Interrater Reliability. The ratings of enunciation exhibited ICC(1, *k*) < .50, so we excluded them from our main analyses. Enunciation ("enunciates clearly"). Notably, when this behavior is included in the regression models (see Tables S12, S13, and S14), the results are consistent with those presented in the main text.

Analytic Approach

Computer-Measured Verbal, Paraverbal, and Nonverbal Behaviors

These details expand on the description of 10 times repeated 10-fold cross-validation provided in the main text. Cross-validation reduces effect size inflation due to overfitting and capitalization on chance variation. *k*-fold cross-validation (of which, 10-fold is commonly used)

provides a nearly unbiased estimate of model error (Varma & Simon, 2006), and repeated *k*-fold cross-validation reduces the influence of chance variation caused by the specific split used during *k*-fold cross-validation (Krstajic et al., 2014). This process involves conducting hyperparameter tuning on nine parts of the data (separately for each dependent variable). Elastic net includes two hyperparameters: alpha and lambda. We tried 11 values of alpha ranging from 0 to 1, stepping by .1, and 15 values of lambda, ranging from .01593 to 50. Alpha determines the extent to which elastic net performs variable selection, where 0 involves no variable selection (i.e., ridge regression), and 1 involves the full variable selection from Least Absolute Shrinkage and Selection Operator (LASSO) regression. Lambda determines how severely regression coefficients are penalized. Then, we trained (fitted) models for each dependent variable on those nine parts using the optimal hyperparameters, and used the models to predict speaker warmth, competence, and morality in the held out part of the data. That process was completed a total of 10 times, holding out each fold exactly once for testing. After doing this for all 10 folds, we calculated the correlation (i.e., Multiple *R*) between y and $\hat{y}$ across all 528 participants. We repeated the entire nested *k*-fold cross-validation procedure 10 times and report the average correlation between y and $\hat{y}$ across the 10 repetitions.

Table S7 reports the results obtained when this same modeling and cross-validation procedure is applied to the human-rated public speaking behaviors. These results concord with those presented in the main text in terms of which types of behaviors are most important for predicting each impression and which provide incremental validity. However, the correspondence of results between the AI and human approaches is lower than in the main text for impressions of competence because models that used only paraverbal behaviors did not cross-validate well when rated by humans.

Online Supplement: Additional Results

In addition to the regression results presented in the main text (Tables S4-S6) and those run with human-rated variables that exhibited low interrater reliability (Tables S12-S14), we also report regression results when including demographic control variables (Tables S15-S17). Overall, the results are consistent with those presented in the main text.

Online Supplement References

Gray, K., Anderson, S., Chen, E. E., Kelly, J. M., Christian, M. S., Patrick, J., Huang, L., Kenett, Y. N., & Lewis, K. (2019). "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist*, *74*(5), 539-554.

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, *6*(1), 1-15.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When it's bad to be friendly and smart: The desirability of sociability and competence depends on morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272-1290.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 1-8.