INNOVATIONS IN SOCIAL PSYCHOLOGY

Six Dimensions Describe Action Understanding: The ACT-FAST Taxonomy

Mark A. Thornton1 and Diana I. Tamir2

1 Department of Psychological and Brain Sciences, Dartmouth College
2 Department of Psychology and Princeton Neuroscience Institute, Princeton University

Humans engage in a wide variety of different actions and activities. These range from simple motor actions like reaching for an object, to complex activities like governing a nation. Navigating everyday life requires people to make sense of this diversity of actions. We suggest that the mind simplifies this complex domain by attending primarily to the most essential features of actions. Using a parsimonious set of action dimensions, the mind can organize action knowledge in a low-dimensional representational space. In seven studies, we derive and validate such an action taxonomy. Study 1 uses large-scale text analyses to generate and test potential action dimensions. Study 2 validates interpretable labels for these dimensions. Studies 3–5 demonstrate that these dimensions can explain human judgments about actions. We perform model selection on data from these studies to arrive at the optimal set of six psychological dimensions, together forming the Abstraction, Creation, Tradition, Food, Animacy, Spiritualism Taxonomy (ACT-FAST). Study 6 demonstrates that ACT-FAST can predict socially relevant qualities of actions, including how, when, where, why, and by whom they are performed. Finally, Study 7 shows that ACT-FAST can explain action-related patterns of brain activity using naturalistic functional MRI (fMRI). Together, these studies reveal the dimensional structure the mind applies to organize action concepts.

Keywords: action, fMRI, machine learning, taxonomy, text analysis

Supplemental materials: https://doi.org/10.1037/pspa0000286.supp

Successful social interactions depend on people’s ability to understand each other’s actions. Reading actions allows people to coordinate with each other, whether to pull in the same direction as their teammates in a tug-o-war or to fulfill complementary roles on a department committee. Reading actions likewise allows people to effectively compete, whether by wrong-footing an opponent on the tennis court or luring them into a chess gambit. Actions give people a key to understanding the hidden movers of the social world, such as emotions, traits, and situations. By carefully observing and understanding actions, people can infer others’ thoughts, feelings, and intentions (Baker et al., 2009; Frijda, 2004), judge others’ dispositions (Ajzen, 1987; Jones & Davis, 1965), or learn the affordances that define a social situation (Bach et al., 2014; Chemero, 2003; Gibson, 1977). The brain reflects the ubiquitous importance of action understanding in daily life: Perceiving agentic action strongly drives neural activity across more than half of cortex (Haxby et al., 2020). Given the manifest value of understanding actions, how do people do it?

Taxonomizing Actions

Humans engage in an immensely diverse range of actions. We perform simple motor behaviors, such as reaching and grasping, complex motor behaviors, such as exercising and dancing, and even abstract, extended activities, such as programming computers and governing nations. If we use verbs as a proxy for actions, there are more than 11,000 actions reflected in the English language alone (Miller, 1995). Understanding this menagerie of actions offers a complex challenge. In this investigation we aim to describe how the mind meets this challenge. We do so by developing a taxonomy that describes action understanding.
For present purposes, we define action as discrete, meaningful events caused by one or more human, living, or nonliving entity. By discrete, we mean that an action must be perceived by humans as separate from other actions. For example, if one sprinted 20 m, most observers would perceive this as one action (sprinting) rather than many repeated instances of separate actions (raising leg, lowering leg, etc.). This criterion helps to distinguish actions from the individual motor impulses which (may) compose them. By meaningful, we mean that actions must be events that humans can and do notice and ascribe value to. This definition helps to exclude phenomena that happen at a spatial or temporal scale that humans cannot perceive: too big or too small, too fast or too slow. These two criteria are deliberately anthropocentric because the goal of the present investigation is not to taxonomize actions themselves, or their objective physical features, but rather how people understand actions. Under this definition, most actions can be encapsulated by a verb or verb phrase, a fact we will capitalize on when deriving candidate dimensions for our taxonomy (Study 1).

Our definition of actions is deliberately broad. It includes both physical actions (reaching for a glass) and mental actions (planning one’s day). It also includes deliberate actions driven by some intention or goal (making a cake), accidental actions performed by agents capable of intention (tripping over a log), and mechanical actions performed by nonsentient entities (a press stamping designs of coins). It may well be that people make major distinctions between these different types of actions. However, our intention is to document such distinctions, rather than to use them to exclude certain actions. That said, our definition is not so broad as to include all aspects of human experience. For example, merely being in a certain state, such as hunger or frustration, would not be considered an action under our definition. Moreover, things that happen without an identifiable entity or set of entities causing them (e.g., the passage of time) would also not be considered an action.

The taxonomy we will construct aims to achieve three goals: First, it should allow us to predict useful, real-world features of actions. This includes what type of person tends to perform a given action, when and where they tend to perform it, why people perform it, and how they perform it. Human action understanding clearly permits these sorts of inferences, so any taxonomy of that understanding should do the same. Second, a successful action taxonomy should capture action concepts. For instance, it should accurately predict which actions people judge to be conceptually similar or different from one another (Goldstone et al., 1991; Tversky, 1977). This goal provides us with a straightforward criterion by which we can assess a taxonomy. Such assessment in turn allows us to optimize any taxonomy we derive and to compare it against potential alternative theories. Third, the ideal taxonomy will be as parsimonious as possible. That is, it will explain as much variance as possible with the fewest dimensions possible. Maximizing parsimony promises to make the taxonomy as interpretable as possible for researchers, and as generalizable as possible across different applications.

As important as is it to describe what a taxonomy aims to achieve, it is equally important to delineate what it does not. Actions may be described at many different levels of abstraction. Some researchers study actions at the level of motor planning and execution, dissecting each movement into the nerve impulses and muscle contractions that compose it (Desmurget & Grafton, 2000; Goldberg, 1985; Rizzolatti et al., 2002; Wessberg et al., 2000). Others study the visual perception of action, trying to understand how we piece together and label actions from visual inputs (Buccino et al., 2001). Still others study verbs, the formalized tokens that frequently represent actions in language (Naigles, 1990). All of these lines of research are crucial to achieving a complete scientific understanding of actions. However, given the vast differences between these levels of abstraction, it seems unlikely that a single tractable taxonomy can successfully embrace them all. We target our investigation specifically at action concepts. That is, we aim to describe how people think about actions using a parsimonious but comprehensive taxonomy of action knowledge.

**Psychological Taxonomies**

Taxonomies—and multidimensional maps in particular—have a long history in social and personality psychology and allied fields. Taxonomies such as the Big Five personality traits (Goldberg, 1990; McCrae & Costa, 1987), the circumplex model of affect (Russell, 1980), relational models theory (Haslam & Fiske, 1999), and the stereotype content model (Cuddy et al., 2008; Fiske et al., 2002) have exerted a profound influence on research for decades. In recent years, researchers have advanced this taxonomic tradition by deriving new taxonomies of domains including situations (Parrigon et al., 2017; Rauthmann et al., 2014), goals (Wilkowski et al., 2019), and major life events (Luhmann et al., 2020). Many of these more recent taxonomies are higher-dimensional than their forebearers, comprising as many as eight dimensions rather than just a two-dimensional map. This might reflect the increasingly complex domains that the field has embarked on studying, or the increasing methodological and computational tools available for dealing with high dimensional data (Jolly & Chang, 2019).

Existing taxonomies differ in many ways, but many are united by their common focus on the hidden movers of the social world. Traits, emotions, stereotypes, situations, goals, and events are all intangible latent constructs. People can often infer these abstractions, but doing requires integration of multiple cues over an extend period of time (Barrett et al., 2011; Jones & Davis, 1965; Zaki et al., 2009). These hidden movers play crucial roles in shaping social behavior. However, observable social behavior also plays a crucial role in shaping these latent constructs. Action is necessary to enact our thoughts, feelings, and intentions or to make stereotypes and situations manifest in the physical world outside our heads. A person could be electrified by excitement or bored out of their mind, but if their actions never betray this inner state then it has minimal impact on the social world.

The connections between the hidden movers of the social world and the manifest actions they cause illustrates the potential value of an action taxonomy. Such a taxonomy would complement our growing understanding of the latent social world by offering a compatible description of the manifest world with which it interacts. This manifest world is full of comparatively concrete stimuli, such as action and objects, that can typically be recognized instantaneously without conscious effort (Hauk et al., 2008; Liu et al., 2009). This concreteness may make the domain of actions seem more straightforward than other domains, but it belies the complexity and importance of carefully describing action understanding.

The pivotal role that actions play in defining other domains of cognition makes them all the more important to explicate. A taxonomy of actions would fill a missing link in social psychological...
theory. Filling it may allow for advances not just with respect to action understanding but also across the field. For example, we know that certain emotions tend to produce certain actions: Anger leads to aggression, love leads to affection, and so on. These emotions are well described by their positions on taxonomic dimensions such as valence: anger is negative and love is positive (Russell, 1980; Thornton & Tamir, 2020a). One could imagine a model that uses emotions to predict actions by taking in the taxonomic coordinates of emotions and outputting the coordinates of likely actions. Such a model would provide a mathematically precise, testable account of emotion-action contingency. Moreover, a relatively low-dimensional taxonomy would make such predictions both easier to interpret, and more likely to generalize. However, to achieve such a useful advance in theory, it would first be necessary to construct a dimensional taxonomy of actions.

The Present Investigation

The present investigation aims to create a comprehensive taxonomy of action understanding. We begin in Studies 1 and 2 by generating a set of candidate action dimensions using large-scale text analyses. This data-driven approach allows us to discover the hidden structure of the action domain without being restricted by preconceived notions of important action dimensions. Next, in Studies 3 and 4, we aim to make our taxonomy as parsimonious as possible by winnowing down the initial set of candidate dimensions. We achieve this by testing the ability of each dimension to contribute to predicting conceptual judgments of actions. Study 5 demonstrates that the resulting set of dimensions are truly a taxonomy of actions, rather than their lexical representations (i.e., verbs), by showing that they can predict judgements about complex actions formed around the same verb (e.g., “run a mile” vs. “run an analysis”). Study 6 shows that the taxonomy can predict many socially important aspects of actions, including when, where, why, how, and by whom actions tend to be performed. Finally, in Study 7 we show that the taxonomy can predict the patterns of brain activity elicited by observing actions in a naturalistic functional MRI (fMRI) experiment. Together, these studies provide strong support for a new taxonomy of action understanding we call “ACT-FAST” – the Abstraction, Creation, Tradition, Food, Animacy, and Spiritualism Taxonomy.

Study 1

The goal of Study 1 was to identify a set of candidate dimensions which might describe the action domain. We adopted a data-driven approach to extract these dimensions from natural language texts. In psychology, the information embedded in natural language has proven helpful in building a taxonomy of social situations (Parrigon et al., 2017) and goals (Wilkowski et al., 2019). Here we analyze the linguistic instantiation of actions—verbs—as proxies for action concepts. We look, in particular, at ways in which verbs co-occur with nouns in natural language because of the meaningful information such co-occurrences reflect about actions. If two verbs tend to co-occur with similar nouns, then this might imply that the corresponding actions are performed on similar objects, that similar agents engage in those actions, or that the actions occur in similar contexts. For example, the verbs bake and boil are likely to appear with similar nouns (e.g., potato), indicating that they are related in function and context. Such co-occurrence patterns are the informational basis of computational approaches to estimating word meaning (Mikolov et al., 2013; Pennington et al., 2014). Sets of dimensions derived from word co-occurrences—known as word embeddings—can perform many natural language tasks, such as accurately solving analogy reasoning questions.

Note that direct co-occurrences between verbs might be similarly informative in some cases (e.g., run and jump co-occurring frequently because of their relevance in sports). However, there are a variety of pragmatic effects that interfere with extracting meaning from such co-occurrences alone. For example, many actions are substitutes for each other: one typically would not bake, boil, and fry the same potato, meaning that these actions are actually less likely to co-occur because of their functional similarity. It is for this reason that we focus on a second-order form of similarity: the extent to which verbs co-occur with nouns, rather than each other.

In this study, we first extracted an initial set of dimensions from a corpus of reddit data (Corpus A) and then applied the same method to a second, different body of text (Corpus B) consisting of approximately 1,000 movie scripts retrieved from IMSDb.com. We then compared the dimensions extracted in each of these data sets to determine how well they generalized across corpora. This procedure helps us to avoid overfitting our initial set of dimensions to a particular dataset.

Finally, we obtained initial evidence that the dimensions we extracted described not only verbs as words but also the conceptual meaning of the actions those verbs represent. To achieve this, we used a third dataset (Corpus C) consisting of reddit comments from a six different activity-related subreddits. We quantified where each of these comments fell on our candidate action dimensions and then used these dimensions to classify each comments subreddit-of-origin. The success of this classification model and its superiority to another well-established language model demonstrate that the candidate Study 1 dimensions offer a meaningful description of actions.

Method

Statistical analyses were carried out using R (R Core Team, 2015). Where applicable, we report how we determined sample size, all data exclusions, all manipulations, and all measures in the studies.

Text Corpora

We assembled three large corpora of text via a process known as “web-scrapping”—that is, programmatically retrieving information from the internet. Two of these corpora consisted of comments from reddit.com, a popular online news aggregator and discussion forum, using Python 2.7 (https://www.python.org/) and the Python Reddit API WRAPPER (PRAW: https://github.com/praw-dev/praw) in the fall of 2013 (Thornton, 2014). Corpora A consisted of 9,357,737 comments made on the “AskReddit” subreddit (forum within reddit). AskReddit is a forum in which one user posts a question and other users answer it. Typical questions focus on broad life experiences, rather than specific technical questions. We selected this subreddit because it covers a wide range of representative topics and elicits a high volume of comments.
Corpus B consisted of movie scripts retrieved from IMSDb.com. We selected this corpus for study because of it differed in a variety of potentially important ways from Corpus A, including the unit of analysis (reddit comment vs. line of spoken dialogue), the medium (online text meant to be read vs. offline text meant to be spoken), and the authors (redditors versus Hollywood screenwriters). Python 2.7 (https://www.python.org/) and Beautiful Soup 4 (https://www.crummy.com/software/BeautifulSoup/) were used to retrieve the texts of all movie scripts publicly available in the July 2017. We successfully retrieved 1,068 scripts. Nearly all of the retrieved scripts shared common formatting features which facilitated automated text analysis. In particular, each line of the script (that is, continuous spoken part, not line in the text file) began with a word in ALLCAPS. There were other fully capitalized words in some scripts, for instance to indicate stage directions, but these were relatively rare in comparison with character prompts. This formatting feature allowed us to split each script into separate spoken lines by using a regular expression to search for fully capitalized words. A small number of scripts were formatted differently or were exceptionally short and thus appeared into separate spoken lines by using a regular expression to search for fully capitalized words. A small number of scripts were formatted differently or were exceptionally short and thus appeared to have 500 or fewer lines. We removed these scripts from the corpus, leaving a final set of 1,036 scripts for further analysis.

Corpus C consisted of comments from subreddits focused on six distinct activities: programming (111,446), relationships (153,211), food (88,709), music (339,090), personal finance (163,633), and sports (39,509). We selected these subreddits on the basis of the following constraints: that they be activity-related, rather than focused on a certain type of content or location (for the basis of the following constraints: that they be activity-related, rather than focused on a certain type of content or location (for example, a TV show’s or city’s subreddit); that they have a large amount of text-based material to scrape and analyze, and were roughly similar in size; that they represented distinctly different activities, so that classification would be possible in principle; and that the subreddits represent activities that are common and important in everyday life.

Measuring Co-Occurrence
To measure verb–noun co-occurrence it was necessary to select sets of verbs and nouns. To do so, we used the SUBTLEXUS frequency norms (Brysbaert & New, 2009). From these norms, we identify initial sets of 16,865 words that were used most often as verbs, and 37,333 words that were used most often as nouns. To minimize redundancy, we collapsed then across different forms of the same word (run versus ran) and across synonyms, as defined by WordNet “synsets” (Miller, 1995). Light and auxiliary verbs (that is, do, is, have, can, will, give, and take, as well as the artifactual wan) were excluded because they possess little or no semantic content of their own and do not describe specific actions. Finally, we eliminated verb groups with 50 or fewer occurrences in SUBTLEXUS and noun groups with 51 or fewer occurrences to yield approximately even set of 1,875 verbs and 1,729 nouns. See the online supplemental materials for further details on word selection.

For in Corpus A, we tabulated how often verbs and nouns co-occurred within the reddit comments. We divided each comment into sentences, then into words, using the Python Natural Language Toolkit (Bird, 2006; https://www.nltk.org/). Each comment contained an average of 2.41 sentences (median = 2) and 34.08 words (median = 18). We checked whether each word belonged to one of the selected verb or noun groups. Different forms of the same verb, or synonymous verbs, counted toward the same group, and likewise for nouns. After all verb/noun group occurrences in a given comment had been counted, these counts were added to a single verb-by-noun co-occurrence matrix. For example, if a comment contained one occurrence each of the verbs “to run” and “to eat” and the noun “shoes,” then a “1” would be added to the “to run” by “shoes” element of the co-occurrence matrix, and the “to eat” by “shoes” element. We tabulated co-occurrences at a relatively broad level (that is, the comment, rather than the sentence or adjacent word) to emphasize semantic associations in the data and suppress short-range syntactic effects. After all comments had been analyzed and added to the co-occurrence table, the table was then normalized to mitigate word-frequency effects. Following recent work (Huth et al., 2016), this normalization consisted of z-scoring the matrix along first rows, then columns. The exact same procedure was then applied to Corpus B. However, because this corpus consisted of movie scripts instead of reddit comments, we tabulated co-occurrences within spoken lines of dialogue.

Principal Components Analysis

Principal component analysis (PCA) was used to perform dimensionality reduction on the normalized verb–noun co-occurrence matrices from Corpora A and B. Verb groups were treated as observations, and noun groups were treated as variables. To determine the optimal number of components to retain, we wished to identify components that generalized across the two corpora, rather than just explaining idiosyncratic variance in one corpus or the other. Our reasoning was that components that fail to generalize between two text corpora are also unlikely to generalize to other measures (for example, human judgments) of action meaning. To identify generalizable components, we computed all pairwise Tucker congruence coefficients between the loadings of the first 200 components. We then computed the root mean square (RMS) of the congruence matrices corresponding to retaining 1–200 principal components. We retained the number of components that maximized this RMS. This allows us to select a component structure that best generalizes from one corpus to the other. The resulting structure was not subjected to rotation (see the online supplemental materials).

Classification Analyses

Using the components extracted from Corpus A, we sought to classify the subreddit-of-origin of comments in Corpus C. Because the subreddits each focused on a different activity, if this classification proved accurate it would indicate that the components encode not just word-level information but high-level action concepts. We retained all comments in Corpus C that had at least one of the verbs used in the Corpus A/B analyses. This yielded a total of 736,102 comments. To eliminate class imbalance, we reduced each subreddit to the number of comments contained in the smallest sample (sports: 30,999). Thus, 30,999 comments were randomly selected from each subreddit for further analysis. These comments were further divided into a training set (80%) and test set (20%). All comments were scored on the eight candidate action dimensions (that is, principal components) from Corpus A. This was accomplished by identifying all verbs in a given comment and averaging their corresponding scores on the dimensions.

We tested the extent to which the candidate action dimensions could successfully classify text as coming from subreddits focused
on distinct activities. We attempted to classify the subreddit-of-origin of the reddit comments with three potential classification algorithms: Gaussian Naive Bayes (GNB), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). GNB is a simple generative classifier, relatively resistant to overfitting. LDA is another linear classifier, which tends to overfit in high-dimensional cases, but can be more powerful than GNB with sufficient training data. QDA employs the same algorithm as LDA but uses a nonlinear (quadratic) boundary to separate classes. This makes it more prone to overfitting, but also capable of outperforming LDA in cases which require a nonlinear solution. We used a fivefold cross-validation procedure within the training set to compare the performances of these classification algorithms. LDA yielded the best performance (42.1% accuracy).

Having selected the optimal classifier, we then trained the LDA classifier on the entire set of training comments. We tested this classifier on the held-out test set to measure the performance. The held-out test set consisted of 37,196 comments, approximately evenly divided across the six subreddits. In addition to overall accuracy, we also calculated the sensitivity (probability that a comment was classified in a particular subreddit, given that it actually came from that subreddit) and specificity (probability that a comment classified in a particular subreddit actually came from that subreddit) of each classifier with respect to each subreddit. We computed chance accuracy as the frequency of the most common class. Because the classes were artificially balanced, this worked out to 17.0% (approximately 1/6).

Null hypothesis significance testing was used to compare the performance of each model’s classifications against chance. This was achieved via a permutation testing procedure. Specifically, we repeatedly (10,000 times) randomized the classified subreddit labels with respect to the actual subreddit labels. On each occasion, we computed the accuracy obtained by this permuted set of classifier labels. Together these permuted accuracy values formed empirical null distributions. We compared the actual accuracy obtained by the LDA model against the empirical null distribution to compute the probability that the actual accuracy was consistent a chance classification rate.

In addition to testing classification accuracy against a chance baseline, we also compared it with an alternative model based on GloVe. GloVe is a word embedding system which represents the semantics of words as 300-dimensional numerical vectors (Pennington et al., 2014). We used a pretrained version of GloVe based on the Common Crawl: a set of 840 billion tokens generated by scraping the entire web. For model comparison, we derived an eight-dimensional verb taxonomy from GloVe. To derive this alternative model, we needed to select a subset of the 300 semantic dimensions that best captured actions. Dimensions that code for action-specific information should show greater variance within verbs than across the full set of words in the corpus. Thus, we compared the standard deviation of each dimension for the entire vocabulary with the standard deviation of each dimension for just the Wordnet verbs describe above. For each dimension, we divided verb-SD by word-SD to measure the extent to which each dimension captured action representation.

Next, we applied a Gaussian mixture model to optimally cluster the resulting quotients. This analysis indicated that a three-cluster solution achieved the best fit, as measured by Bayesian Information Criterion (BIC). This solution produced an eight-dimension cluster which had the highest ratio of verb variance to general word variance, as well as a cluster of 236 dimensions which were substantially less variable among verbs than words in general, and an intermediate cluster with the remaining 56 dimensions. We retained the eight dimensions in the cluster with relatively high verb-SDs versus word-SDs. These eight dimensions served as a point of comparison against which to benchmark the dimensions extracted from Corpus A. Comments in Corpus C were scored on the eight GloVe dimensions in the same way that they were scored on the Corpus A dimensions, by identifying all verbs in a given comment and averaging their corresponding scores on the dimensions. We again examined GNB, LDA, and QDA classification accuracy within the training set to select the best classifier. LDA again yielded the best performance training set accuracy for the GloVe dimensions (36.3% accuracy).

Finally, we compared the Corpus A dimensions with the GloVe alternative. To do so, we permuted classification labels between the two models. For example, if the Study 1 dimensions predicted “food” for a certain comment, and the GloVe dimensions predicted “programming” for the same comment, these two labels might be switched (probability of 50%). After thus permuting which model each classification originated from, we recomputed the accuracy of each model. We then computed the difference between these two accuracies. Because the origin of each classifier label was randomized, the expectation was that the average difference in accuracy between two permuted models would be zero, with some distribution around this mean due to chance. We compared the actual difference in classifier accuracy to this null distribution to compute a value for the difference in performance between the models.

Results

PCA Results

Visual examination of scree plots from both corpora’s PCA solutions suggested that the first two to eight components explained a disproportionate amount of the variance in verb-noun co-occurrences (Figure 1A and 1B). Our measure of component generalizability between the two corpora—RMS Tucker congruence—achieved a maximum at eight components (Figure 1C) with a value of $\Phi = .71$. This result suggests that retaining eight components would extract the most generalizability structure across the two corpora. Together, these dimensions accounted for 25.5% of the variance in verb-noun co-occurrences in Corpus A, and 11.0% in Corpus B. These low proportions of explained variance reflect the complexity of language.

Classification Results

In our classification analysis, we aimed to test whether the eight dimensions derived from Corpus A can distinguish between real-world discussion about different actions—within an independent Corpus C. We tested whether these dimensions could predict which of six activity-related subreddits each of 37,196 comments originated from. We found that an LDA classifier based on the Corpus A dimensions was able to accurately predict the subreddit-of-origin for 42.3% of comments in Corpus C. This accuracy more than doubled what could be expected from chance alone (chance performance = 17.0%; p < .0001). This model was not just accurate but also sensitive and specific across all six subreddits (see Table 1). This result indicates that the semantic information captured by the Corpus A dimensions can successfully distinguish between high-level activities.

The alternative GloVe dimensions were also able to classify Reddit comments far better than could be expected by chance (accuracy =...
The GloVe dimensions also had high sensitivity and specificity for each of the six subreddits (see Table 1). Thus, the GloVe model offers a compelling benchmark to compare Corpus A dimensions. Despite the GloVe model’s robust performance, the Study 1 dimensions were more accurate by 5.5%. This difference was itself statistically significant ($p < .0001$). The Corpus A dimensions also outperformed the GloVe dimensions on sensitivity and specificity, often substantially so (for example, specificity for r/food and r/personalfinance; sensitivity for r/relationships), with the one exception of the sensitivity in the r/food subreddit.

**Discussion**

By analyzing how people discuss actions in text, we were able to extract a set of eight candidate dimensions which might plausibly describe the domain of actions. This set is relatively large by the
standard of many existing dimensional theories in psychology (Jolly & Chang, 2019), yet it is still not large enough to capture the full complexity of natural language. The eight candidate dimensions capture only a fraction of the total variance in verb-noun concurrence. However, language and the psychological concepts that language denotes are not the same. The complexity of human language may outstrip the complexity of the underlying action concepts in certain respects. As such, it is possible that the limited number of dimensions we extracted may provide a reasonable account of actions concepts.

By comparing the dimensions extracted from Corpus A (r/Askreddit comments) with dimensions extracted in the same way from the very different Corpus B (spoken lines of dialogue in movie scripts) we were able to make an initial assessment of the generalizability of these dimensions. These results indicated a high degree of generalization, with most of the prominent dimensions extracted from Corpus A matched by highly correlated dimensions extracted from Corpus B. This suggests that by focusing on a relatively small set of dimensions, we have identified dimensions that are more likely to generalize to new contexts, as opposed to a high dimensional description overfitted to the particular data from which it was generated.

We also examined whether the candidate action dimensions extracted from Corpus A encode meaningful high-level information about actions. Specifically, we tested whether these dimensions distinguish between comments from six activity-specific Internet forums: subreddits devoted to programming, relationships, food, music, personal finance, and sports. We found that the candidate action dimensions were indeed able to accurately classify comments about activities into their subreddit of origin, well above chance levels. This finding provides evidence that these dimensions can indeed be thought of as a potential basis for an action taxonomy rather than a verb taxonomy. We also found that the Corpus A dimensions outperformed a comparable eight-dimensional verb taxonomy derived from the GloVe word vector embedding. This indicates that the text analysis in Study 1—which was specifically tailored to extract verb meaning—outperforms a more generic language model based on a much larger corpus of text.

Study 2

In Study 1, we identified and tested a set of candidate dimensions for describing action understanding. We arrived at these dimensions via quantitative text analyses. This approach allowed us to discover candidate dimensions in a data-driven way. However, this approach has the disadvantage of yielding dimensions that are mathematical abstractions, rather than named, interpretable concepts. To address this limitation, in Study 2, we aimed to match the eight candidate dimensions with interpretable, clearly defined names. This process proceeded through two phases. First, in an iterative exploratory phase, we tested multiple names and definitions for each subreddit. Second, in a confirmatory phase, we performed a preregistered replication (https://osf.io/rjkgg/) of these ratings for the final matched definitions.

Method

Participants

Participants were recruited from Amazon Mechanical Turk using Turkprime (Litman et al., 2017). All participants in this and

Table 1

<table>
<thead>
<tr>
<th>Subreddit</th>
<th>Sensitivity Corpus A dimensions</th>
<th>Sensitivity GloVe dimensions</th>
<th>Specificity Corpus A dimensions</th>
<th>Specificity GloVe dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>r/relationships</td>
<td>62.5%</td>
<td>34.6%</td>
<td>37.4%</td>
<td>32.1%</td>
</tr>
<tr>
<td>r/personalfinance</td>
<td>54.7%</td>
<td>49.6%</td>
<td>53.4%</td>
<td>41.3%</td>
</tr>
<tr>
<td>r/programming</td>
<td>45.8%</td>
<td>44.6%</td>
<td>43.7%</td>
<td>40.7%</td>
</tr>
<tr>
<td>r/music</td>
<td>34.9%</td>
<td>31.3%</td>
<td>36.4%</td>
<td>35.1%</td>
</tr>
<tr>
<td>r/food</td>
<td>30.8%</td>
<td>39.5%</td>
<td>61.9%</td>
<td>37.4%</td>
</tr>
<tr>
<td>r/sports</td>
<td>27.8%</td>
<td>20.2%</td>
<td>30.9%</td>
<td>30.2%</td>
</tr>
</tbody>
</table>

Note. Bolded text indicates the superior sensitivity and specificity in each subreddit.

Figure 2

Comment Classification Accuracy

Note. Bar height indicates the overall accuracy of classifying reddit comments by subreddit achieved by the Study 1 dimensions or the GloVe dimensions. Both models significantly outperformed chance, as indicated by asterisks (**p < .001). The Study 1 dimensions also significantly outperformed the GloVe dimensions.
subsequent studies provided informed consent in a manner approved by the Princeton University Institutional Review Board.

In the exploratory phase of the study, we recruited 516 participants to rate 28 different dimension definitions, targeting a preexclusion sample size of 20 per dimension. Participants were excluded in this and all subsequent studies based on two a priori criteria: if English was not their first language and their command of English was less than excellent (\(N = 3\)), and if they failed to provide more than 10 unique answers on a line scale over the course of the 60-item survey (\(i = 30\)). This yielded a final sample size of 483 (232 female, 246 male, one other, and four who preferred not to answer; mean age = 35.7, range 18–76), across the 28 definitions (sample sizes ranging from eight to 22 per definition).

For the confirmatory portion of the study, a priori power calculations were used to determine that a sample size of 50 per dimension was sufficient to achieve both 90% power, controlling for multiple comparisons, and an interparticipant Cronbach’s \(\alpha\) of .9 or greater for all dimensions. To guard against exclusions and the possibility of stochastic uneven assignment to conditions, we targeted a total sample size 450 participants. We recruited a sample of 456 participants for the confirmatory portion of the study. Participants were excluded due to language (\(N = 3\)), and insufficient unique responses (\(i = 19\)), leaving a final sample of 434 (192 female, 239 male, one other, and two who preferred not to answer; mean age = 35.1, range 19–70). Final sample sizes per dimension ranged from 49 to 58.

**Stimuli**

Two sets of 60 actions—in the form of verbs—were selected for use in the exploratory and confirmatory portions of the experiment, respectively. These sets were selected from the larger set of 1875 verb groups generated in Study 1. The sets were optimized based on the following criteria: maintaining orthogonality between the eight Study 1 dimensions (that is, PCA scores from Corpus A), as well as distance matrices derived from these dimensions; maximizing the average correlation between the Study 1 distance matrices and the correlation matrix between verbs with respect to noun frequency (that is, maximizing how well the dimensions explained verb-noun co-occurrence); minimizing any reduction in this correlation for any dimension; minimizing changes in the means and standard deviations of the Study 1 dimensions relative to the full set of verbs; and minimizing the redundancy of the chosen verbs with respect to one another, as measured by correlation across Study 1 dimensions. Selection was performed via a simulated annealing algorithm. The first set of 60 actions generated by this algorithm were used in the confirmatory portion of the study. These actions were then removed from the total set, and the algorithm as rerun to choose a second set of 60, which was used in the exploratory portion of this study.

**Experimental Procedure**

For each of the eight Study 1 dimensions, we considered what the verbs at one pole of the numerical scores might have in common, what the verbs at the opposite pole might have in common, and what might distinguish the two. Similarly, we examined nouns at each pole of the loadings. Based on an inspection of these poles, we generated candidate names and definitions for each dimension.

At the outset of the experiment, participants were randomly assigned to one dimension, and were provided a definition of that dimension. For example, a participant might be presented with the dimension “animate versus mechanical” defined as actions “done by living agents, such as people or animals” versus actions “done by machines, computer programs, or other automated, artificial processes.” The full set of dimensions are shown in Figure 1 (final set) and Figure S1 in the online supplemental materials (unselected candidates). Each participant then rated a set of 60 actions on that one dimension. For example, they rated “hiking” on a scale from More animate to More mechanical. Participants made these ratings using a continuous line scale (0–100). Actions were presented in a different random order to each participant. At the end of each study, participants reported their demographic information.

The exploratory and confirmatory portions of the study involved identical rating tasks, differing only in the dimensions and sets of actions presented. Across all phases of the exploratory portion of the study, participants rated 28 dimension definitions. We used those data to determine which definitions elicited ratings that were most consistent with each of the eight Study 1 dimensions. The eight definitions that achieved the best matches with the Study 1 dimensions were then rated in the confirmatory portion of the study.

**Statistical Analysis**

In the exploratory study, we aimed to measure the fit of each dimension definition to the Study 1 dimensions. To do so, we first \(z\)-scored each participant’s ratings to mitigate biases in scale use. The resulting values were then averaged across participants to create a single composite for each rated dimension definition. These average ratings were correlated with the PC weights for each of the eight Study 1 dimensions. The definitions in the exploratory phase were rated iteratively across six rounds of data collection. This procedure terminated after we successfully identified definitions matching each of the eight of the Study 1 dimensions. We defined matching as a relative, rather than absolute, property: a rated definition matches a Study 1 dimension if the correlation between them is higher than any correlation of that rated dimension definition with other Study 1 dimension and higher than any correlation of that Study 1 dimension with other rated definitions. The signs of the correlations were ignored for the purpose of calculating matches, because any dimension definition could be subsequently reverse-coded.

We analyzed the confirmatory data using two complementary approaches. First, we used the same approach as in the exploratory portion. That is, these data were \(z\)-scored, averaged within each dimension definition, and then average ratings were correlated with the corresponding dimensions from Study 1. Confidence intervals were produced via percentile bootstrapping. This produced effect sizes that we could compare with the effect sizes on the exploratory data.

Second, we conducted confirmatory analyses using linear mixed effects modeling. For a given dimension definition, unaveraged ratings of each verb were regressed onto the fixed effect of the corresponding Study 1 dimension. To ensure the generalizability of the results, a maximal random effects structure was fit. This included random intercepts for participant and verb, and random slopes for the fixed component within each of these group variables. The statistical significance of the fixed effect was assessed.
via the Satterthwaite approximation for degrees of freedom. The Holm-Bonferroni procedure was used to control for multiple comparisons across the eight dimensions.

Finally, we tested the explanatory power of two sets of dimensions—the Study 1 dimensions (that is, PCA scores) and the named dimensions (that is, human ratings). We test these dimensions in two texts—the reddit text (Corpus A) and IMSDb text (Corpus B)—in an exploratory analysis that allows us to examine how well each set of dimensions generalize to a new corpus (Corpus B). We measured the power of both sets of dimensions to explain variance in verb–noun co-occurrence in each corpus for the 60 action verbs rated in the confirmatory portion of Study 2. Normalized verb–noun co-occurrences from the two corpora were the dependent variable in this analysis. For each of the nouns in Study 1, we extracted that noun’s co-occurrence with the 60 verbs in the confirmatory portion of Study 2. The resulting vector was then regressed onto (a) the Study 1 PCA dimensions and (b) the eight named dimensions from Study 2. We computed the R2 for each such regression and then averaged them across nouns to generate single estimates of the explanatory power of the two sets of dimensions. We expected both the PCA dimensions and the rated dimensions to capture significant variance in both corpora.

### Results

In the exploratory portion of the study, we aimed to find a set of definitions that were well matched to the candidate dimensions derived from text analysis. After six rounds of iterative refinement, we arrived at definitions which matched each of the eight Study 1 dimensions. The absolute correlations between the PCA scores and the matched ratings ranged from $r = .43$ to $r = .67$ (see Figure 3). The names and poles of these eight dimensions, as well as example verbs from each pole (based on the text analysis), are provided in Table 2.

The first dimension—Abstraction—distinguished between abstract actions and concrete actions, and between social actions and (nonsocial) physical actions. The second dimension—Animacy—distinguished between animate actions performed by a living agent (animal or human) and mechanical actions, performed by a machine or force of nature. The third dimension—Spiritualism—recognized the distinction between sacred and profane, separating religious and transcendent actions from mundane, business related activities. The fourth dimension—Food—separated actions related to food and food preparation from other actions. The fifth dimension—Technicality—distinguished between technical actions and more common actions. The sixth dimension—Creation—distinguished between actions involved in the creation and consumption of media from actions involved in crime and its consequences. The seventh dimension—Tradition—distinguished between actions which humans have been performing traditionally for hundreds of years, and innovative actions made possible by modern technology. Finally, the eighth dimension—Wealth—distinguished between actions focused on the acquisition and possession of material wealth, versus actions focused on the acquisition and possession of knowledge.

The confirmatory portion of the experiment further validated these dimensions. As predicted in the preregistration, averaged ratings of all eight dimensions were positively associated with the matched principal components (Figure S2 in the online supplemental materials). In the confirmatory mixed effects models, all eight of these relationships were significant, controlling for multiple comparisons (see Table 3). These results strongly support the appropriateness of the dimension definitions generated in the exploratory portion of the study, by showing that they robustly generalize to a new set of actions.

Finally, we compared the explanatory power of the Study 1 dimensions to the named dimensions in the reddit text (Corpus A) and IMSDb text (Corpus B). The average $R^2$ of the Study 1 dimensions was .36 in Corpus A, and the average $R^2$ of the Study 2 named dimensions was .25 in Corpus A, reflecting the adequate but imperfect nature of the definitions. Importantly, the average $R^2$ for text and human ratings were almost identical in the out-of-sample Corpus B. In Corpus B, the average $R^2$ of the Study 1 dimensions was .18 and the average $R^2$ of the Study 2 named dimensions was .17. Overall, this suggests that the named components can robustly and reliably explain verb usage across different contexts.

### Discussion

In Study 2, we identified interpretable definitions for the eight candidate dimensions of action representation extracted via text analysis in Study 1. An initial exploratory phase considered 28 possible definitions, eventually achieving matches for each of the eight Study 1 dimensions, respectively. In a preregistered follow-up, we confirmed that all eight of these definitions produce ratings that are correlated with the text components in a separate sample of actions. These definitions allowed us to assign names and meanings to the prospective dimensions of the action taxonomy in subsequent experiments.

Additionally, in an exploratory analysis we found that the explanatory power of the Study 1 PCA dimensions and the Study 2 rated dimensions was approximately equal in the separate movie
script text of Corpus B. These results suggest that, in addition to interpretability that comes with names and definitions, turning to human ratings may also bring quantitative benefits in terms of generalization.

**Study 3**

A successful action taxonomy needs to reflect the conceptual structure of human action understanding. Studies 1 and 2 identified and labeled a set of candidate action taxonomy dimensions. In Study 3 we sought to test whether these dimensions capture how humans actually think about actions. To do so, we tested whether these dimensions could explain human judgements of the similarity between actions. Similarity judgements provide a general, face-valid measure of the conceptual content of actions. Participants in this study therefore judged which actions are similar and dissimilar to each other. We expected that actions judged to be similar should have similar dimension ratings, whereas action judged to be dissimilar should have dissimilar dimension ratings. In addition to assessing the contribution of each dimension, we also compared the predictive performance of all of our candidate action dimensions together to an alternative theory.

**Method**

The procedure and analysis of these data were preregistered on the Open Science Framework (https://osf.io/3nzd2/). All deviations from this protocol are described below. Individual difference analyses for Studies 3–5 are reported together in the online supplemental materials.

**Participants**

To estimate an appropriate sample size, we conducted a resampling-based power analysis using data (https://osf.io/gxhkr/) from a previous study of similar design (Thornton & Mitchell, 2017), which featured a similar task. Controlling for familywise error rate of $\alpha = .05$, we estimated that a sample size of 100 participants would provide 96% familywise power, and thus targeted this sample size.

Participants were recruited from Amazon Mechanical Turk using Turkprime (Litman et al., 2017). A total of 101 (35 female, 64 male, two declined to state; mean age = 34.4, range = 21–65) participants completed the experiment. All participants were included in analyses, as no participants met exclusion criteria based on language or by making the same choice on each trial.

**Experimental Procedure**

Participants were linked from Turkprime to a custom JavaScript survey administered through MySocialBrain.org. After completing a demographics questionnaire, participants engaged in an action similarity judgment task. On each trial, participants were presented with three actions: one reference action, and two choice actions. Their task was to judge which of the two choice actions was more similar to the reference action as quickly and accurately as possible. Participants were prompted to respond quickly because we hoped to analyze response times in a future study. However, a preregistered power analysis indicated that this was not feasible (see the online supplemental materials). The actions were sampled randomly from the set of 60 actions from the confirmatory portion of Study 2. Each participant responded to 50 trials over the course of

### Table 2

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Label</th>
<th>Pole 1</th>
<th>Pole 2</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Abstraction</td>
<td>Abstract/social</td>
<td>Concrete/physical</td>
<td>Govern, refute vs. drip, peel</td>
</tr>
<tr>
<td>2</td>
<td>Animacy</td>
<td>Animate</td>
<td>Mechanical</td>
<td>Meow, floss vs. contain, extract</td>
</tr>
<tr>
<td>3</td>
<td>Spiritualism</td>
<td>Work</td>
<td>Worship</td>
<td>Fax, haggle vs. foretell, ascend</td>
</tr>
<tr>
<td>4</td>
<td>Food</td>
<td>Food</td>
<td>Nonfood</td>
<td>Bake, fry vs. detain, testify</td>
</tr>
<tr>
<td>5</td>
<td>Technicality</td>
<td>Technical</td>
<td>Common</td>
<td>Refuel, orbit vs. hint, reassure</td>
</tr>
<tr>
<td>6</td>
<td>Creation</td>
<td>Creation</td>
<td>Crime</td>
<td>Film, sing vs. prosecute, testify</td>
</tr>
<tr>
<td>7</td>
<td>Tradition</td>
<td>Tradition</td>
<td>Innovation</td>
<td>Cook, decorate vs. emit, encrypt</td>
</tr>
<tr>
<td>8</td>
<td>Wealth</td>
<td>Wealth</td>
<td>Knowledge</td>
<td>Sail, gamble vs. decipher, eavesdrop</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Dimension</th>
<th>$b$</th>
<th>df</th>
<th>$t$</th>
<th>$\beta$</th>
<th>$R^2_{\text{marginal}}$</th>
<th>$P_{\text{corrected}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstraction</td>
<td>0.90</td>
<td>85.29</td>
<td>5.46</td>
<td>0.33</td>
<td>0.11</td>
<td>$3.2 \times 10^{-6}$</td>
</tr>
<tr>
<td>Animacy</td>
<td>0.51</td>
<td>16.67</td>
<td>2.79</td>
<td>0.17</td>
<td>0.03</td>
<td>$0.026$</td>
</tr>
<tr>
<td>Spiritualism</td>
<td>1.04</td>
<td>40.50</td>
<td>6.48</td>
<td>0.25</td>
<td>0.06</td>
<td>$7.5 \times 10^{-7}$</td>
</tr>
<tr>
<td>Food</td>
<td>1.07</td>
<td>80.95</td>
<td>4.12</td>
<td>0.24</td>
<td>0.06</td>
<td>$0.0046$</td>
</tr>
<tr>
<td>Technicality</td>
<td>0.89</td>
<td>18.53</td>
<td>3.49</td>
<td>0.19</td>
<td>0.04</td>
<td>$0.010$</td>
</tr>
<tr>
<td>Creation</td>
<td>1.90</td>
<td>50.72</td>
<td>4.84</td>
<td>0.35</td>
<td>0.12</td>
<td>$7.4 \times 10^{-5}$</td>
</tr>
<tr>
<td>Tradition</td>
<td>1.07</td>
<td>4.58</td>
<td>3.65</td>
<td>0.19</td>
<td>0.04</td>
<td>$0.026$</td>
</tr>
<tr>
<td>Wealth</td>
<td>0.84</td>
<td>58.87</td>
<td>2.79</td>
<td>0.13</td>
<td>0.02</td>
<td>$0.021$</td>
</tr>
</tbody>
</table>

*Note.* Here we report confirmatory tests of the appropriateness of eight candidate dimension names. Principal component analysis (PCA) scores from Study 1 dimensions were separately regressed onto matched human ratings of actions in mixed-effects models to estimate the associations between the Study 1 dimensions and their matched definitions. The table relates the statistical parameters of those models, with each coefficient indicating the association between ratings and PCA scores for a given dimension.
the experiment. Responses were excluded if they were made faster than 200 ms. At the end of the study participants completed a brief post survey in which they rated their enjoyment of the study, indicated whether they had previously taken the study or had technical problems, and had the option to provide open-ended feedback.

**Testing Analysis**

**Testing Dimensions.** We first tested whether participants’ similarity decisions could be predicted by similarities between actions on the eight action dimensions described in Studies 1 and 2. A generalized (binomial) mixed effects model was fit to participants’ choices (left vs. right). This model included predictors that reflect each of the eight action dimensions identified in Study 1. We based these predictors on human ratings of the verbs from Study 2. To generate a predictor for one dimension, we computed the absolute difference between the position of each choice action on that dimension and the position of the reference action on that dimension. The predictor for a given dimension was calculated as the difference of these differences.

For example, imagine the reference action was “chop,” the left choice action was “bake,” and the right choice action was “run.” The predictor based on the food dimension would take the value of “chop” on this dimension and subtract it from the values of “bake” and “run.” Because “chop” and “bake” are both food-related, the absolute value difference between them would be small, whereas the absolute value difference between “chop” and “run” on the food dimension would be large. The “chop” minus “bake” difference would then be subtracted away from the “chop” minus “run” difference. A higher value in the resulting difference of differences would indicate that the left choice was more likely—as we would expect in this example. A lower value would indicate that the right choice was more likely. The resulting value would be entered into the mixed effects model as the value for the food predictor for that trial. That is, we would predict that participants would be more likely to select “bake” as similar to “chop” based on its respective positions on the food dimension.

The mixed effects model also included random intercepts for participant, and each of the three actions (left choice, right choice, and reference), as well as a random slope for each fixed effect within participant. Log ratio tests were used to assess the statistical significance of each fixed effect coefficient in the model. Results were corrected for multiple comparisons via the Holm-Bonferroni procedure. We estimated confidence intervals around each fixed effect parameter via bootstrapping, as a secondary method for assessing statistical significance. For computational efficiency, we dropped all random effect terms except the random intercept for participant during this bootstrapping procedure. We preregistered the log ratio tests as our primary inferential procedure for this analysis, and thus insofar as these results differ from the bootstrapping, we default to the former over the latter.

**Model Selection.** Next, we used data from the similarity task to hone our set of eight dimensions down to a smaller set. That is, we aimed to select a subset of the eight action dimensions that best predicted participants’ action decisions. This would allow us to identify the set of dimensions that, together, best capture human action concepts, as measured by behavior. In total, we tested 255 different versions of this model, comprising all possible combinations of the eight dimensions.

To do so, we conducted a cross-validated model selection. The model used the same fixed effects predictors as the mixed effects model described above, but no random effect terms (i.e., it was a binomial logistic regression). The model was fit to only a subset of the data and then tested in the left-out subset. In the preregistration, we planned to use five-fold cross-validation. However, the performances of the best models proved so close to each other that the results were unstable across different random seeds of the cross-validation. Therefore, we used leave-one-participant-out cross-validation instead, to guarantee a unique solution.

Additionally, we initially intended to fit the model separately to the data of each participant in the training set, and then average the coefficients to construct the final model for testing on the held-out data. However, this procedure led to linear separation in numerous cases, which can greatly inflate parameter estimates, leading to spurious poor generalization performance. We instead used a single binomial logistic regression to predict participants’ choices in all of the training data. Performance was measured in terms of the accuracy of these predictions, relative to the held-out participants’ actual choices.

**Model Comparison.** This cross-validation analysis identified the best-performing version of the present taxonomy. We then compared this taxonomy to a preregistered alternative: a 20-category verb taxonomy from WordNet (Miller, 1995). This taxonomy was manually generated by lexicographers sorting verbs into logical groups, such as motion-related verbs, possession-related verbs, and social verbs. If these categories capture the way that people think about actions, then we would expect that verbs within the same category will more often be judged as similar to one another, and verbs in different categories will more often be judged as different. For example, if the reference word in our task was “bake” (category: change) and the choices were “eat” (category: consumption) and “cook” (category: change) then the Wordnet taxonomy would predict that participants should choose “cook” as the more similar word. On this basis, we created a predictor using WordNet. This predictor took on one of three values: 0 if the left choice and the reference are in the same category as each other but a different category than the right choice; .5 neither choice is in the same category as the reference, or if both choices are in the same category; and 1 if the right choice and the reference are in the same category as each other, but a different category than the left choice. We then compared the out-of-sample accuracy of this model to that of the optimal combination of candidate action dimensions returned by the cross-validation procedure.

**Results**

**Parameter Testing**

In Study 3, we examined whether the candidate action dimensions developed and labeled in Studies 1 and 2 could predict human judgements of the similarity between actions. Across all participants, of 5,050 total judgements, 157 were excluded because of short reaction times.

Log ratio tests on generalized (binomial) mixed effects models indicated that five of the eight candidate dimensions made statistically significant unique contributions to explaining which of two actions participants judged to be more similar to a third (see Table 4). The significant dimensions were abstraction, spiritualism, food,
creation, and wealth. Of these, creation and abstraction had the largest effects on participants’ choices. Percentile bootstrapping of the minimal mixed effects model indicated similar results. However, the effect of spiritualism was not significant in this analysis, and the effect of tradition was. This difference may result from the fact that bootstrapping was conducted on a model that did not include a random intercept for participant.

Together these results provide initial validation for a subset of the dimensions derived in Studies 1 and 2. The majority of these candidate dimensions successfully explain explicit human judgments about actions.

Model Selection

The second primary analysis aimed to identify the combination of dimensions that produced the most accurate predictions of participants’ choice behavior, out-of-sample. In this analysis, we fit a binomial regression to the choices of all but one participant, and then tested the predictions of this model in a held-out participant (holding out each participant in turn). We did so for all 255 unique combinations of the eight candidate dimensions (see Figure 4). Across all combinations, cross-validated accuracy in predicting participants’ choices signifi cantly explained human judgments about actions. An optimal combination of these dimensions could predict participants’ choices with 60% accuracy. This model greatly outperformed one of the few existing verb taxonomies. This study thus provides initial validation for the candidate dimensions by demonstrating that their explanatory and predictive power generalizes beyond word co-occurrences to explicit human judgments about action concepts.

Model Comparison

Finally, we compared the best performing combination of our candidate dimensions to an entirely separate theory. This theory consisted of WordNet’s 20-category verb taxonomy. First, we tested whether WordNet was able to successfully predict human judgements. We applied an exploratory mixed effect model paralleling the preregistered analysis conducted with the candidate dimensions to the predictions made by WordNet. Results suggested that it had significant explanatory power (odds = 1.73, \( \chi^2(1) = 12.42, p = .0004 \)). The cross-validated accuracy of the WordNet model in predicting participants’ choices was 52.25%. Thus, the WordNet model offers an above-chance benchmark against which to compare our candidate dimensions (see Figure 4). However, despite this statistically significant performance, 71% of all 255 possible combinations of the candidate dimensions outperformed the WordNet model. The optimal combination of our candidate dimensions outperformed the WordNet model by more than eight percentage points (60.35% vs. 52.25%).

Discussion

The results of Study 3 provide a clear indication that the action dimensions derived in earlier studies may capture the conceptual structure of human action understanding. Five of the eight action dimensions significantly explained human judgments about actions. An optimal combination of these dimensions could predict participants’ judgments with 60% accuracy. This model greatly outperformed one of the few existing verb taxonomies. This study thus provides initial validation for the candidate dimensions by demonstrating that their explanatory and predictive power generalizes beyond word co-occurrences to explicit human judgments about action concepts.

\textbf{Study 4}
specifically tailored to the task of describing action understanding. Rather, it posits that actions can be characterized in terms of the type of person who tends to perform them. This is a useful baseline against which to compare the candidate action dimensions, because the person perception model assumes that actions can be described entirely in terms of their connection to latent mental properties (i.e., traits). We expect that these latent properties do explain a meaningful portion of action judgements, but that they will be outperformed by a taxonomy tailored specifically to the action domain.

**Method**

The procedure and analysis of these data were preregistered on the Open Science Framework (https://osf.io/ra5ec/). Note, in the preregistration we inadvertently refer to the HEXACO dimensions instead of the three-dimensional model of person perception in Data Collection Procedures, although not elsewhere. Additionally, although we included all eight actions dimensions in this study, only the top six dimensions from Study 3 are mentioned in two sections of the preregistration—this was corrected in the rest of the document.

**Participants**

Participants were recruited online via Amazon Mechanical Turk and Turkprime and linked to Qualtrics-based surveys (Litman et al., 2017). In the primary task, participants rated the pairwise similarity between the actions. The target sample size was determined via a resampling-based power analysis using data (https://osf.io/32wrq/) from a previous study (Thornton & Mitchell, 2018). Power was estimated by repeatedly resampling this data and conducting an analysis analogous to the analysis planned for this study. We found that a sample of 200 participants would provide 94.7% family-wise power at a family-wise error rate of .05. A total of 200 participants were recruited to participate in this task. Participants were excluded due to language ($N = 2$) and insufficient unique responses ($N = 11$). This a final sample of 187 (83 female, 103 male, one other; mean age = 35.6, range = 19–72).

We also recruited two separate sets of participants to characterize actions on our candidate action dimensions, and our comparison model. These ratings would be used to assess the predictions made by each model. First, a set of participants rated 60 actions (verbs) on the eight candidate dimensions of action representation, as
defined in Study 2. A target sample size of 450 was determined via the Spearman-Brown Prophecy formula, based on the reliability of the interparticipant composites in Study 2, to provide reliabilities of .90. A total of 451 participants were recruited to rate the 60 actions on the eight candidate dimensions. Participants were excluded because of language (N = 6) and insufficient unique responses (N = 38). This left a final sample size of 407 (184 female, 222 male, one preferred not to indicate gender; mean age = 35.2, range = 19–72). Second, a set of participants rated these same actions on three dimensions of person perception: power, valence, and sociality. A target sample size of 170 was determined via the Spearman-Brown Prophecy formula, based on the reliability of the interparticipant composites in Study 2, to provide reliabilities of .90. A total of 171 participants completed the survey. Participants were excluded for language (N = 1) and insufficient unique responses (N = 2). This left a final sample size of 168 (72 female, 96 male; mean age = 35.9, range = 21–70).

**Experimental Procedure**

Participants in the primary task provided pairwise similarity judgements between the actions. On each trial, participants were presented with two actions, represented by verbs. They used a continuous line scale to indicate how similar they believed the to be to each other. The verbs in this study were selected as the 60 most frequent verb forms—as defined by wordnet—in the SUBTLEXUS frequency norms (Brysbaert & New, 2009), excluding the verb “to fuck” to avoid exposing participants to unnecessary profanity. We focused on only the most frequent verbs because accounting for variance among these verbs would account for an outsized portion of everyday action description. With a set of 60 verbs, 1,770 unique pairings were possible. Each participant responded to a randomly selected subset of 50 of these 1,770, presented in a random order.

To predict people’s ratings of the similarity between actions, we also needed to locate those actions on the dimensions of both our candidate dimensions and the dimensions of the alternative theory. Participants’ action dimension rating tasks used continuous line scales to indicate the position of each verb on one of the eight action dimensions. Participants in the person perception dimension rating tasks made similar ratings with respect to the three dimensions of person perception: power, sociality, and valence. In the case of the person perception dimensions, participants were asked to rate whether someone high or low on these dimensions would have a greater tendency to engage in each action. In both samples, verbs were presented in a different random order to each participant, and each participant rated verbs with respect to only one, randomly selected dimension. At the end of each study, participants in all three samples reported their demographic information.

**Statistical Analysis**

We tested whether participants’ similarity ratings could be predicted by the proximity between actions on the eight candidate action dimensions described in Studies 1–4. In particular, we used participant ratings of each action on each dimension to calculate our primary predictive measure. These ratings were processed by z-scoring within participant and then averaged across participants. The distance between two actions on each dimension should reflect how conceptually dissimilar participants rate the two actions. We tested whether distance on a dimension predicted conceptual dissimilarity using four statistical approaches.

**Dimension Testing.** In the first statistical approach, a linear mixed effects model was fit to the similarity ratings. In this model, individual participants’ similarity judgements were regressed onto the proximities between actions on each action dimension. This model featured eight fixed effects—one for each candidate dimension—and random effects for each participant and each action. Random slopes were also called for in the initial design, but these were dropped due to convergence failure. Individual fixed effects coefficients within the model were tested using Wald tests with Satterthwaite approximation for degrees of freedom.

**Model Comparison.** We fit a second mixed effects model to test the three-dimensional model of person perception. This model had the same structure as the first but featured only three fixed effects—one for each trait dimension. We compared the fits of the two mixed effects models—based on the action dimensions and the person perception dimensions, respectively—using AIC and BIC.

**Model Selection.** We applied cross-validated model selection to identify the best combination of candidate dimensions from among the four indistinguishably high-performing combinations identified in Study 5. For each model, we fit non-negative least-squares regressions to predict similarity ratings based on the dimensional distances. Cross-validation was performed by fitting these regressions to the data of all but one participant in the similarity rating sample, and then testing performance on that left out participant, leaving each participant out in turn. Performance was measured using root mean square error (RMSE) of prediction and the Pearson correlation between predicted and actual similarity ratings.

**Noise Ceiling Analysis.** Finally, we performed an exploratory noise ceiling analysis to assess the performance of the action dimensions. The purpose of this analysis was to estimate how well a perfect model could have done in predicting participants’ responses, given the reliability of the data. The result can be compared with the performance of the actual models to determine how much room remains for improvement. To achieve this, we first conducted a multidimensional scaling (MDS) on the average action similarity ratings to extract a set of dimensions that could best explain these similarities. The dimensionality of the scaling was set to six to match the complexity of the best performing model (described below). The dimensions from the scaling solution were then used to predict individual ratings in a cross-validated regression procedure identical to the one applied to the substantive models. The performance of the MDS model would provide an indication of the noise ceiling of the data: how well any model could hope to perform. In this case, the estimate will be the very upper bound of how much variance can actually be captured because the MDS was based on the same data against which it is ultimately tested.

**Results**

**Dimension Testing**

In this study, we examined whether the candidate action dimensions captured people’s action concepts. Specifically, we tested whether proximity between actions on candidate action dimensions
predicted rated similarity. Results of mixed effects modeling indicated that five of the candidate action dimensions significantly predicted pairwise similarity judgements (see Table 5). The marginal $R^2$ across all eight action dimensions was .082.

**Model Comparison**

A similar mixed effects model was fit to the three-dimensional model person perception. This model indicates that all three of these dimensions significantly predicted the similarity between actions (see Table 5). This suggests that knowledge about personality, or the identity of an actor, offers significant insight into the structure of action knowledge. However, the total marginal $R^2$ of the person perception model was .054, indicating that the action dimensions accounted for approximately 1.5-times more variance. Formal model comparison also favored the action dimensions over the person perception model. The former’s AIC was 85,700.4 and its BIC was 85,783.3 versus the person perception model’s AIC of 85,884.1 and BIC of 85,941.3. These differences of 183.7 and 148.0, respectively, for AIC and BIC suggest that the action dimensions should be very strongly preferred to the person perception model. Together, these outcomes favor the action dimensions developed in the present investigation over the three-dimensional model of person perception.

**Model Selection**

A cross-validated model selection procedure was applied to select the best performing combination of action dimensions, among the four of the best performing combinations from Study 3 (Figure 4, inset). All models continued to perform well: raw RMSEs ranged from 26.482 to 26.454; normalized RMSEs ranged from .3229 to .3237. Out-of-sample correlations ranged from $r = .2806$ to $r = .2860$. All four combinations of action dimensions outperformed the three-dimensional person perception model when it was entered into a similar cross-validated regression (out-of-sample correlation $r = .2545$), providing further evidence that the current action taxonomy outperforms existing alternatives (Figure 4, inset). The best performing variant consisted of six of the eight candidate dimensions: abstraction, animacy, spiritualism, food, creation, and tradition.

**Noise Ceiling Analysis**

An exploratory noise ceiling analysis was used to estimate how well a perfect model would have performed, given the reliability of the data. We found estimated that this perfect model would achieve an RMSE of 25.5, a normalized RMSE of .3078; and an out-of-sample correlation of $r = .37$ (Figure 4, inset). Comparing these values to those of the optimal combination of action dimensions indicates that the optimal combination explained approximately 60% of the reliable variance in participants’ similarity judgements.

**Discussion**

In this study we identified a six-dimensional model which we will hereafter refer to as the Abstraction, Creation, Tradition, Food, Animacy, and Spiritualism Taxonomy, or ACT-FAST. Together, the ACT-FAST dimensions were able to explain well over half of the reliable variance in conceptual judgements of actions, indicating that this taxonomy is not merely a statistically significant, but a compelling description of this domain.

Not all of the ACT-FAST dimensions were statistically significant in a mixed effects model containing all eight of the candidate action dimensions derived and named in Studies 1 and 2. However, these dimensions are not orthogonal, particularly in the sample of highly frequent verbs using in this study. Thus, the significance, or lack thereof, of particular dimensions should not be taken as a valid indicator of the best set of action dimensions. The ACT-FAST dimensions further demonstrated their explanatory power by outperforming an alternative theory. This outcome indicates that our tailored model of action representation outperforms a model which describes actions in terms of the traits of the people who tend to perform them. More broadly, it provides further evidence that ACT-FAST represents the state-of-the art in describing action understanding.

**Study 5**

In Studies 1–4, we derived a six-dimensional action taxonomy using a combination of text analysis and behavioral data. In these

---

**Table 5**

*Psychological Dimensions Predict the Conceptual Similarity Between Actions*

<table>
<thead>
<tr>
<th>Model/Dimension</th>
<th>$b$</th>
<th>$df$</th>
<th>$t$</th>
<th>$\beta$</th>
<th>$p_{corrected}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstraction</td>
<td>5.30</td>
<td>8,153.94</td>
<td>9.13</td>
<td>0.097</td>
<td>$&lt;2.2 \times 10^{-16}$*</td>
</tr>
<tr>
<td>Animacy</td>
<td>0.96</td>
<td>7,719.46</td>
<td>1.50</td>
<td>0.016</td>
<td>0.27</td>
</tr>
<tr>
<td>Spiritualism</td>
<td>0.39</td>
<td>2,334.77</td>
<td>0.45</td>
<td>0.007</td>
<td>$1.9 \times 10^{-11}$*</td>
</tr>
<tr>
<td>Food</td>
<td>5.18</td>
<td>1,706.91</td>
<td>7.00</td>
<td>0.100</td>
<td>0.65</td>
</tr>
<tr>
<td>Technicallity</td>
<td>3.09</td>
<td>2,784.83</td>
<td>3.11</td>
<td>0.041</td>
<td>$0.0074^{*}$</td>
</tr>
<tr>
<td>Creation</td>
<td>9.46</td>
<td>2,310.57</td>
<td>14.05</td>
<td>0.188</td>
<td>$&lt;2.2 \times 10^{-16}$*</td>
</tr>
<tr>
<td>Tradition</td>
<td>2.23</td>
<td>5,239.27</td>
<td>1.69</td>
<td>0.020</td>
<td>0.27</td>
</tr>
<tr>
<td>Wealth</td>
<td>5.74</td>
<td>3,228.29</td>
<td>8.07</td>
<td>0.101</td>
<td>$5.3 \times 10^{-15}$*</td>
</tr>
<tr>
<td>Person perception</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>3.53</td>
<td>4,433.22</td>
<td>4.52</td>
<td>0.051</td>
<td>$6.3 \times 10^{-5}$*</td>
</tr>
<tr>
<td>Valence</td>
<td>3.78</td>
<td>2,513.76</td>
<td>5.43</td>
<td>0.066</td>
<td>$1.2 \times 10^{-7}$*</td>
</tr>
<tr>
<td>Sociality</td>
<td>9.81</td>
<td>2,417.96</td>
<td>14.93</td>
<td>0.201</td>
<td>$&lt;2.2 \times 10^{-16}$*</td>
</tr>
</tbody>
</table>

*Note.* Statistical significance of coefficients was computed using the Satterthwaite approximation for degrees of freedom. Results were corrected for multiple comparisons within model via the Holm-Bonferroni procedure. Asterisks indicated statistically significant dimensions.
studies, we used individual verbs as our primary proxy for different actions. In Study 5, we aimed to test the ability of ACT-FAST to generalize beyond many verbs. To do so, we capitalized on the polysemous nature of many verbs. For instance, the verb “run” means fundamentally different actions in the phrase “to run around the block” and “to run an experiment.” In this study, we constructed multiple action phrases around a set of polysemous verbs. As in Study 4, we tested whether ratings of these action phrases on the ACT-FAST dimensions can explain the human judgments about the actions. If so, this would provide strong evidence that ACT-FAST can serve as a meaningful taxonomy of actions, and not just a taxonomy of verbs.

Method

The experimental procedure and the analysis of these data were preregistered on the Open Science Framework (https://osf.io/93np7/).

Participants

We conducted a power analyses based on the effect size from Study 4 (d = 1.47). This analysis suggested that a sample size of 132 participants would be sufficient to achieve family-wise α = β = .05, divided evenly across the eight verbs. We targeted a sample of 150 to allow for possible exclusions.

A total of 150 participants were recruited to rate the similarity between action phrases, and the positions of those phrases on the ACT-FAST dimensions. Participants were excluded based on language (N = 0) and insufficient unique responses (N = 7). This left a final sample size of 143 (64 female, 79 male; mean age = 36.8, range = 21–70).

Experimental Procedure

In this study, we aimed to assess people’s conceptual understanding of actions, using a similar procedure as in Study 4. In this study, participants made judgements about the similarity between different actions. However, unlike in Study 4, participants in Study 5 rated action phrases, each of which was built around the same polysemous verb. Each participant was randomly assigned to rate phrases based on one of eight verbs: break, build, drive, fight, play, run, see, talk. For each verb, we generated 10 phrases describing different actions based on the same verb (e.g., “to build a fire” vs. “to build up suspense” or “to play chess” vs. “to play a role in a movie”). Each participant rated the pairwise similarity between all pairs of actions within a verb condition (45 comparisons) in random order using a 100-point continuous line scale anchored on “very different” and “very similar.” Next, participants made the same set of similarity judgements a second time, with the pairs presented in the same random order, but with the action phrases reversed (i.e., the first time they would be asked how similar “x” is to “y” and the second time they would be asked how similar “y” is to “x”). After making the similarity ratings, participants rated each of the 10 action phrases on the six ACT-FAST dimensions. The dimensions were presented in random order, and the order of phrases was independently randomized within each dimension. Participants provided their demographic information at the end of the experiment.

Statistical Analysis

The primary statistical analysis aimed to test whether the distances between action phrases on the action dimensions predict participants’ ratings of the similarities between those action phrases. To do this, we conducted a cross-validated non-negative least-squares regression nested within an intercept-only linear mixed effects model. Within each participant, we fit a non-negative least-squares regression in which we regressed similarity ratings (reverse-coded as non-negative dissimilarities) upon absolute differences (i.e., distances) between ratings of the action phrases on the ACT-FAST dimensions. We predicted that actions that were further apart on these dimensions would be rated as less similar to each other. Similarity ratings were averaged across both presentations of each pair of phrases to increase reliability. The predictive performance of this regression was assessed using a leave-one-phrase-out cross-validation scheme. The regression was first fit to the 36 pairs of action phrases that did not include a given phrase and then tested on the nine pairs of phrases that included the left-out phrase. The performance of the model was estimated as correlation between the actual pairwise similarity ratings of the left-out phrase, and the predictions based on the regression fit to the other pairs. The RMSE between these predicted and actual values was also calculated. Performance metrics were averaged across folds of the cross-validation to produce a single correlation and RMSE value for each participant.

The correlations were Fisher transformed and then entered into a second-level mixed effects analysis. This model consisted of a single random intercept—for the verb condition assigned to each participant—and no other fixed or random terms. The crucial test was whether the fixed intercept of this model (i.e., the grand mean of cross-validated accuracy) was greater than zero. The mixed effects model thus essentially amounted to a one-sample t test across participants but with the random intercept for verb included to allow for generalizable inference to other verbs. The statistical significance of this term was computed using the Satterthwaite approximation for degrees of freedom. We also analyzed the results from each verb condition separately. This was achieved via eight one-sample t tests applied to the Fisher transformed correlations from each participant. Results were corrected for multiple comparisons using the Holm-Bonferroni procedure.

Results

The primary analyses tested whether the ACT-FAST dimensions reflect people’s action concepts. Specifically, we evaluated whether the proximity of action phrases on the ACT-FAST dimensions could predict judgements of the similarity between those actions. In cross-validated regressions, the average correlation between proximity and similarity was r = .30 (SD = .22, 95% bootstrap CI [.26, .33]) with an RMSE of 23.47 on a 100-point scale. An exploratory noise ceiling analysis indicated that the ACT-FAST explained 41% of the reliable variance in action phrase judgements (see the online supplemental materials). We tested the statistical significance of this performance using two different approaches. First, mixed effects modeling indicated that the average correlation was significantly greater than zero across all verbs, t(8.53) = 11.48, p = 1.8 × 10^−6. Second, we tested this effect separately within each of the eight verb conditions using one-sample
t-tests (Figure 5; Table 6). The accuracy of the cross-validated regression was significantly greater than chance for all eight of verb conditions, controlling for multiple comparisons.

The ACT-FAST's ability to predict action similarity had a very large effect across participants: $d = 1.50$. The effect was large, by conventional standards, for all verbs, ranging from $d = .8$ for Break to $d = 1.9$ for Play. The average effect size was similar to the $d = 1.47$ previously observed in Study 4, when actions were operationalized as verbs. This result thus suggests that ACT-FAST applies equally well to actions, whether they are operationalized as individual verbs or as extended phrases predicated on the same verb.

Discussion

In Study 5, we tested whether ACT-FAST could capture how people conceptualize actions, even when all of the action phrases relied on the same verb. We found that the proximity on the action dimensions was robustly correlated with judgments of the similarity between action phrases. This result thus provides a strong indication that ACT-FAST captures how people conceptualize actions in general, and not just verbs. The present results also provide further evidence for the robustness and predictive power of this taxonomy by showing that it can explain much of the perceived similarities and differences between a set of 80 action phrases.

The repeated measurement of action similarities in Study 5 also allowed us to carry out several informative individual difference analyses (see the online supplemental materials). We found that 18% of the total variance in these judgements was idiosyncratic: reliable within, but not between, participants. This compared with 30% of the variance that was shared between participants. These results indicate that there are substantial idiosyncrasies in the way people think about actions, although these idiosyncrasies are still outweighed by shared understanding. Both derivation of the ACT-FAST, and the results of the individual difference analyses in Study 3 (see the online supplemental materials) suggest that ACT-FAST is best thought of as a description of the common

Table 6

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$n$</th>
<th>$r$</th>
<th>RMSE</th>
<th>$t$</th>
<th>$p_{corrected}$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Break</td>
<td>21</td>
<td>.22</td>
<td>20.18</td>
<td>3.67</td>
<td>.015*</td>
<td>.80</td>
</tr>
<tr>
<td>Build</td>
<td>21</td>
<td>.39</td>
<td>20.48</td>
<td>6.12</td>
<td>$3.4 \times 10^{-5}$*</td>
<td>1.34</td>
</tr>
<tr>
<td>Drive</td>
<td>18</td>
<td>.38</td>
<td>22.03</td>
<td>4.99</td>
<td>.0034*</td>
<td>1.18</td>
</tr>
<tr>
<td>Fight</td>
<td>16</td>
<td>.23</td>
<td>23.55</td>
<td>5.65</td>
<td>.0023*</td>
<td>1.41</td>
</tr>
<tr>
<td>Play</td>
<td>16</td>
<td>.32</td>
<td>31.78</td>
<td>7.63</td>
<td>$1.2 \times 10^{-5}$*</td>
<td>1.91</td>
</tr>
<tr>
<td>Run</td>
<td>19</td>
<td>.28</td>
<td>22.50</td>
<td>6.55</td>
<td>$2.6 \times 10^{-5}$*</td>
<td>1.50</td>
</tr>
<tr>
<td>See</td>
<td>16</td>
<td>.26</td>
<td>23.29</td>
<td>5.40</td>
<td>.0030*</td>
<td>1.35</td>
</tr>
<tr>
<td>Talk</td>
<td>16</td>
<td>.27</td>
<td>26.27</td>
<td>4.80</td>
<td>.0047*</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Note. Correlations coefficients reflect cross-validated results. $p$ values were corrected for multiple comparisons via the Holm-Bonferroni procedure. Asterisks indicate statistically significance results. RMSE = root mean square error.
understanding of actions shared across people, rather than a model of these individual differences.

**Study 6**

Studies 1–5 showed that a set of six dimensions, ACT-FAST, capture how people think about actions. Together these studies derived, named, and provided basic validation for this taxonomy of action understanding. We next aimed to test whether these dimensions are useful for predicting socially relevant features of actions. A socially useful taxonomy of action understanding must predict features such as who tends to perform an action, when and where they tend to do it, why they are motivated to do so, and how they perform the action. The answers to these questions inform how people think about actions in everyday life, and solve problems such as action prediction, trait attribution, and planning. In Study 6, we test whether ACT-FAST can accurately predict such features. To this end, we measure 16 socially relevant features of actions, based around the who, when, where, why, and how questions posited above. We then evaluate the ability of ACT-FAST to predict these features.

**Method**

The experimental procedure and analysis of these data were preregistered on the Open Science Framework (https://osf.io/7wem/).

**Participants**

We conducted a power analyses based on the smallest effect size from Study 5 (d = .8). This analysis suggested that a sample size of 329 participants would be sufficient to detect such an effect in a one-sample t test with 95% power and α = .05, controlling for family-wise type I and II error rates. We targeted a sample size of 350 to allow for exclusions.

We recruited 350 participants from Amazon Mechanical Turk using TurkPrime (Litman et al., 2017). Fifteen of those participants did not provide ratings on all trials and were thus excluded because of our preregistered plan to eliminate data listwise. Participants were excluded because of language (N = 4) and insufficient unique responses (N = 2). The final sample size was thus 329 (169 female, 158 male, 2 other; mean age = 35.9, range = 19–71).

**Experimental Procedure**

In Study 6, participants rated the 80 action phrases from Study 5 on a set of socially relevant features: the who, why, when, where, and how of these actions. We aimed to test whether one could infer these practical qualities based on an action’s coordinates on the ACT-FAST dimensions (based on ratings from Study 5). Participants were randomly assigned to answer one of six different surveys: who performs each action, why people perform the actions, when people perform the actions, where people perform the actions, and how much effort they require. Each survey contained multiple feature ratings. In the “who” survey, participants rated the personality of the people who tend to engage in each action on three dimensions: the extent to which people who tend to engage in each action are powerful versus powerless, good versus bad, and social versus asocial people tend to engage in each action (Thornton & Mitchell, 2018). In the “why” survey, participants rated the extent to which a person engages in each action because of approach motivations and avoidance motivations (Carver & White, 1994). In the “when” survey, participants rated how likely people would be to engage in each action at “Morning/midday” “Afternoon/evening” and/or “Night/early morning” (a coarse coverage of all possible times). In the “where” survey, participants rated whether people tend to engage in the action in private inside, in public inside, or in outside locations (a coarse coverage of possible locations). There were two “how” surveys. In the first “how” survey, participants rated the extent to which people use their “head, face and eyes,” “arms and hands” and “legs and feet” to engage in each action. The body parts were clustered into these three groups based on physical proximity. In the second “how” survey, participants rated how much mental effort and how much physical effort each action required. These “how” questions were inspired by other recent research on action understanding (Tarhan & Konkle, 2020).

In each version of the survey, participants rated all 80 action phrases from Study 5. For a given action phrase, continuous line scales were displayed simultaneously for ratings. The order of action phrases was independently randomized for each participant. Across all six survey versions, we obtained ratings of the placement of all 80 states on all 16 features. All participants provided their demographic information at the end of the study.

We used the ratings that participants in Study 5 had provided for the same set of actions to determine the positions of each action on the action dimensions. Averaging these ratings across participants allowed us to estimate where each of the 80 action phrases fell on the six ACT-FAST dimensions.

**Statistical Analysis**

Statistical analyses examined whether ratings of action features could be predicted from the positions of those actions on the ACT-FAST dimensions. The primary statistical analysis consisted of a cross-validated regression, in which we predicted feature ratings based on ACT-FAST ratings. Each of 16 feature dimensions was regressed onto the six ACT-FAST dimensions. These regressions were trained and tested via a double-split-half cross-validation scheme, which split the data by both participant and action. By comparing performance of the regression across these two “boundaries,” we could learn how well the ACT-FAST predictions could be expected to generalize to either new people, new actions, or both.

First, for each feature dimension, participants were randomly divided into two halves. Second, each half of participants was then split in half by actions. That is, within each half of participants we divided the data into two halves consisting of 40 action phrases each. The two orthogonal split-halves—participants and action phrases—divided the data into four total subsets. We trained regressions on each of these four segments separately. We then used each of these regressions to predict feature ratings in the other three subsets of data. We measured the performance of these predictions in terms of the correlation between predicted and actual ratings, as well as raw and normalized RMSE.

In addition to this primary analysis, we also performed a cross-validated regression within each participant. This allowed us to combine data across independent participants for the purposes of null hypothesis significance testing. In this analysis, each
participants’ ratings of a given feature dimension were regressed onto the average ratings of the ACT-FAST dimensions from Study 5. This regression was trained and tested using a leave-one-verb-out cross-validation procedure. Because the 80 action phrases were constructed around eight verbs, this amounted to an eightfold cross-validation with 10 phrases held out on each iteration. Performance was measured using the correlation between predicted and actual ratings in the test set. These correlations were Fisher-transformed and averaged across folds to produce a single value for each participant. These means were then entered into one-sample t tests across participants. The results were corrected for multiple comparisons across the 16 feature dimensions via the Holm-Bonferroni correction.

We also repeated this analysis at the item level (i.e., with the feature ratings averaged across participants) for the purposes of visualization only. Specifically, this item analysis allowed us to produce scatter plots illustrating the relationship between ACT-FAST predictions and feature dimension ratings, with each point representing an action. This model used the same leave-one-verb-out cross-validation scheme but was conducted only once per feature—on the average ratings—rather than once for each participant.

Results

In Study 6, we examined whether the ACT-FAST dimensions capture practical, socially-relevant features of an action—who, why, when, where, and how it is performed. ACT-FAST proved capable of predicting all 16 of the features we examined (Figure 6 and Table 7, and Table S1 in the online supplemental materials). When cross-validated, the taxonomy successfully generalized across participants, (mean r = .37; Table 7 and Table S1 in the online supplemental materials S1, participant column), across actions (mean r = .24; Table 7 and Table S1 in the online supplemental materials, Action column), or across both simultaneously (mean r = .24; Table 7 and Table S1 in the online supplemental materials, Both column).

Similar results were obtained when features were regressed on ACT-FAST dimensions within each participant, in a leave-one-verb-out cross-validation scheme. Null hypothesis significance testing indicated that the ACT-FAST dimensions were able to significantly predict all 16 features out-of-sample, even after adjusting for multiple comparisons (all corrected ps < 1.4 × 10^-6). Effect sizes ranged across features from d = .72 to d = 3.01, with an average d = 1.50. These results further indicate that the ACT-FAST dimensions could inform accurately inform judgements about practically important features of actions.

Discussion

Study 6 reveals that the ACT-FAST dimensions reflect meaningful information about how people understand actions. Specifically, these dimensions predict properties of actions that people need to know to navigate their daily social lives. People rely on knowledge about when or where an action typically takes place to decide if an action was appropriate in a particular context; people rely on knowledge about how an action is performed to perform it oneself or to prevent someone else from performing it; people must understand who performs actions and why they are performed to make predictions about others’ likely future actions. The ACT-FAST dimensions can inform all of these crucial real-world inferences. These dimensions thus offer a practical and beneficial scheme for organizing action concepts, in addition to an efficient one. The present results also provide further evidence that ACT-FAST can generalize well to new people and to new actions. However, it should be noted that these results are descriptive: they do not provide evidence that drawing on ACT-FAST dimensions—explicitly or implicitly—is necessary to answer real world questions about actions.

Study 7

In Studies 1–6, we derived and validated a six-dimensional taxonomy of action concepts. In Study 7 we examined whether ACT-FAST is a biologically plausible model for representing actions. Specifically, we tested whether the brain encodes actions in a way that can be predicted by ACT-FAST. To test this hypothesis, we analyzed the neural responses of participants watching a naturalistic stimulus: the first half of the first episode of the popular TV show, Sherlock (Chen et al., 2017).

Study 7 offers a rich assay of how actions are represented. First, we use neural activity as an implicit measure, allowing us to test how people spontaneously perceive actions with minimal response bias. Second, participants in this study perceive actions in a video stimulus, allowing us to test whether ACT-FAST generalizes from text to visual stimuli. Third, Sherlock is a relatively realistic stimulus, emulating many aspects of natural experience. This stands in contrast to the decidedly less naturalistic paradigms employed in the previous behavioral studies. As a result, if ACT-FAST predicts neural activity associated with watching actions in the movie, it would provide a strong indication that they apply in a natural context, as people perceive actions in real-time. We also replicated our findings with Sherlock in a separate fMRI dataset in which participants watched Forrest Gump (see the online supplemental materials). In addition to providing further evidence to support our conclusion, this replication demonstrates generalization to a different stimulus, MRI scanner, and culture (German).

Method

FMRI Data

We drew upon open data from a previous fMRI investigation (Chen et al., 2017). These data are publicly available on the Princeton DataSpace (http://arks.princeton.edu/ark:/88435/dsp01nz8062179). In the original study, a sample of 22 participants (12 male, 10 female, ages 18–26, mean age = 20.8) who had previously not viewed the TV show Sherlock were recruited to watch half of the first episode of the series while undergoing fMRI scanning. Five participants were ultimately excluded—two because of excessive head motion, two because of poor recall, and one for falling asleep—leaving a final sample of N = 17. Participants viewed the audio-video stimulus in two segments of 23 and 25 minutes, respectively. Each segment was preceded by a 30-s cartoon unrelated to the movie (“Let’s All Go to the Lobby”). Participants were simply asked to attend to the movie, with no behavioral responses required. After the movie, participants engaged in a verbal recall procedure. However, we do not draw upon
the recall data in the present investigation, so we will not discuss it further.

Imaging data were acquired using a 3T Siemens Skyra scanner with 20-channel head coil. Functional images were acquired from the whole brain (TR = 1.5 s, TE = 28 ms, flip angle = 64°, 27 ascending interleaved slices of 4mm thickness, 3 mm² in-plane resolution, 192 × 192 mm FOV). Anatomical images were obtained using a T1 MPRAGE protocol with .89 mm³ voxels. In the present investigation, we used already-preprocessed data from the original study. This preprocessing was carried out using
accounting for hemodynamic lag. 3 TRs to align functional activity with the movie time stamps, along the time-course, and 6 mm FWHM Gaussian smoothing z resampling to 3 mm isometric voxels, fii Statistical Analysis FSL (http://fsl.fmrib.ox.ac.uk/fsl) and included slice-time and troling for multiple testing. The HHow (effort) Physical effort .44 .28 .28 2.30 5.7 HHow (body) Leg/Foot .47 .21 .21 1.80 2.5 HHow (body) Arm/Hand/Finger .47 .19 .19 0.89 1.4 HWhere Inside (private) .39 .32 .33 2.36 1.5 HWhere Inside (public) .33 .26 .26 0.92 1.2 HWhere Outside .36 .20 .20 1.41 2.7 HHow (body) Head/Face/Eyes/Mouth .32 .14 .13 0.72 1.4 HHow (body) Arm/Hand/Finger .50 .16 .16 1.44 2.0 HHow (body) Leg/Foot .47 .21 .21 1.80 2.5 HHow (effort) Physical effort .44 .28 .28 2.30 5.7 HHow (effort) Mental effort .44 .33 .33 3.01 6.0 Note. Feature dimensions (left) were regressed onto the six dimensions of ACT-FAST. The taxonomy significantly predicted all feature dimensions, controlling for multiple testing. The r values in the table represent the correlation between the ACT-FAST predictions and human ratings, assessed in three different cross-validated models: (a) Participants: regressions were trained and tested on different participants but the same actions; (b) Actions: regressions were trained and tested on different actions but the same participants; (c) Both: regressions were trained and tested on different sets of both participants and actions. The (Cohen’s) d and p values were computed from a separate analysis in which the predictive model was trained and tested separately within each participant. Table S2 in the online supplemental materials provides root mean square error (RMSE) and normalized RMSE values.

FSL (http://fsl.fmrib.ox.ac.uk/fsl) and included slice-time and motion corrections, linear detrending, high-pass (140 s cutoff) filtering, alignment to a common template brain (MNI), spatial resampling to 3 mm isometric voxels, z-scoring of each voxel along the time-course, and 6 mm FWHM Gaussian smoothing Jenkinson et al., 2012. The fMRI time-course was also shifted by 3 TRs to align functional activity with the movie time stamps, accounting for hemodynamic lag.

Statistical Analysis

Automated Action Annotation. To investigate the representation of actions during movie-viewing, we first needed to know which actions were occurring at each timepoint in the video. To do so, we automatically annotated all actions present in the video using a temporal relation network (Zhou et al., 2017) pretrained on the Moments in time dataset (Monfort et al., 2018). The training dataset consists of 1 million videos, each 3 s in length, which represent 332 different actions. We applied this pretrained neural network to annotate these actions within the Sherlock video stimulus. We split the video into 3-s increments to match the length of actions in the training set and used the network to estimate the probability of each action. We subsequently averaged several actions that differed only in the agent performing them (e.g., male vs. female singing). This process resulted in estimates for the likelihood of the presence of 332 actions across the entire video stimulus, in 3-s (2 TR) increments.

Voxel Selection. We focused our analyses on brain regions which represent actions. To identify the voxels sensitive to action perception we performed a voxel-selection analysis using action regressors based directly on the 332 annotated actions. To produce the action regressors, we convolved the annotated probabilities for each action at each point in the video with a canonical hemodynamic response function (HRF) at 1-s resolution. We then resampled the predicted time-course to TR (1.5 s) resolution and shifted it back in time 4.5 s to align it to the fMRI data (which had been similarly shifted in preprocessing). This process produced a set of 332 regressors encoding the probability with which each action was present in each TR across the scan.

We correlated each participant’s whole brain activity across the time-course with each of these 332 action regressors. This produced 332 maps—one for each action—representing the extent to which that voxel responded to the presence or absence of an action. Across participants, we computed the reliability of these responses within each voxel. To assess reliability, we extracted a vector with the correlation values representing the strength of all 332 action-specific responses within a given voxel. These vectors reflect how each voxel activates to different degrees across all of the actions. We correlated these vectors with each other across participants within each voxel. This allowed us to compute a Cronbach’s alpha estimating the reliability of action-related activity for each voxel in the brain (Figure S3A in the online supplemental materials).

Most voxels expressed some degree of reliability, but the voxels sorted into a clearly bimodal distribution (Figure S3B in the online supplemental materials). We divided voxels into two classes via a Gaussian mixture model and retained only those voxels with reliability above the class boundary (Figure S3C in the online supplemental materials). These voxels activated to different degrees for different actions, with a high level of reliability across participants.

To ensure that the voxel selection procedure did not bias subsequent analyses (e.g., via circularity), we conducted this procedure in a cross-validated fashion. Specifically, we divided viewed portions of Sherlock into five approximately equal contiguous parts, based on the official scene boundaries in the episode. We then performed the feature selection five times, leaving out each of the five

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>Participants</th>
<th>Actions</th>
<th>Both</th>
<th>d</th>
<th>Pcorrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>Power</td>
<td>.28</td>
<td>.21</td>
<td>.21</td>
<td>1.64</td>
<td>3.3 × 10⁻¹⁸</td>
</tr>
<tr>
<td>Who</td>
<td>Valence</td>
<td>.44</td>
<td>.36</td>
<td>.36</td>
<td>1.94</td>
<td>1.9 × 10⁻²¹</td>
</tr>
<tr>
<td>Who</td>
<td>Sociality</td>
<td>.38</td>
<td>.22</td>
<td>.22</td>
<td>1.35</td>
<td>8.9 × 10⁻¹⁵</td>
</tr>
<tr>
<td>Why</td>
<td>Approach</td>
<td>.36</td>
<td>.21</td>
<td>.21</td>
<td>1.16</td>
<td>5.1 × 10⁻⁹</td>
</tr>
<tr>
<td>Why</td>
<td>Avoidance</td>
<td>.38</td>
<td>.28</td>
<td>.28</td>
<td>1.25</td>
<td>7.7 × 10⁻¹⁰</td>
</tr>
<tr>
<td>When</td>
<td>Morning/Midday</td>
<td>.33</td>
<td>.30</td>
<td>.30</td>
<td>1.49</td>
<td>1.9 × 10⁻⁸</td>
</tr>
<tr>
<td>When</td>
<td>Afternoon/Evening</td>
<td>.18</td>
<td>.14</td>
<td>.13</td>
<td>0.80</td>
<td>8.7 × 10⁻⁸</td>
</tr>
<tr>
<td>When</td>
<td>Night/Early Morning</td>
<td>.28</td>
<td>.19</td>
<td>.19</td>
<td>0.89</td>
<td>5.6 × 10⁻⁹</td>
</tr>
<tr>
<td>Where</td>
<td>Inside (private)</td>
<td>.39</td>
<td>.32</td>
<td>.33</td>
<td>2.36</td>
<td>1.5 × 10⁻²³</td>
</tr>
<tr>
<td>Where</td>
<td>Inside (public)</td>
<td>.33</td>
<td>.26</td>
<td>.26</td>
<td>0.92</td>
<td>1.2 × 10⁻⁸</td>
</tr>
<tr>
<td>Where</td>
<td>Outside</td>
<td>.36</td>
<td>.20</td>
<td>.20</td>
<td>1.41</td>
<td>2.7 × 10⁻¹⁴</td>
</tr>
<tr>
<td>How (body)</td>
<td>Head/Face/Eyes/Mouth</td>
<td>.32</td>
<td>.14</td>
<td>.13</td>
<td>0.72</td>
<td>1.4 × 10⁻⁶</td>
</tr>
<tr>
<td>How (body)</td>
<td>Arm/Hand/Finger</td>
<td>.50</td>
<td>.16</td>
<td>.16</td>
<td>1.44</td>
<td>2.0 × 10⁻¹⁵</td>
</tr>
<tr>
<td>How (body)</td>
<td>Leg/Foot</td>
<td>.47</td>
<td>.21</td>
<td>.21</td>
<td>1.80</td>
<td>2.5 × 10⁻¹⁹</td>
</tr>
<tr>
<td>How (effort)</td>
<td>Physical effort</td>
<td>.44</td>
<td>.28</td>
<td>.28</td>
<td>2.30</td>
<td>5.7 × 10⁻²⁵</td>
</tr>
<tr>
<td>How (effort)</td>
<td>Mental effort</td>
<td>.44</td>
<td>.33</td>
<td>.33</td>
<td>3.01</td>
<td>6.0 × 10⁻³¹</td>
</tr>
</tbody>
</table>
parts in turn. This same cross-validation procedure was applied during the encoding model fitting and evaluation (below). Thus, the voxel selection did not allow the encoding model an inadvertent peek at the held-out test data. Figure S3 in the online supplemental materials illustrates the voxel selection procedure using all of the data simultaneously.

**Action Ratings.** To determine whether ACT-FAST could predict neural activity, we needed to place each of the 332 automatically annotated actions on the taxonomic dimensions. Of the 332 annotated actions, 52 had been rated on the ACT-FAST dimensions in previous studies. Using Amazon Mechanical Turk and TurkPrime (Litman et al., 2017), we recruited a sample of 662 participants to rate the remaining 280 additional actions. Each participant in the rating study was randomly assigned to rate 70 actions on one of the six ACT-FAST dimensions. At the end of the survey, participants provided their demographic information. We excluded participants for language (N = 16) and insufficient unique responses (N = 46). This left a final sample of 600 (294 female, 301 male, two other, three preferred not to indicate gender; mean age = 36.2, range = 18–19).

Following data collection, we averaged ratings across participants to create a single set of ratings for each of the 280 actions on each of the ACT-FAST dimensions. We then combined these action ratings with ratings of other actions from previous studies. Averaged ratings were z-scored on each dimension. This process resulted in six coordinates—one for each of the ACT-FAST dimensions—for each of the 332 action the temporal relation network could annotate.

**Neural Encoding Model.** The primary goal of Study 7 was to evaluate whether ACT-FAST characterizes how people spontaneously think about the actions they perceive. If this is the case, then we would expect to see it reflected in the brain activity elicited by watching *Sherlock*. Specifically, if we know “where” an action is—in terms of its six-dimensional coordinates on the ACT-FAST dimensions—then we should be able to accurately predict the pattern of brain activity elicited by perceiving that action. To test this hypothesis, we used a method known as a voxelwise encoding model (Mitchell et al., 2008; Naselaris et al., 2011). This class of model predicts neural activity—typically at the level of individual voxels—based on a set of hypothesized features of the stimuli. In the present case, we sought to predict neural activity as participants perceived actions in *Sherlock*, based on the coordinates of those actions on the ACT-FAST dimensions.

The first step in this process was to produce regressors reflecting the ACT-FAST coordinates of the actions on screen at each time-point, over the course of the video. The ACT-FAST regressors were also based on the annotated actions. We could determine the location of each time point on the ACT-FAST dimensions by averaging across the action ratings, with each action’s weighted by the likelihood that it was on screen during that time point. For example, if several food-related actions were judged to be very likely at a given time point, and nonfood related actions were unlikely, then the ultimate coordinate of that time point would be high on the Food dimension.

The second step was to fit an encoding model within the action-sensitive voxels. This model predicted the activity in each voxel of the brain using the regressors for the six ACT-FAST dimensions. The activity of each voxel across the video was regressed onto coordinates of the actions presented in the video on the ACT-FAST dimensions. To estimate the performance of this model, we employed a bicross-validation scheme. That is, we trained and tested out-of-sample with respect to both the video and the participants. Thus, the voxelwise regressions were fit to data from four-fifths of the movie from 16 of 17 participants and tested on the left-out one-fifth of the movie in the left-out participant (leaving out each segment of the movie and each participant in turn).

Third, we measured the performance of the encoding in terms of how well it could predict patterns of brain activity based on the ACT-FAST coordinates of the corresponding moment in *Sherlock*. We used the position of each TR on the ACT-FAST regressors to predict what pattern of brain activity should have been elicited at that moment in time and then correlated this predicted pattern with the actual pattern that was expressed in each participant’s brain at that moment. These correlations were averaged across all TRs in the left-out portion of the data (and then across folds in the cross-validation) to produce a single estimate of pattern reconstruction accuracy for each participant.

Finally, we computed the statistical significance of the model performance separately for each participant. This allows us to treat each participant as a separate replication of the effect of the encoding model, rather than pooling a single summary statistic across participants. We applied permutation testing to compute statistical significance: on each fold of the cross-validation we phase-randomized the training data prior to fitting the encoding model (Nastase et al., 2019). Phase-randomization destroys the ordering of the data, while preserving the temporal structure (autocorrelations). This allows permutation of time-courses without violating the interchangeability assumption of permutation testing, as would happen if the order of individual time points was randomized. This phase-randomization procedure allowed us to compute an empirical null distribution for the performance of the encoding model with respect to each participant. We could then compare the actual performance of the model to this empirical null distribution to compute *p* values for each participant or average across participants to produce a single *p* value for group.

**Results**

**Voxel Selection**

The voxel selection procedure implicated a wide variety of regions in action representation (Figure 7A), including large portions of visual and auditory cortex—presumably involved in seeing and hearing actions in naturalistic stimuli—as well as portions of premotor cortex, medial prefrontal cortex, and the anterior temporal lobes. The widespread nature of this action-related activity throughout visual cortices is consistent with recent findings that suggest that perceiving actions plays a dominant role in visual representation (Haxby et al., 2020). Consistency in which voxels were selected was high across folds of the cross-validation procedure, particularly in posterior regions. This indicates that actions mapped onto brain activity in similar regions and similar ways across the entirety of the video.

**Encoding Model**

Within the selected regions, we tested whether the brain spontaneously encodes perceived actions using the ACT-FAST dimensions. The results strongly supported this hypothesis: encoding
models significantly predicted patterns of brain activity in each of the 17 participants (Figure 7B). The average correlation between model-predicted patterns of brain activity and actual patterns of brain activity was $r = .040 (p < .001)$. The fact that the encoding model were statistically significant predictors in each participant, even when it was trained and tested across different segments of video and different participants’ brains, provides strong evidence that ACT-FAST is a highly generalizable account of the spontaneous brain activity elicited by perceiving actions.

**Discussion**

The results from Study 7 provide neural evidence that ACT-FAST describes how the brain represents observed actions. As participants watched *Sherlock*, their brain activity could be predicted from the ACT-FAST coordinates of the actions they viewed. ACT-FAST was significantly predictive of brain activity in each and every participant analyzed. The predictive capability of the ACT-FAST model generalized across different portions of the video, and across the brains of different participants. Thus, despite the small sample size available in this open fMRI dataset, these results provide strong evidence of the generalizability of the ACT-FAST model. Together these results indicate that the ACT-FAST dimensions capture how the brain spontaneously encodes perceived actions.

Although the size of the effects observed in Study 7 appear small, this reflects the data analysis context more than the explanatory power of ACT-FAST. The correlation between model-predicted and actual brain activity was performed separately for each individual time point (1.5 s) of fMRI data. FMRI data are typically aggregated over many such TRs prior to pattern analyses to increase reliability, but this option was not available to us due to the naturalistic nature of the stimulus. The current effect size thus appears quite large, considering that studies that average across nearly 2 hr of fMRI data often produce correlations that are only two to five times larger (Tamir et al., 2016).

The regions of the brain predicted by ACT-FAST have previously been implicated in the perception of both biological motion and action representation (Culham & Valyear, 2006; Gottlieb, 2007; Saxe et al., 2004; Thompson et al., 2005). Our findings thus add to this work on action understanding by suggestion that these regions encode actions in ways that can be described by ACT-FAST.

In addition to supporting ACT-FAST with evidence from a new measure (i.e., neural activity), the present results also demonstrate that the explanatory power of ACT-FAST generalizes to more naturalistic stimuli, in which participants view video of real people, rather than words on a screen. This indicates that ACT-FAST can explain the audio-visual perception of real actions, in addition to amodal action concepts. It also provides further evidence that
ACT-FAST may generalize beyond the laboratory to help explaining how people represent actions in the real world.

Additionally, we replicated our findings from the Sherlock dataset in all 15 participants in a separate dataset in which participants watched the movie Forrest Gump (Hanke et al., 2016; see the online supplemental materials). This replication increases the amount of evidence supporting the conclusion that ACT-FAST can predict action-related brain activity. Moreover, the encoding model trained on Sherlock accurately predicted action-related brain activity in Forrest Gump even though the two samples were drawn from different populations (United States and Germany), and the films audio was presented in different languages (English and German). This provides direct evidence of generalization along several important axes, including stimuli, fMRI scanners and acquisition parameters, and language and culture. The language difference also helps to rule out the possibilities that discussion or subvocalization of actions could explain the results.

General Discussion

People must act to interact. This makes the ability to understand actions crucial for navigating everyday social life. People must understand others’ actions so that they can anticipate others’ goals and help or hinder them accordingly (Baker et al., 2009). Despite the importance of the action domain to understanding the social mind, this territory has so far gone largely un mapped. This may be because it presents unique challenges to mental cartographers, with its sheer breadth, rapidly evolving landscape, and hierarchies of abstraction. Here we have overcome these challenges, using a combination of linguistic, behavioral, and neuroimaging data to map human action understanding. The result is the Abstraction, Creation, Tradition, Food, Animacy, and Spiritualism Taxonomy, or ACT-FAST. Figure 8 illustrates the locations of a variety of actions on each of the ACT-FAST dimensions.

Goals Achieved by ACT-FAST

This taxonomy of action understanding achieves several important goals. First, ACT-FAST allows us to predict important real-world features of actions, such as how, when, where, why, and by whom actions are performed (Study 6). These are the key questions with which we must all grapple on a daily basis when thinking about others’ behavior. The answers to these questions could help us engage in essential activities such as understanding others’ motivations, following social norms, planning our days, or learning new skills. Such inferences could be useful for familiar actions: reducing an action to its coordinates on ACT-FAST would provide a way to make a complex stimulus like an action more tractable. However, such inferences may prove particularly useful when it comes to generalization. If one encounters a novel action, how can one know when, where, or why to perform it? If the mind could localize the action to a particular region of ACT-FAST space—for instance, by assessing how food-related it is—then it could use this location to make inferences about important real-world features of the action.

Second, ACT-FAST captures action concepts. In Studies 3–5, participants judged the similarity between actions. In all three studies, we found that the proximity between actions on the ACT-FAST dimensions robustly predicted judgements of their similarity. These findings applied to a wide sample of actions (Study 3), to the most common actions (Study 4), and to extended action phrases (Study 5). The success of ACT-FAST in predicting conceptual judgements across this diverse set of tasks indicates that it can serve as a reliable basis for describing how people think about actions.

Third, ACT-FAST is parsimonious: The taxonomy describes much of the incredibly rich domain of actions with just six dimensions. This simplicity makes the taxonomy accessible to a wide audience and useful for researchers seeking to gain deeper insight into how people think about actions. Although higher dimensional models may provide more explanatory power for any given data set, our results suggest that this may come at the cost of generalizability, as well as simplicity. Specifically, results from Study 1 suggest that focusing on a relatively small set of highly predictive dimensions may generalize well from one text corpus to another. Such generalization is typically elusive in text analyses, because highly complex models overfit to the idiosyncrasies of particular data sets. Moreover, although machine-learning algorithms might be able to make use of high-dimensional data, it can be very difficult for human researchers to interpret or practitioners to apply. Parsimony is therefore desirable on the basis of practicality as well as generalizability.

Validity and Generalizability of ACT-FAST

The studies presented in this investigation provide strong convergent evidence for the validity of ACT-FAST across a wide variety of methods, measures, stimulus modalities. We observe that the ACT-FAST dimensions explain the co-occurrences of verbs and nouns in natural language use and the linguistic differences involved in discussing different activities (Study 1), explicit human judgements of actions (Studies 3–6), and action-related neural activity (Study 7). The examination of actions in naturalistic contexts such as text from the web (Study 1) and narrative movies (Study 7) provides particularly strong evidence that our findings will generalize beyond the lab. The extension of Study 7 to predict the neural responses of participants from a different language and culture (Germany vs. United States) provides further evidence for the generalizability of the present findings (see the online supplemental materials). Moreover, except for the explicitly exploratory portion of Study 2, all behavioral studies in this investigation were preregistered in detail. We adhered to these plans with only a few minor exceptions, which we carefully noted and justified. This practice helps to rule out the possibility that flexible data analysis aligned the results of these studies in favor of ACT-FAST.

All of the preregistered studies involved a priori specifications of, and rationales for, the sample sizes we targeted. Studies 3–6 in particular used formal power analyses to justify the number of participants we recruited. These power analyses set uncommonly high standards for power. Specifically, rather than the popular 80% power threshold, we targeted 95% power in these studies. Moreover, in the cases where multiple comparisons were involved, our power simulations took into account the effects of these comparisons on both false positive and false negative rates. That is, we targeted a .05 family-wise type I error rate together with a 95% family-wise power. Setting this high standard likely contributed to
the consistent validation of the ACT-FAST’s predictive capacity across a wide range of outcomes.

This investigation drew upon several different types of samples. The text data from which the ACT-FAST were initially derived were retrieved from public discussions on reddit. The users who wrote this text are anonymous, and so we cannot detail the composition of this sample. However, reddit is an extremely widely used website—at time of writing, the 17th-most-visited website in the world, and the seventh-most-visited in the United States—with more than 300 million registered users. A representative survey of users (Agrawal, 2013) suggests that redditors are relatively male (59%), young (95% between 18 and 49), and educated (46% have a college degree or higher level of education). Thus, it is likely that these data represent a relatively broad—if nonrepresentative—slice of the global population of English-speakers. As for the IMDb scripts, these were created primarily by Hollywood screenwriters. Although we do know the exact demographics of those who produced the scripts in our sample, industry statistics indicate they were likely to be disproportionately White and men (Women and Hollywood, 2019). Studies 2–6 relied on samples of US Americans recruited via Amazon Mechanical Turk. Although potentially more diverse than student samples, Mechanical Turk
samples are not nationally representation. Study 7 relied on a small university sample for the fMRI portion the study. However, the secondary fMRI dataset—although also small—was drawn from a different culture (see the online supplemental materials). Although each of these samples has its limitations, the convergence between them provides stronger evidence for the generalizability of these results. That said, caution should be used in applying these results in subgroups and cultures outside of those currently studied until these results are replicated in those populations.

ACT-FAST is not a complete theory of action representation. We view ACT-FAST as a promising first step toward taxonomizing how people think about actions, rather than the final word in this matter. Much remains to be explained, both in the quantitative sense of capturing more of the reliable variance in participants’ behavior, and the conceptual sense of explaining more deeply why the particular dimensions we identify should be the most important for understanding actions. Unlike other domains, such as person perception, there were few existing theories against which to compare a new taxonomy of action representation. Nevertheless, we compared ACT-FAST to several existing theories across the course of the present investigation, including models of natural language, such as the GloVe word vector embedding (Pennington et al., 2014) and WordNet (Miller, 1995), as well as a model of person perception (Thornton & Mitchell, 2018). The ACT-FAST taxonomy outperformed all alternative models. These results suggest that ACT-FAST is the best theory currently available for explaining how people thinking about action concepts. In this capacity, it may serve as a compelling baseline against which to compare new behavioral and neural benchmarks. For example, several such hypothesis-driven investigations. For example, several such studies identified sociality as a key dimension of action representation (Tarhan & Konkle, 2020; Tucciarelli et al., 2019; Wurm et al., 2017). The fact that we spontaneously recovered the same dimension from text analysis—and subsequently validated it against new behavioral and neural benchmarks—lends further credence to the importance of this dimension. Not all of the dimensions identified here overlap with those examined in previous studies. For example, although spiritualism (vs. materialism) has been identified as an important dimension for organizing other psychological domains—such as beliefs (Leshinskaya et al., 2017) —to our knowledge it has not previously been linked to actions. However, sociologists at least as far back as Durkheim have posited the sacred-profane dichotomy as a primary—if perhaps culturally circumscribed—dimension of religion (Durkheim, 1912). Rediscovering this dimension in this new context helps to demonstrate the value of our data-driven approach.

Limitations

One limitation of the present research is that it relies heavily on between-person studies. In Study 1, we pooled text across many different writers; in Studies 3 and 4, separate participants rated ACT-FAST dimensions and action similarity. This design limits the conclusions which can be drawn. For example, actions that place near the “middle” of a dimension could wind up there because everyone agrees that they are in the middle, or because two subgroups have opposite views on which pole an action should be near. That said, Study 5 ameliorates this concern, as participants in this study made both dimension and similarity judgements and results nonetheless replicate those of the prior studies. Study 7 also divided data collection between participants, with separate participants providing imaging and rating data. However, all analyses in this study were conducted within-participant. The 100% replication rate at the individual level across two samples of imaging participants helps to further address the limitations of between-subjects designs by providing evidence for the high prevalence of the ACT-FAST’s descriptive powers in individuals.

Importantly, the evidence that we marshal to validate the ACT-FAST is primarily correlational. Although several studies featured random assignment of some kind, we did not manipulate the availability of ACT-FAST information and observe consequent changes in the ability to understand actions. Doing so would help to determine whether the ACT-FAST dimensions were necessary for understanding actions, rather than simply being a good description for action understanding. Unfortunately, manipulating the availability of this information in a naturalistic way would be very difficult. Although we could have manipulated people’s knowledge about novel actions (e.g., “daxing”), the applicability to real-world actions would have been dubious. For example, if all one knows about “daxing” is that it is a spiritual action, then one might use that information to judge the similarity of daxing to other actions regardless of whether one typically used spiritualism to understand real actions. For this reason, we restricted ourselves to a descriptive approach in the present investigation.

Future Directions

Future research can characterize the causal, rather than descriptive, function of ACT-FAST by using stronger experimental design—that is, tasks which more tightly constrain participant behavior. Many of the current paradigms were relatively unconstrained. For example, participants in Studies 3–5 judged the similarity between actions, but we did not instruct them upon what basis to make these judgements. Similarly, in Study 7, participants engaged in a passive viewing task which posed few constraints on their mental processes as they watched Sherlock or Forrest Gump. These unconstrained tasks have the virtue of allow participants to easily display their natural, spontaneous tendencies. However, they have the drawback of reducing the interpretability of their judgements or brain activity. Rather than manipulating the availability of action knowledge, constructing tasks that constrain participants to use a certain dimension of the ACT-FAST may provide a way to whether this taxonomy can extend beyond description to causal explanation.

One advantage of representing actions dimensions is that it could facilitate generalization to new actions. That is, the coordinates of an action on the ACT-FAST dimensions could provide an indication of how to generalize from previously known actions to new actions. Future research could leverage naturally occurring differences in action knowledge to test this hypothesis. For example, over development, children learn about different actions at different times—food-related actions are obviously important from the beginning, but abstract social behaviors may only become important when approaching adolescence. Studying action word learning might thus allow researchers to make differentiated
predictions about when children will start to generalize along each ACT-FAST dimension. The acquisition of new action concepts when entering a new culture, or when acquiring technical training, might provide equally promising testbeds for studying the generalization of action understanding.

Another promising future direction for ACT-FAST research is to consider its role in social prediction. In recent research, it was found that a dimensional taxonomy of mental states (Tamir et al., 2016) could help explain how people accurately predict each other’s thoughts and feelings (Thornton & Tamir, 2017). This research found that mental states transition from one to another with considerable statistical reliability, and that perceivers’ had accurate mental models of which transitions were more or less likely. They appear to achieve this in part by attending to which states were closer together on the psychological dimensions of rationality (cognitive vs. emotional states), social impact (states that impact others more or less), and valence (positive vs. negative states). The closer to states were together, the more likely transitions were between them, both in reality and perceivers’ perceptions.

This perspective suggests that a similar mechanism might support action prediction (Tamir & Thornton, 2018). The action which someone is currently engaged in can often signal their likely future actions. For example, someone who is currently cooking is likely to soon be eating. Attending to these reliable transitional probabilities could inform accurate prediction of others’ actions. However, rather than trying estimate likelihoods of the millions of transitions between thousands of actions, perceivers’ minds could simplify this process considerably by relying on the proximity between actions on the ACT-FAST dimensions. Returning to the example above, eating and cooking are close together on the food dimension in particular, so this proximity could allow perceivers to accurately predict that this likely transition. Indeed, recent behavioral and neural investigations have found evidence for precisely this hypothesis: proximity on the ACT-FAST dimensions mediates the accuracy of judgements about action transitions (Thornton & Tamir, 2020b, 2021). These findings demonstrate a novel mechanism for action prediction, as well as further evidence for the psychological usefulness of the ACT-FAST.

One other important direction for future study is to identify the cultural and/or evolutionary pressures that made the ACT-FAST dimensions good descriptors of action understanding. For example, all people have a biological need to eat, so it is perhaps unsurprising that this dimension plays a role in action understanding. At the same time, different cultures have provided us with many different ways to obtain, prepare, and consume food, necessitating the description of food-related actions in terms of other conceptual dimensions as well. Consideration of the factors that shaped ACT-FAST may reveal how to taxonomize not only actions, but the dimensions of action understanding as well. For example, this effort could help to quantify the relative abstraction of each dimension, how domain general versus specific they are, and what aspects of action they typically describe.

Although we have taken initial steps toward characterizing individual differences in action concepts (see the online supplemental materials), this topic should be further investigated in future research. We speculate that there may be many highly specific dimensions of action understanding that are idiosyncratic to particular people or contexts. For example, an expert metalworker might think about actions related to their craft—heating, quenching, soldering, and so forth—using a much higher-dimensional space than a novice. In previous research on emotion representation, we found evidence consistent with this hypothesis: People represented their own thoughts and feelings using a representational space that was more granular than the ones they applied to understand others’ mental states (Thornton et al., 2019). In the domain of actions, studying the size and shape of individuals’ representational spaces might provide novel insight into the development of expertise.

Conclusion

In conclusion, in the present investigation we derived and validated a six-dimensional taxonomy of action concepts. This action taxonomy—ACT-FAST—explains a wide variety of action-related phenomena, ranging from natural language use to behavioral judgments of actions to the brain activity elicited by action observation. Although ACT-FAST may be the first explicit taxonomy of actions per se, it outperforms a number of related theories, such as taxonomies of verbs, in capturing human action concepts. ACT-FAST provides a solid basis for answering some of the most common and important questions people face every day about actions. Together, these findings make ACT-FAST a powerful new tool for understanding human behavior.

References


This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.