DOI 10.15622/ia.21.5.3

I. SUROV

OPENING THE BLACK BOX: FINDING OSGOOD'S SEMANTIC FACTORS IN WORD2VEC SPACE

Surov I. **Opening the Black Box: Finding Osgood's Semantic Factors in Word2vec Space.**

Abstract. State-of-the-art models of artificial intelligence are developed in the black-box paradigm, in which meaningful information is limited to input-output interfaces, while internal representations are not interpretable. The resulting algorithms lack explainability and transparency, requested for responsible application. This paper addresses the problem by a method for finding Osgood's dimensions of affective meaning in multidimensional space of a pre-trained word2vec model of natural language. Three affective dimensions are found based on eight semantic prototypes, composed of individual words. Evaluation axis is found in 300-dimensional word2vec space as a difference between positive and negative prototypes. Potency and activity axes are defined from six process-semantic prototypes (perception, analysis, planning, action, progress, and evaluation), representing phases of a generalized circular process in that plane. All dimensions are found in simple analytical form, not requiring additional training. Dimensions are nearly orthogonal, as expected for independent semantic factors. Osgood's semantics of any word2vec object is then retrieved by a simple projection of the corresponding vector to the identified dimensions. The developed approach opens the possibility for interpreting the inside of black box-type algorithms in natural affective-semantic categories, and provides insights into foundational principles of distributive vector models of natural language. In the reverse direction, the established mapping opens machine-learning models as rich sources of data for cognitive-behavioral research and technology.

Keywords: semantics, dimension, Osgood, affective meaning, interpretation, word2vec, language, black box.

1. Introduction. Machine-learning approaches to natural language processing suffer from the interpretability problem. Popular models represent words by states of a hidden layer of a neural network, trained for various linguistic tasks on large corpora of texts [1–4]. Relevant regularities of natural language are then encoded in high-dimensional vectors, components of which say nothing to their developers and users. Although efficient for many tasks [5–8], this "black box" paradigm is problematic for applications of artificial intelligence (AI), requested to be transparent and explainable [9, 10].

Compared to the dimensionality of machine-learning (ML) representations, interpretable factors of human thought are much less in number. They are *evaluation* (pleasantness-unpleasantness), *activity* (active-passive, external-internal), and *potency* (strength, dominance, openness, freedom) [11, 12]. As established by Charles Osgood, these dimensions account for the majority of judgment variance for various objects in semantic-differential scales like good-bad, heavy-light, soft-hard, straight-curved, etc.

(ibid). Although widely known in cognitive science, these results are largely ignored in modern AI and ML research.

Previous attempts to establish the missing link achieved only partial success. *Evaluation*, *potency*, and *activity* are found to correlate with 208, 187, and 175 out of 300 raw word2vec dimensions, respectively [13]. Out of 280 principal components of the same model, these numbers are 38, 35, and 37 (ibid.), indicating that the achieved correspondence is far from useful. Nevertheless, word2vec and similar (distributional-vector) representations of language are known to encode specific types of semantic information that can be extracted by additional methods [14, 15].

An inspiring example is provided by self-organized semantic maps of natural languages [16, 17]. These maps, built from synonym-antonym relations by a specially designed algorithm (minimizing an energy-like cost function of word configuration), consistently identify Osgood's dimensions as the main principal components of English, German, Spanish, French, and Russian. Although expected to take a central place in biologically-inspired cognitive architectures [18], the present form of these maps is yet of demonstrative character. The combination of semantic interpretability with practicality of applied machine learning remains a challenge.

The present work addresses this problem for the word2vec model [1], baseline in ML approach to natural language processing. The paper describes a novel method to identify directions in the multidimensional word2vec space, responsible for Osgood's semantic factors. In contrast to the aforementioned attempts, the method does not rely on special variance properties of semantic dimensions, presupposed in principal component analysis. Instead, it finds unique dimensions based on a small sample of supervised data and a simple mathematical procedure. The theory and realization of the algorithm are described in Sections 2 and 3. Tests of stability and predictive potential of the developed method are reported in Section 4. Section 5 discusses the implications of the result.

2. Theory. The theory is developed in the following steps. After introducing the source data, Section 2.1 describes a method for finding the evaluation axis (Z) in the word2vec space. Section 2.2 then elaborates this logic to find the potency-activity (XY) plane. Finally, Section 2.3 shows how to use the obtained dimensions to extract Osgood's semantics of arbitrary word2vec representation.

Source data. This work uses a word2vec model of the English language, trained to predict skipped words based on their surroundings on Google News dataset of about 100 billion words [19]. The model encodes ~ 3 million English words w in 300-dimensional vectors $\vec{w}$. These vectors reflect regularities of

linguistic practice in algebraic relations of type:

$$\overrightarrow{Greece} - \overrightarrow{Athens} \approx \overrightarrow{Russia} - \overrightarrow{Moscow}, \tag{1a}$$

$$\overrightarrow{king} - \overrightarrow{man} \approx \overrightarrow{queen} - \overrightarrow{woman}, \tag{1b}$$

useful in natural language analysis [1, 20]. Neither individual dimensions of vectors (1), nor principal components of the model as a whole, however, have understandable meanings, leading to the aforementioned interpretability problem.

2.1. Evaluation Z axis. The finding of Osgood's evaluation axis in 300-dimensional word2vec space is suggested by the above identities. Both sides of (1a), for example, refer to the concept close to *country*, while sides of (1b) encode the notion of *mightiness*. In the same way, *pleasantness* can be defined as:

$$\begin{gathered}\overrightarrow{good} - \overrightarrow{bad}, \\ \overrightarrow{joy} - \overrightarrow{grief}, \\ \overrightarrow{well} - \overrightarrow{poor},\end{gathered} \tag{2}$$

and similar differences. To get a unitary definition, the sides of these pairs are combined in averaged vectors:

$$\vec{W}_{\mathrm{good}} = \frac{1}{N}\sum_{j=1}^{N} \vec{w}^{j}_{\mathrm{good}}, \qquad \vec{W}_{\mathrm{bad}} = \frac{1}{N}\sum_{j=1}^{N} \vec{w}^{j}_{\mathrm{bad}}, \tag{3}$$

where N in the number of individual words in "good" and "bad" *prototypes*. Difference between them then defines Osgood's *evaluation* axis:

$$\vec{Z} = \vec{W}_{\mathrm{good}} - \vec{W}_{\mathrm{bad}}, \tag{4}$$

averaging out semantic variations among lines of (2) as illustrated in Figure 1(a).

2.2. Potency – Activity XY plane. Osgood's *activity* and *potency* dimensions could be found by the same opponent-prototype method as used for the *evaluation* axis above. The XY plane then would be defined via "active-passive" and "strong-weak" prototype pairs analogous to (4). Four points, however, are excessive to define a plane; on the other hand, it is desirable to be able to use more prototypes to achieve the necessary precision of the resulting dimensions $\vec{X}$ and $\vec{Y}$. The rest of this section describes a method for this.

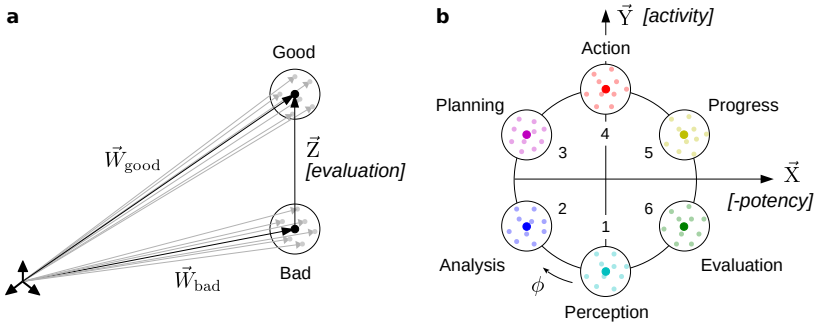

Fig. 1. Finding Osgood's semantic dimensions in the word2vec space: a) evaluation axis $\vec{Z}$ is defined as a difference between 300-dimensional "good" and "bad" prototypes (4); individual word vectors $\vec{w}^i$ and central prototypes (3) are shown in black and gray, respectively, b) finding the XY plane via six process-semantic prototypes, introduced in [21]

2.2.1. General idea. Suppose we have k word2vec prototypes of type (3):

$$\vec{W}_i = \frac{1}{N_i} \sum_{j=1}^{N} \vec{w}_i^j, \quad i = 1 \dots k, \tag{5}$$

that should have coordinates $(x_1, y_1) \dots (x_k, y_k)$ in the XY plane to be found. If these prototypes are indeed coplanar, their vectors $\vec{W}_i$ must be representable as linear superpositions of its basis vectors $\vec{X}$ and $\vec{Y}$. In the matrix form, illustrated in Figure 2, this is expressed by decomposition:

$$W = A * \Omega_{xy}, \tag{6}$$

where two rows of Ω_{xy} are the basis vectors $\vec{X}$ and $\vec{Y}$, k rows of W are word2vec prototypes $\vec{W}_i$, and k rows of A are the expected coordinates (x_i, y_i) of these prototypes in the XY plane.

If the coordinate matrix A is invertible, Ω_{xy} is obtained by multiplying the sides of (6) by A^{-1} from the left, so that:

$$\Omega_{xy} = A^{-1} * W. \tag{7}$$

Being rows of this matrix, the sought dimensions $\vec{X}$ and $\vec{Y}$ are then obtained as linear combinations of the prototype vectors $\vec{W}_i$ in 300-dimensional word2vec space.

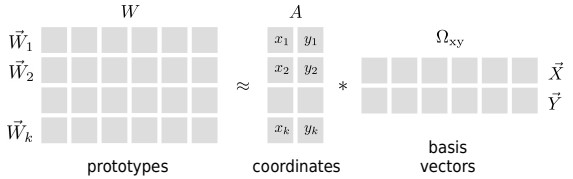

Fig. 2. Finding the potency-activity XY plane from several prototypes. The corresponding word2vec vectors $\vec{W}_i$ are represented by linear superposition of the basis vectors $\vec{X}$ and $\vec{Y}$ with coordinates $\{x_i, y_i\}$. Precise equality (6) is reached if the prototypes are coplanar. Basis vectors forming matrix Ω_{xy} are found by multiplying both sides by inverted coordinate matrix A^{-1} (7)

2.2.2. Specification of the prototypes. To find the XY plane, this study used the set of $k = 6$ prototypes close to that developed in [21]. These prototypes are *perception*, *analysis* (of novelty), *planning*, *action*, *progress*, and *evaluation* (of the result), forming a circular template for process representation in natural thinking (ibid)[1]:

1. **Perception** prototype describes observations, leading to the discovery of novelty. In the process of writing a paper this could include, for example, the accumulating new knowledge or demands for clarification of previous results. Realization of the fact that a new paper is needed turns the process to the next stage.

2. **Analysis** prototype accounts for the understanding of a newly identified factor. In the paper example, this is the thinking process revealing what exactly needs to be explained, and which problem will be resolved. The formalization of these requirements turns the process to the next stage.

3. **Planning** prototype describes a prospective vision of how the novelty will be handled. In the same example, this includes conceptualization of the paper and specification of its structure, approximate content of the sections, and selection of an appropriate journal for submission. The finalization of this project moves the process to the next stage.

4. **Action** prototype describes an implementation of the project. This is the process of making a paper with all accompanying activities: writing the text, producing the figures, selecting the references, and formatting the

[1]These prototypes are subsequent stages of a generalized cyclical process, most obviously derived from day-night and seasonal cycles. This progression is found in any process from taking a shower to running a space mission, with particular granularity chosen according to the desired precision. Listed prototypes are obtained from doubling the resolution of a basic three-stage sequence *analysis - action - evaluation*, identical to the baseline cybernetic *sense (evaluation) - think (analysis) - act (action)* control loop [22].

manuscript according to the journal's requirements. Submission of the paper finalizes this stage of the process.

5. **Progress** stage situates the obtained prototype in real circumstances based on the received feedback. The paper-making process includes answering to reviewer's questions, correcting the contents, and resubmission to another journal if necessary. The stage is finalized by the journal's decision.

6. **Evaluation** prototype accounts for subjective estimation of the result. In the case of a positive decision, it describes how well the paper achieved the initial goals, and estimates intended and unintended consequences. Alternatively, implications of a negative result, e.g. abandoning the project based on the received feedback, are summarized and recorded.

Individual words forming the prototypes were selected to represent their function in the abstract cyclical process. For example, various aspects and types of perception are accounted by observation, cognition, feedback, feeling, reflection, intuition, and the like; analysis, similarly, is strongly associated with novelty as its object, attention, thinking, reasoning, questioning, forming hypotheses and theories. As compared e.g. to *arm* or *table*, such concepts with definite process functions, stable across the majority of contexts, are small in number. Manually formed lists of such process-functional English terms for each prototype are given in Table 1.

2.2.3. Coordinate matrix. According to the cyclical topology of the process template, it can be visualized as a circular trajectory in the XY plane. Uniform discretization of this trajectory to six process stages is then expected to form a regular hexagon as shown in Figure 1(b). With the radius of the circle set to unity, corresponding coordinates of the prototype centers become:

$$\begin{aligned} x_i &= -\sin\Phi_i, \\ y_i &= -\cos\Phi_i, \qquad \Phi_i = 60^\circ * (i-1), \end{aligned} \tag{8}$$

where the process phase ϕ starts from the *perception* prototype with $\Phi_1 = 0^\circ$ and increases in steps of 60°. (Pseudo)inverse of the resulting coordinate matrix A (6), remarkably, is proportional to its transposition:

$$A^{-1} = \frac{1}{3}A^{\mathrm{T}} = \frac{1}{6}\begin{bmatrix} 0 & -\sqrt{3} & -\sqrt{3} & 0 & \sqrt{3} & \sqrt{3} \\ -2 & -1 & 1 & 2 & 1 & -1 \end{bmatrix}, \tag{9}$$

with coefficient $1/3$[2]. Substituting (9) in (7) produces the requested $\vec{X}$ and $\vec{Y}$ vectors in 300-dimensional word2vec space.

[2]Obtained as $2/k$. In this form, the analytical part of (9) holds for any number of prototypes, dividing the process circle into k equal sectors analogous to Figure 1(b).

Table 1. Individual terms, forming two Z-axis (3) and six XY-plane prototypes (5) shown in Figure 1

Prototype	**Individual terms**
Good	good light well fine yes
Bad	bad dark poor vice no
Perception	perception observation cognition feedback feeling reflection intuition insight introspection monitoring sensing data forecast prediction expectation contemplation anticipation
Analysis	analytics novelty hypothesis theory problem reason mystery question attention factor issue query puzzle challenge think
Planning	plan aim goal model concept intent purpose project principle plot motive strategy design map vision solve
Action	action act work duty develop implement manage deal compete cooperate execute accomplish produce construct engage fulfill
Progress	progress regress advance growth attainment agreement negotiation gain bargain increase decrease output yield completion profit return
Evaluation	evaluation estimation result end summation summary victory defeat conclusion final outcome aftermath expiration termination record score

2.2.4. Complex-valued form. The latter procedure can be made intuitive by rendering it in complex-valued form, with $\vec{Y}$ and $\vec{X}$ becoming real and imaginary axes of the complex plane. Values (8) then become real and imaginary parts of a single complex-valued coordinate $c_i = -\exp \mathrm{i}\phi_i$, so that (9) transforms to a single complex-valued row:

$$A^{-1} = \frac{-1}{3}\left[e^{\mathrm{i}\Phi_1} \quad e^{\mathrm{i}\Phi_2} \quad \dots \quad e^{\mathrm{i}\Phi_6}\right]. \tag{10}$$

For arbitrary k, the product (7) then takes form:

$$\vec{\Omega} = -\frac{2}{k}\sum_{i=1}^{k} \vec{W}_i e^{\mathrm{i}\Phi_i}, \tag{11}$$

with $\vec{Y}$ and $\vec{X}$ being real and imaginary parts of this vector. Up to the normalization factor $2/k$, (11) is then recognized as a complex-valued generalization of (4), combining relevant prototypes weighted by their theoretically-expected positions in the corresponding subspace.

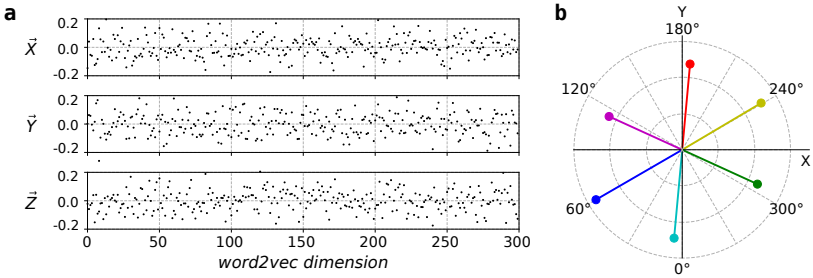

Fig. 3. a) components of Osgood's potency $\vec{X}$, activity $\vec{Y}$, and evaluation $\vec{Z}$ dimensions in 300-dimensional word2vec space, obtained from (4) and (11), b) projection (12) of six process-semantic prototypes (5) to the (uncorrected) XY plane. In the same color encoding, this layout closely aligns with the theoretical scheme in Figure 1(b). The difference is corrected by stretching the plane along the 150° - 330° direction as described in Section 3.1

2.3. Extraction of Osgood's semantics of arbitrary words. Any word of natural language is located in the XYZ space by projecting the corresponding word2vec representation $\vec{w}$ to its basis vectors:

$$\vec{s} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \vec{w} \cdot \vec{X} \\ \vec{w} \cdot \vec{Y} \\ \vec{w} \cdot \vec{Z} \end{bmatrix}, \tag{12}$$

where $\cdot$ denotes scalar product. This mapping of 300-dimensional vector $\vec{w}$ to 3-dimensional vector $\vec{s}$ finalizes the algorithm. The obtained components z, $-x$, and y are Osgood's semantic factors *evaluation*, *potency*, and *activity*, extracted from the word2vec data.

3. Experiment. The XYZ axes are obtained according to Sects. 2.1 and 2.2 are explicitly shown in Figure 3. Each axis appears to be a broad superposition of word2vec dimensions, indicating a non-trivial relation between word2vec model and Osgood's semantics.

To verify the above theory, it is necessary to check (i) if the prototypes (5) are actually located in the obtained XY plane according to the theoretical expectation shown in Figure 1(b), and (ii) if the Z prototypes (3), Figure 1(a), fall on different sides of this plane above and below the origin.

The first check is done by applying the first two lines of (12) to the prototypes (5). The obtained vectors $\vec{S}_{xy}$, shown in Figure 3(b), align with Figure 1(b) up to the following differences:

– prototypes *perception* (cyan) and *planning* (purple) deviate from their expected phases 0° and 120° towards the *analysis* prototype (blue) by 5.3° and 5.5°, respectively;

– prototypes *action* (red) and *evaluation* (green) deviate from their expected phases 180° and 300° towards the *progress* prototype (yellow) by 5.2° and 5.5°, respectively;

– *analysis* and *progress* prototypes have the largest vector lengths 0.25 and 0.23, compared to the other four being 0.21 on average.

3.1. Adjusting the X and Y axes. These differences might result from the asymmetry of the chosen prototypes, or of the word2vec model itself, possibly reflecting the process-semantic anisotropy of natural language. In any case, all of them are eliminated by squeezing the XY plane along the *analysis-progress* 60° - 240° direction by ≈ 1.2, or by stretching it along the orthogonal 150° - 330° direction by the same factor. The procedure also decreases the scalar product between the $\vec{X}$ and $\vec{Y}$ axes nearly twice, making them more orthogonal. Three pairwise scalar products become:

$$\vec{X} \cdot \vec{Y} = 0.036, \quad \vec{Y} \cdot \vec{Z} = 0.009, \quad \vec{Z} \cdot \vec{X} = 0.006,$$

showing nearly perfect orthogonality as expected for independent semantic dimensions. This correction is included in the following calculations.

3.2. Full map and features. The full semantic map of the prototypes $\vec{W}_i$ and their individual words $\vec{w}_i^{\,j}$ is obtained by projection (12) of the corresponding vectors to the adjusted XY plane:

$$\vec{S}_i = \vec{W}_i \cdot \vec{\Omega}, \qquad \vec{s}_i^{\,j} = \vec{w}_i^{\,j} \cdot \vec{\Omega}. \tag{13}$$

Figure 4 shows the resulting XY components of these vectors in the same color coding as in Figure 1(b). Unlike Figure 3(b), the prototypes are now centered at their theoretic angles (8) with a standard deviation of 0.6°. Individual terms of each prototype (Table 1) scatter around their means with standard angular deviations:

$$\Delta\phi_i = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} \left(\phi_i^j - \Phi_i\right)^2}, \tag{14}$$

amounting to 17° on average. The standard distance of individual words from each prototype center is indicated by semi-transparent circles of the corresponding radii and color.

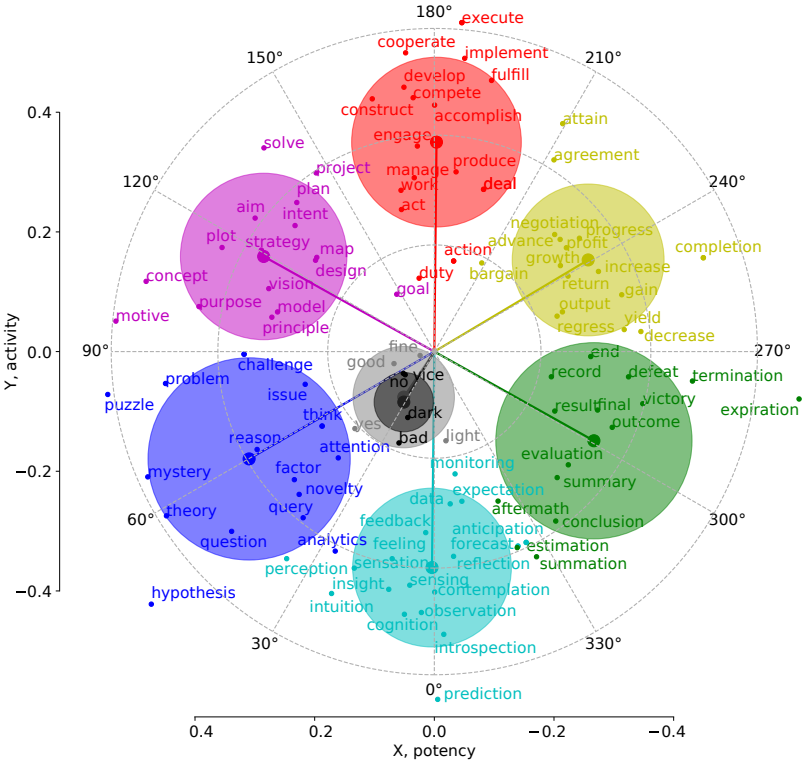

Fig. 4. Prototypes and their individual words (Table 1), projected to the potency-activity XY plane (11) via (12). Gray and black: “good” and “bad” prototypes (3). Process-stage prototypes (5): cyan - *perception*, blue - *analysis*, magenta - *planning*, red - *action*, yellow - *progress*, and green - *evaluation*, located in agreement with theoretical expectations shown in Figure 1(b). Semi-transparent circles indicate the scattering of individual terms in each prototype

With lengths in the XY plane 0.33 ± 0.025 and Z coordinates 0.024 ± 0.012, six XY prototypes are nearly coplanar as expected in theory, approximating the regular hexagon shown in Figure 1(b).

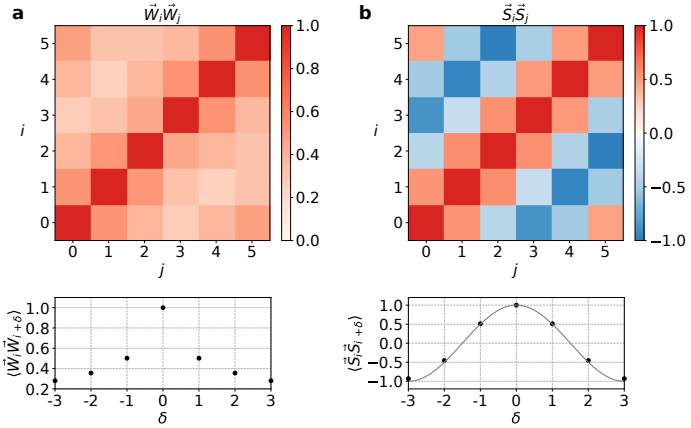

Fig. 5. Transformation of similarities among six XY prototypes (5) due to projection (12) from 300-dimensional word2vec space (a) to the semantic XYZ space (b). Top: pairwise scalar products of the prototype vectors. Bottom: the same as an averaged function of a stage-number difference $\delta = i - j$. All similarities among the original vectors (a) are positive, while in the process-semantic projection (b) similarities follow the harmonic function of the angular difference, as expected for circular layout (gray line)

"Good" and "bad" evaluation prototypes (3) are shown in gray and black. As expected, they are projected close to the origin of the XY plane with a displacement of ≈ 0.1 in the *perception* - *analysis* direction $\phi \approx 30°$. Z coordinates of these prototypes are 0.29 and -0.31, respectively. Both verification conditions, indicated at the beginning of this section, are thereby fulfilled.

Transformation of similarity Mapping of the XY prototypes from 300-dimensional word2vec representation $\vec{W}_k$ to the process plane (13) is illustrated by the transformation of their mutual similarity. According to their circular arrangement in the process plane, pairwise scalar products of three-dimensional prototype projections S_k follow the harmonic pattern shown in Figure 5(b). Initial prototype vectors $\vec{W}_k$ in 300-dimensional word2vec space, in contrast, all have positive scalar products shown in panel (a).

4. Testing. This section verifies the statistical significance of the obtained map and quantifies the ability of the method to predict semantic scores of individual words.

4.1. Randomization test. To verify the significance of the obtained map, the above procedure was performed for the same number of prototypes,

but composed of individual words sampled from Table 1 in a random way. The prototypes were ascribed with the same theoretical phases (8) and used to find the XY plane as before (11). The resulting map does not show the regularity demonstrated above. In contrast to Figure 4, prototypes are located at random phase angles and distances from the origin, while the scattering of individual terms (14) within prototypes is much larger than in the original map.

Finding a plane in which random word2vec vectors would take the prescribed angular positions Φ_i, therefore, does not seem to be possible. The regular structure appears only for semantically-coherent prototypes, composed in agreement with objective regularities of the word2vec data. The map shown in Figure 4 thereby reflects a non-trivial feature of natural language, rather than a mathematical artifact brought in by the construction method.

4.2. Mapping of novel words. In this test, 15 terms populating each context class according to Table 1 were divided into n "seed" and $15 - n$ "probe" items. The process-semantic plane (3) was then identified based on $6n$ seed terms, while the remaining $6(15 - n)$ probe terms were mapped to this plane by the procedure described above.

With n seed terms per semantic class randomly selected from Table 1, this procedure was repeated $m = 200$ times. For $n = 10$, the resulting scattering of $6m(15 - n) = 6000$ probe terms is shown in Figure 6(a). The mean of standard angular deviations (14) amounting to 35° indicates the ability of the method to correctly map novel words to the process-semantic plane, extracting the necessary information from the word2vec model.

Figure 6(b) shows the dependence of standard angular deviation (14) and mean length of the prototype vectors $\langle S_i \rangle$ (13) on the seed size n. With increasing n, semantic features of individual words average out more efficiently, suppressing angular noise $\Delta\phi$ and increasing process-semantic coherence of the prototypes $\langle S_i \rangle$. Angular discrimination threshold of $\Delta\phi \approx 30^\circ$ is reached near $n = 10$ when the mean scattering radius $\langle R \rangle$ drops below one-half on the mean length $\langle S_i \rangle$. The map in panel (a) is close to this borderline regime.

5. Discussion.

5.1. Natural semiotics in ML. The theoretical structure of the prototypes used in Section 2 is not constructed specifically for the purpose of this paper, but was previously identified from a cognitive-semiotic perspective. In particular, the process-stage prototypes, shown in Figure 1(b), form an abstract template for causal-predictive simulation of behavior, structurally isomorphic to periodic processes in Nature like year- and day-night cycles [21]. Along with the evaluation dimension, the resulting spherical space is considered as a core semiotic system of living Nature, underlying natural sense-making from single-cell organisms to humans (ibid.).

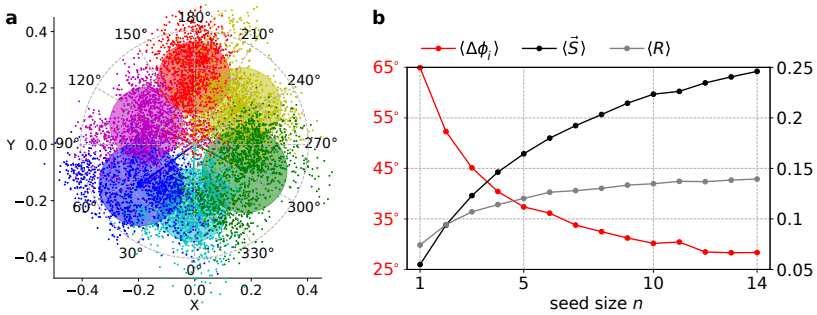

Fig. 6. Prediction of the process-semantics phase of individual words. Lists of individual terms for six XY prototypes in Table 1 are split into two parts so that the "seed" part of the words is used to find the XY plane, while the remaining "probe" part is used for testing. a: In each class of 15 individual words, 10 randomly selected ones are used as seed and the rest is used for testing. The result of 200 runs is shown. b: a mean of the standard angular deviations (14), a mean of the lengths of the prototype vectors $\langle\vec{S}\rangle$, and a mean of the scattering radii $\langle R\rangle$ as a function of the number of seed words

Observing the expected theoretical structure in Figure 4 provides insight into the general principles of ML and AI. Although aimed at a simple predictive task (recovery of skipped words based on their surroundings), efficiency required word2vec to accurately reflect the spatial structure of this highly abstract semiotic alphabet. Machine-learning image of natural language thus extends to the level of universally interpretable affective meaning, achieving deeper correspondence with cognitive semantics than was previously known [23]. Central features of the interpretable semantic map, recognized as a key element of next-generation biologically-inspired AI [18], thus appear to be already in place.

The established isomorphism can also be harnessed in novel architectures of AI and ML. If an efficient model is bound to reflect fundamental regularities of Nature, knowing the latter facilitates the design of the former. Instead of converging to such regularities in a blind search, they could be hardwired into the neural architectures from the start, reducing the dimensionality of the training optimization without loss in quality. The resulting economy of computational resources opens prospects for better replication of natural cognition and learning [24–26], as advocated e.g. in the field of information retrieval [27]. Generalization of the developed approach to other vector representations [14, 28–30] and machine-learning models [15, 31] allows going beyond the simplest case considered in this paper.

5.2. Affectively interpretable AI. The obtained mapping contributes to solving the interpretability problem noted in the Introduction, that is, the inability to express internal states of the black box-type algorithms in sensible categories like that of natural language. Evaluation, potency, and activity dimensions are such categories, unique in their affective nature and cross-linguistic universality [32–34]. Due to the centrality of affective meaning in human cognition [35, 36], remarkably overlooked in recent surveys on explainable AI [37–44], these dimensions are basic for making the internal workings of such algorithms available for human inspection. Further, this might be used to align AI and ML algorithms with the principles of commonsense reason, enabling computing in natural categories of human thought [45, 46].

By finding Osgood's semantics in the baseline machine-learning model, the reported approach also opens a prospect for developing explainability tests for other complex algorithms. In the decision-support systems, for example, it could be used to "look inside" the black box [47] and observe or correct the affective state it simulates towards a target entity[3]. The concurrence of such a state with human ethics and reason could supplement other certification criteria [53].

5.3. Use for cognitive modeling. In the reverse direction, the established link between Osgood's semantics and word2vec data allows using the latter for cognitive modeling and research. According to Osgood's original method [12] and its successors, 50 to 80 percent of judgment data variance, defined by *evaluation*, *potency*, and *activity* factors, could be extracted directly from word2vec representation of situations and things. Based on that, judgment and decision probabilities of interest could be predicted without performing real-world experiments as envisioned by [54–56].

Besides speed and cost advantages, this approach is also expected to be higher in precision, since word2vec models (trained on huge corpora of texts) accumulate much more information, than usually collected in old-style experiments. Through subtle regularities of natural language, this also includes implicit knowledge, hardly observable in laboratory conditions.

6. Conclusion. The possibility to retrieve Osgood's semantics from the word2vec data shows that the most agnostic models of data science converge to the basic principles of natural thinking, previously revealed in cognitive and semiotic studies. After such validation, these principles facilitate finding of nature-inspired solutions for hard problems in computer science. The interpretability problem of AI, for example, might be not as hard as seen

[3]This is not to be mistaken with an affective state of a machine itself, sometimes ascribed to it within a so-called intentional stance [48] exemplifying cognitive fallacy to ensoul complex systems [49–52].

from a brute-force computational paradigm, dominating the field today. If the reported method could be extended to other ML models, the expainability of the present black-box AI might be approached by a minor add-on, analogous to the projection procedure described in this paper.

References

1. Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations. Proceedings of NAACL-HLT. 2013. pp. 746–751.
2. Pennington J., Socher R., Manning C.D. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing. 2014. pp. 1532–1543.
3. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018.
4. Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.V. XLNet: Generalized autoregressive pretraining for language understanding. Proceedings of 33rd Conference on Neural Information Processing Systems. 2019.
5. Mikolov T., Joulin A., Baroni M. A Roadmap Towards Machine Intelligence. *Computational Linguistics and Intelligent Text Processing.* Cham: Springer, 2018. pp. 29–61.
6. Al-Saqqa S., Awajan A. The Use of Word2vec Model in Sentiment Analysis: A Survey. ACM International Conference Proceeding Series. 2019. pp. 39–43.
7. Dhar A., Mukherjee H., Dash N.S., Roy K. Text categorization: past and present. *Artificial Intelligence Review.* 2021. vol. 54. no. 4. pp. 3007–3054.
8. Konstantinov A., Moshkin V., Yarushkina N. Approach to the Use of Language Models BERT and Word2vec in Sentiment Analysis of Social Network Texts. *Recent Research in Control Engineering and Decision Making.* Cham: Springer, 2021. pp. 462–473.
9. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.Z. XAI—Explainable artificial intelligence. *Science Robotics*. 2019. vol. 4. no. 37.
10. Suvorova A. Interpretable Machine Learning in Social Sciences: Use Cases and Limitations. *Proceedings of Digital Transformation and Global Society 2021. Communications in Computer and Information Science, vol. 1503*. Cham: Springer, 2022. pp. 319–331.
11. Osgood C.E. The nature and measurement of meaning. *Psychological Bulletin*. 1952. vol. 49. no. 3. pp. 197–237.
12. Osgood C.E. Studies on the generality of affective meaning systems. *American Psychologist*. 1962. vol.17. no.1. pp. 10–28.
13. Hollis G., Westbury C. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin and Review*. 2016. vol. 23. no. 6. pp. 1744–1756.
14. Lenci A. Distributional Models of Word Meaning. *Annual Review of Linguistics*. 2018. vol. 4. no. 1. pp. 151–171.
15. Günther F., Rinaldi L., Marelli M. Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science*. 2019. vol. 14. no. 6. pp. 1006–1033.
16. Samsonovich A.V., Ascoli G.A. Principal Semantic Components of Language and the Measurement of Meaning. *PLoS ONE*. 2010. vol. 5. no. 6.
17. Eidlin A.A., Eidlina M.A., Samsonovich A.V. Analyzing weak semantic map of word senses. *Procedia Computer Science*. 2018. vol. 123. pp. 140–148.

18. Samsonovich A.V. On semantic map as a key component in socially-emotional BICA. *Biologically Inspired Cognitive Architectures*. 2018. vol. 23. pp. 1–6.
19. Pretrained word2vec model “GoogleNews-vectors-negative300.bin.gz”. Google Code Archive. https://code.google.com/archive/p/word2vec/. 2013.
20. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 2013.
21. Surov I.A. Natural Code of Subjective Experience. *Biosemiotics*. 2022. vol. 15. no. 1. pp. 109–139.
22. Siegel M. The sense-think-act paradigm revisited. Proceedings of the 1st International Workshop on Robotic Sensing. 2003.
23. Gastaldi J.L. Why Can Computers Understand Natural Language? *Philosophy & Technology*. 2021. vol. 34. no. 1. pp. 149–214.
24. Jensen A.R. The relationship between learning and intelligence. *Learning and Individual Differences*. 1989. vol. 1. no. 1. pp. 37–62.
25. Sowa J.F. The Cognitive Cycle. Proceedings of the 2015 Federated Conference on Computer Science and Information Systems. 2015. vol. 5, pp. 11–16.
26. Wang Y., Yao Q., Kwok J.T., Ni L.M.: Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*. 2021. vol. 53. no. 3. pp. 1–34.
27. Hoenkamp E. Why Information Retrieval Needs Cognitive Science: A Call to Arms. 2005.
28. Turney P.D., Pantel P. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*. 2010. vol. 37. pp. 141–188.
29. Wang B., Buccio E.D., Melucci M. Representing Words in Vector Space and Beyond. *Quantum-Like Models for Information Retrieval and Decision-Making* (eds: Aerts D., Khrennikov A., Melucci M., Toni B.). Cham, Springer. pp. 83–113. 2019.
30. beim Graben P., Huber M., Meyer W., Römer R., Wolff M. Vector Symbolic Architectures for Context-Free Grammars. *Cognitive Computation*. 2022. vol. 14. no. 2. pp. 733–748.
31. Coenen A., Reif E., Kim A.Y.B., Pearce A., Viégas F., Wattenberg M. Visualizing and measuring the geometry of BERT. Proceedings of the Advances in Neural Information Processing Systems. 2019.
32. Tanaka Y., Oyama T., Osgood C.E. A cross-culture and cross-concept study of the generality of semantic spaces. *Journal of Verbal Learning and Verbal Behavior*. 1963. vol. 2. no. 5-6. pp. 392–405.
33. Tanaka Y., Osgood C.E. Cross-culture, cross-concept, and cross-subject generality of affective meaning systems. *Journal of Personality and Social Psychology*. 1965. vol. 2. no. 2. pp. 143–153.
34. Osgood C.E., May W.H., Miron M.S. *Cross-cultural universals of affective meaning*. Champaign, University of Illinois Press. 1975.
35. Zajonc R.B. Feeling and thinking: Preferences need no inferences. *American Psychologist*. 1980. vol. 35. no. 2. pp. 151–175.
36. Duncan S., Barrett L.F. Affect is a form of cognition: A neurobiological analysis. *Cognition and Emotion*. 2007. vol. 21, no. 6. pp. 1184–1211.
37. Lipton Z.C. The Mythos of Model Interpretability. *Queue*. 2018. vol. 3. pp. 31–57.
38. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2019.
39. Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. 2019. vol. 51. no. 5.

40. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019. vol. 1. no. 5. pp. 206–215.
41. Vassiliades A., Bassiliades N., Patkos T. Argumentation and explainable artificial intelligence: A survey // *Knowledge Engineering Review*. 2021. vol. 36. pp. 1–35.
42. Borrego-Díaz J., Galán-Páez J. Explainable Artificial Intelligence in Data Science. *Minds and Machines*. 2022.
43. Chou Y.L., Moreira C., Bruza P., Ouyang C., Jorge J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*. 2022. vol. 81. pp. 59–83.
44. Tian L., Oviatt S., Muszunski M., Chamberlain B.C., Healey J., Sano, A. Applied Affective Computing. ACM Books. 2022.
45. Michelucci P. (ed.) Handbook of Human Computation. New York, Springer. 2013.
46. Samsonovich A.V. (ed.) Biologically Inspired Cognitive Architectures. Advances in Intelligent Systems and Computing vol. 948. Cham, Springer. 2020.
47. Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018. vol. 6. no. 52. pp. 138–152.
48. Dennett D.C. The intentional stance. Cambridge, MIT Press. 1998.
49. Caporael L.R. Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*. 1986. vol. 2. no. 3. pp. 215–234.
50. Guthrie S.E. Anthropomorphism: A definition and a theory. *Anthropomorphism, anecdotes, and animals*. (eds. Mitchell R.W., Thomson N.S., Miles H.L.), chap. 5, pp. 50-58. State University of New York Press, New York. 1997.
51. Watson D. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines.* 2019. vol. 29, no. 3. pp. 417–440.
52. Salles A., Evers K., Farisco M. Anthropomorphism in AI. *AJOB Neurosci*. 2020. vol. 11. no. 2. pp. 88–95.
53. Maclure J. AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Minds and Machines*. 2021.
54. Arnulf J.K., Larsen K.R., Martinsen Ø.L., Bong C.H. Predicting survey responses: How and why semantics shape survey statistics on Organizational Behaviour. *PLoS ONE*. 2014. vol. 9. no. 9.
55. Jones M.N., Gruenenfelder T.M., Recchia G. In defense of spatial models of semantic representation. *New Ideas in Psychology*. 2018. vol. 50. pp. 54–60.
56. Arnulf J.K. Wittgenstein's Revenge: How Semantic Algorithms Can Help Survey Research Escape Smedslund's Labyrinth. *Respect for Thought* (eds. Lindstad T.G., Stänicke E., Valsiner J.), chap. 17, pp. 285–307. Springer, Cham. 2020.

Surov Ilya — Ph.D., Associate Professor, Senior researcher, ITMO University. Research interests: cognitive-behavioral modeling, quantum semiotics and semantics. The number of publications — 25. ilya.a.surov@itmo.ru; 49A, Kronverksky Av., 197101, St. Petersburg, Russia; office phone: +7(812)480-0000.

Acknowledgements. This research is supported by RNF (grant № 20-71-00136).

УДК 004.8 **DOI** 10.15622/ia.21.5.3

И.А. Суров

ОТКРЫТИЕ ЧЁРНОГО ЯЩИКА: ИЗВЛЕЧЕНИЕ СЕМАНТИЧЕСКИХ ФАКТОРОВ ОСГУДА ИЗ ЯЗЫКОВОЙ МОДЕЛИ WORD2VEC

Суров И.А. **Открытие чёрного ящика: Извлечение семантических факторов Осгуда из языковой модели word2vec.**

Аннотация. Современные модели искусственного интеллекта развиваются в парадигме чёрного ящика, когда значима только информация на входе и выходе системы, тогда как внутренние представления интерпретации не имеют. Такие модели не обладают качествами объяснимости и прозрачности, необходимыми во многих задачах. Статья направлена на решение данной проблемы путём нахождения семантических факторов Ч. Осгуда в базовой модели машинного обучения word2vec, представляющей слова естественного языка в виде 300-мерных неинтерпретируемых векторов. Искомые факторы определяются на основе восьми семантических прототипов, составленных из отдельных слов. Ось оценки в пространстве word2vec находится как разность между положительным и отрицательным прототипами. Оси силы и активности находятся на основе шести процессно-семантических прототипов (восприятие, анализ, планирование, действие, прогресс, оценка), представляющих фазы обобщённого кругового процесса в данной плоскости. Направления всех трёх осей в пространстве word2vec найдены в простой аналитической форме, не требующей дополнительного обучения. Как и ожидается для независимых семантических факторов, полученные направления близки к попарной ортогональности. Значения семантических факторов для любого объекта word2vec находятся с помощью простой проективной операции на найденные направления. В соответствии с требованиями к объяснимому ИИ, представленный результат открывает возможность для интерпретации содержимого алгоритмов типа “чёрный ящик” в естественных эмоционально-смысловых категориях. В обратную стонону, разработанный подход позволяет использовать модели машинного обучения в качестве источника данных для когнитивно-поведенческого моделирования.

Ключевые слова: аффект, семантика, пространство, Осгуд, смысл, язык, word2vec, чёрный ящик, объяснимость, интерпретация

Литература

1. Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations. Proceedings of NAACL-HLT. 2013. pp. 746–751.
2. Pennington J., Socher R., Manning C.D. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing. 2014. pp. 1532–1543.
3. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018.
4. Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.V. XLNet: Generalized autoregressive pretraining for language understanding. Proceedings of 33rd Conference on Neural Information Processing Systems. 2019.
5. Mikolov T., Joulin A., Baroni M. A Roadmap Towards Machine Intelligence. *Computational Linguistics and Intelligent Text Processing*. Cham: Springer, 2018. pp. 29–61.

6. Al-Saqqa S., Awajan A. The Use of Word2vec Model in Sentiment Analysis: A Survey. ACM International Conference Proceeding Series. 2019. pp. 39–43.
7. Dhar A., Mukherjee H., Dash N.S., Roy K. Text categorization: past and present. *Artificial Intelligence Review.* 2021. vol. 54. no. 4. pp. 3007–3054.
8. Konstantinov A., Moshkin V., Yarushkina N. Approach to the Use of Language Models BERT and Word2vec in Sentiment Analysis of Social Network Texts. *Recent Research in Control Engineering and Decision Making.* Cham: Springer, 2021. pp. 462–473.
9. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., Yang G.Z. XAI—Explainable artificial intelligence. *Science Robotics*. 2019. vol. 4. no. 37.
10. Suvorova A. Interpretable Machine Learning in Social Sciences: Use Cases and Limitations. *Proceedings of Digital Transformation and Global Society 2021. Communications in Computer and Information Science, vol. 1503*. Cham: Springer, 2022. pp. 319–331.
11. Osgood C.E. The nature and measurement of meaning. *Psychological Bulletin*. 1952. vol. 49. no. 3. pp. 197–237.
12. Osgood C.E. Studies on the generality of affective meaning systems. *American Psychologist*. 1962. vol.17. no.1. pp. 10–28.
13. Hollis G., Westbury C. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin and Review*. 2016. vol. 23. no. 6. pp. 1744–1756.
14. Lenci A. Distributional Models of Word Meaning. *Annual Review of Linguistics*. 2018. vol. 4. no. 1. pp. 151–171.
15. Günther F., Rinaldi L., Marelli M. Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science*. 2019. vol. 14. no. 6. pp. 1006–1033.
16. Samsonovich A.V., Ascoli G.A. Principal Semantic Components of Language and the Measurement of Meaning. *PLoS ONE*. 2010. vol. 5. no. 6.
17. Eidlin A.A., Eidlina M.A., Samsonovich A.V. Analyzing weak semantic map of word senses. *Procedia Computer Science*. 2018. vol. 123. pp. 140–148.
18. Samsonovich A.V. On semantic map as a key component in socially-emotional BICA. *Biologically Inspired Cognitive Architectures*. 2018. vol. 23. pp. 1–6.
19. Pretrained word2vec model “GoogleNews-vectors-negative300.bin.gz”. Google Code Archive. https://code.google.com/archive/p/word2vec/. 2013.
20. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 2013.
21. Surov I.A. Natural Code of Subjective Experience. *Biosemiotics*. 2022. vol. 15. no. 1. pp. 109–139.
22. Siegel M. The sense-think-act paradigm revisited. Proceedings of the 1st International Workshop on Robotic Sensing. 2003.
23. Gastaldi J.L. Why Can Computers Understand Natural Language? *Philosophy & Technology*. 2021. vol. 34. no. 1. pp. 149–214.
24. Jensen A.R. The relationship between learning and intelligence. *Learning and Individual Differences*. 1989. vol. 1. no. 1. pp. 37–62.
25. Sowa J.F. The Cognitive Cycle. Proceedings of the 2015 Federated Conference on Computer Science and Information Systems. 2015. vol. 5, pp. 11–16.
26. Wang Y., Yao Q., Kwok J.T., Ni L.M.: Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*. 2021. vol. 53. no. 3. pp. 1–34.

27. Hoenkamp E. Why Information Retrieval Needs Cognitive Science: A Call to Arms. 2005.
28. Turney P.D., Pantel P. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*. 2010. vol. 37. pp. 141–188.
29. Wang B., Buccio E.D., Melucci M. Representing Words in Vector Space and Beyond. *Quantum-Like Models for Information Retrieval and Decision-Making* (eds: Aerts D., Khrennikov A., Melucci M., Toni B.). Cham, Springer. pp. 83–113. 2019.
30. beim Graben P., Huber M., Meyer W., Römer R., Wolff M. Vector Symbolic Architectures for Context-Free Grammars. *Cognitive Computation*. 2022. vol. 14. no. 2. pp. 733–748.
31. Coenen A., Reif E., Kim A.Y.B., Pearce A., Viégas F., Wattenberg M. Visualizing and measuring the geometry of BERT. Proceedings of the Advances in Neural Information Processing Systems. 2019.
32. Tanaka Y., Oyama T., Osgood C.E. A cross-culture and cross-concept study of the generality of semantic spaces. *Journal of Verbal Learning and Verbal Behavior*. 1963. vol. 2. no. 5-6. pp. 392–405.
33. Tanaka Y., Osgood C.E. Cross-culture, cross-concept, and cross-subject generality of affective meaning systems. *Journal of Personality and Social Psychology*. 1965. vol. 2. no. 2. pp. 143–153.
34. Osgood C.E., May W.H., Miron M.S. *Cross-cultural universals of affective meaning*. Champaign, University of Illinois Press. 1975.
35. Zajonc R.B. Feeling and thinking: Preferences need no inferences. *American Psychologist*. 1980. vol. 35. no. 2. pp. 151–175.
36. Duncan S., Barrett L.F. Affect is a form of cognition: A neurobiological analysis. *Cognition and Emotion*. 2007. vol. 21, no. 6. pp. 1184–1211.
37. Lipton Z.C. The Mythos of Model Interpretability. *Queue*. 2018. vol. 3. pp. 31–57.
38. Molnar C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2019.
39. Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. 2019. vol. 51. no. 5.
40. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019. vol. 1. no. 5. pp. 206–215.
41. Vassiliades A., Bassiliades N., Patkos T. Argumentation and explainable artificial intelligence: A survey // *Knowledge Engineering Review*. 2021. vol. 36, pp. 1-35.
42. Borrego-Díaz J., Galán-Páez J. Explainable Artificial Intelligence in Data Science. *Minds and Machines*. 2022.
43. Chou Y.L., Moreira C., Bruza P., Ouyang C., Jorge J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*. 2022. vol. 81. pp. 59–83.
44. Tian L., Oviatt S., Muszunski M., Chamberlain B.C., Healey J., Sano, A. Applied Affective Computing. ACM Books. 2022.
45. Michelucci P. (ed.) Handbook of Human Computation. New York, Springer. 2013.
46. Samsonovich A.V. (ed.) Biologically Inspired Cognitive Architectures. Advances in Intelligent Systems and Computing vol. 948. Cham, Springer. 2020.
47. Adadi A., Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018. vol. 6. no. 52. pp. 138–152.
48. Dennett D.C. The intentional stance. Cambridge, MIT Press. 1998.

49. Caporael L.R. Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*. 1986. vol. 2. no. 3. pp. 215–234.
50. Guthrie S.E. Anthropomorphism: A definition and a theory. *Anthropomorphism, anecdotes, and animals*. (eds. Mitchell R.W., Thomson N.S., Miles H.L.), chap. 5, pp. 50-58. State University of New York Press, New York. 1997.
51. Watson D. The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines*. 2019. vol. 29, no. 3. pp. 417–440.
52. Salles A., Evers K., Farisco M. Anthropomorphism in AI. *AJOB Neurosci*. 2020. vol. 11. no. 2. pp. 88–95.
53. Maclure J. AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Minds and Machines*. 2021.
54. Arnulf J.K., Larsen K.R., Martinsen Ø.L., Bong C.H. Predicting survey responses: How and why semantics shape survey statistics on Organizational Behaviour. *PLoS ONE*. 2014. vol. 9. no. 9.
55. Jones M.N., Gruenenfelder T.M., Recchia G. In defense of spatial models of semantic representation. *New Ideas in Psychology*. 2018. vol. 50. pp. 54–60.
56. Arnulf J.K. Wittgenstein's Revenge: How Semantic Algorithms Can Help Survey Research Escape Smedslund's Labyrinth. *Respect for Thought* (eds. Lindstad T.G., Stänicke E., Valsiner J.), chap. 17, pp. 285–307. Springer, Cham. 2020.

Суров Илья Алексеевич — канд. физ.-мат. наук, доцент, старший научный сотрудник, Университет ИТМО. Область научных интересов: когнитивно-поведенческое моделирование, квантовая семиотика и семантика. Число научных публикаций — 25. surov.i.a@yandex.ru; Кронверкский проспект, 49А, 197101, Санкт-Петербург, Россия; р.т.: +7(812)480-0000.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 20-71-00136).