Computational Brain & Behavior
https://doi.org/10.1007/s42113-021-00121-2

ORIGINAL PAPER

Representing and Predicting Everyday Behavior

Malhar Singh[1] · **Russell Richie**[2] · **Sudeep Bhatia**[1]

Accepted: 7 October 2021

Abstract
The prediction of everyday human behavior is a central goal in the behavioral sciences. However, efforts in this direction have been limited, as (1) the behaviors studied in most surveys and experiments represent only a small fraction of all possible behaviors, and (2) it has been difficult to generalize data from existing studies to predict arbitrary behaviors, owing to the difficulty in adequately representing such behaviors. Our paper attempts to address each of these problems. First, by sampling frequent verb phrases in natural language and refining these through human coding, we compile a dataset of nearly 4000 common human behaviors. Second, we use distributed semantic models to obtain vector representations for our behaviors, and combine these with demographic and psychographic data, to build supervised, deep neural network models of behavioral propensities for a representative sample of the US population. Our best models achieve reasonable accuracy rates when predicting propensities for novel (out-of-sample) participants as well as novel behaviors, and offer new insights for modeling psychographic and demographic differences in behavior. This work is a first step towards building predictive theories of everyday behavior, and thus improving the generality and naturalism of research in the behavioral sciences.

Keywords Transformer models · Machine learning · Distributed semantics · Decision-making

Introduction

People engage in thousands of complex actions and behaviors over the course of the day. They may read the news in the morning, send emails in the afternoon, play with their children in the evening, and worry about the future at night. These behaviors are the causes and the consequences of mental activity, of social, economic, and political reality, and of human well-being and flourishing. For this reason, the study of everyday behavior is of special interest to cognitive, behavioral, and social scientists, and a central focus of academic disciplines such as psychology.

However, established theories and methodologies in psychology and other fields have difficulty predicting the occurrence of everyday behaviors, and are unable to formally relate these behaviors to the abstracted variables observed in artificial laboratory environments (see Bhatia & Stewart, 2018; Bhatia et al., 2019 for a discussion). Of course, many survey-based methods and theories do use common behavioral patterns as stimuli, for example, items in personality (Goldberg, 1990) and risk (Blais & Weber, 2006) questionnaires. However, these stimuli are hand-picked by experimenters and restricted to narrow domains of human psychology. Thus, results from questionnaire-based studies cannot easily be generalized to the thousands of everyday behaviors that could be of interest to researchers.

Ultimately, the complexity and wide scope of naturalistic behavior makes it especially difficult to study. We do not currently have a way of formally representing the nearly infinite set of everyday behaviors, and are thus unable to formulate scientific theories capable of predicting and explaining these behaviors.

In this paper, we propose and test a new approach to quantifying naturalistic behavior. Specifically, we suggest that common behaviors can be represented as verb phrases (e.g., *read the news*, *send email*, *play with children*, or *worry about the future*), and that recent advances in natural language processing, such as transformer networks and other deep language models (Cer et al., 2018; Devlin et al., 2018), can be used to give these phrases high-dimensional vector representations that preserve their meanings. Such

✉ Sudeep Bhatia
bhatiasu@sas.upenn.edu

Malhar Singh
malhars@sas.upenn.edu

Russell Richie
drussellmrichie@gmail.com

[1] Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

Published online: 25 March 2022

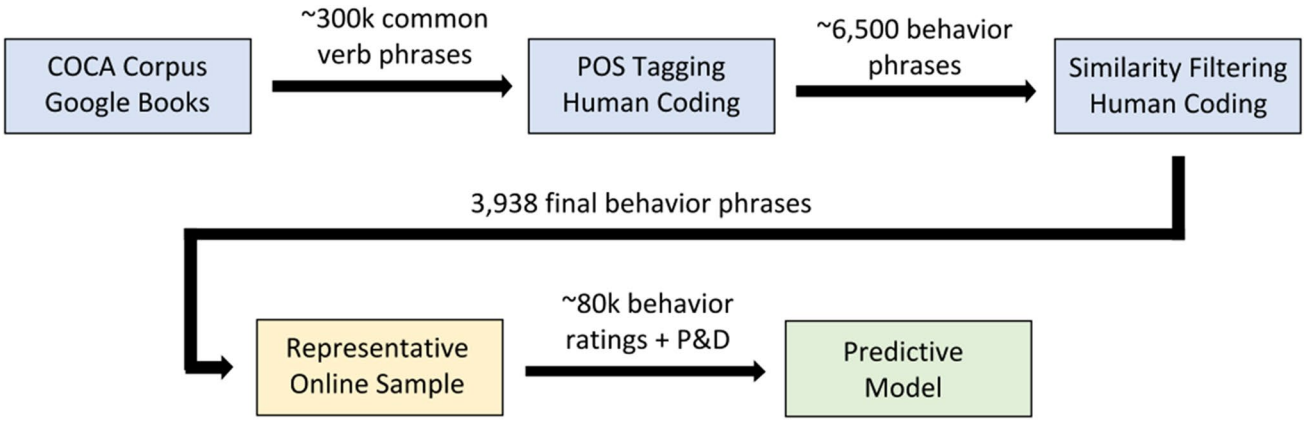

Fig. 1 Core components of our study. Blue boxes refer to the analysis performed in the section titled "Corpus Analysis to Obtain Common Behaviors," the yellow box refers to data collection described in the section titled "Survey of Behavioral Propensities," and the green box refers to the analysis in the section "Predictive Modeling of Behavioral Propensities." P&D refers to participant psychographic and demographic data

representations can be obtained for nearly any natural language phrase, which implies that it is possible to develop formal models that can take arbitrary human behaviors (in the form of vector representations) as inputs or alternatively produce these behaviors as outputs, facilitating more naturalistic behavioral theorizing.

Although we consider a number of ways in which researchers can use vector representations of behaviors, our focus in this paper is on the predictive modeling of behavioral propensities, that is, on building machine learning models capable of predicting how likely different people are to perform thousands of everyday behaviors. To facilitate such an analysis, we first compile a very large dataset of common behaviors based on the natural language occurrence frequencies of hundreds of thousands of verb phrases. We then offer a subset of these phrases to human participants to measure self-reported behavioral propensities. Finally, we use the vector representations of the verb phrases (obtained from deep language models) as inputs in machine learning models, to predict the behavioral propensities of our participants. Our aim is to make such predictions for out-of-sample behaviors as well as for out-of-sample participants (i.e., participants, behaviors, and participant-behavior combinations, that our model is not trained on), in order to test the generalizability of our approach. We also examine the ability of this approach to predict group-level (psychographic or demographic) differences in behavioral propensities. Figure 1 outlines the key computational and empirical steps performed in the current paper.

The study of behavioral propensity, and individual differences in behavior propensity, is a key focus of research in psychology, especially in subfields like judgment and decision-making, moral psychology, personality research, and clinical psychology (e.g., Bruine de Bruin et al., 2007; Cacioppo & Petty, 1982; Blais & Weber, 2006; Goldberg, 1990; Lovibond & Lovibond, 1995; Patton et al., 1995; Rushton et al., 1981; Schwartz et al., 2002). For these reasons, scholars in other fields, such as marketing, management, policy, and economics, are also interested in describing and understanding how likely people are to engage in different behaviors. Our paper will test the applicability of deep language models, such as transformer networks, to the study of naturalistic behavior in these diverse domains, and form the basis of future research that uses these models in order to better understand behavior and its correlates.

Transformer Models of Language

The past few years have seen impressive technological breakthroughs in computational linguistics: Computer models are now able to achieve unprecedented levels of performance in question answering, semantic entailment, machine translation, sentiment analysis, and other natural language understanding tasks. Perhaps the most impressive advances have come from a new type of deep neural network language model known as the transformer (Cer et al., 2018; Devlin et al., 2018; Radford et al., 2018; Vaswani et al., 2017; Brown et al., 2020). The details of transformer models are complex, but in brief, a transformer is a stack of encoders followed by a stack of decoders, where inside each encoder/decoder is a feed-forward neural network and a self-attention mechanism, with decoder modules having one additional self-attention mechanism (Fig. 2). The self-attention mechanism itself is also sophisticated, but essentially, as a transformer processes each word in a phrase or sentence, self-attention enables the transformer to look at other words in the input sequence for information about how to best encode

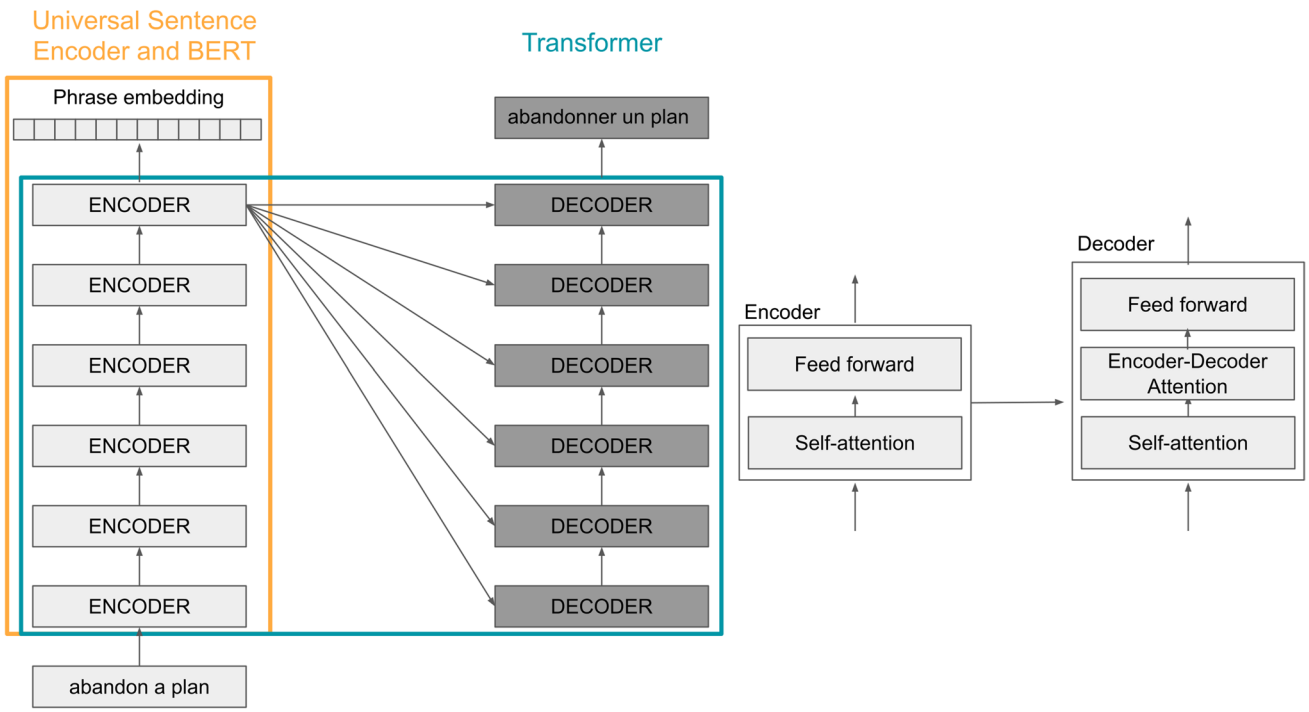

Fig. 2 Transformer model architectures. The transformer contains a stack of encoders and a stack of decoders. Inside each encoder/decoder is an attention mechanism (or two, for decoders) and a feed-forward network. While a typical transformer, with both encoders and decoders, can be used for sequence-to-sequence prediction, as in the visualized English-French translation example, BERT and USE make use of only the encoder stack (in different ways) to obtain phrase representations. Figure adapted from Alammar (2018)

the current word. For example, in a sentence like *Russell*$_i$ *likes his*$_i$ *cat*, attention allows the internal representations for *his* to be influenced by the representation for *Russell*, since there is a co-reference relation between these words (see supplemental materials for a technical walkthrough of the attention mechanism, and see Alammar, 2018 for an accessible, illustrated introduction to the transformer and especially attention). When trained on appropriately large amounts of text data, transformers can produce vector representations that approximate key elements of sentence meaning, and can subsequently be used as inputs in secondary machine learning models that fine-tune the vector representations for down-stream tasks.

Transformer models that encode phrases and sentences as vectors are, in a sense, an evolution of older models that produce vectors for words, like LSA (Landauer & Dumais, 1997), BEAGLE (Jones & Mewhort, 2007), Word2Vec (Mikolov et al., 2013), or GloVe (Pennington et al., 2014), based on the distributional statistics of words in large collections of texts. In both word vector models and phrase and sentence encoders, vectors for linguistic units are obtained such that similar words, phrases, or sentences occupy nearby positions in semantic space. In addition, our application of transformer-derived sentence vector representations to predicting complex, real-world behaviors is largely inspired by various applications of word vector models in psychology. These applications include list and category recall, similarity and relatedness judgments, and free association (for review, see Bhatia et al., 2019; Jones et al., 2015; Lenci, 2018; Mandera et al., 2017), but perhaps most relevant for the present work are applications of word vector models to judgments about the psychological properties of words and phrases, e.g., the "riskiness" of potential risk sources like *smoking* or *skydiving* (Hollis et al., 2016; Bhatia 2019; Bhatia et al., 2021; Richie et al., 2019; Utsumi, 2020; Zou & Bhatia, 2021). In this work, ratings for a particular kind of judgment (e.g., riskiness) are directly linearly regressed onto the vectors for words (e.g., potential risk sources). This approach can explain about half of the variation in out-of-sample subject-averaged judgment ratings, and strongly outperforms an association/similarity baseline that merely measures the similarity between a target word (e.g., *smoking*), and words representing the judgment dimension (e.g., *risky* or *unsafe*; Richie et al., 2019). We will take a similar approach when predicting propensities of behavior phrases. The advantage of using transformer models to obtain phrase vectors, over simply, say, averaging GloVe or Word2Vec vectors in a phrase, is that such models will take into account the order and identity of all words within a phrase when computing a vector. Obviously, the order of words within a phrase or sentence, and not just their identity, is a critical component of meaning (cf *dog bites man* vs *man bites dog*).

Transformers have grown in popularity and variety since their introduction, but in this work, we will focus on two prominent transformers, USE (Universal Sentence Encoder, Cer et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2018), due to their accessibility via off-the-shelf tools and their high performance on semantically nuanced natural language understanding tasks. We describe each briefly.

Whereas a complete transformer model is a sequence-to-sequence model which is used for tasks like machine translation or POS tagging, USE utilizes only the encoding subgraph of the transformer architecture. Thus, the final output of this subgraph is a real-valued vector representation for each word of a phrase or sentence, which are simply summed,[1] element-wise, to obtain a fixed-length vector for the entire phrase or sentence. The model used in the current paper has 512-dimensional vectors which can be obtained from the TensorFlow implementation of USE (Abadi et al., 2015; Cer et al., 2018). This model was trained using next sentence prediction on text from Wikipedia, web news, web question–answer pages, and discussion forums. This model also received training on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015). Pre-trained on these tasks, USE generalizes very well to related tasks, including sentiment analysis, question classification, and sentence similarity (Cer et al., 2018).

A major shortcoming of USE and similar transformers is that it is "unidirectional," in the sense that every token can only attend to the previous tokens in the self-attention layers of the transformer. To resolve this problem, Devlin et al. (2018) developed BERT, the Bidirectional Encoder Representations from Transformer, which does allow the representation for a token to vary by what comes before and after it. As with USE, fixed-length representations of sentences can be obtained by aggregating (e.g., summing) over the token representations at various hidden layers of the network. The BERT model used in this paper was trained using masked word prediction (in which the modeler randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary ID of the masked word based only on its context) and next sentence prediction, on text from Wikipedia and Google Books. After being fine-tuned on additional task-specific data, this model demonstrated (at the time) state-of-the-art performance in many NLP tasks, including question answering, sentiment analysis, and sentence acceptability. The primary application in this paper will not be fine-tuning the full BERT model but rather use the 768-dimensional out-of-the-box vectors offered by the bert-as-a-service Python package (Xiao, 2018).

[1] The sum is also divided by the square root of the length of the phrase or sentence, to normalize for sequence length.

We acknowledge that, of course, vector representations obtained from the above models are not always able to accurately capture sentence meaning: They sometimes generate errors in syntactically complex sentences and fail at common sense reasoning (e.g., McCoy et al., 2019). There is also a philosophical debate about whether semantics can be inferred purely from the statistics of natural language (Lake & Murphy, 2021, Marcus, 2020, or Bender & Koller, 2020). Nonetheless, the success of transformer models in tasks involving simpler sentence structure and limited high-level reasoning implies that these models may have practical utility for quantifying simple phrases and sentences corresponding to common human behaviors. We would expect phrases and sentences that pertain to similar behaviors to be given similar vector representations by these models.

Consider, for example, the verb phrases $p_1 =$ *paint a house*, $p_2 =$ *decorate a room*, and $p_3 =$ *rent a room*. p_1 and p_2 are highly similar behaviors despite having different verbs and nouns: both would likely be involved in home renovation. p_2 and p_3 share a word (*room*) but are otherwise quite different, as they concern different events (decorating vs renting; in linguistic terminology, the verbs are the "head" of the verb phrases and thus typically contribute more to its meaning than the direct object or other dependents of the verb). Transformer models are useful for quantifying behaviors as they are able to correctly represent the emergent meanings of the word combinations in such phrases. To illustrate this, we passed these phrases through the Universal Sentence Encoder (USE) (Cer et al., 2018), to generate representations that preserve the semantic similarity of sentences. The USE model gave us 512-dimensional vector representations $\mathbf{x_1}$, $\mathbf{x_2}$, and $\mathbf{x_3}$, for the three phrases. We found that there is a cosine similarity of 0.77 between $\mathbf{x_1}$ and $\mathbf{x_2}$, but only 0.64 between $\mathbf{x_2}$ and $\mathbf{x_3}$, indicating that the USE model judges p_1 and p_2 to be more similar despite these phrases not sharing any words. Note that the previous generation of vector representation models, like Word2Vec (Mikolov et al., 2013), are unable to capture this pattern, as they cannot represent novel[2] multi-word phrases except by averaging, which does not respect the centrality of the verb phrase head that we indicated above. Indeed, performing the above tests with a Word2Vec bag-of-words model gives a cosine similarity of 0.71 between $\mathbf{x_1}$ and $\mathbf{x_2}$, and 0.77 between $\mathbf{x_2}$ and $\mathbf{x_3}$, suggesting that this model incorrectly judges p_2 and p_3 to be more similar.

[2] It is possible to detect strong collocations like *New York City* in a collection of texts, and then tokenize such collocations as a single unit, and learn vectors for that unit. Of course, this does not help the generation of vectors for novel phrases that were not treated as a single unit in the tokenization (like *paint a house*).

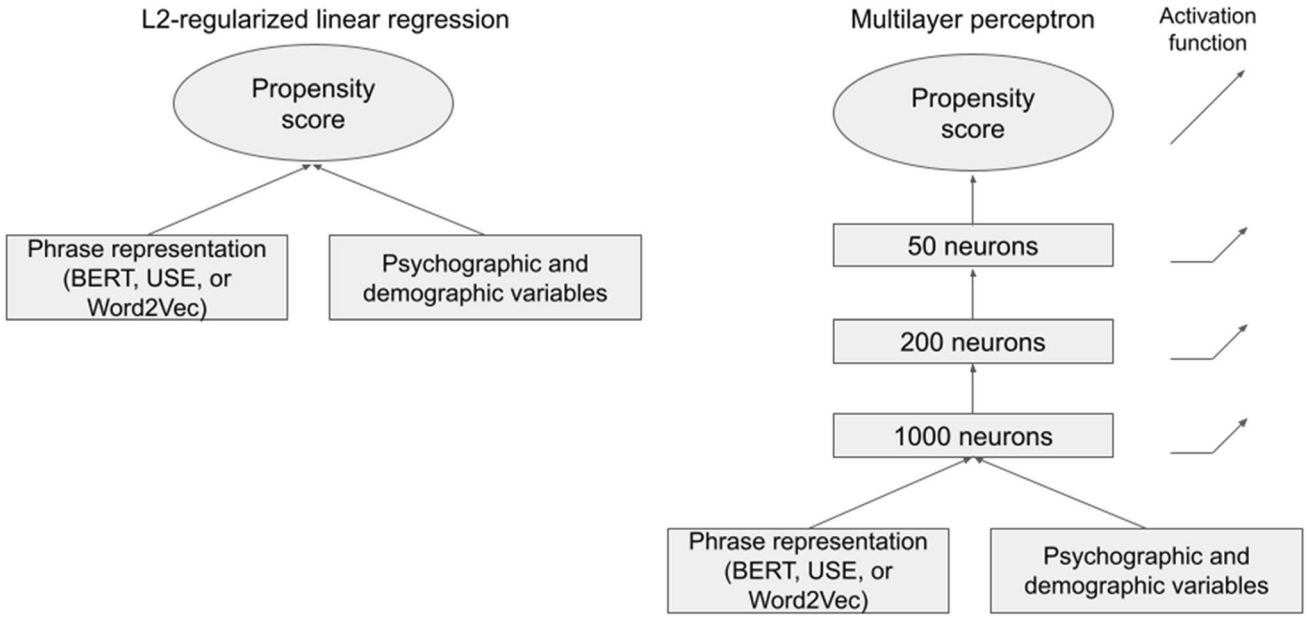

Fig. 3 Behavioral propensity predictive models. To predict behavioral propensity ratings, we used phrase representations—from BERT, USE, or Word2Vec—and participant demographic and psychographic variables. We tried L2-regularized linear regression (left), as well as multilayer perceptrons, which can capture interactions among our features (right)

Predictive Modeling of Behavior

If verb phrases that describe naturalistic behaviors can be quantified with vector representations, then it is possible to build predictive models that take vector representations of behavior as inputs and produce, as outputs, predictions regarding other variables associated with these behaviors.

One such variable could be an individual's propensity to engage in the behavior. Consider, for example, a dataset with a set of behaviors as well as (self-reported or observed) measurements of how likely an individual is to engage in the behaviors relative to others. We can use a standard linear regression to regress the behavioral propensity variable on the vectors for the behaviors obtained from transformer models like BERT or USE. Such a regression will learn a relationship between points in the vector space of behaviors and the behavioral propensity variable, and thus implicitly characterize how different behaviors vary in terms of behavior propensity. Such a model would also be able to make predictions when given a novel out-of-sample behavior; i.e., a behavior that is not in its training dataset. If such predictions are accurate, then the model could, in principle, be applied to thousands of additional behaviors that can be expressed as verb phrases and be given vector representations, allowing us to extrapolate behavioral propensities from the training data, in order to better understand the individual in consideration.

We could also use a similar approach on a dataset with behavioral propensities of multiple individuals. Such an approach may also benefit from individual-level variables (e.g., those involving demographics and psychographics), which could be introduced as covariates into the above regression. Of course, more sophisticated machine learning techniques may yield better predictions. One promising approach is a multilayer perceptron, that projects the input variables (in our case, the vector representations for behaviors and possibly demographic and psychographic variables for individuals) onto one or more intermediary, hidden layers. Hidden layers of this type can be used to learn interactions between behaviors and various individual-level characteristics, thus describing behavioral propensities on the group level, as well as sources of individual-level variability. Thus, a predictive model that accommodates interactions between individual-level characteristics and aspects of the behavior would be able to predict that an extraverted individual is more likely to engage in sociable behaviors (*go to a party*) and less likely to engage in solitary behaviors (*read a book*), while an introverted individual is likely to display the reverse pattern. Such models may also succeed at making predictions for out-of-sample individuals (in addition to out-of-sample behaviors). In this paper, we use both (regularized) linear regression models and neural network models to map behavior vectors and individual-level data onto behavioral propensity ratings from large numbers of participants. These models are summarized in Fig. 3.

The approach introduced here is not just limited to behavioral propensities. Any variable associated with a behavior could be predicted in a similar manner. For example, a

dataset of human ratings of the riskiness of different behaviors (e.g., Blais & Weber, 2006) can be used to train the above models, and subsequently predict how (potentially out-of-sample) individuals would evaluate the riskiness of (potentially out-of-sample) behaviors. Similar techniques would also work for other judgments, e.g., those involving the moral appropriateness of behaviors or the gender stereotypicality of behaviors. These, and other extensions of our framework, are examined in detail in the discussion section of this paper.

Building a Set of Common Behaviors

Of course, any predictive modeling analysis that uses large numbers of variables to make predictions requires a large amount of training data. In our case, we require not only ratings from a large and diverse group of participants (the details of which we will provide in the subsequent section), but also ratings of a large and diverse set of common behaviors. In this section, we describe the collection of a novel dataset of thousands of phrases describing human behaviors.

Obtaining Initial Dataset of Verb Phrases

We began by extracting the 1000 most frequent verbs in the Corpus of Contemporary American Literature (COCA; Davies, 2009). We then used Google Books' n-gram dataset (Michel et al., 2011) to construct verb phrases to populate our dataset of human behaviors. For each of the 1000 COCA verbs, we obtained the 100 most frequent 3-g, 4-g, and 5-g phrases from the n-gram dataset beginning with the given verb. This resulted in the creation of a list of ~ 300,000 n-grams. Notably, many of the resulting n-grams in the dataset were not valid verb phrases. For example, *say that the* is the second most common 3-g beginning with the verb *say*, likely because it is a common prefix of other complete phrases that have the verb *say*.

As an initial attempt to prune these cases from our dataset, we performed part-of-speech (POS) tagging with the natural language processing package spaCy (Honnibal & Montani, 2017), to produce a POS sequence for each verb phrase. We then examined the 150 most frequent POS sequences. Although these 150 sequences accounted for only about 1.5% of all unique POS sequences in our 300,000 n-grams, they accounted for a majority of n-grams. From these 150 POS sequences, we then manually selected 16 sequences that consistently produced complete and grammatically correct verb phrases (see our OSF repository for a complete list). We selected these POS sequences by first randomly sampling 20 verb phrases from the 100 most frequent POS sequences, and then randomly sampling 10 verb phrases from the 101st to 150th most frequent POS sequences. We reviewed the sampled verb phrases and chose 16 POS sequences whose sampled verb phrases were valid at least 50% of the time, and that we did not expect would consistently produce invalid behaviors. Using only n-grams with POS sequences matching these 16 sequences, we reduced the dataset to 31,942 n-grams. While POS tagging significantly helped reduce the number of invalid verb phrases, many n-grams that either did not constitute complete verb phrases or were not valid human behaviors remained in the dataset. To solve these issues, we turned to human coding.

Human Coding and Validation Study

To ensure that the dataset of n-grams contained valid behaviors, we needed to remove all phrases that were not (1) complete and grammatically correct verb phrases or (2) plausible for an individual to perform. We thus designed an annotation study where participants were tasked with coding the phrases from our dataset based on these criteria.

Participants We recruited 438 participants (51% female, $M_{age} = 36$, $SD_{age} = 12$) through Prolific Academic to participate in coding our list of phrases. Data collection was limited to participants from the USA whose first language was English. Participants were only allowed to participate once in this task and were paid approximately $10 per hour.

Procedure Participants were given a set of instructions explaining our criteria for behavioral plausibility and grammatical correctness. Participants were provided multiple examples of complete phrases that are valid human behaviors, as well as strategies that could be used to evaluate how well a phrase met these criteria. For example, one strategy for testing whether or not a verb phrase is complete is by checking whether or not it can be said in response to a question like, "What does the person/animal/thing/etc. do?" Further details of these instructions, examples, strategies, and criteria can be found in our OSF repository.

Participants then moved to a training section to develop a stronger sense of how phrases might or might not meet the validation criteria. Participants were given eight predetermined phrases and asked to rate these phrases on a scale from 1 (definitely not a valid human behavior) to 5 (definitely a valid human behavior) on the criteria provided. After rating a phrase, an explanation would appear on the screen explaining why the participant's rating was correct or incorrect. Following this training section, participants moved to the main portion of the study. Further details of the training section can be found in our OSF repository.

For the main portion of the study, participants were randomly assigned to evaluate a subset of approximately 250 phrases from the 31,942 n-grams remaining from the POS-based filtering. On average, each phrase received 3.125

ratings ($SD = 0.77$). We also utilized attention checks: randomly placed, researcher-generated phrases that obviously met or did not meet the criteria listed in the instructions. For example, *kick the ball* meets the validation criteria, while *eat the very* does not.

Results We ignored data from participants that did not correctly evaluate at least 75% of the attention checks, leading to 406 of 438 participants being retained. Inter-annotator agreement was measured by taking a phrase's average rating, noting the direction (> 3 or < 3), and then dividing the number of annotators that rated the phrase in that direction by the total number of annotators for that phrase. If the phrase rating did not have a direction (i.e., total number of ratings was even between < 3 and > 3, or all $= 3$), then the rating given was a 0. We thus observed an average inter-annotator agreement score of 0.71 across all phrases, indicating that participants were effectively evaluating phrases. Using the data from these 406 human coders, we removed all phrases from our dataset that received an average rating below 4.5. This reduced the total size of the dataset to ~ 6500 n-grams constituting complete verb phrases that describe valid human behaviors.

Additional Cleaning and Similarity Reduction

The human validation study was useful in removing the majority of phrases that did not meet our criteria for a valid human behavior. However, due to the difficulty of this task, some phrases that did not fully meet the criteria remained in the dataset. Human error along with inconsistent structure between the phrases, specifically with determiners and pronouns, prompted a need to further clean and code the dataset.

To address both of these issues, we developed a data coding procedure, the details of which can be found in our OSF repository. Phrases were deemed valid if they were complete verb phrases, had a clear direct object (if one existed), were plausible for an individual to perform, and were able to inform us about a clear and meaningful behavior an individual would likely engage in. All pronouns in valid phrases were replaced with the appropriate form of *someone* or *my* such that the phrase made sense from the participant's perspective (e.g., *kiss her cheek* became *kiss someone's cheek*). To fix phrases that were missing a direct object, either *something* or *someone* was inserted in the appropriate location in the phrase (e.g., *push over the edge* became *push someone over the edge*). To ensure consistency among determiners, all instances of *the* were replaced by *a* or *an* or were removed entirely (e.g., *arrange the flowers* became *arrange flowers* and *drink the soda* became *drink a soda*) unless the *the* was necessary for the phrase's meaning to remain the same (e.g., *live in the wilderness*). The phrases in the dataset were first coded and cleaned by the researchers, and then reviewed by a research assistant to ensure all phrases were coded correctly.

Furthermore, many phrases were nearly synonymous with each other (e.g., *throw the ball* vs *throw the balls*). Therefore, we used the pre-trained BERT model to extract feature vectors for our phrases and used these vectors to cluster semantically similar phrases. We clustered phrases whose vectors had cosine similarities of over 0.9, looked through each cluster of phrases, and kept any phrase within a cluster that had a unique meaning. If multiple phrases were synonymous in a cluster, we chose the phrase that had the most general meaning as the one to keep. Two researchers performed this task separately and all disagreements were resolved by consensus. This procedure led to the removal of 951 phrases yielding a final dataset of 3938 verb phrases describing plausible human behaviors.

While human coding was useful in further refining the set of valid human behavior phrases, it also provided valuable annotations of tens of thousands of n-grams on their grammatical correctness and whether or not they described plausible human behaviors for an individual to perform. Using these annotations, we can train classifiers on vectors of these n-grams to predict whether or not an n-gram is a complete verb phrase referring to a plausible human behavior. As predicting grammaticality and behavioral plausibility of our verb phrases is not the primary aim of this paper, details of this analysis are left to supplemental materials, and we only report here that, using BERT to derive phrase representations, we were able to achieve accuracy rates over 90% and F-scores over 0.85. Thus, these classifiers could be used to validate the addition of thousands of potential behavior phrases to our dataset, improving its comprehensiveness. We return to this issue of the comprehensiveness of our behavior phrases in the discussion.

Describing the Content of Behavior Phrases

As our final set of valid behavior phrases is very large and very rich, it is worthwhile exploring the distribution of syntactic structures and semantic content within them, especially with an eye towards detecting types of behaviors that are over- or under-represented in our dataset. Table 1 contains the ten most frequent POS sequences contained with our final set of valid behavior phrases, as well as example phrases of each sequence, and the frequency of the POS sequence. It is apparent that our phrases span a rich variety of syntactic structures, ranging from relatively simple VERB-DETERMINER-NOUN sequences like *avoid a collision* to more complex structures like VERB-ADPOSITION-DETERMINER-ADJECTIVE-NOUN as in *cook in a double boiler*. Although we have not performed any automatic

Table 1 The ten most frequent part-of-speech sequences contained within our final set of valid behavior phrases. Also indicated are randomly selected examples of each sequence, and the frequency of the part-of-speech sequence

Part-of-speech sequence	Example phrase	Frequency
VERB-DET-NOUN	*avoid a collision*	1285
VERB-ADP-DET-NOUN	*complain to the police*	589
VERB-ADP-NOUN	*die in battle*	337
VERB-ADJ-NOUN	*spread my wings*	314
VERB-NOUN	*assemble equipment*	251
VERB-DET-NOUN-ADP-NOUN	*restore some semblance of order*	206
VERB-DET-ADJ-NOUN	*embrace the christian faith*	158
VERB-ADP-DET-ADJ-NOUN	*cook in a double boiler*	113
VERB-NOUN-PART-NOUN	*quote someone's words*	107
VERB-PART-DET-NOUN	*squeeze out a tear*	102

semantic parsing or semantic role labeling of our phrases to derive a structured semantic representation or label words in our phrases for semantic roles like agent, patient, and theme, we strongly suspect that, to the extent that syntactic structure often follows semantic structure, our behavior phrases also span a great range of semantic structures.

To explore semantic content (as opposed to semantic structure) in a more direct way, we performed two analyses. First, we conducted a dictionary-based analysis with the well-known LIWC dictionary (Pennebaker et al., 2015), which contains lists of words in various categories, like work, home, and leisure. For example, LIWC's list of leisure-related words includes *alcohol*, *mall*, and *yoga*. For each of these categories (excluding syntactic categories like articles or conjunctions; see our OSF repository for all LIWC categories), we counted the total number of words, across all behavior phrases, falling into a given category.

Figure 4 displays the 15 most and least frequent categories across all of our phrases. Again, our phrases span a variety of categories, but some appear (much) more frequently than others. The most common categories include space (with words like *above*, *map*, *within*), cognitive processing (words like *think*, *know*, *believe*), drives (words like *accomplish*, *command*, *motivate*), and social (words like *help*, *together*, *talkative*). Other categories that strike us as central to human life, like death (words like *alive*, *grieve*, *war*) and sex (words like *nude*, *abortion*, *womb*), are vanishingly rare, with only seven phrases containing sexual words according to LIWC.[3] It is also notable that the male category appears nearly three times as often (43 times) as the female category (15 times), despite LIWC having more words for female than male.[4] This gender bias, and the absence of sexual words, may be a result of our usage of the Google Books n-grams

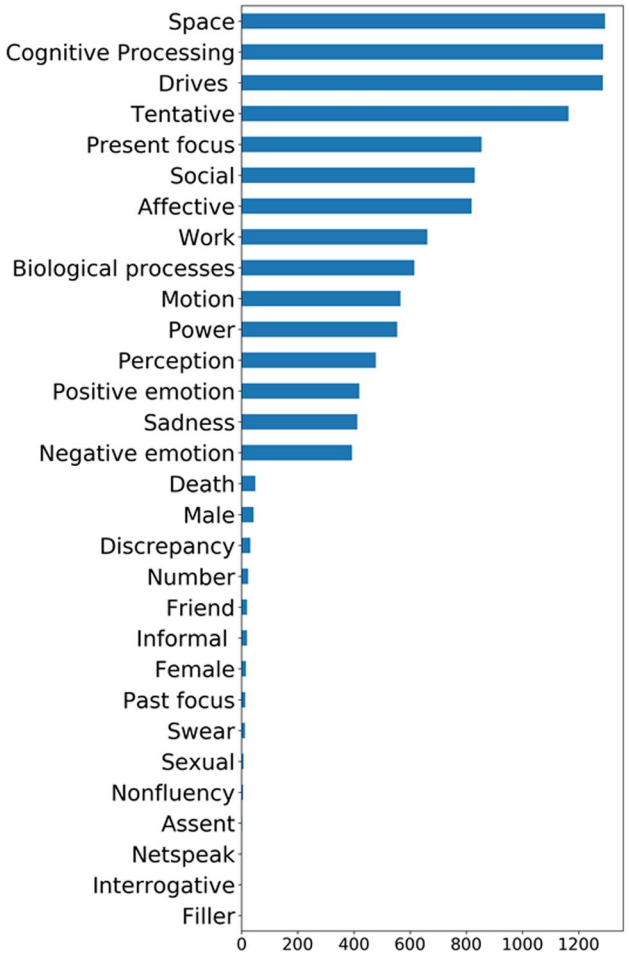

Fig. 4 Frequency of 15 most and 15 least common LIWC categories in the final set of valid behavior phrases. (Bars can exceed the total number of phrases because a category can appear multiple times in a phrase.) The figure excludes syntactic categories

[3] We do acknowledge that it is not altogether clear how often we should expect phrases of a certain topic to appear. Moreover, it is not clear that the token frequency of a topic in the phrase set has to reflect the centrality or frequency of a type of behavior in human life. Still, having only 7 of ~ 4000 phrases concern sexuality strikes us as severe underrepresentation.

[4] As a particularly striking example of this bias, the phrase *admire a man* is in our dataset, but *admire a woman* is not.

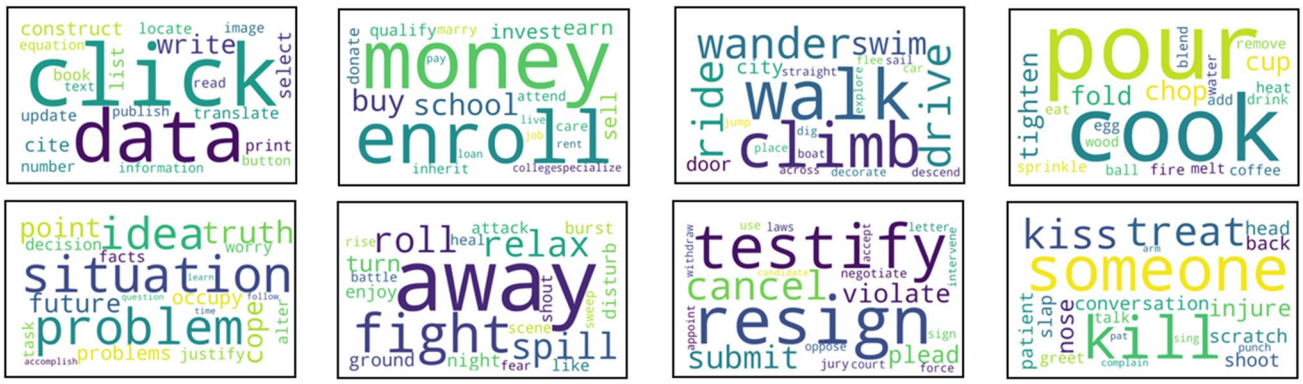

Fig. 5 Word clouds describing *k*-means clusters in our set of 3938 behaviors

dataset in particular, or even generic corpora in general. We return to this issue in the discussion.

For our second analysis to better understand the semantic content of our phrase set, we performed clustering of phrase vectors. First, we extracted 512-dimensional vectors for each behavior phrase using the Universal Sentence Encoder. We used this language model instead of, e.g., BERT, because USE obtains state-of-the-art performance on sentence similarity without fine-tuning, and clustering is similarity-driven (Cer et al., 2018). We then performed *k*-means clustering on all 3938 behavior phrase vectors, with $k = 8$. To determine a word's importance to a cluster, we performed the following procedure. First, we lower-cased all words and removed stop words and non-alphabetic tokens. Then, we counted the frequency of all words, and divided a word's frequency in a cluster by the sum of its frequency in all clusters, to obtain a word's relative importance to a cluster. Figure 5 shows word clouds for each cluster of the foregoing analysis, with words sized according to their relative importance to a cluster. These clusters appear to span diverse domains including digital actions, money and career-related behaviors, travel and physical activities, household tasks, problem-solving, and social activities, suggesting that our approach is able to uncover and quantify a wide range of common human behaviors.

Survey of Behavioral Propensities

In this section, we describe the methodology for our survey that was used to collect data on participants' propensity to commit certain behaviors, as well as their psychographic and demographic data. Our aim was to use this latter data, along with vectors from transformer models of the behavior phrases, to predict propensities to perform behaviors, both for out-of-sample behaviors and out-of-sample participants.

Participants

We recruited 319 participants on Prolific Academic. Our sample was chosen to be representative of the age, gender, and race distribution of the USA. Participants were only allowed to participate once and were paid approximately $10 per hour.

Psychographic Measures

We collected the following questionnaires to measure psychographic features of our participants: Ten Item Personality Inventory (Gosling et al., 2003), Domain-Specific Risk-Taking Scale (Weber et al., 2002), Barratt Impulsiveness Scale (Patton et al., 1995), Self-Report Altruism Scale (Rushton et al., 1981), Grit Scale (Duckworth & Quinn, 2009), Satisfaction With Life Scale (Arrindell et al., 1999), and Maximization Scale short (Nenkov et al., 2008).

Design and Procedure

In the first section of our study, participants were randomly assigned to a single block containing a subset of approximately 247 behaviors from the finalized dataset of human behaviors. There were 16 blocks of phrases—15 contained 247 behaviors and 1 contained 233 behaviors. Note that while the sample of participants was representative across multiple demographic variables, each block was not guaranteed to be evaluated by a representative sample. In total, there were 78,116 evaluations of individual phrases collected in the study.

Participants were told that we were interested in understanding how much they agreed with the statement "Relative to others, I am likely to *X*" where *X* was one of our

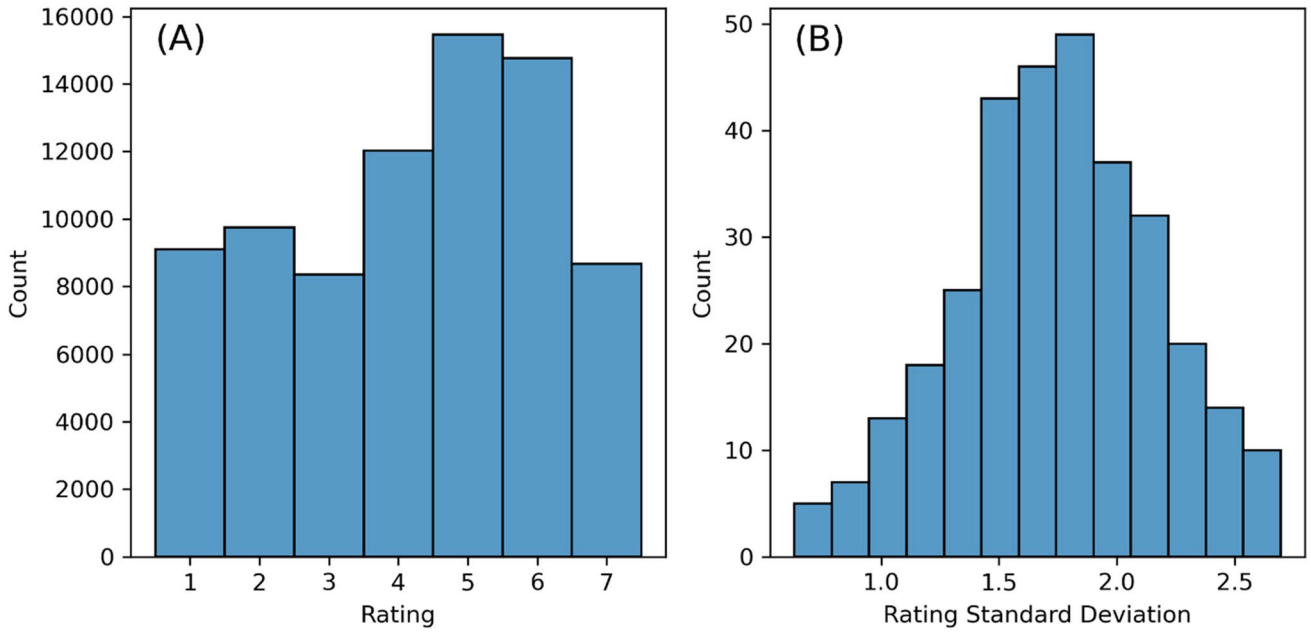

Fig. 6 Histogram of raw subject ratings (**A**), and histogram of standard deviations in ratings for each behavior phrase (**B**)

Table 2 Top: Phrases with the lowest (left) and highest (right) mean ratings. Bottom: Phrases with the lowest (left) and highest (right) standard deviation in rating across subjects

Phrase	**Mean rating**	**Phrase**	**Mean rating**
kill my father	1.05	*finish high school*	6.63
suck on a pacifier	1.16	*participate in a study*	6.50
burn houses	1.16	*answer a question*	6.47
abuse my children	1.19	*breathe again*	6.45
burn patients	1.20	*reflect on a subject*	6.42
steal from the poor	1.25	*think of someone*	6.37
commit violent crimes	1.25	*answer questions*	6.36
assist in an execution	1.25	*search for information*	6.35
cast someone into a furnace	1.30	*open the door for myself*	6.32
explode a bomb	1.30	*save a document*	6.32
Phrase	**Rating std**	**Phrase**	**Rating std**
kill my father	0.22	*grab a cup of coffee*	2.61
burn houses	0.37	*travel the world*	2.55
suck on a pacifier	0.37	*thank my wife*	2.55
abuse my children	0.40	*enroll in high school*	2.54
assist in an execution	0.44	*scare easily*	2.49
explode a bomb	0.47	*pray for myself*	2.49
commit a felony	0.48	*embrace the christian faith*	2.48
burn patients	0.52	*behave like a lady*	2.44
engage my attention	0.54	*proclaim the message of salvation*	2.44
commit violent crimes	0.55	*sing in a choir*	2.42

behavior phrases. Participants rated how much they agreed with this given statement on a Likert scale from 1 (strongly disagree) to 7 (strongly agree). Participants were told to compare themselves against the general population, rather than solely their peers, while evaluating the statements. Participants completed this task for all behaviors in their assigned block before moving on to the next portion of the study.

In the second section of the study, participants completed the psychographic questionnaires mentioned above. After

completing the psychographic questionnaires, participants indicated their education level, race, gender, income level, age, marital status, and employment status, in that order.

Summary of Phrase Ratings

We briefly report some descriptive summaries of the propensity ratings for our behaviors. First, Fig. 6A contains a histogram of all behavior ratings. The mean propensity rating for all behavior phrases was 4, and more importantly, the standard deviation of this distribution was 1.9, suggesting substantial variability that might be modeled. The top of Table 2, in contrast, shows phrases with the highest and lowest propensity ratings when averaged across subjects. Sensibly, the lowest rated phrases tend to be criminal (e.g., *kill my father*) behaviors while highly rated phrases are extremely mundane, universal behaviors (e.g., *open the door for myself*). To the extent that some behaviors generally tend to be rated low and others rated high, predictive models that only rely on phrase representations could suffice.

However, different subjects often rated the *same* phrase rather differently. Figure 6B shows the distribution of phrase-level standard deviations: on average, phrases had a standard deviation of 1.7 in their ratings across subjects. Phrases with this level of variation include behaviors like *yield to temptation* (1.8) and *abandon a plan* (1.5). It is plausible that variation in ratings for these behaviors might relate to individual differences in, say, self-control, grit, and/or impulsiveness. Similarly, the bottom of Table 2 shows the phrases with the most and least variability across subjects. Again, subjects tended to rate criminal behaviors similarly, but showed substantial variability in behaviors reflecting differences in personal taste (*grab a cup of coffee*), marital status (*thank my wife*), and religious affiliation (*embrace the Christian faith*). To account for individual differences in these types of behaviors, predictive models would need to rely on phrase representations, psychographic/demographic information, and their interaction.

Predictive Modeling of Behavioral Propensities

Methods

The primary goal of this paper was to evaluate the effectiveness of predicting behavioral propensities of individuals from phrase vectors provided by transformer models (BERT and USE). By collecting measurements of how likely individuals are to engage in certain behaviors, we were able to use different machine learning models to regress the behavioral propensity variable onto the vectors for the behaviors obtained from the transformer models. These regressions allowed us to learn the relationships between the vector space of the behaviors and the behavioral propensity ratings. Because this regression was being calculated using the behavioral propensities of multiple individuals, we hoped that individual-level psychographic and demographic variables might be useful covariates in this regression in order to predict behavioral propensity on the individual level.

We evaluated the success of both BERT and USE vectors for this task by training regularized ridge regressions with Scikit-learn (Pedregosa et al., 2011), and multilayer perceptions (MLPs) with Keras (Chollet et al., 2015) and TensorFlow (Abadi et al., 2015). As mentioned in the introduction, the hidden layers in the MLP may allow us to model interactions between participant characteristics and behavior phrase characteristics, e.g., the tendency for extraverted individuals to be more likely to perform social behaviors (*go to a party*) than solitary behaviors (*read a book*), and introverted individuals to do the reverse.

We used the phrase vectors, alongside psychographic and demographic data, as input features for the behavioral propensity prediction task. Individual-level psychographic data were given as either aggregated (i.e., each participant received a single scalar score for Grit, Agreeableness, Openness to Experiences, Satisfaction with Life, Risk Taking, Conscientiousness, Altruism, Impulsiveness, Maximization, Extraversion, and Emotional Stability using the scoring methods described in the sources of the questionnaires) or non-aggregated (i.e., each participant's score for each questionnaire item was used individually in the feature set). All ridge regressions were run with a grid search over alpha using 20 evenly spaced values on a log scale between e^5 and e^{-5}. All multilayer perceptrons (MLP) contained 4 layers: the first with 1000 neurons and a ReLU activation function, a hidden layer with 200 neurons and a ReLU activation function, another hidden layer with 50 neurons and a ReLU activation function, and a final layer with one neuron with a linear activation function to provide a propensity rating. Additionally, dropout was set to 50% between each layer. Each MLP was trained using tenfold cross-validation with 100 epochs per fold, where each epoch trained the MLP in batches of 20 items at a time. Ten percent of the training data was preserved as a validation set in order to avoid overfitting. We also introduced an early stopping method where the model in the current fold would end training early if validation loss did not improve for 10 consecutive epochs to avoid overfitting. As a baseline, we also tested a Word2Vec model with phrase vectors obtained from averaged word vectors (continuous bag-of-words; Mikolov et al., 2013) using the regularized ridge regression and MLP techniques.

In the following section, we show the results of these models under 3 different methods of splitting the dataset. The first is a true random split over all the data, meaning that neither behaviors nor participants are guaranteed to

Fig. 7 **A–I** Out-of-sample performance for every (model type, [non-] aggregate psychographic data, vector source) triple for each split of the dataset. Cells in the heatmaps indicate the Pearson correlation between individual evaluations of behavioral propensity and model predictions of behavioral propensity

be exclusively represented in either the training or testing dataset. The second splits the data by participants, guaranteeing that the models make behavioral propensity predictions for the test set with out-of-sample participants. The final method splits the data by behaviors, guaranteeing that the models make predictions on out-of-sample behaviors. In order to normalize the data for each participant, each rating was z-scored with the given participant's other ratings in the current testing or training set of the data and this rating was used in the given set instead of the raw rating.

Results

Figure 7A–I shows the out-of-sample correlation between actual propensity ratings and predicted propensity ratings for the ridge regression and multilayered perceptron models, using aggregated vs non-aggregated psychographic data, USE vs BERT vs averaged Word2Vec vectors, and the 3 different types of splits of the dataset.

In all cases, except when using averaged Word2Vec word vectors as phrase representations, MLPs had more accurate predictions of behavioral propensity ratings than ridge regressions trained on the same dataset with the same input features. For every MLP model—except the models utilizing Word2Vec vectors, which saw similar behavioral propensity ratings between both types of psychographic data—non-aggregated psychographic data yielded slightly improved performance (average $r = 0.441$) over aggregated psychographic data (average $r = 0.438$) as an input in the feature space. This effect did not appear

Table 3 Top phrases for groups based on the five psychographic and demographic variables with the greatest inter-rater reliability of group-based differences

Variable	Group	Top phrases				
Emotional stability	Below median Emotional Stability	*enroll in graduate school*	*worry a great deal*	*explode from pressure*	*vanish into the night*	*remain in school*
Emotional stability	Above median Emotional Stability	*dance with someone*	*strengthen my family*	*drive in the country*	*buy a suit of clothes*	*swim in a pool*
Marriage	(Formerly) Married	*qualify for a mortgage*	*decorate a room*	*wake up every morning*	*travel across the country*	*accompany a child*
Marriage	Never married	*criticize the actions of others*	*enroll in graduate school*	*ruin my life*	*stay up all night*	*sink to the bottom*
Grit	Below median Grit	*remember a name*	*attend school*	*enjoy life*	*solve an equation*	*quit a job*
Grit	Above median Grit	*publish a story*	*rub against someone*	*smoke in a room*	*wear a sword*	*struggle for recognition*
Gender	Female	*wear a dress*	*marry some man*	*behave like a lady*	*decorate a room*	*write in a book*
Gender	Male	*negotiate a price*	*marry someone's daughter*	*become a man*	*marry a woman*	*resemble a man*
Risk taking	Below median Risk Taking	*qualify for financial aid*	*diagnose a problem*	*pay the full amount*	*hug my mother*	*enroll in first grade*
Risk taking	Above median Risk Taking	*shoot a gun*	*spend a night with someone*	*spend a lot of money*	*flee a scene*	*seize my prey*

for the ridge regression models (average $r = 0.417$ vs $r = 0.417$). Psychographic and demographic data as a whole did not seem to impact model performance (average $r = 0.409$ vs average $r = 0.410$) across all models with and without psychographic or demographic data, respectively. Our best performing model, an MLP trained over a random test/train split of the dataset, performed only slightly worse when trained without psychographic data ($r = 0.516$) than with the psychographic and demographic data ($r = 0.524$). Without demographic or psychographic data, this model achieved a correlation of 0.506. It is notable that models trained with the vectors from transformer models alone still achieved high correlation values.

Excluding the models trained with averaged Word2Vec vectors, the models trained on a random test/train split of the dataset (Fig. 7A and B) outperformed models trained on splits over participants or behavior phrases. Of course, it must be noted that in a random test/train split of the dataset, the same phrase(s) or participant(s) (but not both) could appear in both the training and testing dataset. The models evaluated on ratings from behaviors that were entirely out-of-sample (Fig. 7D and E) performed worse than models evaluated on ratings from participants that were entirely out-of-sample (Fig. 7G and H) indicating that is harder to extrapolate to new behaviors than it is to new participants.

BERT vectors also yielded equal or improved model performance over USE vectors for all models except for the ridge regressions in Fig. 7D and E. This difference in performance is most notable in the MLP models in Fig. 7G and H where the MLP models trained with USE vectors had correlations that were, on average, 0.037 lower than the MLP models trained with BERT vectors.

Predictive Modeling of Group Differences

Methods

The previous analysis shows that we can predict individual-level responses, but that, contrary to expectations, psychographic and demographic measures do not improve predictive accuracy. One explanation of this result is that these measures are simply not predictive of propensities for the behaviors in our dataset. However, an alternative explanation is that our models are unable to learn the (likely very complex) interactions between phrase meaning and psychographic/demographic variables. To distinguish these possibilities, we conducted another analysis that simplifies the modeling problem. First, for each demographic or psychographic variable, we constructed two groups (e.g., men and women for gender, or participants above the median Grit score and participants below the median Grit score). Then, for each behavior, we calculated the mean propensity within each half, and subtracted the mean of one half from the mean of the other. For example, the mean propensity of *wear a dress* was 1.4 for men and 4.2 for women, yielding a difference of −2.7. The analogous difference of *wear a tie*,

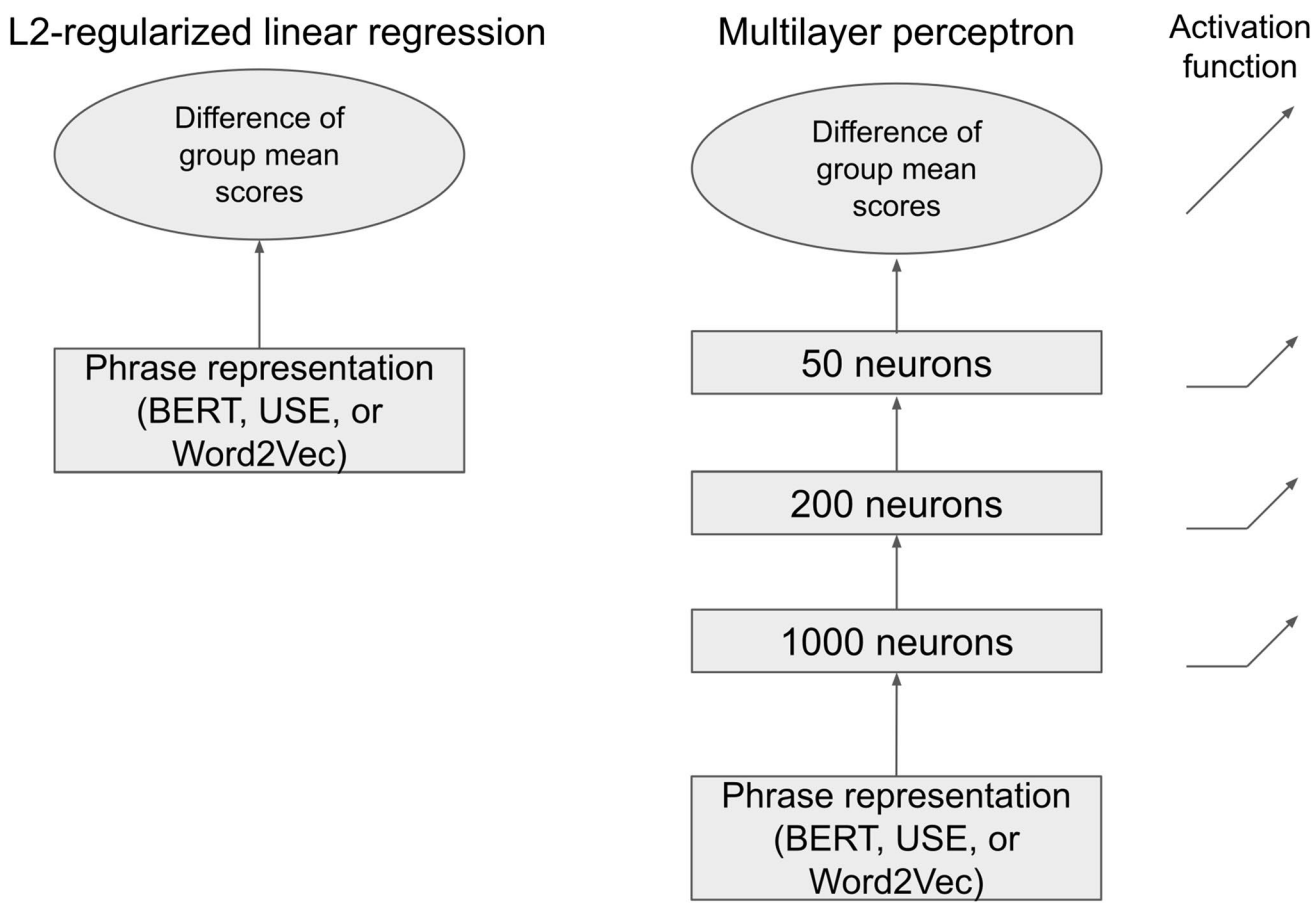

Fig. 8 Models of group differences in behavioral propensity. In contrast to the models of Fig. 3, reported in section "Predictive Modeling of Behavioral Propensities," these models predict group differences (e.g., female mean propensity minus male mean propensity for *wear a dress*), based *only* on phrase representations from BERT, USE, or Word2Vec

by contrast, was 2.9, reflecting the fact that men reported greater propensity for that behavior than did women.

Table 3 contains such phrases for the five psychographic and demographic variables with the highest reliability as calculated by split-half correlation (calculation described in the next section). As can be seen, sensible phrases appear for other groups, e.g., *explode from pressure* was rated higher by emotionally unstable participants than by emotionally stable participants, while *struggle for recognition* was rated higher by participants high on grit than by participants low on grit. At the same time, some phrases seem unrelated to their group, e.g., *enroll in first grade* for participants low on risk taking (unless participants interpreted the phrase to refer to enrolling someone *else*, like a child, in first grade). As we shall see in the "Results," odd top phrases may be due to issues of inter-rater reliability.

For each demographic or psychographic variable, we then attempted to predict these mean differences by training ridge regressions and MLP's similar to those in the previous section (see Fig. 8). However, instead of constructing test-train splits by random, participant, or behavior, we now just trained and tested models in tenfold cross-validation across all behaviors. In particular, we trained each ridge or MLP on a train set of 90% of the data (including gridsearch to find alpha, in the case of ridge regression), generated predictions of the mean differences for the remaining 10%, and repeated this for each of other nine train-test splits.

Results

Figure 9 shows the Pearson correlations between true and predicted mean differences, for each psychographic and demographic variable, for ridge and MLP, and for USE, BERT, and Word2Vec. First, we note that the ridge model typically does better than the MLP. This is likely due to the fact that there are not nonlinearities or interactions in this prediction exercise (by contrast, the individual behavior propensity prediction exercise involved potential interactions between the vector representations of the behavior and the vector representations for the individual). There are also fewer observations in the current prediction exercise (less than 4000 group differences for behaviors) vs the previous

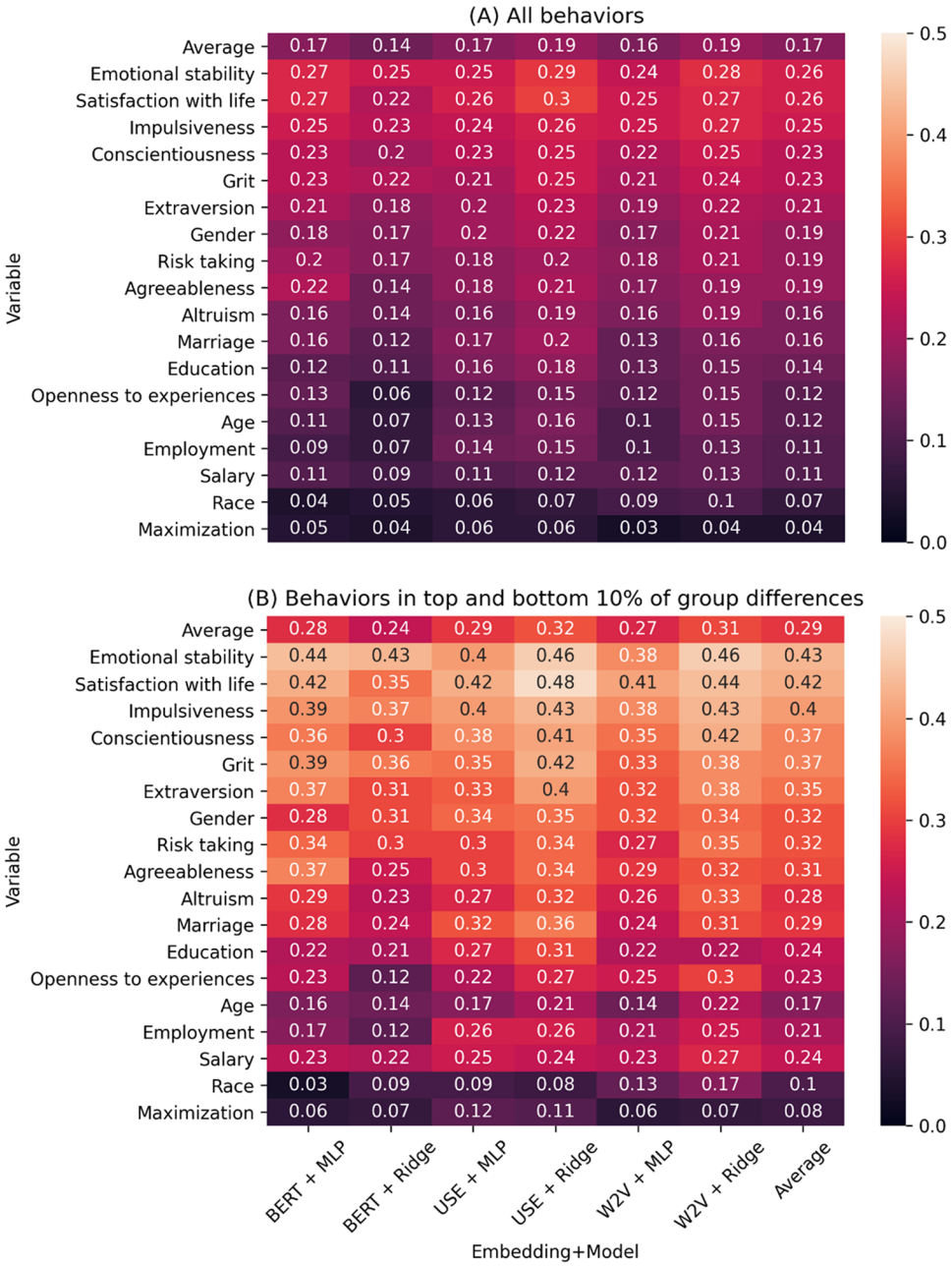

Fig. 9 Pearson correlations between predicted and true mean group differences in behavior propensity, for **A** all behaviors, and **B** behaviors showing large mean differences for a given variable. (Note that the color scale is different from Fig. 7, which also reports correlations.)

prediction exercise (which had more than 75,000 ratings). This implies that more complex models, like MLP, may have a harder time fitting the data in the current analysis.

We focus next on the results in Fig. 9A, which reports correlations based on all behaviors. As can be seen, modest correlations are generally possible, with a mean correlation of $r = 0.17$ across all embeddings and supervised model types, and minimal differences among these. Of course, this performance is lower than what we found when predicting behavior propensity ratings at the individual level in the previous section ($\sim r = 0.40$). Furthermore, we found large differences between psychographic and demographic variables, with Maximization-based differences predicted at only $r < 0.05$ and Satisfaction with Life predicted at nearly $r = 0.30$.

We therefore sought explanations of (a) the generally modest performance in predicting group-based differences, and (b) the large differences in predictability of different group differences. Because we only collected 20 ratings per phrase such that each group (e.g., low Maximization) might only have about 10 ratings per phrase (or fewer for minority groups like Black or African-American respondents), we suspected low inter-rater reliability was the primary culprit. We therefore calculated the split-half correlation

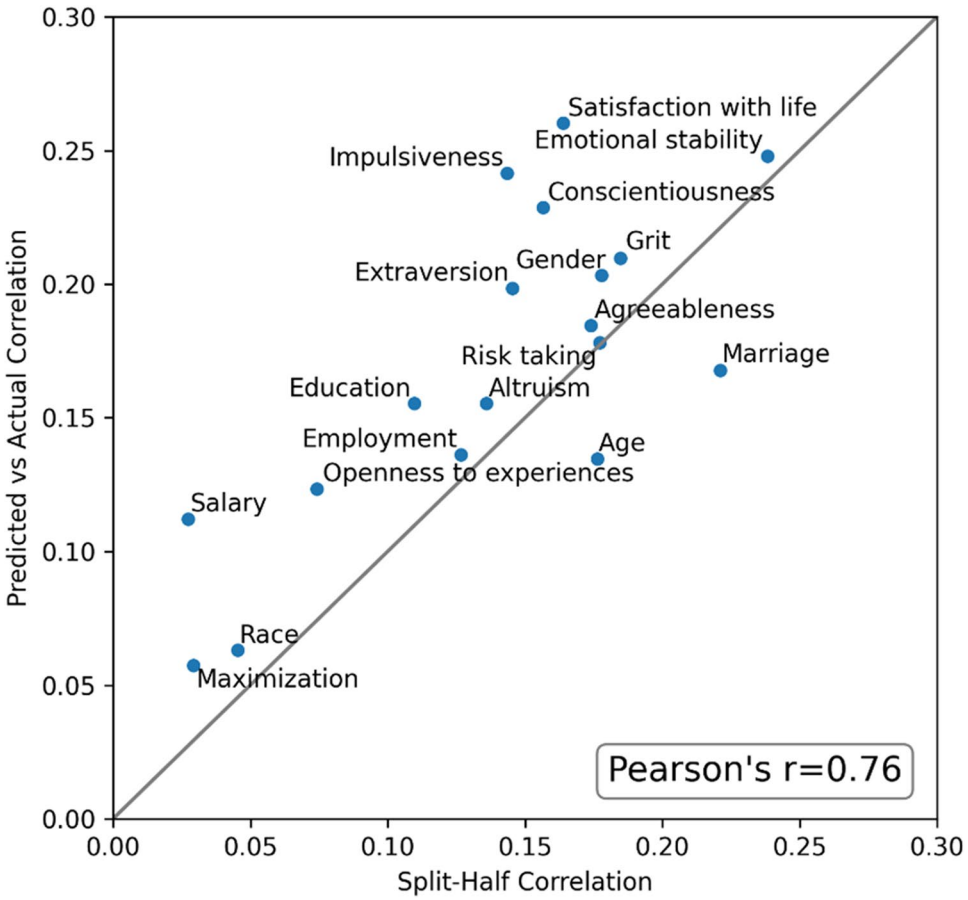

Fig. 10 Split-half correlation (*x*-axis) against the correlation between predicted and actual mean differences of behavior propensity for USE embeddings combined with MLP (*y*-axis), for every dimension

in group-based mean differences for every variable, as follows. First, for every phrase and group (e.g., low maximizers rating *invest in stocks*), we split the set of ratings in half and calculated means for each half. We then carried out analogous splitting and averaging for the complementary group (e.g., high maximizers rating *invest in stocks*), calculated differences of means between the first halves of each group and between the second halves of each group, and correlated these differences of means between the first and second halves. Each correlation was finally adjusted with the Spearman-Brown (Brown, 1910; Spearman, 1910) correction for split-half reliability, $r_{\text{corrected}} = (2*r)/(1+r)$. We conducted this process 10 times with different random splits into halves and averaged the correlations to obtain a single estimate of split-half correlation for every psychographic and demographic variable.

Figure 10 plots these split-half correlations against model performance (specifically, correlations between predicted and actual group differences, for the MLP trained on USE embeddings for behavior phrases). As can be seen, each variable's predictability of group-based differences is well-explained by its inter-rater reliability, reflected in the strong correlation between split-half correlation and model correlation ($r = 0.76$, $p < 0.001$). Furthermore, not only are reliability and performance strongly correlated, they are in fact nearly identical (as indicated by the tight clustering around the line $y = x$), suggesting that our models are generally performing as well as the (low) limits of inter-rater reliability will allow.

To give a sense of what performance might be possible if we had higher reliability in our data, we restricted attention to only those behaviors that showed strong group-based differences. To extract these behaviors, for each variable, we ranked behaviors by mean group difference, and then retained just the top 10% and bottom 10% of this ranking. This retains, for example, *wear a dress* and *wear a tie* for the gender variable, as these are among the behaviors at the top and bottom of the ranking of female-male mean differences, respectively, but drops *sleep in a bed* as this is in the middle 80% of the behaviors for gender. Figure 9B reports correlations only among these top 10% and bottom 10% behaviors. As comparing to Fig. 9A shows, performance is naturally higher when restricting analysis to only behaviors showing large group differences ($r = 0.29$ in behaviors with large differences, vs $r = 0.17$ in all behaviors). Of course, we believe the power of our approach is in automatically constructing a large, comprehensive set of behaviors, so manually curating behaviors in this way is both post hoc, and limits the generalizability of our approach.

Overall, this pattern of results suggests that psychographic- and demographic-dependent variations in behavioral propensities are predictable from phrase representations, at least when predicting group-level propensities (as opposed to individual-level propensities, as in the last section). Furthermore, our reliability analysis suggests our modeling approach might achieve better performance with more reliable data, which we suspect would emerge simply with collecting more ratings per phrase, or even through alternative scaling methods like Best–Worst Scaling (Kiritchenko & Mohammad, 2017). Hence, we suspect we have only scratched the surface of what is possible with our approach.

Discussion

Vector Representations of Behavior

The space of naturalistic human behavior is vast, and thus nearly impossible to comprehensively quantify and analyze. This is why most theories in the cognitive and behavioral sciences are parametrized and tested using highly stylized experimental tasks or surveys. However, in order to develop formal scientific theories of naturalistic human cognition and behavior, researchers need to be able to quantitatively represent the nearly limitless set of behaviors that people engage in on a day-to-day basis.

This project attempts to address this important conceptual and technical challenge. The core insight underlying our approach is as follows: Many naturalistic human behaviors can be described with simple natural language verb phrases and sentences. Using deep language models, such as transformers, the meanings of these phrases and sentences can be quantified as vectors in high-dimensional semantic spaces. Importantly, semantic vectors can be obtained for nearly any phrase or sentence, which implies that quantified representations are feasible for thousands of common human behaviors.

The ability to quantify naturalistic behaviors using high-dimensional vector representations opens up many new avenues of research in psychology and related disciplines. Specifically, it is possible to use quantified representations of behaviors as inputs into formal models that attempt to predict important psychological variables associated with behaviors. To facilitate such an analysis, we collected a dataset of naturalistic behaviors by observing the frequencies of verb phrases in natural language. We extracted hundreds of thousands of such phrases from the Google Books dataset, and then, through part-of-speech tagging, human coding, and manual editing, distilled this dataset into a subset of 3938 verb phrases that describe common human behaviors. We also trained a machine learning model—reported in detail in supplemental materials—that is capable of accurately predicting whether a given phrase describes a common behavior, automating this process for future research.

Predicting Behavior

The main test in this paper involved using our dataset of behavior phrases to predict people's propensities in engaging in these behaviors. For this, we collected a large dataset of individual-level behavior propensity ratings, as well as associated psychographic data (e.g., responses to personality surveys) and demographic data. We then used both the vector representations of behavior as well as psychographic and demographic variables for our participants as inputs in machine learning models trained to predict the individual's propensity rating. We found that our models achieved reasonable accuracy rates when predicting out-of-sample behavior propensities, including propensities for individuals not in the training data, and behaviors not in the training data, showing that transformer-based vector representations of behavior can be used to make behavioral predictions for truly out-of-sample individuals and behaviors. We considered both regularized linear regressions and multilayer perceptrons and found that the best performing model turned out to be the multilayer perceptron that used the BERT vectors as inputs. This is not surprising given the computational power of deep neural networks and recent successes of the BERT model in natural language understanding tasks.

However, we were somewhat surprised to see that phrase representations obtained by simply averaging word-level Word2Vec representations were not that far behind BERT (e.g., out-of-sample correlations of 0.52 for BERT vs 0.37 for Word2Vec on the random train/test split with non-aggregated psychographic measures). We suspect that BERT (and USE) present only moderate advantages over Word2Vec because our behavior phrases are rather short (no more than five words), and hence, there is little word order and grammatical information that needs to be manipulated. Instead, it may be adequate for a supervised model to predict behavioral propensity purely based on the topics or domains that the phrase represents, which Word2Vec easily represents through averaging. For example, for phrases like *sew a dress* or *invest in stocks*, vector averages would likely reflect feminine clothing actions/objects for the former, and financial behavior for the latter. Still, the superiority of the transformer-based representations suggests that Word2Vec is unable to represent at least some relevant information. Moreover, for behavior phrases and sentences longer than those tested here, i.e., those common to surveys measuring psychological and behavioral individual differences, we suspect that transformer-based representations will present even greater advantage over bag-of-words representations.

We were also surprised that our MLP did not grossly outperform a purely linear model (by no more than approximately $r = 0.05$), which is surprising to the extent that we think behavior propensity is an interactive and not merely additive function of the behavior and the individual (that is, we would expect extraverted individuals to endorse going to a party over reading a book, and introverted individuals to do the opposite). Similarly, we were surprised to see that the addition or removal of psychographic and demographic information from the inputs to the models did not have much impact on predictive accuracy (difference in $r < 0.01$). As mentioned in the "Results," one explanation of these null effects is that psychographic and demographic variables simply carry no information relevant for predicting behavior. However, we found that *differences* among psychographic and demographic groups—men vs women, or high impulsiveness vs low impulsiveness subjects—could be predicted with modest accuracy ($r = 0.17$) by building separate supervised models for each psychographic or demographic variable. Some variable-based differences could even be predicted with moderate correlation ($r = 0.3$, as in Emotional Stability and Satisfaction with Life), and what seemed to hold back greater accuracy for all variables was not impoverished representations or predictive models, but low inter-rater reliability of group differences.

Taken together, these results suggest to us that phrase representations can combine with psychographic or demographic information to predict behavior propensity, but that our primary behavior modeling approach is limited in some fashion. Our MLP may be overly flexible, with too many hidden layers and neurons, relative to the amount of data we have (78,116 participant-behavior combinations), and/or our input representations (phrase vectors and psychographic/demographic survey responses) may be too high-dimensional. Whereas the models predicting individual-level propensities were tasked with learning how *all* psychographic/demographic variables influenced propensity, the group-level models needed to only learn the effect of one variable at a time. Or, it may be that our number of participants per phrase (~20) was simply inadequate for effectively learning how our individual-level characteristics impacted propensity ratings, especially given that, as stated above, we expected interactive and not additive effects, with the former generally being more difficult to learn and requiring more data. We certainly suspect that the inter-rater reliability of the group-based differences was hampered by the fact that, in the best case, each behavior phrase would be rated by only 10 subjects from one group (e.g., men) and only 10 from the other group (e.g., women) of a particular variable split. Owing to random assignment of subjects to phrases, of course, a phrase will often be rated by (far) fewer than 10 members of a group.

Finally, it may also just be inherently difficult to attain higher accuracy rates than those obtained in our analysis. Eisenberg et al. (2019), for example, present evidence that surveys predict self-reported real-world behavioral outcomes only modestly, and with substantial heterogeneity. On the other hand, our behavioral propensity ratings and our psychographic and demographic measures are all self-reported survey measures, and many of the psychographic measures contain items that are very similar to our behavior phrases. For example, one of the items on the DOSPERT asks participants to rate their likelihood of performing the behavior *going camping in the wilderness*. Furthermore, to the extent that our psychographic scales generally have internal consistency, responses on one item predict responses on other items from the same (sub)scale. It is perhaps therefore surprising that responses to the psychographic measures do not help predict responses to our behavior phrases. One possible explanation for this may be that the behavior phrases simply concern domains of behavior that are generally unrelated to, and therefore cannot be predicted from, the domains of behavior, personality, and demography reflected in our psychographic and demographic surveys.

In any case, we suspect that the implementation of our approach could be improved. Further refining our phrase representations (though, e.g., dimensionality reduction) or predictive models, increasing the number of participant ratings per phrase, restricting modeling of individual-level ratings to phrases with strong (a priori expectations of) individual-level variation, or collecting additional or different psychographic and demographic information may all be directions for future research.

New Applications in the Study of Behavior

Our results show that transformer models of language can provide useful vector representations of behavior phrases. These representations may not capture the entirety of the meaning of the behavior phrase or all of the richness of the physical instantiation of the behavior, but they are a good first step towards modeling people's behavior propensities. Importantly, transformer language models can be used to quantitatively represent a wide range of naturalistic human behaviors, allowing for novel applications of cognitive and behavioral research that taxonomize, predict, and explain naturalistic human behavior, using formal computational models.

One such application could involve a more detailed analysis of BERT and USE vector representations of behavior. Our preliminary tests involving the *k*-means clustering of behavior phrases (shown in Fig. 5) reveal that our vector representations capture some intuitive distinctions between different behavioral domains. Further work could examine the dimensions of the vector space of behaviors in more detail,

and thus better understand how vector representations of behavior obtained from natural language data represent the content of behavior and the meaning of verb phrases depicting behavior. For example, it might be useful to conduct more systematic tests of the influence of different elements of a verb phrase on the BERT or USE vector, as we did in a preliminary fashion in the introduction with the phrases *paint a house*, *decorate a room*, and *rent a room*. That is, the verb, as the head of a verb phrase, ought to determine the location in vector space more than other parts of the phrase, except possibly in the case of light verbs in phrases like *do a review*, in which case *do a review* perhaps ought to be closer to *revise a paper* than it is to *do the cleaning*.

It may also be possible to use our approach to study sequences of behavior, specifically behaviors performed one after another over the course of the day (and perhaps observed using diary studies). Such sequences can be used to understand complex behavioral schemas and scripts that guide human action (e.g., Abelson, 1981). We can analyze these dynamics using transformer models calibrated for "sequence-to-sequence" prediction, as in the French–English translation example of Fig. 2. Such models use vector representations of sentences to learn dependencies between different sentences, and have been shown to be successful at next sentence prediction, machine translation, and other tasks in which an input sentence must be mapped onto an output sentence (Devlin et al., 2018). In our case, sequence-to-sequence models can be used to learn how behaviors performed at one point in time determine behaviors in the subsequent point in time, providing analytical rigor in the study of behavioral dynamics and cognitive schemas.

Finally, as we have discussed earlier in this paper, the general paradigm introduced in this paper can be applied to other variables of interest to psychologists. For example, instead of predicting people's propensities for different behaviors, it may be possible to predict people's judgments of behaviors. Such judgments are a key topic of study in domains such as risk perception and moral psychology, and our paradigm offers the promise of extending theories in these fields to the nearly unbounded set of behaviors that could be judged by individuals in the world.

Limitations

Our approach is of course not without limitations. Perhaps chief among these are the biases in our set of verb phrases resulting from their generation from corpora. For one, we used the Google Books n-gram corpus to extract phrases, which is known to over-represent certain text genres, like scientific publications (Pechenick et al., 2015). In turn, this may mean that certain scientific behaviors like *generate a table* are over-represented in our corpus, while more informal behaviors like *take a selfie* or *have sex* are under-represented. And of course, since we have used the English version of the Google Books n-gram corpus, our generated behavior phrases may under-represent behaviors important to non-English-speaking populations (which, of course, is most of the human population). Finally, we found that male words appeared in our behaviors almost three times as often as female words, despite the set of male words being smaller in our dictionary (LIWC), which may be a bias in not just the Google Books corpus, but many generic corpora (Johns & Dye, 2019). Therefore, obtaining more general and representative sets of human behavior phrases is a crucial goal for future research. Diary studies, in which participants write out the sets of behaviors they engaged in over the course of the day, may be one way to manually augment our automatically constructed set of behavior phrases. It may also be possible to use phrase structure grammars (or even probabilistic variants thereof), combined with a lexicon of common words with their grammatical class (and possibly semantic features, e.g., POSSIBLE-AGENT), to generate new verb phrases (which could then be filtered down into a set of valid human behaviors with the classifier we briefly reported in Building a Set of Common Behaviors and expand on in supplemental materials). As the number of possible phrases can be impractically vast with even (a) a relatively small lexicon and grammar and (b) limits on the length of the phrase or number of phrase structure rule applications, it would be important to intelligently sample from the possible productions such that the space of behaviors (e.g., as represented in USE or BERT space, or in terms of LIWC constructs) is efficiently covered with a relatively small number of phrases.

Conclusion

We have proposed a novel approach to studying naturalistic behavior. Our approach is not limited by artificial experimental tasks or narrow aspects of cognition and behavior pre-selected by psychologists. Rather it embraces the complexity of the real world, and attempts to study this complexity using novel techniques taken from machine learning and natural language processing. When applied to a large dataset of behavioral propensity ratings, our approach is able to achieve reasonable predictions for how likely individuals are to engage in behaviors in an out-of-sample manner.

The reader may note that our approach is not grounded in an established theoretical paradigm. The reason for this is that there is no current psychological theory that can accommodate the richness and variety of everyday behaviors. By quantifying and predicting everyday behaviors, and by extracting insights regarding naturalistic behavior in a data-driven manner, this paper lays the groundwork for such a theory (see Yarkoni & Westfall, 2017; Hofman et al.,

2017 for discussions of the value of prediction in social and behavioral science). In doing so, it shows the power of computational models trained on large-scale digital data for analyzing and predicting behavioral phenomena (Griffiths, 2015; Harlow & Oswald, 2016). We look forward to the use of such an approach in the development of a new scientific paradigm, one that is capable of quantitatively describing the naturally occurring and free-flowing behaviors humans engage in over the course of their everyday lives.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42113-021-00121-2.

Funding Funding was received from the National Science Foundation grant SES-1847794.

Availability of Data and Material Additional materials can be found at https://osf.io/93nfb/.

Code Availability Code is available from authors upon request.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 2015. URL https://www.tensorflow.org.

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36*(7), 715–729.

Alammar, J. (2018). *The illustrated transformer*. Retrieved from https://jalammar.github.io/illustrated-transformer/.

Arrindell, W. A., Heesink, J., & Feij, J. A. (1999). The satisfaction with life scale (SWLS): Appraisal with 1700 healthy young adults in The Netherlands. *Personality and Individual Differences, 26*(5), 815–826.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition, 179*, 71–88.

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science, 65*(8), 3800–3823.

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

Bhatia, S., Olivola, C., Bhatia, N., & Ameen, A. (2021). Predicting leadership perception with large-scale natural language data. *Leadership Quarterly*.

Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making, 1*(1), 33–47.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprintarXiv:1508.05326.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296–322.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*(5), 938–956.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Strope, B. (2018). Universal sentence encoder for English. In *Proceedings of EMNLP* (pp. 169–174).

Chollet, F., et al.. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–190.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment, 91*(2), 166–174.

Eisenberg, I. W., Bissett, P. G., Enkavi, A. Z., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications, 10*(1), 1–13.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37*, 504–528.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21–23.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science, 355*(6324), 486–488.

Hollis, G., Westbury, C., & Lefsrud, L. (2016). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology, 70*(8), 1603–1619.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Johns, B. T., & Dye, M. (2019). Gender bias at scale: Evidence from the usage of personal names. *Behavior Research Methods, 51*(4), 1601–1618.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 232–254). Oxford University Press.

Kiritchenko, S., & Mohammad, S. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 465–470).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latentsemantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*(2):211.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy, 33*(3), 335–343.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL* (pp. 3428–3448).

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Nenkov, G. Y., Morrin, M., Schwartz, B., Ward, A., & Hulland, J. (2008). A short form of the Maximization Scale: Factor structure, reliability and validity studies. *Judgment and Decision Making, 3*(5), 371–388.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51*(6), 768–774.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE, 10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of liwc2015 (Tech. Rep.).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP* (pp. 1532–1543).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.

Rushton, J. P., Chrisjohn, R. D., & Fekken, G. C. (1981). The altruistic personality and the self-report altruism scale. *Personality and Individual Differences, 2*(4), 293–302.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*(5), 1178–1197.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science, 44*(6), e12844.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*(4), 263–290.

Xiao, H. (2018). bert-as-a-service. Retrieved from https://github.com/hanxiao/bert-as-service.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.

Zou, W. & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition*, 104647.