Distribution Agreement

Matthew G. Rubin December 1, 2010

Predicting Brain Activity Associated with Complex Nouns: Designing an Incentive Compatible Mechanism

by

Matthew G. Rubin

Dr. Gregory Berns
Adviser

Department of Economics

Dr. Gregory Berns

Adviser

Dr. Michael Prietula

Committee Member

Dr. Emily Hamilton

Committee Member

December 1, 2010

Predicting Brain Activity Associated with Complex Nouns: Designing an Incentive Compatible Mechanism

By

Matthew G. Rubin

Dr. Gregory Berns

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Economics

2010

Abstract

Predicting Brain Activity Associated with Complex Nouns: Designing an Incentive Compatible Mechanism

By Matthew G. Rubin

New methods of neuroimaging and machine learning have recently been utilized to reveal underlying neural dimensions of simple noun representation. One bottom-up procedure, multi-voxel pattern analysis (MVPA), predicts cognitive states by detecting spatial patterns in brain data and correlating differences in neural activity with behavioral responses. This technique has proved successful with simple nouns like tools and vegetables, but has never been attempted with complex nouns which incentives may exist for subjects not to be truthful. We used functional MRI (fMRI) to investigate the neural representation of Identities, or labels that describe people, and whether an Identity's meaning can be characterized by its association with actions or attributes. We then trained 3 classifiers to differentiate between Identities based on subjects' ratings of an Identity's actions and attributes, as well as subjects' sentiments about how good or bad each Identity was. Although the classifier results varied tremendously by Identity word and had slightly inflated accuracy level significance because the training and testing data were not independent, the many innovations that were introduced foreshadow an optimistic future for pattern analysis classification.

Predicting Brain Activity Associated with Complex Nouns: Designing an Incentive Compatible Mechanism

By

Matthew G. Rubin

Dr. Gregory Berns

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Economics

2010

TABLE OF CONTENTS

FIGURES AND TABLES

INTRODUCTION

Economic theory provides a conceptual structure for the way people make decisions based on the relative desirability of each choice. These models traditionally assume that the world is inhabited by *homo economicus*: rational, unemotional, self-interested people who know their preferences. Foundational assumptions have led to many fundamental economic theories; however some of these concepts, such as subjective utility theory and Nash equilibrium, often fail to explain human behavior. (Breiter et al. 2001, Gigenreizer and Selten 2002, and Luce 2000 are just some examples) In the early 1970s, economists began working with psychologists to develop the field of behavioral economics, which led to the creation of several new decision-making models. (Camerer 2003, Rabin 2002) More recently, neuroscientists have joined them, and the emerging discipline of neuroeconomics has attempted to explain *why* humans make decisions, not just *how*. (for a nice introduction, see Glimcher and Rustichini 2004) Neuroeconomics offers these explanations by adding neurological evidence to support its conclusions; some neuroeconomic hypotheses justify behaviors which violate traditional economic theory. (Camerer, Lowenstein, and Prelec 2005) Testing existing economic models with these new strategies and techniques could lead to the development of a new global theory of choice.

Neuroeconomics explores many traditional economic concepts such as preferences, risk aversion, game theory and strategic choice (Prisoner's Dilemma, Ultimatum Game, Stag Hunt, etc.), reward, altruism, trust, and utility. (for an overview, see Kenning and Plassman 2005) Whereas economic theory maintains that every decision an individual makes is utilitarian (a person weighs the costs and benefits), recent neuroeconomic research by Berns et al. (2010) suggests that decisions regarding sacred values are processed in a rule-based fashion. A personal value is defined to be sacred when it trumps other values, particularly material or economic incentives. This means that individuals resist trade-offs between sacred values and other values. (Baron 1997) Examples of sacred values include moral norms, cultural (religious, ethnic, national) identities, and fundamental religious beliefs. Individuals do not “choose” between a sacred value and another option, as is done in utilitarian decision-making when relative value is computed and compared; instead, people follow sacred values because “it is the right thing to do,” irrespective of the anticipated outcome of the situation. (Tetlock 2003)

In addition to driving many decisions in life, these values can potentially impact our impressions of any situation through something as subtle as the framing of a sentence. To further investigate how the brain interprets sacred values, we decided to deconstruct the value statements that Berns used in his experiment. The statements in Berns’ experiment were complete sentences in the second person – e.g. “You believe in God” and “You are willing to kill an innocent human being.” Because the subject of the sentence was always “you,” the sacred value was conveyed through the predicate. The neural representation of the entire predicate, however, is still far too complex; we therefore decided to explore the representation of a single word: the object of the sentence. We adopted some of the conceptually complex objects that Berns used and then added some similar nouns as well.

To reveal the dimensions of noun representation in the brain, new methods of neuroimaging and machine learning have recently been applied. This growing literature has made it clear that concrete nouns’ neural representations require multiple brain areas that are specialized for different types of information. (Just et al. 2010) Conceptually however, the literature is all quite similar: it all deals with simple nouns (usually physical objects) which are easily imaginable. Until recently, these experiments have also always used pictures of the objects as stimuli. Just (2010) was the first to demonstrate that the representation of a concrete noun can be accurately identified in the absence of a picture (participants were shown a word and told to imagine its properties and characteristics).

No one has yet investigated the neural representation for complex nouns, or nouns whose meanings vary depending on an individual's interpretation and beliefs. In this study, we focus on "Identity" words, which we define to be nouns that group people into categories (such as racial, sexual, familial, religious, and occupational). Identities are thus much more complex than physical objects because everyone has unique mental pictures of other people.[1] Whereas simpler, concrete nouns have been shown to be represented similarly by different people, (Just et al. 2010) the underlying dimensions that represent an Identity (which are unknown) can vary significantly depending on what an individual includes when imagining that Identity.

This imagination process is subject to an individual's preferences. Preferences affect not only the interpretation of a specific Identity (i.e. does the participant like a "Muslim"), but also what becomes a part of the imagination process (i.e. what is important in conjuring an image of a "Muslim," such as personality traits, physical attributes, potential actions or lifestyle choices, etc.). Preferences have traditionally been discovered because they arise among choices: when the values of various options are compared, pursuing the option of greatest relative value reveals an individual's preference. Recent research however, suggests that preferences also exist in the absence of a choice paradigm and that these passive, "choiceless" brain responses are still predictive of future behavior. (Wunderlich 2010, Berns et al. 2010) Furthermore, even though choices can vary, the underlying preferences may remain consistent. (Berns et al. 2008)

Because influencing preferences would impact how an individual interprets an Identity, entirely uninhibited imagination was required while participants were in the scanner. Our experiment thus could not involve a choice framework. Instead, participants were just asked to conjure a mental picture of each

[1] Person A views "Muslim" differently than person B for a plethora of psychological, historical, and neurological reasons. Similarly, Identities can overlap, such as the representation of a "Muslim" and an "Arab" (and then a person could be defined by other Identities too, such as that "Muslim Arab" also being a "Man," "American," and "Lawyer." Compared to an "apple," identifying the neural representation for such Identity words is thus far more complex.

Identity. To ensure that we did not constrain or assist in the creation of participants' representations in any way, our stimuli were not even pictures; the word-only design that Just (2010) pioneered was implemented.

The complex neurological mechanisms that represent Identities may be unknown, but certain underlying neural dimensions have been shown to exist in the representation of simple nouns. (Just et al. 2010) This dimensional searching requires a bottom-up analytic procedure to discover distinct spatial patterns across the brain, which have then been used to ascribe "meaning" to physical objects; (Mitchell et al. 2008, Just et al. 2010) "meaning" has been characterized by verbs of perception and action that relate to the object. But what does "Muslim" "mean?" We hypothesized that an individual interprets the meaning of an Identity based on a combination of the Identity's actions and attributes. We therefore devised semantic questionnaires to rate the impact of both dimensions (actions and attributes).

An alternative explanation offered to assign meaning is founded in Affect Control Theory (ACT), which models social behavior and its relation to personal attitudes and feelings. According to ACT, people generate emotional responses specific to the situations they are in, and if these feelings do not appropriately reflect the situation, then the individual can update them (Heise 1979). Work by Osgood (1957) shows that three basic dimensions are involved in generating affective responses: Evaluation, Potency, and Activity (EPA); the presence of these three aspects has been shown to exist around the world (Osgood, May, and Miron 1975). Although they are psychological measurements, the EPA sentiments relate to social life by quantifying status, power, and expressivity (Heise 1987). Evaluation can be classified on a scale between opposites such as good-bad, nice-awful, happy-sad, beautiful-ugly, and pleasant-unpleasant. Potency corresponds to contrasts like powerful-powerless, strong-weak, big-little, and high-low. Activity, the final dimension, measures affective arousal and rates words between opposing pairs such as fast-slow and lively-quiet.

The novel design of this experiment presents methodological problems, an obvious one being the issue of self-reported ratings. Although insightful, these reports can be contextually influenced; if a belief is perceived to be socially unacceptable, for example, an individual has an incentive to be dishonest. A new approach to overcoming the potential issues with self-reporting utilizes functional brain imaging. (Krajbich et al. 2009) Known as a neurally informed mechanism (NIM), this design makes payment for participating in the experiment depend on the successful comparison of participants' self-reported values to an algorithmic prediction based on their brain imaging data. Participants were informed that payment was dependent on this comparison, and this induced truth-telling nearly 100% of the time. The only requirement for such levels of truthful reporting is for participants to believe that the NIM can work with a certain degree of accuracy.[2] Krajbich's experiment involved the free-rider problem, a scenario in which dishonesty could entitle participants to more than their allotted portion of a public resource. Unlike this incentive compatible design,[3] in which honest self-reporting is a utilitarian decision, it is possible that sacred values influence an individual's beliefs about an Identity. If sacred values do affect Identity perception, not only would a participant's view be steadfast, but she may inherently understand the cost of being truthful, making the "choice" to hide her true feelings not really a choice at all. Participants would not evaluate self-reporting in a utilitarian fashion, but would instead process this decision in a deontic manner. If this is the case, not only would truth telling be suboptimal, but the introduction of a NIM may not prove to be an incentive compatible design.

Neural representation is not limited to transforming honest self-reporting into a dominant strategy and differentiating between utilitarian and deontic decision-making; it is the principal component in all neuroeconomic experiments that enables the field to step outside the bounds of traditional and even

[2] Their classifier algorithm had a 60% accurate prediction level, which is only slightly higher than chance (50%), especially when compared to their 97% accurate self-reporting rate.

[3] A mechanism is considered incentive compatible if the dominant strategy for a player with private information is to tell the truth. In the free-rider situation, the optimal strategy for a player is to reveal less than one's true value; only after the NIM is introduced does being honest become dominant.

behavioral economics. Neuroeconomics utilizes functional Magnetic Resonance Imaging (fMRI) technology to assess the roles of various brain areas during decision making. fMRI estimates local blood flow from tens of thousands of distinct neuroanatomical locations (voxels), and this blood-oxygen-level-dependent (BOLD) signal has been shown to reflect stimulus driven neural activity. (Logothetis et al. 2001) By focusing on the activity of specific regions of interest (ROIs) during a task, conventional activation-based analysis aims to infer the involvement of those regions during certain mental functions (by detecting regional-average activation differences). Pattern-information analysis, by contrast, infers representational content by detecting activity-pattern differences, which can still occur in the absence of changes in regional-average activation. (Mur 2009) To do this, a classifier algorithm is trained to indicate or predict different cognitive states from the combinatoric patterns that arise among the multi-voxel responses. (O'Toole et al. 2007) A key benefit of this method, known as multi-voxel pattern analysis (MVPA), is the ability to correlate activity pattern estimates with behavioral measures. (Norman 2006)

One of the foundational assumptions of the MVPA approach is that specific cognitive states consist of multiple underlying dimensions which are represented by different neuronal firing patterns. (Norman 2006) Because this pattern analysis can be undertaken without specifically knowing how many or what dimensions exist (Just et al. 2010), MVPA is particularly appropriate for an exploratory examination of the representation of Identities. Whereas traditional ROI analysis would ask if the brain encodes different Identities in similar, distinct regions (and if so, which), MVPA allows us to explore the general process of Identity representation. Using this, we could then assess the truthfulness of participants' responses. To find out if there are meaningful fluctuations in how participants represent different Identity words, we must ask: How well do participants' neural data of imagining an Identity correlate with their respective ratings of that Identity's actions and attributes, or their perception of how good or bad the Identity is?

To answer these questions, we labeled time periods of participants' scans by which Identity was displayed, and trained different classifiers to discriminate based on participants' various ratings for each Identity. In the first step of MVPA analysis, *feature selection*, the voxels are reduced to just those that will be included in a classifier. Experimenters typically look at regions that have traditionally related to the experimental paradigm, then select certain voxels within that set that are significantly active. (select examples include Haxby et al. 2001, Hanson 2004, Polyn et al. 2005, Quamme 2010) We instead wanted to see if a classifier could be trained on a set of voxels, regardless of specific location. To do this, we selected voxels maximally active throughout the entire brain (over the course of all time points when stimuli were shown). A step-by-step description of the entire MVPA analysis is given in the Methods section.

Three classifiers were then tested. First, using the most powerful sentiment of ACT, Evaluation, a classifier was trained to predict if participants considered words "Good" or "Bad." Our hypothesis that an Identity is represented by a combination of its actions and attributes was then put to the test. To do this, we trained a different classifier to see which dimension, action or attributes, better predicts neural activity. Then, we trained a final classifier to see if the richness of our semantic ratings correlated better with neural data than the sentiment of Evaluation. Ideally, the semantic rating scales we devised would be stronger predictors than Affect Control Theory.

In summary: In the absence of any pictures, participants imagined various Identities. To investigate the neural representation of such complex words, classifiers were trained to discriminate between patterns among the fMRI data and their correlating behavioral responses. Because potential incentives exist for participants to hide their true feelings, a monetary compensation will be included to induce higher levels of truthful self-reporting.[4] This experiment introduced many innovative methods, such

[4] Because this experiment finished as pilot data that consisted of three subjects, the monetary incentive for successful matching of neural and behavioral data was never introduced. All subjects were lab members, not individuals from the greater Emory community.

as the passive nature of the task,[5] a stimulus set of complex nouns, displaying stimulus words instead of pictures, feature selection on whole-brain data, and testing the incentive compatibility of a NIM on stimuli potentially influenced by sacred values. The findings here thus constitute several types of advances, even though the methods applied will require further calibration in future studies.

MATERIALS AND METHODS

Experimental Paradigm and Task

Three adults (two females) from the Center for Neuropolicy participated in the task. The handpicked stimuli were 40 Identity words, each a label that describes a person.[6] Identities included racial, sexual, familial, occupational, and religious groups in order to represent a broad categorical range of people. The 40 Identities were each presented once. Each word was shown for 7 seconds, during which period the participants were instructed to imagine the Identity (Fig. 1). The sole task of a participant was to create a vivid mental picture for each Identity, without any experimenter instruction on what this image should include; participants were free to choose any properties/characteristics for each Identity. Participants were also instructed to do their best to remember the specific images they conjured because they would later be asked to rate each Identity on questionnaires. The specific rating systems were intentionally withheld from the participants so that the images conjured would not be biased towards or by those scales. This 'contemplative duration' was followed by a jittered rest period (time $\in \{1,2,\ldots,9\}$, mean =

[5] Imagination is quite an active mental process. The "passive nature" refers to the "choiceless" design.
[6] See Appendix for a list of all Identity words selected.

4s). After the rest period, participants were asked to solve a basic arithmetic problem[7] to clear their mind and ensure that their image of an Identity was not carried over into the subsequent Identity's images. After the math problem was answered, another jittered rest period occurred (same distribution and mean) before the next Identity was shown. After the entire scanner task was completed, participants rated each Identity on three questionnaires, to be discussed below.

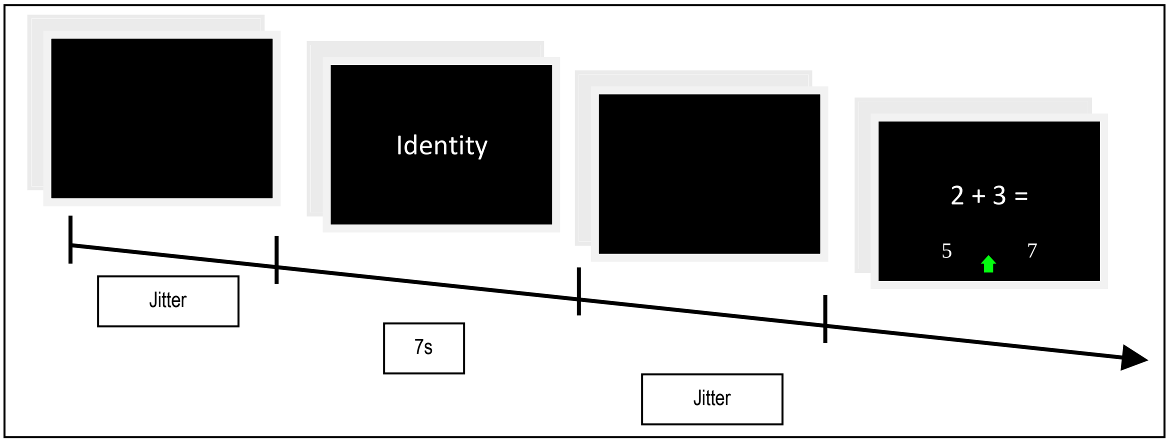

Fig 1 | Experimental design in the scanner. Subjects were instructed to imagine the Identity, then solve a basic arithmetic problem to clear their minds between words.

fMRI Data

Neuroimaging data were collected using a 3 Tesla Siemens Magnetron Trio whole body scanner (Siemens Medical Systems, Erlangen, Germany). A three dimensional, high-resolution anatomical data set was acquired using Siemens' magnetization prepared rapid acquisition gradient echo (MPRAGE) sequence (TR of 2600 ms, TE of 3.02 ms, TI of 900 ms, 1mm isotropic voxels and a 256mm FOV). A DTI scan was obtained with diffusion-sensitizing gradient encoding applied in 12 directions with a diffusion-weighting factor of b = 1000 s/mm^2, and one (b0) image was acquired without a diffusion gradient (b = 0 s/mm^2). Four

[7] For example, "2 + 3 =" Problems were in the form of the sum of two randomly selected numbers, m,n, such that $m,n \in \{1,2,\ldots,9\}$.

sets of each image were acquired, to be subsequently averaged. Functional data consisted of thirty-three axial slices that were sampled with a thickness of 3.5 mm and encompassing a field of view of 192 mm with an inplane resolution of 64 x 64 (T2* weighted, TR = 2000ms, TE = 30ms). The task was presented with Presentation software (Neurobehavioral Systems, Albany, CA), and visual stimuli were projected onto a frosted glass screen, which the participant viewed through an angled mirror mounted to the head coil. Inhomogeneities in the magnetic field introduced by the participant were minimized with a standard two-dimensional head shimming protocol before each run and the anatomical data acquisition.

fMRI Analysis

fMRI data were analyzed using SPM5 (Wellcome Department of Imaging Neuroscience, University College London). Data were subjected to standard preprocessing, including motion correction, slice timing correction, normalization to an MNI template brain and smoothing using an isotropic Gaussian kernel (full-width half-maximum = 8mm). Although not always used in MVPA analyses, spatial smoothing can increase the signal-to-noise ratio and has been found to increase classification accuracy by improving large-scale spatial pattern detection. (Rissman 2010)

Voxel Selection

The analyses below focused on a small subset of all the voxels in the brain. Before the MVPA classification began, brain maps were created for each analysis using the first-level model contrasts of each subject's semantic or Evaluation ratings. Using each rating as a parametric modulator, voxels were restricted to those that passed a test for joint significance (on either 'good' and 'bad' Evaluation scores or ‘positive’ and ‘negative’ scores representing action and attributes[8]). These subject-specific maps thus focused on only voxels that were significant at a threshold level of $p<0.05$ with $k\geq 60$.

Creation of the Questionnaires

In order to create contrasts with parameters based on each rating score, the questionnaires had to be created to describe semantic and EPA ratings. Participants rated each Identity on the three sentiments detailed in many Affect Control Theory experiments: Evaluation (good vs. bad), Potency (powerful vs. powerless), and Activity (active vs. inactive). (Osgood 1975, Heise 1987, Heise 2002) Each rating was on a 1-9 scale, with 1 being infinitely negative on each dimension, 9 being infinitely positive, and 5 being neutral. Because no standard rating system exists to describe the actions and attributes of an Identity, a series of 8 verbs and 8 adjectives were selected (see Questionnaire Word Selection) to represent Identity action and attributes, respectively. Participants were asked how often they would associate each verb and adjective with every Identity word. The rating scale for measuring both verb (actions) and adjective (attributes) association was 1 to 5, with each respective number equating to an association of “not applicable,” “rarely,” “slightly,” “somewhat,” and “frequently.”

Questionnaire Word Selection

Text Corpus Data

The text corpus data contains approximately 15 million words and was provided by the American National Corpus project. (http://www.americannationalcorpus.org/index.html) It consists of sets of *n*-grams

[8] The scores that represent action and attributes were from the semantic questionnaires we designed. See the “Creation of the Questionnaires” section below.

(sequences of *n* words) ranging from unigrams (single words) up to five-grams (sequences of five tokens). The database also includes counts of the number of times each *n*-gram appears in the large corpus.

Verbs

The 600 most common verbs were extracted (excluding “was” and “were”), and the database was parsed for combinations of those 600 verbs with the 40 chosen Identity words. These combinations were from either the 2-gram datasets, in the form “noun verb,” or 3-gram datasets, in the form “noun * verb,” where * was not “is,” “was,” or “were.” 2-grams and 3-grams which had the verb precede the noun were ignored because these would make the Identity word the object instead of the subject of a sentence in the corpus. The total occurrences of phrases (regardless of capitalization) were summed, then normalized by dividing these sums by the noun occurrences (again regardless of capitalization), which were tabulated from the 1-gram dataset. The data was imported into SPSS (IBM corporation, Somers, NY) in 8 sets of 75, and a stepwise discriminant analysis (Wilks’ Lambda, F to enter = 3.84, F to remove = 2.71) was run. 77 verbs met these criteria, and the “best” 15 verbs were hand-picked from this list of 77. 9 participants rated the 40 identities on these 15 verbs, and an exploratory factor analysis based on a principal component analysis with varimax rotation was performed in SPSS. The list was reduced to the 8 verbs which accounted for 82% of the total variance.

Adjectives

The 600 most common adjectives were extracted separately from the ANC database. Comparative and superlative adjectives were ignored (“green” was included, but “greener” and “greenest” were not). The corpus was parsed for 2-grams of the form “adjective noun” and “noun adjective,” and was similarly parsed for the 3-grams of the form “noun * adjective” and “adjective * noun.” The total occurrences were again summed and normalized by dividing the sums by the noun occurrences (again from the 1-gram dataset). The data was imported into SPSS in 4 sets of 150, and a stepwise discriminant analysis (Wilks' Lambda, F to enter = 3, F to remove = 2.71) was again run. Of the 37 adjectives that made these criteria, the "best" 15 were hand-picked. The same 9 participants also rated the 40 identities on these 15 verbs, and a similar factor analysis showed that 8 adjectives accounted for 77% of the total variance. These were the 8 adjectives selected.

Word Check

10 other participants were asked to rate all of the Identities on the final 8 adjectives and 8 verbs that had been selected to validate that the words did not need to be further reduced. Discriminant analyses were performed (Wilks' Lambda, F to enter = 3, F to remove = 2.25) and the rating scales were confirmed.

Machine Learning Methods

Overview

Pattern classification analyses were implemented in MATLAB using code from the Princeton MVPA Toolbox (www.csbmb.princeton.edu/mvpa). There were three stages in the machine learning techniques: algorithmic selection of voxels to be used in classification, training of a classifier on a subset of the data, and lastly, testing the classifier on another subset of the data. As is typical in MVPA experiments, the pattern analysis was trained on a particular participant's neural data—having been classified by that participant's behavioral responses—then tested on data from that same participant. Translating brain data between participants (training a classifier on participant A and testing on participant B) with MVPA is suboptimal (Norman 2006), so classifier strength was determined on each participant separately.[9] In this

[9] Cross-subject models of classification are rare because they are still quite in their infancy. For one of the most successful applications, see Just et al. (2010)

experiment, an across-subject classification attempt would have been quite difficult to run[10] and even more difficult to interpret,[11] so classifiers were not compared across subjects. Significance implied that the null hypothesis—that the fMRI data contains no information about the variable being predicted—was rejected. A leave-one-out cross-validation procedure was used to iterate through each word during the training and testing stages. Three classifiers were tested: (i) binary good/bad classification, based on the "extreme" scores of participants' Evaluation ratings; (ii) regression on continuous variables, comparing participants' verb and adjective scores; (iii) regression on continuous variables, comparing participants' Evaluation scores to the combined semantic ratings (both verb and adjective scores were included).[12]

Feature Selection

The participant-specific brain maps reduced the voxels of interest to the 500-1250 voxels that showed the most activation averaged over the time of the entire experiment. Unlike other experiments that limit analysis to specific anatomical regions (Haxby et al. 2001, Mitchell et al. 2004), voxel-wise statistics (an ANOVA with a threshold of $p<0.05$) were used to discriminate between the most consistently active voxels. The activation values over the entire map were then normalized (mean=0, SD=1). Before each TR was labeled, we compensated for the time-lag from the hemodynamic response through a convolution operation with a model hemodynamic response function. Rest TR's and TR's that occurred during math problems were then removed, and the remaining TR's were labeled by their corresponding Identity (TR's during the first word, "American," were labeled as 1, and so forth). TR's for N-minus-one words were used

[10] Each participant conjured his/her own mental image and was not instructed as to what that should include. This would activate different voxels for different participants. Because voxels of interest were limited to the most active voxels across the entire brain, these varied tremendously across participants.

[11] The classifier was only trained and tested on words which a participant did not rate as "Neutral." Because these ratings were also subject-specific, words used by any participant's classifier may not have been used by a different participant's classifier.

[12] In the original research on the three dimensions that were used in ACT (Osgood 1957), Evaluation was the first factor and thus counted for the most variance. Subsequent work (Osgood 1975) has shown Evaluation as the most dominant sentiment. Knowing this, preliminary exploratory first level models were created for the 3 dimensions of ACT to investigate distinct neural pathways. As a sociological theory without any biological evidence, the discovery of these pathways could have provided neurological support to corroborate ACT. Because Evaluation was the only sentiment to show any contrast significance, we decided not to train classifiers on the other sentiments.

to train the classifier to associate fMRI data patterns with the respective scores (either the Evaluation or the semantic ratings or both) of the Identity words.

Classifier Training and Testing

(i) Binary Good/Bad Classification

From a participant's EPA ratings, the Evaluation dimension ("Good-Bad") scores for each word were paired with the respective TR's (i.e. All of the TR's during the showing of "American" were assigned with the participant's Evaluation rating for "American."). Of the 40 words, any with a "Neutral" rating (4 or 5 on the 1-9 scale) was removed before the N-minus-one cross-validation was run. All words with a 1-3 rating were labeled "Bad" and 6-9 were labeled "Good." Each word lasted for approximately 5 scans, and about 30 words per participant remained.

A neural network classifier was trained on the preprocessed imaging data to recognize patterns of brain activity elicited by "Good" and "Bad" words. This was implemented with the MVPA package. The two-layer (input and output, no hidden layers) classifier was trained using the conjugate gradient descent variant of the backpropagation algorithm ("train_bp" in Matlab). The input layer contained one unit for every voxel that passed the ANOVA feature selection process. The output layer contained two units: one for "Good" and one for "Bad" words. A specific weight connects every input unit to each of the outputs.

The trained classifier tells how well a given input (i.e., voxel values for a given scan) matches the patterns of activity corresponding to the two categories of words. The backpropagation algorithm is an error-driven learning algorithm which measures the difference between the actual activity pattern at the output layer, and the target activity pattern for the output layer. (Polyn et al. 2005 *s*) This error signal is used to modify the weights; it adjusts them in a direction which will reduce the error signal when the next pattern is displayed.

After the classifier is trained on the TR's for all of the words except the test word, the weights of the network are fixed. The network was then presented with the set of brain patterns from the test Identity word (note that a different classifier was made for each word). For each TR, a classification of 1 meant that the test pattern correctly matched the characteristics of the specific state, and 0 meant it missed. For each Identity, all TR classification accuracy levels were averaged; the mean classifier accuracy for each participant was then calculated.[13]

(ii) Continuous Verb/Adjective Regression

Whereas classification attempts to predict discrete, categorical conditions, a regularized linear regression was performed here because the data were continuous variables. Instead of a classification accuracy level, the metric of interest is the cross-correlation coefficient, τ, between a participant's various ratings and activation patterns.[14] The data of interest were the first factor loadings in the two factor analyses discussed above: one on participants' verb ratings and another on their adjective ratings of the Identity words. The actual scores for each word were not used because the variable of interest was the amount of variance explained by the primary factor. Of concern was whether an Identity's actions or attributes help conjure its image, not whether the brain maps specific pathways for the particular verbs and adjectives chosen. Unlike in (i), no Identity words were discarded.

The classifier was trained using a ridge regression algorithm ("train_ridge" in Matlab), which adds a linearly increasing parameter, the ridge penalty term, as the number of voxels increases. Because the size of the training set (Identity words) is small relative to the number of input dimensions (the activation pattern

[13] Let n be the number of non-neutral Identity words and let $i=1,\ldots,n$ so that Identity_i has y_i TRs, with $y_i=1,\ldots,m$. If x_i TRs were classified correctly, with $0 \le x_i \le m$, then $ClassificationAccuracy_i = x_i/y_i$ and $\mu = (\sum x_i/y_i)/n$ where μ is the mean accuracy level across n Identities.

[14] This paper refers to τ as the cross-correlation coefficient, the correlation score, as well as simply coefficient and score. The cross-correlation coefficient, $\tau \,\epsilon\, (-1,1)$, such that $\tau = 0$ means no correlation, $\tau = 1$ means perfect positive correlation, and $\tau = -1$ means perfect negative correlation.

of voxels), the penalty term prevents overfitting.[15] (Ng 2004) The ridge penalty (L2 = 0.05*N) provides for L2 regularization, which encourages the sum of squares of the input to be small.

Although nonlinear classifiers have been shown to be more powerful than linear classifiers, they do not have a better performance record (for a direct comparison, see Cox and Savoy 2003). It has also been argued that good performance in nonlinear classification is harder to interpret than in a linear classifier. (Kamitani and Tong 2005)

After the classifier is trained, it is presented with the TRs from a test Identity word. Similar to the binary classification in (i), a different classifier was run for every Identity word. Because this test is regression, not classification, the classifier in (ii) calculates the correlation between a participant's respective ratings (in this case, either verb or adjective) and a predicted input (the participant's predicted voxel patterns).

(iii) Evaluation/Semantic-Combination Regression

Another regularized linear regression was used to compare the predictive capabilities of our semantic scales with those of the Evaluation dimension of ACT. Whereas there were two classifiers with one regressor each in (ii), one for verb and one for adjective scores, in (iii) there was a single classifier to measure combined semantic richness (now with two regressors, verbs and adjectives) and another to measure Evaluation. Unlike in (i), Evaluation was not categorized as a binary condition; instead, it was measured as a continuous variable (ranging from 1-9). Similar to (ii), a ridge penalty parameter was incorporated and the resulting classifier measured cross-correlation.

[15] Overfitting occurs when the model explains noise instead of the underlying relationship.

RESULTS

Behavioral Statistics

The summary statistics for the behavioral data are shown in Table 1. These statistics represent the ratings of 10 participants on 8 verbs, 8 adjectives, and the three dimensions of ACT over all 40 Identity words (for the summary statistics of each individual Identity word, see the Appendix).

Table 1 | Mean behavioral statistics

Ratings	Overall Identity Word Statistics		
	Median	Mean	Std. Deviation
Ethnic	2.00	2.36	1.292
Violent	2.00	2.35	1.248
Academic	4.00	3.33	1.373
Corporate	3.00	2.92	1.497
Critical	4.00	3.35	1.428
Nice	3.00	3.07	1.336
Criminal	2.00	2.26	1.194
Funny	3.00	2.70	1.317
Learns	4.00	3.51	1.330
Votes	4.00	3.51	1.393
Leads	4.00	3.60	1.359
Creates	3.00	3.18	1.334
Examines	4.00	3.24	1.389
Embraces	3.00	3.18	1.409
Lies	3.00	2.92	1.315
Hurts	3.00	2.78	1.296
Evaluation	6.00	5.90	2.208
Potency	7.00	6.37	2.166
Activity	7.00	6.50	1.947

Adjective ratings (orange) and verb ratings (blue) are on a 1-5 scale. EPA ratings (grey) are on a 1-9 scale.

All mean ratings were between 2 and 4, which translates into the set of verbs and adjectives “rarely” to “somewhat” associating with the Identities. This suggests that the Identities were able to be thought of in terms of the chosen verbs and adjectives and that no verb or adjective was inappropriate overall (no association was made too frequently or infrequently). The only dimension of ACT that was incorporated into the machine learning was Evaluation. It's mean and median equate to participants' perceptions that the set of words as a whole was "slightly positive."

Classifier Performance

(i) Binary Good/ Bad Classification

The binary classifier based on the neural network algorithm showed promising results for 2 of the 3 participants (Fig. 1), with an average accuracy rate for all participants of 64%. Although the significance of this average accuracy was not tested, the classifier was tested for significance on each participant separately.

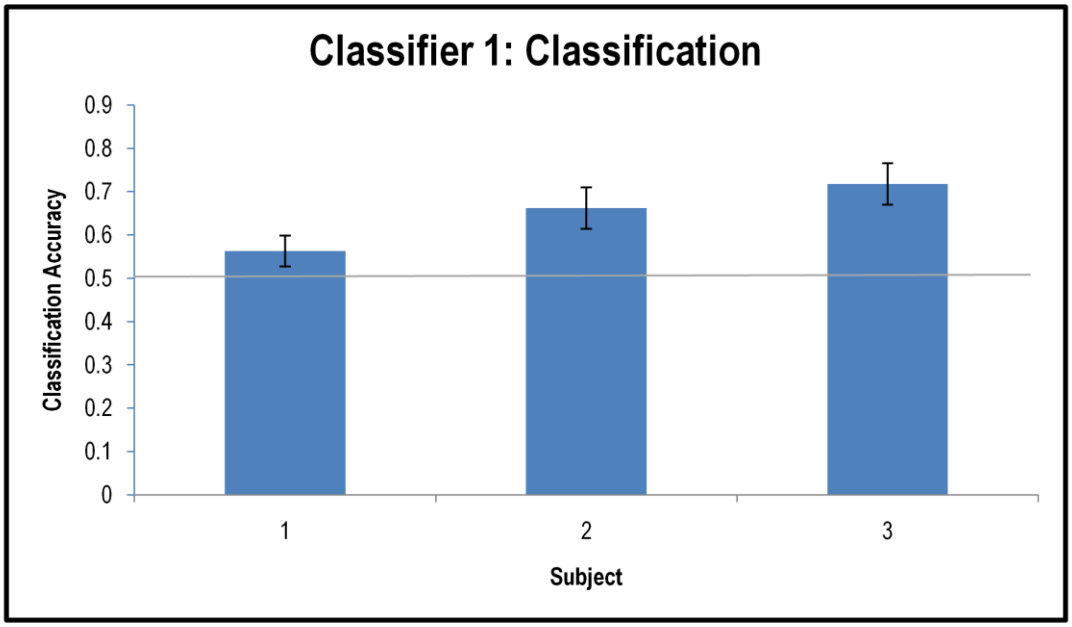

Fig. 2 | "Good" vs. "Bad" leave-one-out classification for each subject.

The probability of a successful classification can be modeled as a Bernoulli trial; with probability of success *p*, in *n* independent trials, the binomial distribution gives the probability of *k* successes. By defining *k* to be the number of test set examples (out of *n*) that are labeled correctly, under the null hypothesis, the probability of obtaining a test statistic as extreme as the one observed (the *p*-value) is P(X≥*k*), where X is a random variable with a *n*-trial binomial distribution. (Pereira et al. 2009) In this classification, because there are only two categories, the probability of success is 0.5. In order to reject the null hypothesis, therein declaring the classifier significant, the *p*-value must be below a certain threshold, α. Because the *p*-value for subjects 2 and 3 fell below a very low threshold (α = 0.01), our classifier proved extremely significant for both of them, but insignificant (even at α = 0.05) for subject 1 (Table 2).

Table 2 | Mean classification rates across reduced list of Identities and their significance.

	Subject 1	Subject 2	Subject 3
Mean	0.563	0.662	0.718
	(0.036)	(0.048)	(0.047)
T-statistic	1.766	3.396	4.584
P-value	0.096	0.003‡	0.000‡
Identity Count	27	25	31

‡ denotes statistically different from 0.5 at α = 0.01.
Standard errors are presented in parentheses. For a complete listing of classification rate by word, as well as which words were removed for which subjects, see the Appendix.

An analysis of the individual words classified highlights the uniqueness of each participant's mental images (and their corresponding neural patterns). Of the words that were correctly classified best (Table 3), only "Mother" and "Woman" are consistent for all participants, and even this does not imply that participants had similar (yet alone identical) representations. Further, even though these two words were among those classified best for each participant, neither of them had perfect classification rates across all participants.

Table 3 | Identities classified best: above 75% classifier accuracy level

Subject 1	Subject 2	Subject 3
Mother*†	American*	American
Muslim	Atheist*	Arab*
Victim	Celebrity*	Celebrity
Woman*†	Cop*	Christian*
	Daughter	Cop
	Doctor*	Daughter
	Father	Doctor*
	Man	Farmer*
	Mother†	Feminist*
	Priest	Heterosexual
	Son	Homosexual*
	Woman†	Immigrant*
		Mother*†
		Soldier
		Teacher
		Victim*
		Woman*†

* denotes a perfect (100%) classification rate.
† denotes words common to all 3 subjects.

(ii) Continuous Verb/Adjective Regression

The classifiers in (ii) used verb and adjective ratings to predict activation. The cumulative scores are shown in Fig. 2, and at first do not appear particularly predictive. Looking at the range of scores, however, helps explain this; the minimum iteration scores all exceed -0.85, with one as low as -0.99 (Table 4). This means that the classifiers' predictive capabilities are occasionally far worse than random, and for some words, it essentially made perfectly incorrect correlations between neural data and ratings of action/attributes. In spite of this, the classifiers in (ii) had some extraordinarily high correlations, with the maximum correlation scores across participants exceeding 0.9 for both verb and adjective scores. Furthermore, in all six regressions, at least 25% of the Identity words had coefficients greater than or equal to 0.5. Of the Identities that were in this high group ($\tau > 0.5$), when correlating with verb scores, only

“Muslim,” “Man,” and “Slut” had predictive power across all participants; with adjectives, however, the only cross-participant Identity was “Executive.” Our measure of characteristics attributed to an Identity never statistically correlated with predictive neural activation (α = 0.05). At the same level of significance however, our measure of action by an Identity did correlate significantly with predicted neural patterns for both subjects 2 and 3.

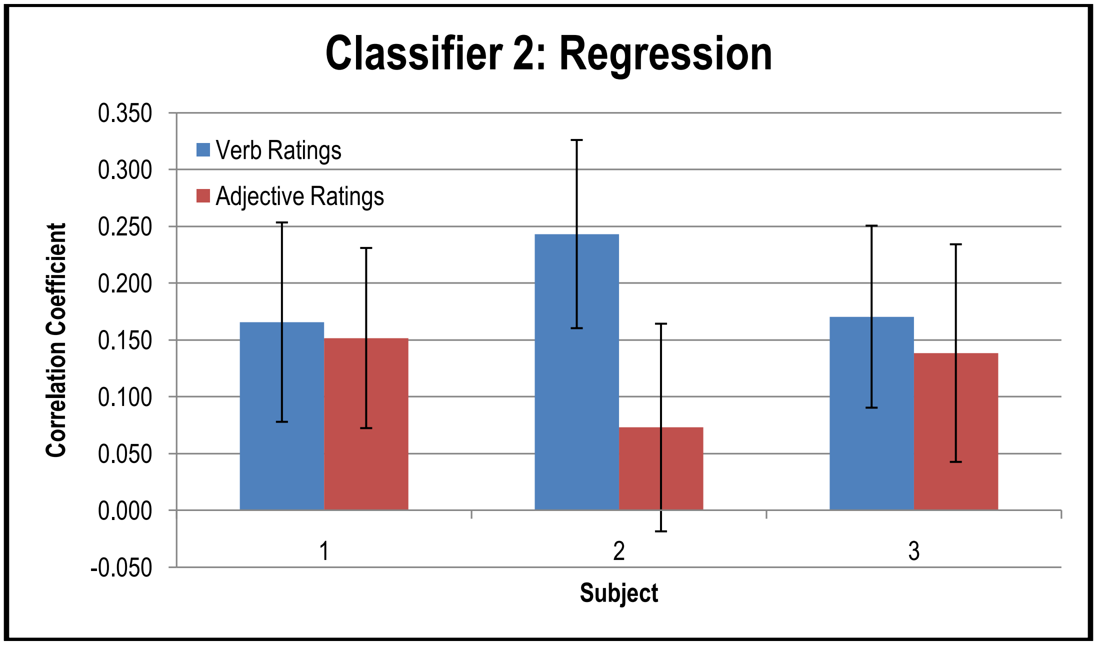

Fig. 3 | Mean correlation coefficient between predicted voxel activation patterns and subjects’ ratings. The y-axis ranges from 0 (no correlation) to 1 (perfect positive correlation).

Table 4 | Mean classifier-2 power and significance across all Identity words

	subj1_verb	*subj1_adj*	*subj2_verb*	*subj2_adj*	*subj3_verb*	*subj3_adj*
Mean	0.166 (0.088)	0.152 (0.079)	0.243 (0.083)	0.073 (0.091)	0.171 (0.080)	0.1385 (0.096)
P-value	0.066	0.063	0.006‡	0.429	0.040‡	0.156
Minimum τ	-0.900	-0.920	-0.990	-0.950	-0.908	-0.870
Maximum τ	0.970	0.930	0.980	0.950	0.979	1.000
Identity Count	40	40	40	40	40	40

‡ denotes statistically different from 0 at α = 0.05
Standard errors in parentheses. τ ϵ (-1,1), such that τ = 0 means no correlation and τ = 1 means perfect positive correlation.
See Appendix for the classifier’s power with each word as well a separate list of the Identities with top correlation.

Instead of comparing the two-classifiers with a two-sample t-test, the most appropriate[16] test is the Wilcoxon signed-rank test because the classifiers are trained on the same neural data. (Pereira et al. 2009) This nonparametric test was run for each participant under the null hypothesis that correlation scores based on verbs were equal to those based on adjectives (H_0: V_i - A_i = 0 for i = 1,…,40). For all participants, the null hypothesis was accepted; there was no significant difference (α = 0.05) between classifiers using action and attribute scores.[17] It should be emphasized, however, that the lack of difference is a less pertinent result. The more important conclusion was that training a classifier on verb scores had significant predictive strength for 2 of the 3 participants, and was quite close to being significant for the last, whereas the predictive power using adjective ratings was not significant for a single participant.

(iii) Continuous Evaluation/Semantic-Combination Regression

The final classifier compared the predictive capabilities of our semantic questionnaires to the Evaluation dimension of ACT. The former differed from (ii) because the classifier incorporated a bivariate regression which combined the first factor loadings *both* from participants' verb and adjective ratings (versus (ii) which was a comparison *between* the effects of verb and adjective ratings). The latter differed from (i) because it treated all Evaluation ratings as continuous variables, on which it regressed predicted activation (in (i) the continuous variable Evaluation scores were transformed into a condition and classified as either "Good" or "Bad" after removing the "Neutral" words). Similar to (ii), the data in both models were from all 40 Identities and utilized a cross-correlation coefficient (Table 5).

[16] The Wilcoxon signed-rank test can be used when the distributional assumptions for the t-test cannot be satisfied. Although a paired t-test could be used for dependent samples, such as this one, it is feasible so long as the number of examples is sufficient to invoke the Central Limit Theorem (Lowry 1999). This is not the case in this experiment.

[17] Wilcoxon test statistics. α = 0.05

	Subject 1	Subject 2	Subject 3
Z-statistic	-0.042	-1.586	-0.363
P-value	0.967	0.113	0.717

Table 5 | Mean classifier-3 power and significance across all Identity words

	subj1_2reg	*subj1_eval*	*subj2_2reg*	*subj2_eval*	*subj3_2reg*	*subj3_eval*
Mean	0.1593 (0.0652)	0.0970 (0.0946)	0.1573 (0.0654)	-0.0373 (0.0891)	0.1543 (0.0647)	0.1855 (0.1027)
P-value	0.0192‡	0.3114	0.021‡	0.6782	0.0221‡	0.0785
Minimum τ	-0.6800	-0.8700	-0.8300	-0.9800	-0.6800	-0.8600
Maximum τ	0.8200	0.9900	0.9200	0.9700	0.9300	0.9700
Identity Count	40	40	40	40	40	40

‡ denotes statistically different from 0 at α = 0.05
Standard errors in parentheses. τ ϵ (-1,1), such that τ = 0 means no correlation and τ = 1 means perfect positive correlation. "_2reg," short for two regressors, refers to each participant's bivariate semantic combination. See Appendix for the classifier's power with each word as well a separate list of the Identities with top correlation.

There were also high maximum correlation scores in (iii), with both classifiers having at least one Identity word with a maximum coefficient over 0.8 for each subject. At least 25% of the Identity words for participants 1 and 3 had coefficients over 0.5 in both classifiers, while 20-25% of the words in both of participant 2's classifiers had correlation coefficients in the same range. Unlike in (ii), not a single Identity had high predictive power ($\tau > 0.5$) across all participants.

Despite the high maximum scores and the number of highly predictive words, the Evaluation classifier had much higher variance than the bivariate classifier,[18] which explains why it was not significantly different from zero (α = 0.05) for any subjects. Fig. 3 shows the mean predictive power for both classifiers across participants. The inclusion of both ratings of action and attributes into a single regression in (iii) made the combined semantic rating classifier statistically significant for all 3 participants. The bivariate classifier was so significant because it had less variance than either individual semantic classifier (and much less than the Evaluation classifier). This differed from (ii), in which training a classifier on attribute ratings was insignificant for all participants, and only two of the participants' verb-trained classifiers were significant. The semantic combination in (iii) was thus significantly stronger (t-test, H_0: $\mu = 0$, $\alpha = 0.05$) than either classifier in (ii), even though it was not statistically different (Wilcoxon signed rank, α = 0.05)

[18] The Results section has abbreviated tables. For the complete descriptive statistics, see Appendix.

from either. Similarly, even though the semantic combination did not differ from the Evaluation classifier significantly, it was statistically significant, and thus much stronger than this classifier as well (same tests and alpha levels).[19]

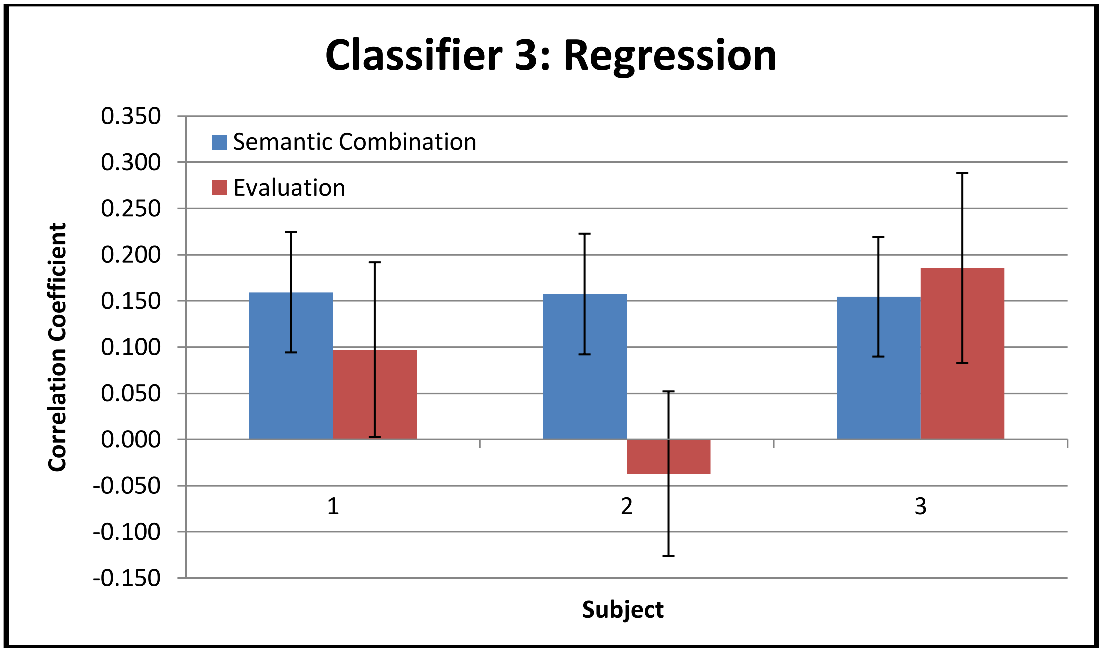

Fig. 4 | Mean correlation coefficient between predicted voxel activation patterns and subjects' ratings.

DISCUSSION

This work is one of a growing number of fMRI studies which demonstrate the benefits of multi-voxel pattern classification techniques. (Haxby et al. 2001, Cox et al. 2003, Mitchell et al., 2004, Kamitani and

[19] Wilcoxon signed rank test statistics: (α = 0.05)

		Subject 1	Subject 2	Subject 3
Verb Ratings v Semantic Combination	Z-statistic	-0.477	-0.719	-0.363
	P-value	0.633	0.472	0.717
Adjective Ratings v Semantic Combination	Z-statistic	-0.215	-0.614	-0.356
	P-value	0.83	0.539	0.722
Evaluation v Semantic Combination	Z-statistic	-1.096	-1.768	-0.363
	P-value	0.273	0.077	0.717

Tong 2005, Hanson 2010) Unlike other studies, the goal here was to investigate the neural representation of a complex noun. To do this, we wanted to see if the information we could extract from multi-voxel patterns could detect subtle distinctions in correlating behavioral responses about which the participant may have had incentives not to be truthful. We hypothesized that an Identity is represented based on its actions and attributes, and the correlating neural patterns would thus fluctuate accordingly. Instead of localizing for involved brain regions like in other MVPA experiments, we selected voxels that were maximally active throughout the entire experiment.

Our results varied tremendously across participants, but that was particularly due to our sample size (N=3). The most important conclusion to be drawn from these findings is that an optimistic future exists for training classifiers to accurately identify complex nouns using verbal stimuli. Using a backpropagation neural network, our classifier showed significant performance when trained to differentiate between good and bad Identities (average accuracy = 64%). Another classifier that used a ridge regression algorithm showed significant average correlation between neural data and participants' ratings of action, but not attribute ratings (average correlations = 0.15 vs. 0.12). Because the neural representation of Identities was better predicted by ratings that describe action than those that characterize attributes, future research should focus on representing an Identity by its actions; if this work utilizes MVPA, and there is a different feature selection process, voxels could be reduced by concentrating on areas of the brain that support motor planning. (Willems 2009) A different classifier that used the same ridge algorithm found that brain imaging scans correlated significantly with participants' semantic ratings of Identities, but not with Evaluation scores (average correlations = 0.15 vs. 0.08). This classifier showed the simplicity of the Evaluation scale: the rich neural representation of an Identity correlated better with our more complex semantic ratings. Given the designs of the experiment and MVPA implementation (to be discussed below), these findings are quite promising.

Because of the exploratory nature of this experiment, there were several changes that could have yielded different—and possibly improved—results. Each Identity word, for example, was only shown for 7 seconds. Although participants reported that this was enough time to imagine the Identity, a longer length of time (perhaps accompanied with a shorter jittering period before and after) could have potentially enabled participants to create richer representations of the stimuli. This may have improved the neural data and would have increased the number of TR's that were analyzed. With so few data points (there were 40 Identities which had on average 5 data points), increasing the length of time that each stimulus was displayed would have trained the classifier on more points and might have improved its accuracy. Another possible idea would be to use less Identity words (still being displayed for longer periods of time), but have multiple repetitions of each. Most other MVPA experiments (e.g. Haynes and Rees 2006, Misaki et al. 2010, Wolbers et al. 2010) have many trials for each stimulus, whereas we only displayed each Identity once.[20] This would have affected the "shock value" of each stimulus because the second (or third) time that the Identity was displayed, participants would get increasingly bored. However, this issue impacted peeking, which will be discussed below.

Another complication involved voxel selection. The brain maps created for this study consisted of the most maximally active voxels throughout the brain (n=500-1250), which should have been further reduced. Just (2010) claimed that voxel sets greater than 80 do not significantly improve classifier performance, and Pereira (2009) showed decreasing accuracy rates after 400 voxels. One possibility that Pereira suggested is to run a nested cross-validation (NCV), which would allow the training set to pick the *n*-most maximally active voxels. In our study, if Identity 1 was our test set (and words 2-40 were the training set), it would have run a leave-one-out cross-validation on words 2-40 (test on word 2, train on 3-40, then test on 3, etc.) to determine how many voxels yields the highest accuracy rate; then, Identity 1 is

[20] For participant 3, a second run was attempted. The data was thrown out because it was determined to be inconclusive, leading to the classifiers predicting far below chance. However, it is possible that this was participant specific.

tested on that many voxels. In addition to reducing the number of voxels, some form of voxel selection (or *unselection*, such as specifically ignoring ventricular activation[21]) may have improved classifier accuracy, as opposed to just taking the maximally active voxels from the entire brain. One possible alternative to the classical activation-based approach is the use of a multivariate searchlight. (Kriegeskorte 2006) In this method, a moving spherical spotlight (centered on each voxel with an optimal radius of 4mm) combines signals from all voxels falling within the searchlight and computes a multivariate effect statistic that marks informative regions. The spatial activity pattern would then be used to define a mask (as opposed to our feature selection which just declared a mask by the most maximally active voxels). After this, it would be possible to further reduce the number of voxels by selecting the *n*-most active voxels with a NCV.

One of the most fundamental criteria for any MVPA study was specifically excluded in this experiment. The idea of peeking, or using the test data to help with voxel selection, is traditionally prohibited because it illegitimately improves classification.[22] (Mur 2009) The training set and test set should be entirely independent, usually with the sets being based on different scanner runs (for example, train the classifier on runs 1-9, then test it by seeing if the classifier can predict each stimulus in run 10). However, our experiment only had one run, so we trained the classifier on 39 words then tested on the fortieth. Although this means that the impressive results of this study were artificially inflated, they should be viewed as “preliminary” and are encouraging for future research that utilizes these innovative methods.[23]

The final innovation that we hoped to pursue, the incentive compatibility of monetary rewards, was not implemented due to time constraints. Because all three participants were members of the Center for Neuropolicy, they were not financially compensated for participating in the experiment. A key problem that might have arisen, had this last step been executed, incorporates the sacredness of each Identity. A

[21] At least one participant’s mask had significant ventricular activation.

[22] If the test dataset is used to help select voxels, those voxels will have already been known to be active. This will bias the testing results.

[23] Any further research will of course have to prevent peeking.

mechanism would need to be implemented to tell if any particular Identity was more sacred than another (in the current design); otherwise, there would be no way to tell if being truthful is a deontic or utilitarian decision. Although the NIM could still prove successful, the reasoning behind why would be hidden. It would be possible to test for left TPJ and VLPFC activation in each Identity, an activation pattern that is consistent with sacred values (Berns et al. 2010);[24] even a high classifier correlation that was paired with significant TPJ/VLPFC activation, however, would not be enough to identify for which words the NIM proved effective. A question that then arises involves the appropriateness of payment: would participants be rewarded for honest reporting on all Identities, or only Sacred Identities? And then there could be potential classifier errors, in which the participant is honest (which would not be known) but the classifier fails. Further classifier modification is still necessary before a monetary incentive for honesty could be successfully utilized in such a task. The work completed, however, can be thought of as the first step, a "calibration stage" essentially, in the future implementation of such a mechanism.

[24] The TPJ has been associated with evaluating rights and wrongs and the VLPFC with semantic rule retrieval. Together this pattern encodes sacred value processing.

REFERENCES

Baron, J. & Spranca, M. (1997). Protected Values. *Organizational Behavior and Human Decision Processes* 70, 1-16.

Berns, G.S., Bell, E., Capra, C.M., Prietula, M.J., Moore, S., Anderson, B., et al. (2010). *Neural evidence for the deontic processing of personal sacred values*. Article submitted for publication.

Berns, G.S., Capra, C.M., Chappelow, J., Moore S., & Noussair, C. (2008) Nonlinear neurobiological probability weighting functions for aversive outcomes. *NeuroImage* 39, 2047-2058.

Breiter, H., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 30(2), 619-39.

Camerer C.F. (2003). *Behavioral game theory experiments in strategic interaction*. Princeton: Princeton University Press.

Camerer, C.F., Loewenstein, G.D., & Prelec, D. (2005). Neuroeconomics: how neuroscience can inform economics. *Journal of Economic Literature* 43, 9-64.

Cox, D.D., & Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19, 261–270.

Gigerenzer, G., & Selten, R. (Eds.). (2002). *Bounded rationality: the adaptive toolbox*. Boston: MIT Press.

Glimcher, P.W., & Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. *Science* 306(5695), 447-452.

Hanson S.J., & Schmidt, A. (2010). High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. *NeuroImage* (epub ahead of print), doi: 10.1016/j.neuroimage.2010.08.028

Hanson, S.J., Matsuka, T., & Haxby, J.V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *NeuroImage* 23(1), 156-166.

Haxby J.V., Gobbini M.I., Furey M.L., Ishai A., Schouten J.L., & Pietrini P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539), 2425-2430.

Haynes J.D., Rees G. (2006). Decoding mental states from brain activity in humans. *National Review of Neuroscience* 7: 523-534.

Heise, D.R. (1979). *Understanding events: affect and the construction of social action*, New York: Cambridge University Press.

Heise, D.R. (1987). Affect control theory: concepts and model. *Journal of Mathematical Sociology* 13(1-2), 1-33.

Heise, D.R. (2002). Understanding social interaction with affect control theory. In J. Berger, & M. Zelditch (Eds.), *New Directions in Sociological Theory* (Chapter 2). Boulder: Rowman and Littlefield.

Just, M.A., Cherkassky, V.L., Aryal, S., & Mitchell, T.M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE* 5(1): e8622, 1-15.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *National Review of Neuroscience* 8, 679–685.

Kenning, P., & Plassmann, H. (2005). Neuroeconomics: overview from an economic perspective. *Brain Research Bulletin* 67(5), 343-354.

Krajbich, I., Camerer, C., Ledyard, J., & Rangel, A. (2009). Using neural measures of economic calue to solve the public goods free-rider problem. *Science* 326(5952), 596-99.

Kriegeskorte N., Goebel R., & Bandettini P. (2006). Information-based functional brain mapping. *PNAS* 103 (10), 3863-3868.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412(6843), 150-157.

Lowry, R. (1999). Concepts and Applications of Inferential Statistics. Retrieved November 20, 2010, from http://faculty.vassar.edu/lowry/webtext.html

Luce, D.R. (2000). *Utility of gains and losses: measurement-theoretical and experimental approaches*. In *Scientific Psychology Series* (Vol. 8). Cambridge: Psychology Press.

Misaki, M., Kim, Y., Bandettini, P.A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage* 53(1), 103-118.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., Mason, R., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880), 1191-1195.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning* 57(1-2), 145-175.

Mur, M., Bandettini, P.A., Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI – an introductory guide. *SCAN* 4, 101-109.

Ng, A.Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the 21st International Conference on Machine Learning*. Morgan Kauffmann.

Norman, K.A., Polyn, S.M., Detre, G.J., & Haxby, J.V. (2007). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Neuroscience* 10(9), 424-430.

O'Toole, A.J., Jiang, F., Abdi, H., Penard, N., Dunlop, J.P., & Parent, M.A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience* 19(11), 1735-1752.

Osgood, C. E., May, W.H., & Miron, M.S. (1975). *Cross-cultural universals of affective meaning.* Urbana: University of Illinois Press.

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

Pereira, F., Mitchell, T.M., Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* 45,s199-s209.

Polyn S.M., Natu V.S., Cohen J.D., & Norman K.A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science* 310(5756), 1963-1966.

Quamme, J.R., Weiss, D.J., & Norman, K.A. (2010). Listening for recollection: a multi-voxel pattern analysis of recognition memory retrieval strategies. *Frontiers in Human Neuroscience* 4(61), 1-17.

Rabin, M. (2002). A perspective on psychology and economics. *European Economic Review* 46, 657–686.

Rissman, J., Greely, H.T., & Wagner, A.D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. *PNAS* 107(21), 9849-9854.

Tetlock, P.E. (2003). Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences* 7, 320-324.

Willems, R.M., Hagoort, P., & Casasanto, D. (2009). Body-specific representations of action verbs: neural evidence from right- and left-handers. *Psychological Science* 21(1), 67-74.

Wolbers, T., Zahorik, P., & Giudice, N.A. (2010). Decoding the direction of auditory motion in blind humans. *NeuroImage* (epub ahead of print), doi: 10.1016/j.neuroimage.2010.04.266

Wunderlich, K., Rangel, A., & O'Doherty, J.P. (2010) Economic choices can be made using only stimulus values. *PNAS* 107(34), 15005-15010.

APPENDIX

Behavioral Instructions

Thank you for participating in our experiment. This should take approximately 30 minutes to one hour to complete. You will be compensated $20 for your time. On these forms, you will be shown a series of 40 social identities. Each of these nouns names different kinds of individuals (example: Asian). You will be asked to rate each identity. **Please do not leave any questions blank.** The first rating will be based **on how often you associate 8 adjectives with each identity**. If you would not use an adjective with the specific identity, please rate it as "Not Applicable." Otherwise, select whether you would relate the adjective to the identity "rarely," "slightly," "somewhat," or "frequently." After completing this first form, please click "continue" to move onto the next rating form. This **next form asks how often you associate 8 verbs with each identity.** Make sure that the Identity is the subject of each sentence when thinking about the rating (example: "Asian hurts [someone]" not "[someone] hurts an Asian"). Similar to the first sheet, if you would not associate a verb with the specific identity, please rate it as "Not Applicable." Otherwise, select whether you would relate the verb to the identity "rarely," "slightly," "somewhat," or "frequently." Unlike the first two forms ratings, the third one deals with **how you feel about the identity, and is on a bi-polar scale.** When the box on top does not contain another identity but is blank instead, the experiment is complete.

A1 | 15 Hand Selected Verbs and Adjectives after Corpus Reduction Process

Adjectives	Verbs
Ethnic	Learns
Violent	Votes
Academic	Leads
Corporate	Annoys
Attractive	Creates
Critical	Abandons
Free	Examines
Soft	Embraces
Nice	Dies
Thick	Demonstrates
Criminal	Promises
Funny	Plays
Low	Advises
Nasty	Lies
Empty	Hurts

A2 | Behavioral Statistics by each Identity Word: Adjective Ratings

Identities		Ethnic	Violent	Academic	Corporate	Critical	Nice	Criminal	Funny
Alcoholic	Median	2.00	4.00	1.50	2.50	2.50	1.50	3.00	2.50
	Mean	1.90	3.70	2.10	2.50	2.60	1.70	3.20	2.40
	StdDev	.876	.823	1.370	1.354	1.430	.823	.919	1.430
American	Median	2.00	3.00	4.00	5.00	3.00	4.00	2.00	4.00
	Mean	2.40	2.80	4.00	4.80	2.90	3.60	2.40	3.40
	StdDev	1.350	1.033	.943	.422	1.197	1.075	.843	1.430
Arab	Median	5.00	3.00	3.00	2.50	4.00	3.50	2.50	2.00
	Mean	4.67	2.80	3.30	2.50	3.60	3.10	2.70	2.10
	StdDev	.500	1.317	.949	.850	1.350	1.101	1.252	1.101
Atheist	Median	1.50	1.50	4.50	2.00	5.00	3.00	2.00	3.00
	Mean	2.10	1.90	4.30	2.30	4.80	2.80	1.70	3.30
	StdDev	1.370	1.101	.949	1.337	.422	1.317	.675	1.059
Banker	Median	1.50	1.00	4.00	5.00	3.00	3.00	3.00	2.50
	Mean	1.70	1.40	3.70	4.80	2.70	2.90	2.80	2.50
	StdDev	.823	.516	1.160	.422	1.337	1.449	1.135	1.179
Celebrity	Median	1.50	1.50	1.50	4.00	2.50	3.00	2.00	3.50
	Mean	1.80	1.50	2.00	3.50	2.60	2.90	1.80	3.50
	StdDev	.919	.527	1.333	1.650	1.506	1.287	.789	1.354
Cheater	Median	2.00	2.00	3.00	3.00	3.00	1.50	3.00	2.00
	Mean	1.90	2.10	2.40	3.00	2.60	2.00	2.90	2.10
	StdDev	.876	.994	1.075	1.700	1.350	1.333	1.287	1.287
Christian	Median	2.50	2.50	3.00	4.00	3.50	4.00	2.00	3.00
	Mean	2.50	2.60	3.00	3.50	3.20	3.70	2.10	2.60
	StdDev	1.354	.966	.943	.972	1.476	.949	1.197	1.174
Cop	Median	1.50	3.00	2.50	2.00	4.00	3.00	2.00	2.00
	Mean	1.70	3.10	2.70	2.20	4.10	2.80	2.10	2.30
	StdDev	.823	.994	1.418	1.135	.738	1.398	1.101	1.337
Daughter	Median	2.00	2.00	3.00	2.50	3.00	4.00	1.00	3.00
	Mean	2.00	1.60	3.20	2.50	2.90	3.70	1.50	2.80
	StdDev	1.054	.516	1.317	1.080	1.370	1.252	.707	1.317
Democrat	Median	3.00	2.00	4.00	3.00	4.00	4.00	2.00	3.00
	Mean	2.50	1.70	4.20	3.20	3.70	3.50	1.70	3.00
	StdDev	1.080	.675	.632	1.229	1.337	1.080	.483	1.333
Doctor	Median	3.00	1.00	5.00	2.50	4.00	4.00	1.00	3.00
	Mean	2.70	1.20	4.70	2.40	3.50	3.80	1.20	2.70
	StdDev	1.337	.422	.675	1.174	1.434	.789	.422	1.252
Entrepreneur	Median	3.00	1.00	4.00	4.00	4.00	4.00	2.00	3.00
	Mean	2.44	1.33	4.00	4.11	3.44	3.33	1.78	2.78
	StdDev	1.333	.500	.866	1.054	1.333	1.225	.833	1.202
Environmentalist	Median	1.00	2.00	4.00	1.00	4.50	3.50	2.00	3.50
	Mean	1.80	2.00	4.20	1.90	4.20	3.30	2.10	3.10
	StdDev	1.033	.943	.919	1.449	.919	1.160	.994	1.287
Executive	Median	2.00	2.00	4.00	5.00	4.00	3.00	3.00	2.00
	Mean	1.90	2.20	4.30	5.00	3.80	2.60	2.70	2.10
	StdDev	.876	1.229	.483	.000	1.229	1.350	1.160	1.287

Farmer	Median	2.00	1.50	2.00	2.00	2.00	4.00	1.00	3.00
	Mean	1.90	1.50	2.20	2.10	2.40	3.70	1.30	2.70
	StdDev	.876	.527	1.229	1.101	1.506	.675	.483	1.252
Father	Median	2.00	2.00	4.00	4.00	4.00	4.00	2.00	3.50
	Mean	2.00	2.40	4.00	3.60	3.40	3.70	1.70	3.60
	StdDev	1.054	1.075	1.155	1.075	1.174	1.252	.483	1.265
Feminist	Median	2.00	1.50	4.50	1.50	5.00	3.50	1.50	3.00
	Mean	2.10	1.80	4.10	1.90	4.60	3.10	1.60	2.90
	StdDev	1.101	.919	1.287	1.197	.699	1.370	.699	1.197
Heterosexual	Median	2.00	2.00	3.00	4.00	3.00	4.00	1.50	3.00
	Mean	2.00	2.10	3.10	3.60	3.00	3.70	1.80	3.30
	StdDev	1.054	.738	1.287	1.578	.943	1.252	.919	1.059
Homosexual	Median	2.50	1.00	4.00	2.50	4.00	4.00	1.50	4.00
	Mean	2.50	1.40	3.50	2.50	3.50	3.50	1.60	3.50
	StdDev	1.354	.516	1.434	1.269	1.434	1.080	.699	1.080
Immigrant	Median	5.00	2.00	3.00	2.00	2.00	3.50	2.00	3.00
	Mean	4.80	2.10	2.90	2.10	2.70	3.30	2.50	2.90
	StdDev	.422	.568	1.370	1.287	1.252	1.418	1.179	1.197
Jew	Median	4.00	2.00	4.50	4.00	4.00	4.00	2.00	3.50
	Mean	3.70	1.90	4.40	4.40	3.70	3.70	2.00	3.30
	StdDev	1.160	.738	.699	.516	1.337	.823	.667	1.418
Latino	Median	4.50	3.00	2.50	2.00	2.50	3.50	3.00	3.00
	Mean	4.40	2.60	2.70	1.90	2.40	3.30	2.70	3.00
	StdDev	.699	.843	1.160	.994	1.265	1.160	.949	1.247
Lawyer	Median	1.50	2.00	5.00	5.00	4.50	3.00	2.50	3.00
	Mean	1.90	1.90	4.90	4.70	3.90	2.90	2.90	3.00
	StdDev	.994	.568	.316	.483	1.449	1.370	1.524	1.155
Man	Median	2.50	3.50	4.00	4.50	3.00	3.00	3.00	3.50
	Mean	2.20	3.40	3.90	4.20	3.20	3.20	2.50	3.30
	StdDev	.919	.699	.738	1.229	1.398	1.033	.972	1.418
Mother	Median	2.50	2.00	4.00	2.50	4.00	4.00	1.00	3.00
	Mean	2.20	1.60	3.80	2.70	3.50	4.30	1.20	3.10
	StdDev	1.135	.516	1.317	1.567	1.509	.675	.422	1.370
Muslim	Median	5.00	3.50	4.00	2.50	4.00	3.50	2.50	2.00
	Mean	4.30	3.30	3.40	2.30	4.00	3.20	2.70	2.30
	StdDev	1.252	1.337	1.430	1.337	1.247	1.398	1.337	1.418
Politician	Median	1.00	2.00	4.50	5.00	4.50	2.00	2.00	3.00
	Mean	1.50	2.50	4.10	4.50	4.20	2.70	2.50	2.70
	StdDev	.707	1.269	.994	.707	1.229	1.494	.707	1.252
Priest	Median	2.00	1.50	3.00	1.00	2.50	4.00	2.00	2.00
	Mean	2.10	1.80	3.30	1.90	2.80	3.50	2.00	2.40
	StdDev	.876	.919	.949	1.287	1.814	1.509	1.054	1.506
Protestor	Median	3.00	3.00	4.00	1.00	5.00	3.00	2.00	2.50
	Mean	2.20	3.00	3.50	1.70	4.20	2.60	2.40	2.60
	StdDev	1.033	.816	1.269	1.059	1.317	1.174	.843	1.265
Racist	Median	1.50	4.00	1.00	2.50	5.00	1.00	4.00	1.00
	Mean	1.80	4.40	1.40	2.60	4.20	1.20	3.50	1.30
	StdDev	1.033	.516	.699	.966	1.398	.422	1.179	.483

Rapist	Median	1.50	5.00	1.50	1.00	1.50	1.00	5.00	1.00
	Mean	1.80	4.90	1.60	1.70	2.10	1.10	5.00	1.10
	StdDev	.919	.316	.699	1.059	1.370	.316	.000	.316
Republican	Median	2.00	2.00	3.00	5.00	4.00	3.00	2.50	2.50
	Mean	2.10	2.50	3.20	4.30	3.90	2.90	2.50	2.60
	StdDev	1.197	1.179	1.317	1.252	1.370	1.449	1.080	1.430
Slut	Median	1.50	2.00	2.00	2.00	2.50	2.50	1.50	2.50
	Mean	1.90	1.60	2.10	2.20	2.70	2.50	1.60	2.70
	StdDev	.994	.516	1.370	1.229	1.418	1.354	.699	1.337
Soldier	Median	2.00	4.50	2.50	3.00	3.00	4.00	2.00	2.50
	Mean	1.90	4.10	2.50	2.50	3.20	3.50	2.00	2.40
	StdDev	.876	1.101	1.179	1.269	1.476	1.269	1.054	1.265
Son	Median	1.00	2.00	3.50	3.00	3.00	4.00	2.00	3.00
	Mean	1.80	2.20	3.40	2.90	3.00	3.60	1.90	3.10
	StdDev	1.033	.919	1.075	1.449	1.333	1.174	.876	1.287
Teacher	Median	1.50	1.00	5.00	2.00	4.00	4.00	1.00	3.00
	Mean	1.90	1.20	4.60	1.80	3.60	3.80	1.40	3.10
	StdDev	.994	.422	.699	.789	1.174	.632	.516	1.287
Terrorist	Median	4.00	5.00	1.50	1.00	5.00	1.00	5.00	1.00
	Mean	3.20	4.90	1.90	1.40	4.00	1.10	4.90	1.20
	StdDev	1.619	.316	1.101	.699	1.633	.316	.316	.422
Victim	Median	3.00	1.00	2.50	1.00	1.50	3.00	1.00	2.00
	Mean	2.30	1.20	2.50	1.80	2.00	2.80	1.50	2.20
	StdDev	.949	.422	1.354	1.317	1.333	1.476	.707	1.398
Woman	Median	2.50	2.00	3.50	3.00	4.00	4.00	2.00	3.00
	Mean	2.10	1.80	3.80	3.30	3.50	4.10	1.80	3.10
	StdDev	.994	.422	.919	.675	1.269	.738	.422	1.101

A3 | Behavioral Statistics by Identity Word: Verb Ratings

Identities		Learns	Votes	Leads	Creates	Examines	Embraces	Lies	Hurts
Alcoholic	Median	1.50	2.00	2.00	2.00	1.50	1.50	4.00	4.00
	Mean	1.80	1.80	1.80	2.10	1.80	1.80	4.10	4.10
	StdDev	.919	.789	.789	1.287	.919	.919	.568	.738
American	Median	4.00	4.00	5.00	4.50	4.00	4.00	4.00	4.00
	Mean	3.80	4.10	4.50	4.40	3.50	3.70	3.50	3.60
	StdDev	.919	.876	.707	.699	.972	1.252	1.179	1.075
Arab	Median	3.00	3.00	3.00	3.00	3.50	2.50	3.00	4.00
	Mean	3.20	2.60	2.90	3.00	3.20	2.60	3.30	3.50
	StdDev	1.317	1.075	1.370	.816	1.476	1.430	1.059	1.269
Atheist	Median	5.00	4.50	3.50	4.00	5.00	2.50	2.00	2.50
	Mean	4.80	4.00	3.60	3.40	4.60	2.60	2.20	2.50
	StdDev	.422	1.333	1.265	1.578	1.265	1.430	1.033	1.080
Banker	Median	3.50	3.50	3.50	2.00	2.50	3.00	3.50	3.00
	Mean	3.70	3.50	3.40	2.40	2.50	2.40	3.20	2.80
	StdDev	.823	1.179	1.075	1.075	1.080	1.350	1.398	1.317
Celebrity	Median	2.00	3.50	4.00	3.00	2.00	3.00	3.00	2.00
	Mean	2.30	2.90	3.20	3.20	2.20	3.30	3.10	2.00
	StdDev	1.418	1.287	1.549	1.317	1.229	1.494	1.449	.816
Cheater	Median	2.00	3.00	2.00	2.00	2.00	1.50	5.00	3.50
	Mean	1.70	2.60	2.20	1.90	2.30	1.80	4.80	3.70
	StdDev	.675	.966	1.033	.738	1.160	.919	.632	1.337
Christian	Median	3.50	4.50	4.00	3.00	3.00	4.50	3.00	3.00
	Mean	3.20	4.00	4.10	2.90	3.10	4.30	3.30	2.80
	StdDev	1.229	1.333	1.197	1.197	.876	.823	1.418	1.135
Cop	Median	3.00	3.50	4.00	2.50	4.00	3.00	2.00	3.00
	Mean	2.70	3.20	3.60	2.40	3.40	2.70	2.60	3.00
	StdDev	1.059	1.687	1.075	1.430	1.174	1.494	1.174	1.333
Daughter	Median	4.00	2.00	3.00	3.50	3.00	4.00	3.00	2.00
	Mean	4.20	2.70	3.00	3.20	3.10	3.50	2.70	2.10
	StdDev	.789	1.418	1.333	1.317	1.287	1.434	1.160	.738
Democrat	Median	4.50	5.00	5.00	4.00	4.00	4.00	2.50	2.00
	Mean	4.40	4.90	4.60	3.60	4.10	3.70	2.50	2.10
	StdDev	.699	.316	.516	.843	.876	1.160	.850	.738
Doctor	Median	5.00	4.00	4.50	4.00	5.00	3.00	2.00	2.00
	Mean	4.90	3.70	4.10	3.80	5.00	3.10	1.90	1.80
	StdDev	.316	1.494	1.287	1.033	.000	.994	.738	.789
Entrepreneur	Median	5.00	4.00	5.00	5.00	4.00	4.00	3.00	2.00
	Mean	4.67	3.78	4.33	4.67	4.22	3.33	2.56	2.11
	StdDev	.500	1.302	1.323	.500	.667	1.658	1.130	.782
Environmentalist	Median	5.00	5.00	4.00	4.00	4.00	4.00	2.00	2.00
	Mean	4.50	4.10	4.10	4.20	4.00	4.10	2.20	1.70
	StdDev	.707	1.663	.738	.632	1.054	.738	.789	.483
Executive	Median	5.00	4.00	5.00	4.00	4.00	3.00	3.50	2.00
	Mean	4.60	4.10	4.80	4.20	4.00	3.00	3.20	2.50
	StdDev	.516	.568	.422	.789	.667	1.155	1.398	1.179

Farmer	Median	3.00	3.50	2.50	4.00	3.00	2.50	1.50	2.00
	Mean	3.00	3.20	2.50	3.90	3.20	2.60	1.50	1.80
	StdDev	1.054	1.033	1.179	.994	1.033	1.265	.527	.632
Father	Median	4.00	4.50	5.00	4.00	4.00	3.50	2.00	2.00
	Mean	3.40	4.40	4.60	3.70	3.40	3.60	2.20	2.00
	StdDev	1.578	.699	.699	.675	1.350	.966	.789	.943
Feminist	Median	5.00	5.00	5.00	3.00	5.00	4.00	2.00	2.00
	Mean	4.20	4.40	4.20	3.30	3.80	3.80	2.40	2.20
	StdDev	1.317	1.265	1.476	1.059	1.687	1.229	1.174	1.135
Heterosexual	Median	4.00	4.00	4.00	4.00	3.50	4.00	2.00	2.50
	Mean	3.40	3.20	3.70	3.40	3.00	3.60	2.60	2.50
	StdDev	1.506	1.549	1.567	1.430	1.491	1.174	1.506	1.080
Homosexual	Median	4.00	4.00	3.50	4.00	4.00	5.00	2.00	2.00
	Mean	3.50	4.10	3.10	3.90	3.80	4.00	2.10	2.50
	StdDev	1.509	1.197	1.370	1.287	1.229	1.633	.994	1.269
Immigrant	Median	4.50	2.00	2.00	2.50	3.00	4.00	2.00	2.00
	Mean	4.10	2.00	2.30	2.60	3.20	3.90	2.40	2.50
	StdDev	1.101	1.247	1.337	1.506	1.135	1.370	1.075	1.179
Jew	Median	5.00	4.50	4.00	4.00	4.00	4.00	2.50	2.50
	Mean	4.60	4.00	4.00	4.00	3.70	3.50	2.30	2.50
	StdDev	.516	1.414	.667	.816	1.494	1.269	.823	.850
Latino	Median	3.00	3.50	2.00	2.50	2.50	4.00	2.50	2.00
	Mean	3.40	3.00	2.70	2.90	2.70	3.70	2.50	2.20
	StdDev	.843	1.155	1.252	1.729	1.337	1.337	.850	.919
Lawyer	Median	5.00	5.00	5.00	4.00	5.00	3.00	5.00	2.00
	Mean	4.60	4.40	4.70	3.50	4.80	2.70	4.20	2.60
	StdDev	.699	.966	.483	1.269	.422	1.252	1.135	1.075
Man	Median	4.00	4.00	5.00	4.50	4.00	4.00	3.50	4.00
	Mean	4.00	4.40	4.90	4.30	3.30	3.20	3.40	3.40
	StdDev	.471	.516	.316	.949	1.337	1.549	.966	1.174
Mother	Median	4.00	4.00	4.00	4.00	3.50	5.00	2.00	2.00
	Mean	3.80	3.70	4.10	3.70	3.10	4.40	2.00	1.90
	StdDev	1.229	1.059	.568	1.337	1.524	.843	.816	.738
Muslim	Median	4.00	3.00	3.50	3.00	3.00	2.50	3.00	3.50
	Mean	3.90	3.00	3.20	2.80	2.80	2.70	3.20	3.50
	StdDev	1.101	1.155	1.229	1.398	1.317	1.567	1.476	1.269
Politician	Median	3.50	5.00	5.00	3.50	3.50	2.50	4.00	3.00
	Mean	3.50	4.90	4.90	3.30	3.70	3.00	4.00	3.00
	StdDev	1.354	.316	.316	1.059	1.252	1.414	1.054	1.333
Priest	Median	4.00	4.00	4.50	3.00	4.00	4.50	2.00	2.00
	Mean	3.60	3.40	4.40	3.00	3.50	4.30	2.40	2.50
	StdDev	1.075	1.075	.699	1.491	1.434	.949	1.430	1.080
Protestor	Median	3.00	4.50	4.00	3.50	4.00	4.00	3.00	2.00
	Mean	3.10	4.40	3.50	3.10	3.70	3.10	2.70	2.20
	StdDev	1.101	.699	1.354	1.370	1.494	1.524	.949	.919
Racist	Median	2.00	4.00	3.00	2.00	2.00	1.50	5.00	5.00
	Mean	1.90	3.70	2.90	2.20	2.10	1.80	4.30	4.90
	StdDev	.738	1.337	1.197	1.229	1.101	1.229	1.252	.316

Rapist	Median	2.00	1.00	1.50	1.00	1.50	2.00	5.00	5.00
	Mean	1.78	1.60	1.70	1.50	1.70	2.20	4.80	4.90
	StdDev	.972	1.075	.823	.707	.949	1.229	.422	.316
Republican	Median	3.00	5.00	5.00	3.00	3.50	3.00	3.00	3.00
	Mean	3.20	4.90	4.50	2.90	3.00	3.10	3.30	3.10
	StdDev	1.317	.316	.707	1.197	1.414	1.287	1.252	1.197
Slut	Median	2.00	2.00	2.00	2.00	2.00	3.50	3.00	3.00
	Mean	2.20	2.10	2.10	2.30	2.30	3.00	2.90	2.90
	StdDev	1.398	.994	1.287	1.418	1.418	1.491	1.287	1.287
Soldier	Median	3.00	4.00	4.50	2.50	3.00	3.00	2.00	4.00
	Mean	3.20	4.30	4.30	2.50	2.90	3.00	2.10	4.00
	StdDev	.919	.675	.949	1.179	1.370	1.155	1.101	.943
Son	Median	4.00	3.50	4.00	3.50	4.00	3.00	3.00	2.00
	Mean	4.00	3.20	3.80	3.10	3.30	3.00	3.00	2.30
	StdDev	.667	1.317	1.135	1.370	1.337	1.333	.943	1.160
Teacher	Median	5.00	4.00	4.00	4.00	4.00	4.00	2.00	2.00
	Mean	4.50	4.00	4.20	3.80	3.80	3.90	1.60	1.70
	StdDev	.972	.943	.789	.632	1.135	.876	.516	.483
Terrorist	Median	2.50	2.00	4.00	2.00	2.00	2.00	5.00	5.00
	Mean	2.40	2.20	3.30	2.50	2.20	1.90	4.70	4.90
	StdDev	.966	1.229	1.252	1.269	.919	1.197	.483	.316
Victim	Median	2.50	1.00	1.50	2.00	2.00	2.50	2.00	2.00
	Mean	2.50	1.70	2.10	2.40	2.20	2.40	2.10	2.20
	StdDev	1.509	1.160	1.370	1.506	1.398	1.430	.738	1.317
Woman	Median	4.00	4.00	4.00	3.50	4.00	5.00	3.00	2.50
	Mean	4.10	4.00	3.80	3.40	3.40	4.60	3.00	2.60
	StdDev	.876	.667	.789	1.265	1.430	.699	.943	1.075

A4 | Behavioral Statistics by Identity Word: EPA Ratings of Affect Control Theory

Identities		Evaluation	Potency	Activity
Alcoholic	Median	3.50	5.00	5.00
	Mean	3.70	4.10	4.90
	StdDev	1.418	1.595	2.331
American	Median	6.50	8.00	7.50
	Mean	6.20	7.60	7.50
	StdDev	1.687	1.776	.850
Arab	Median	5.00	6.50	5.00
	Mean	5.30	6.40	5.50
	StdDev	2.312	1.955	2.677
Atheist	Median	5.50	6.00	7.00
	Mean	5.60	6.00	6.10
	StdDev	1.897	.816	2.283
Banker	Median	5.00	8.00	5.50
	Mean	5.40	7.50	6.10
	StdDev	1.430	1.269	1.595
Celebrity	Median	5.00	8.50	7.50
	Mean	5.90	8.00	7.20
	StdDev	1.853	1.333	1.398
Cheater	Median	3.00	5.00	5.00
	Mean	3.40	5.20	4.50
	StdDev	1.075	1.033	.707
Christian	Median	7.00	8.50	6.50
	Mean	6.70	7.90	6.60
	StdDev	1.947	1.287	1.838
Cop	Median	6.50	8.00	8.00
	Mean	6.40	7.90	7.50
	StdDev	1.838	1.524	1.269
Daughter	Median	7.00	4.50	5.00
	Mean	7.00	4.40	5.40
	StdDev	1.414	1.713	1.647
Democrat	Median	7.00	8.00	7.00
	Mean	7.10	7.60	6.90
	StdDev	1.101	1.506	1.287
Doctor	Median	8.00	7.50	7.00
	Mean	7.60	7.40	7.10
	StdDev	1.174	1.430	1.729
Entrepreneur	Median	7.00	6.00	8.00
	Mean	7.11	6.56	7.89
	StdDev	1.453	1.236	.928
Environmentalist	Median	8.00	4.50	7.00
	Mean	7.60	4.70	7.00
	StdDev	1.350	1.567	1.333
Executive	Median	5.00	9.00	7.50
	Mean	5.50	8.50	7.40
	StdDev	2.121	.707	1.578

Farmer	Median	7.00	5.00	4.50
	Mean	6.80	5.00	4.60
	StdDev	1.476	1.563	2.413
Father	Median	7.50	8.00	6.50
	Mean	7.40	7.70	6.90
	StdDev	1.265	1.418	1.101
Feminist	Median	7.00	6.00	8.00
	Mean	6.40	5.40	7.40
	StdDev	2.271	2.413	2.319
Heterosexual	Median	7.00	8.00	7.00
	Mean	6.80	7.80	7.30
	StdDev	1.687	1.398	1.636
Homosexual	Median	7.00	3.00	7.00
	Mean	6.40	4.30	6.50
	StdDev	2.319	2.497	2.224
Immigrant	Median	5.00	3.00	5.00
	Mean	5.70	4.00	5.00
	StdDev	2.214	2.867	1.886
Jew	Median	7.00	6.50	6.00
	Mean	6.90	6.20	6.70
	StdDev	1.287	1.989	1.703
Latino	Median	7.00	4.00	5.50
	Mean	6.50	4.50	5.50
	StdDev	1.434	2.068	1.716
Lawyer	Median	5.00	8.00	8.00
	Mean	5.60	8.20	7.90
	StdDev	1.506	.789	1.197
Man	Median	6.50	8.00	7.00
	Mean	6.60	7.70	7.40
	StdDev	1.647	1.418	.843
Mother	Median	8.00	6.00	6.50
	Mean	7.90	6.20	6.80
	StdDev	1.101	1.751	1.476
Muslim	Median	6.00	6.50	5.50
	Mean	5.80	6.30	6.30
	StdDev	1.989	2.312	1.829
Politician	Median	5.00	9.00	8.00
	Mean	5.10	8.70	7.60
	StdDev	1.792	.483	1.506
Priest	Median	7.00	7.50	6.50
	Mean	6.40	7.50	6.40
	StdDev	2.319	1.354	2.119
Protestor	Median	6.00	6.00	8.00
	Mean	6.00	5.60	8.30
	StdDev	1.333	1.647	.675
Racist	Median	2.00	6.50	6.50
	Mean	1.80	6.90	5.40
	StdDev	.919	1.101	2.547

Rapist	Median	1.00	8.00	6.00
	Mean	1.50	7.90	6.10
	StdDev	.707	.994	2.283
Republican	Median	5.00	8.00	7.50
	Mean	5.40	7.90	7.10
	StdDev	1.776	1.101	1.663
Slut	Median	4.50	5.00	5.50
	Mean	4.80	4.90	5.80
	StdDev	1.989	2.025	1.814
Soldier	Median	6.50	7.00	8.50
	Mean	6.50	6.90	8.10
	StdDev	2.068	1.287	1.101
Son	Median	6.50	6.50	7.50
	Mean	6.40	6.40	7.10
	StdDev	1.430	1.776	1.524
Teacher	Median	7.50	5.50	5.50
	Mean	7.50	5.50	6.10
	StdDev	1.179	2.224	1.729
Terrorist	Median	1.00	7.50	8.00
	Mean	1.60	7.10	7.50
	StdDev	.843	1.449	1.509
Victim	Median	6.50	2.00	3.50
	Mean	6.40	1.90	3.60
	StdDev	1.897	.876	1.430
Woman	Median	8.00	4.50	5.00
	Mean	7.30	5.00	5.20
	StdDev	1.703	1.247	1.317

A5 | Classifier 1 Predictive Accuracy Rate by Identity Word

Number	Identity	Subj1	Subj2	Subj3
1	American	0.6667	1	0.8333
2	Arab	NaN	0.25	1
3	Atheist	NaN	1	0.4286
4	Christian	NaN	0.4	1
5	Immigrant	NaN	0.6667	1
6	Jew	NaN	0.6667	0.5
7	Latino	0.5	0.5	0.3333
8	Muslim	0.75	NaN	NaN
9	Racist	0.6	0.2	0.2
10	Heterosexual	0.6667	NaN	0.8889
11	Homosexual	NaN	NaN	1
12	Man	NaN	0.8333	0.6667
13	Rapist	0.625	0.25	0.1667
14	Slut	NaN	NaN	NaN
15	Woman	1	0.75	1
16	Daughter	0.3333	0.8333	0.8333
17	Father	0.4444	0.7778	0.4444
18	Mother	1	0.75	1
19	Son	0.5556	0.875	0.5
20	Cop	0.4	1	0.8571
21	Doctor	0.3333	1	1
22	Farmer	0.3333	0.6667	1
23	Lawyer	NaN	NaN	NaN
24	Politician	0.5714	NaN	NaN
25	Priest	0.7143	0.75	0.4286
26	Soldier	0.6	0.6667	0.8
27	Teacher	0.5	0.625	0.8
28	Democrat	NaN	NaN	0.7143
29	Environmentalist	NaN	NaN	0.7143
30	Feminist	NaN	NaN	1
31	Protestor	NaN	NaN	NaN
32	Republican	0.4	NaN	NaN
33	Terrorist	0.5	0.5	0.5
34	Alcoholic	0.4286	NaN	NaN
35	Celebrity	NaN	1	0.75
36	Cheater	0.5	0.1667	0.3333
37	Victim	0.75	NaN	1
38	Banker	NaN	0.625	NaN
39	Executive	0.4	0.5	NaN
40	Entrepreneur	0.5	0.625	0.5556

NaN means the Identity was rated as neutral and thus not involved in classification.

A6 | Complete descriptive statistics for classifier 2

	subj1_verb	*subj1_adj*	*subj2_verb*	*subj2_adj*	*subj3_verb*	*subj3_adj*
Mean	0.1658	0.1518	0.2433	0.0730	0.1706	0.1385
Standard Error	0.0877	0.0793	0.0828	0.0914	0.0801	0.0958
Median	0.3600	0.2450	0.2550	0.2350	0.2527	0.3100
Mode	0.3600	-0.0700	0.8200	-0.2100	--	-0.2000
Standard Deviation	0.5550	0.5014	0.5237	0.5779	0.5066	0.6056
Sample Variance	0.3080	0.2514	0.2743	0.3339	0.2566	0.3668
Kurtosis	-1.2495	-0.8986	-0.7949	-1.2243	-0.9165	-1.3873
Skewness	-0.3104	-0.1745	-0.3943	-0.1874	-0.3375	-0.2406
Range	1.8700	1.8500	1.9700	1.9000	1.8872	1.8700
Minimum	-0.9000	-0.9200	-0.9900	-0.9500	-0.9078	-0.8700
Maximum	0.9700	0.9300	0.9800	0.9500	0.9794	1.0000
Sum	6.6300	6.0700	9.7300	2.9200	6.8223	5.5400
Count	40	40	40	40	40	40

A7 | Complete descriptive statistics for classifier 3

	subj1_2reg	*subj1_eval*	*subj2_2reg*	*subj2_eval*	*subj3_2reg*	*subj3_eval*
Mean	0.1593	0.0970	0.1573	-0.0373	0.1543	0.1855
Standard Error	0.0652	0.0946	0.0654	0.0891	0.0647	0.1027
Median	0.1300	0.1250	0.1450	-0.0500	0.0750	0.4150
Mode	0.1400	-0.0400	0.2100	0.0100	0.0200	0.8900
Standard Deviation	0.4124	0.5981	0.4133	0.5635	0.4091	0.6492
Sample Variance	0.1701	0.3578	0.1709	0.3175	0.1674	0.4215
Kurtosis	-0.7911	-1.2666	-0.0348	-1.2042	-0.7547	-1.4769
Skewness	-0.2094	-0.0876	-0.2626	0.0541	0.0788	-0.3987
Range	1.5000	1.8600	1.7500	1.9500	1.6100	1.8300
Minimum	-0.6800	-0.8700	-0.8300	-0.9800	-0.6800	-0.8600
Maximum	0.8200	0.9900	0.9200	0.9700	0.9300	0.9700
Sum	6.3700	3.8800	6.2900	-1.4900	6.1700	7.4200
Count	40	40	40	40	40	40

A8 | Classifier 2 Predictive Correlation Rate by Identity Word

Number	Identity	Subj1_verb	Subj2_verb	Subj3_verb	Subj1_adj	Subj2_adj	Subj3_adj
1	American	-0.5	0.44	-0.2829	0.38	-0.13	0.34
2	Arab	-0.36	0.82	0.0205	-0.28	0.88	-0.75
3	Atheist	0.03	0.19	0.2465	-0.07	0.3	-0.21
4	Christian	0.36	0.85	0.5062	-0.12	0.68	0.94
5	Immigrant	0.66	0.16	0.7844	0.69	-0.21	0.39
6	Jew	-0.32	0.98	0.5764	0.03	0.83	-0.87
7	Latino	0.85	0.77	0.2588	-0.45	0.69	-0.2
8	Muslim	0.79	0.82	0.7515	0.38	-0.55	0.35
9	Racist	0.53	-0.23	0.843	0.43	0.95	0.72
10	Heterosexual	0.42	0.24	-0.5102	0.29	0.31	0.7
11	Homosexual	0.6	0.4	0.1389	-0.02	-0.42	-0.18
12	Man	0.5	0.65	0.7296	0.43	-0.66	0.93
13	Rapist	-0.67	-0.11	0.153	-0.08	0.36	-0.62
14	Slut	0.83	0.77	0.6076	0.33	0.25	-0.43
15	Woman	-0.12	0.84	0.5328	0.4	-0.21	0.41
16	Daughter	0.71	-0.73	0.3366	0.93	-0.94	0.66
17	Father	0.15	0.06	0.8384	-0.13	-0.35	0.04
18	Mother	-0.44	0.42	0.8559	-0.92	0.81	1
19	Son	0.54	0.39	0.1508	-0.3	0.5	0.72
20	Cop	0.57	0.73	-0.9078	0.74	0.39	-0.24
21	Doctor	-0.45	0.19	-0.122	0.75	-0.06	0.58
22	Farmer	0.5	0.27	0.3247	-0.33	-0.24	-0.2
23	Lawyer	-0.56	0.88	-0.4386	-0.56	-0.84	0.48
24	Politician	-0.56	-0.99	-0.6714	-0.42	0.78	0.7
25	Priest	-0.04	-0.5	-0.1217	0.2	-0.33	-0.71
26	Soldier	0.64	-0.39	0.1261	0.82	0.22	-0.76
27	Teacher	0.36	-0.15	-0.5719	-0.33	0.79	-0.2
28	Democrat	0.85	0.9	0.2688	0.49	0.44	0.81
29	Environmentalist	-0.17	0.9	-0.503	0.29	-0.33	0.93
30	Feminist	0.54	-0.25	0.9794	0.84	-0.29	0.53
31	Protestor	-0.16	-0.19	-0.0727	-0.07	0.41	0.28
32	Republican	-0.9	0.17	0.6729	0.35	-0.24	-0.68
33	Terrorist	0.74	0.12	-0.5088	0.85	0.52	-0.84
34	Alcoholic	0.42	-0.51	0.0319	0.9	-0.66	-0.14
35	Celebrity	-0.53	0.36	-0.6001	0.81	0.28	0.86
36	Cheater	0.95	-0.34	0.7117	-0.51	-0.95	0.69
37	Victim	-0.84	-0.28	0.2702	-0.31	0.56	-0.58
38	Banker	0.97	0.78	-0.4212	-0.07	-0.82	0.03
39	Executive	-0.32	0.65	0.3549	0.5	0.81	0.7
40	Entrepreneur	0.06	-0.35	0.4831	-0.79	-0.61	-0.64

Identity words that have correlations above 0.5 have been highlighted in red for convenience.

A9 | Classifier 3 Predictive Correlation Rate by Identity Word

Number	Identity	Subj1_2reg	Subj2_2reg	Subj3_2reg	Subj1_eval	Subj2_eval	Subj3_eval
1	American	-0.56	-0.27	0.03	-0.42	0.01	-0.35
2	Arab	-0.68	0.25	-0.36	-0.87	-0.98	-0.69
3	Atheist	-0.49	0.71	0.02	-0.8	-0.11	0.08
4	Christian	-0.38	-0.69	0.72	-0.84	-0.66	0.91
5	Immigrant	0.29	0.21	0.59	0.01	-0.54	0.36
6	Jew	0.14	-0.03	-0.15	0.16	0.4	0.16
7	Latino	0.36	-0.37	0.03	0.26	0.64	-0.77
8	Muslim	0.14	0.4	0.55	0.81	0.37	-0.66
9	Racist	0.01	0.5	0.78	-0.76	-0.18	0.85
10	Heterosexual	0.66	-0.02	0.09	0.59	-0.45	0.37
11	Homosexual	0.67	0.14	-0.02	0.87	0.56	0.16
12	Man	0.08	0.41	0.83	-0.35	-0.18	0.83
13	Rapist	0.69	-0.04	-0.24	0.82	-0.56	0.79
14	Slut	0.2	0.47	0.09	-0.71	0.74	0.76
15	Woman	0.67	0.87	0.47	0.32	0.47	-0.16
16	Daughter	0.09	0.21	0.5	-0.04	-0.33	-0.82
17	Father	0.58	0	0.44	0.1	-0.64	-0.31
18	Mother	0.79	-0.83	0.93	0.94	-0.35	-0.53
19	Son	-0.14	0.39	0.44	-0.04	0.17	0.65
20	Cop	0.58	0.33	-0.57	0.86	-0.86	0.89
21	Doctor	-0.27	0.06	0.23	0.36	0.37	0.52
22	Farmer	0.22	-0.19	0.06	-0.17	-0.81	0.89
23	Lawyer	0.82	-0.46	0.02	0.85	0.62	-0.01
24	Politician	0.73	0.14	0.02	0.95	-0.13	-0.69
25	Priest	0.47	0.12	-0.42	0.15	-0.45	-0.71
26	Soldier	0.09	-0.17	-0.32	0.23	0.52	0.91
27	Teacher	-0.12	0.15	-0.39	-0.77	0.01	0.48
28	Democrat	0.48	-0.54	0.54	0.35	-0.6	-0.75
29	Environmentalist	-0.57	0.24	0.21	0.65	0.28	-0.86
30	Feminist	0.45	0.92	0.76	-0.56	-0.97	0.89
31	Protestor	0.02	0.1	0.1	-0.16	0.05	0.56
32	Republican	0.12	0.39	0	-0.11	0.82	0.73
33	Terrorist	0.15	-0.13	-0.68	0.51	0.18	0.55
34	Alcoholic	-0.36	-0.01	-0.05	-0.49	0.27	-0.72
35	Celebrity	0.66	0.19	0.13	0.8	-0.56	0.93
36	Cheater	0.12	0.73	0.7	0.67	-0.44	0.46
37	Victim	-0.02	0	-0.16	0.99	-0.69	0.73
38	Banker	-0.06	0.68	-0.2	-0.24	0.84	0.61
39	Executive	0.06	0.81	0.53	-0.7	0.97	0.97
40	Entrepreneur	-0.32	0.62	-0.08	-0.34	0.71	-0.59

Identity words that have correlations above 0.5 have been highlighted in red for convenience.

A10 | List of Identities with Classifier 2 Correlations Greater than 0.5 by Subject

Subj1_verb	Subj2_verb	Subj3_verb	Subj1_adji	Subj2_adji	Subj3_adji
Immigrant	Arab	Christian	Immigrant	Arab	Christian
Latino	Christian	Immigrant	Daughter	Christian	Racist
Muslim	Jew	Jew	Cop	Jew	Heterosexual
Racist	Latino	Muslim	Doctor	Latino	Man
Homosexual	Muslim	Racist	Soldier	Racist	Daughter
Man	Man	Man	Feminist	Mother	Mother
Slut	Slut	Slut	Terrorist	Son	Son
Daughter	Woman	Woman	Alcoholic	Politician	Doctor
Son	Cop	Father	Celebrity	Teacher	Politician
Cop	Lawyer	Mother	Executive	Terrorist	Democrat
Farmer	Democrat	Feminist		Victim	Environmentalist
Soldier	Environmentalist	Republican		Executive	Feminist
Democrat	Banker	Cheater			Celebrity
Feminist	Executive				Cheater
Terrorist					Executive
Cheater					
Banker					

A11 | List of Identities with Classifier 3 Correlations Greater than 0.5 by Subject

Subj1_2reg	Subj2_2reg	Subj3_2reg	Subj1_eval	Subj2_eval	Subj3_eval
Heterosexual	Atheist	Christian	Muslim	Latino	Christian
Homosexual	Racist	Immigrant	Heterosexual	Homosexual	Racist
Rapist	Woman	Muslim	Homosexual	Slut	Man
Woman	Feminist	Racist	Rapist	Lawyer	Rapist
Father	Cheater	Man	Mother	Soldier	Slut
Mother	Banker	Daughter	Cop	Republican	Son
Cop	Executive	Mother	Lawyer	Banker	Cop
Lawyer	Entrepreneur	Democrat	Politician	Executive	Doctor
Politician		Feminist	Environmentalist	Entrepreneur	Farmer
Celebrity		Cheater	Terrorist		Soldier
		Executive	Celebrity		Feminist
			Cheater		Protestor
			Victim		Republican
					Terrorist
					Celebrity
					Victim
					Banker
					Executive