AI & SOCIETY
https://doi.org/10.1007/s00146-021-01167-3

OPEN FORUM

Word vector embeddings hold social ontological relations capable of reflecting meaningful fairness assessments

Ahmed Izzidien[1]

Received: 12 October 2020 / Accepted: 18 February 2021

Abstract

Programming artificial intelligence (AI) to make fairness assessments of texts through top-down rules, bottom-up training, or hybrid approaches, has presented the challenge of defining cross-cultural fairness. In this paper a simple method is presented which uses vectors to discover if a verb is unfair (e.g., slur, insult) or fair (e.g., thank, appreciate). It uses already existing relational social ontologies inherent in Word Embeddings and thus requires no training. The plausibility of the approach rests on two premises. That individuals consider fair acts those that they would be willing to accept if done to themselves. Secondly, that such a construal is ontologically reflected in Word Embeddings, by virtue of their ability to reflect the dimensions of such a perception. These dimensions being: responsibility vs. irresponsibility, gain vs. loss, reward vs. sanction, joy vs. pain, all as a single vector (FairVec). The paper finds it possible to quantify and qualify a verb as fair or unfair by calculating the cosine similarity of the said verb's embedding vector against FairVec—which represents the above dimensions. We apply this to Glove and Word2Vec embeddings. Testing on a list of verbs produces an F1 score of 95.7, which is improved to 97.0. Lastly, a demonstration of the method's applicability to sentence measurement is carried out.

Keywords Fairness · Word Meaning · Morality · Legal Philosophy · Responsibility · Policy

1 Introduction

A common feature of human-centric artificial intelligence design is the necessity of using humans to assess, where fundamental rights and responsibilities lie in a situation. From which, rules are introduced into the AI to mitigate any potential harm (Bauer 2020a). We argue that this bottle-necks AI, and forgoes the power afforded by the technology. We put forward the suggestion that the AI itself ought to have the capacity to perceive the action space-state and make rights and responsibilities allocations. Such a perception would allow an AI to draw on a wealth of information, to include precedent, and prior outcomes to solve multi-factorial ethical conundrums in real world settings.

This contrasts with top-down rule-based systems which, to a degree, replicate the modus operandi of non-AI computing (Cervantes et al. 2020). On the other hand, bottom-up programming uses machine learning (ML) algorithms to learn from patterns in a prepared set of data to infer the next move (Bauer 2020a). Such a methodology considers normative values as being inherent in the activity of the agents but not explicitly defined in terms of a general theory (Wallach et al. 2008). This paper's approach can be thought of as bottom-up but uses a universal fairness rule that is inherent within Word Embeddings, as will be expanded on.

For an AI system to be able to perceive engaged contexts to assess whether the description of an act, or instruction, is fair, a fairness metric by which it can measure such activity is required. Currently, metrics to assess human qualities such as sentiment and personality have been well validated in the literature (Boyd et al., 2015; Hai-Jew 2017; Youyou et al. 2015). However, a valid and reliable measure of fairness has yet to be developed.

Our work in this paper will focus on delivering the first step in the development of such a measure, one that focuses on interpreting human readable texts and assessing the fairness of the social power interactions described therein. As documents can be broken into their constituent paragraphs, sentences and words, this paper will concentrate on analysing singular words, specifically, verbs.

✉ Ahmed Izzidien
ai297@cam.ac.uk

[1] The Psychometrics Centre, Cambridge Judge Business School, The University of Cambridge, Trumpington Street, Cambridge CB2 1AG, US

Published online: 24 March 2021

A certain limitation exists in using singular words, being devoid of context. The sentence ‘The man killed the taxi driver’ vs. ‘The man killed the weeds in his garden’ carries both qualities of being unfair and fair, respectively, for the same word ‘killed’. The same can be said of homonyms. However, sentences such as: ‘The boy thanked the teacher for his help’ is easily classifiable as a fair compared to ‘The boy used a slur against the teacher’. Here the two verbs: ‘thank’ vs. ‘slur’ are typically considered as fair and unfair acts, respectively, even devoid of context. We accept this limitation for this stage of the research. We will be testing the verb list used by (Jentzsch et al. 2019) who incorporate a list of ‘Do’ and ‘Don’t’ verbs into their pipeline as training data. However, our methodology differs from their own, as we do not use any training data, but rely on inherent social ontologies. Our methodology will be covered following an introduction to the background used in the design of the measure, which focuses on the social anatomy of the human mind and social discourse.

1.1 Principles behind the fairness measure

The human mind is able to build rich causal models, perform generalizations, and assemble powerful abstractions despite sparce and incomplete input (Tenenbaum et al. 2011). Modeling how the mind uses abstract knowledge to guide inferences has been attempted with Bayesian statistics. Abstract knowledge is seen as being encoded in a probabilistic generative model. One that describes the causal processes of the world in a way that facilitates the analysis of perceived spaces and their latent variables. Causal learning data can be gained from co-occurrences between events, whereby causal relations are hypothesized. Likelihoods favor causal links that make such co-occurrence more probable, whereas priors favor links that fit background event knowledge of likely causes (Tenenbaum et al. 2011).

It has been proposed that such abstract knowledge provides essential constraints for learning. Developmentalists posit that humans innately hold a set of principal abstract concepts such as “agent”, “object” and “cause” to provide a fundamental ontology for qualifying experience (Carey, 2011a, b). Indeed, there is a growing trend in the literature for multiple representation views, whereby abstract concepts are grounded in an array of inputs: linguistic, emotional, sensorimotor, internal experiences and social (Andrews et al. 2014; Borghi et al. 2018). It has been suggested that the divergence between abstract and material concepts may be best modeled in terms of multidimensional space, in which concepts varying both in their level of abstraction and along other content dimensions are distributed (Borghi et al. 2018).

This form of representation of the abstract, in multidimensional space, one that incorporates probability learning and co-occurrence statistics is reminiscent of the ontological features of a form of neural network computation known as Word Embeddings. These embeddings are able to capture rich features of human language, language that inherently reflects society and its values (Boyd and Richerson 2009; Smith 2010; Drozd et al. 2016).

1.2 The social mind, human language, and Word Embeddings

Word Embeddings use a process known as co-occurrence probability to represents words. As such, these words are no longer represented by their dictionary definitions, but by their relations to other words. The approach uses word context to represent meaning. Oft captured by the saying ‘you shall know a word by the company it keeps!’ (Firth 1958; Nerbonne and Hinrichs 2006).

Vectors are used to capture how frequent each word occurs in a particular context. Each vector consists of a list of numbers, whereby each number reflects a probability. As a list of co-occurrences is built up, probability patterns begin to emerge. Thus, terms such as ‘dog’ and ‘cat’ would be seen to have a higher probability of co-occurring with each other than words that do not occur together as often, such as ‘dog’ and ‘pipe’. As the list of vectors grows, more useful information on word meanings form. For example, the word ‘ice’ would be found to co-occur more frequently with the word ‘solid’ than the word ‘gas’. Whereas the word ‘steam’ co-occurs more frequently with ‘gas’ than the word ‘solid’. Of note is that both words co-occur frequently with water, as it is their shared property while infrequently with unrelated words (Pennington et al. 2014).

These vectors can then be represented in multi-dimensional space. Each word in the document is given a set of coordinates that represents its location in a geometric space in respect to every other word. The setting of these words is based on their context. Those sharing many contexts are found to be situated next to each other, compared to words which have different contexts (Kozlowski et al. 2019). Thus, words such as ‘pain’ and ‘pleasure’ may be found to be distant to each other while being closer to ‘abuse’ and ‘love’, respectively.

An advantage in using vector notations lies in their arithmetic properties. Two vectors can be compared, added and scaled, allowing for a number of calculations to be made. A highly cited example is that of manipulating a vector which represents the word ‘King’. In subtracting the vector for ‘man’ from it, then adding the vector for ‘woman’ from it, the result is the word ‘Queen’. This happens, because the representation of ‘King’ contains a representation of ‘man’ due to co-occurrence. When this quality is removed using a subtraction, the word is no longer closely associated with ‘King’ yet remains closely associated with royalty. As such, replacing ‘man’ with ‘woman’ allows for a new vector to be closely matched to a word that represents royalty and women, i.e., a 'Queen' (Chen et al. 2017; Drozd et al. 2016).

It is also possible to consider how similar or dissimilar two vectors are by measuring their cosine similarity. From trigonometry, Cos (0) = 1, Cos (90) = 0, and 0 < = Cos (θ) < = 1. Vectors maximally similar are parallel (i.e., at 0 degrees to each other) and minimally similar if they are perpendicular (i.e., at 90 degrees to each other). This feature allows for a straightforward comparison of words. Singular words such as 'slur' and 'irresponsible' may be compared using this method, for example, with the expectancy that similar words will hold a higher cosine score than dissimilar words. The power and flexibility offered by this method has seen it reinforce much of the work done in natural language processing (NLP) (Almeida and Xexéo 2019; El-Amir 2020).

Such a conception of semantics has been described as the distributional hypothesis (Clark and Pulman 2007). This approach represents, in part, how the mind operates through parallel processes and weighted connections (Mikolov et al. 2013).

1.3 An epistemology of Word Embeddings

One of the discoveries made with Word Embeddings is their ability to validly reflect meaningful patterns from the data they have learnt. Capturing the statistics inherent in language using this method and projecting it into multidimensional space has allowed for subtle relations to be reflected in arithmetic terms. For example, when Word Embeddings are derived from documents that describe the sociology of a country over several decades, dimensions induced by word differences such as (rich – poor) are found to correspond to dimensions of cultural meaning. A projection of words onto these dimensions has been shown to reveal widely shared associations that are validated with survey data (Kozlowski et al. 2019). This ability of Word Embedding to concurrently locate objects on multiple cultural dimensions, to include classes such as race, gender, socio-economic class, has been found to make them a powerful tool for research on intersectionality, for example (Kozlowski et al. 2019).

This finding is not specific to the social sciences. The natural sciences have also gained from their use. For example, materials science knowledge present in published literature was encoded using Word Embeddings without any explicit addition of chemical knowledge. The embeddings were found to capture intricate materials science concepts such as the underlying structure of the periodic table and structure–property relationships in materials. Utilizing this implicit information held in the vector space, researchers proposed materials for functional applications several years before their discovery. Suggesting that latent knowledge regarding future discoveries is to an extent embedded in past academic papers (Tshitoyan et al. 2019). Embeddings have also been successful at capturing latent concepts such as ideology, providing an integrated framework for an indirect study of political language (Rheault and Cochrane 2020).

Yet Word Embeddings have their limitations. One of which is that they can reflect the biases contained within the texts they represent. Words such as 'doctor' and 'engineer' have been found to co-occur more often with 'man' than 'woman' in contemporary writing and reporting. Thus, a vector space constructed using such documents will also represent such a bias (Caliskan et al. 2017; Garg et al. 2018). These biases have been seen as a hindrance to the effectiveness of using embeddings for social interaction applications, such as their use in candidate selection (Köchling and Wehner 2020). However, other biases inherent in Word Embeddings can in some instances be useful for extracting an underlying concept that has caused such a bias to manifest. In this paper we will demonstrate the existence of a fairness bias within Word Embeddings and leverage it to our advantage to design a fairness metric.

1.4 The fairness bias, a pro-social propensity

Just as Word Embeddings have been found to contain gender and ethnic biases (Caliskan et al. 2017; Brunet et al. 2019), we put forward the case that humans are biased against conducting acts which provide them with no sense of gain. That is, humans are instinctively averse to gainless activity. That in being a social species, humans are biased to favour social acts. Acts that provide a sense of gain and joy as opposed to harm and pain to themselves. We instinctively class acts that we would be happy to have done to ourselves as positive, and acts which we would not wish to happen to ourselves as negative. Such a bias, we posit is universal in humans. To expand on this bias, as it is a central point in this paper, we consider the social psychology and moral psychology literature on this topic.

1.5 An ontology of fairness

Despite impulses for survival, acts of cooperation have been seen as central to human behavior (Trivers 1971; Milinski et al. 2002), generating senses that facilitate cooperation (Nowak 2006). One of the prime senses when it comes to deciding about an act towards another, is a realisation of how the other person will react to the said act (Civai 2013). Individuals are evolutionarily deterred from acting in a harmful manner, avoiding possible sanction. They are concomitantly evolutionarily encouraged towards cooperation, gaining possible benefits and reward, direct, or indirect. This sense of calculation that carries with it considerations towards group accountability, be it thorough reward or sanction, has been seen as one that facilitates cooperation and social bonds (Fehr et al. 2002; Fehr and Rockenbach 2004).

This evolved sense of cooperative behavior has the effect of generating a sense of an ought in the person. We argue that a sense of ought has the same connotations of a responsibility: Feeling deterred generates an inherent sense of responsibility not to harm the other, as well concomitantly assigning the other an inherent right not to be harmed (van Dijk and Vermunt 2000). While these cannot be said to be generated as explicit social values, the senses have the same consequential qualities. For despite the evolutionary origins of the sense of being deterred from and encouraged to act in a manner that aids social survival, the outcome is inherently frameable as one that generates these meta-qualities of rights and responsibilities. Meta-qualities that are produced as corollaries of an evolved sense of cooperative behavior, of feeling one ought to, or ought not to. Responsibility becoming guided by a sense of concern (Berkowitz and Daniels 1963; Cremer and Lange 2001).

These cognitions, can be frameable as the perceptions that form the basis for the golden rule (George Duke and George 2017, p. 44), since to be able to assess if an act is one that 'I would wish for myself', I have to perceive the context in terms of qualities which suggests a course of action. One that I would wish for myself, even when acting socially will not, or cannot, be reciprocated (van Dijk and Vermunt 2000).

Even a Machiavellian, seeing harming others as justified, would not wish to be on the receiving end of their acts. An inherent cross-cultural aversion to treating others as one would wish not to be treated remains, even if they proceed to act it out. This feeling contrasts with organisms that do not process the capacity for such senses, such as viruses and bacteria, for example. Such an aversion to inequity has been characteristic of species that cooperate regularly even with non-kin (Brosnan and Bshary 2016), and forms the basis of a social bias, that is, a bias to act socially.

Based on this, it would be a measure of a person's responsibility and their perception of the frame as one that warrants such qualification (Handgraaf et al. 2008) that would reflect the starting point for an ethical evaluation.

In each context, a measure of the perception of the frame allows a person to consider the relevant dimensions. When a context is evaluated as harmful to one actor, for example, such as murder, there will be a higher salience to it. Feelings have been found to be an integral part of the analysis by which individuals measure decisions in complex judgmental situations (Sadler-Smith 2012). Here context perception plays a qualifying role (Decety et al. 2012; Fessler and Haley 2003) and such salience can be thought of through emotions, negative and positive, such as that of pain and joy.

It may be objected that war and cruelty emanate from cognitions that point towards anti-sociality (Kahane 2016, p. 285). However, this objection may be countered by the observation that prosocial acts are desired by oneself, anti-social acts are not. Even a Machiavellian, as mentioned, seeing the usurpation of power as justified, would not wish the same for themselves. An aversion to such acts persists, characterizing humans as socially aware agents (Izzidien and Chennu 2018).

This sense of ought is not to be confused by any normative statement. The paper is not inferring a moral course of action due to the presence of such social cognitions. Rather, the paper argues that due to perceptions that aid in social survival, humans are socially biased towards being social. The elicitation of this sense in humans can be seen as one that inherently encourages acts of cooperation and who's continued survival incorporates cognitions of not just themselves, but of other agents (Simon 1990; Brewer 2004). Each individual is deterred from acting in a manner that would be detrimental to each's survival, while at the same time concomitantly promoting them towards cooperative behavior, encouraging prosocial action, supporting an ultra-cooperative lifestyle (Tomasello 2014).

It has been shown that the perception of others who depend on us for gaining needed benefits evokes such feelings of responsibility, incentivizing us to help further their interests (van Dijk and Vermunt 2000). With an interdependency of relations for survival, individuals can be found to have a propensity – or positive social bias—to come to the aid of other individuals the more dependent these others are (Berkowitz and Daniels 1963; Berkowitz 1972; Schwartz and Howard 1982). With such calculations having repercussions on survival, some have held that social behavior has biological roots (Hewstone et al. 2012, p. 184) and in shared neurological processes such as theory of mind, a comparison heuristic and empathy (Tabibnia et al. 2008; Civai 2013; Corradi-Dell'Acqua et al. 2013).

Furthermore, studies find that correlations between actual behavior and expectations leads itself to qualify expectations as a significant factor in cooperative behavior or generous acts (Brañas-Garza et al. 2017) and have been associated with herding behavior, affecting a development of social norms (Brunnermeier 2001; Castelfranchi et al. 2003; Bicchieri 2006).

As such, we posit that when humans perceive a social context that demands a fairness assessment, they instinctively generate a sense of an ought. One that can be construed as a sense of responsibility. This is coupled, or tempered, by the measure of the salience of the act and its effect: harm/benefit, pain/joy, and its outcome: sanction/reward.

Thus, to mark an act as fair or unfair, it appears that an AI ought to consider these primary cognitions. These may allow an AI to begin to make human like assessments that incorporate the relevant dimensions needed. Perceptions that are arguably required to make fairness assessments.

1.6 Using Word Embeddings to extract the human pro-social bias

We posit that based on this human propensity – or social bias—to survive as a social species (Burkart et al. 2014; Peysakhovich et al. 2014) human language presents a medium by which such a bias is reflected (Boyd and Richerson 2009; Smith 2010). Furthermore, just as social acts are relations between agents and patients, we put forward the case that one manner in which this characterization can also be captured is through Word Embeddings. This is because in such embeddings, given the human social bias to be social, certain acts will be more closely associated to concepts of responsibility than irresponsibility. Acts that are imbued with a sense of responsibility, that is, a duty towards others, will also be associated with positive emotional, material, and social-outcome dimensions. These dimensions will be shown to be the prime perceptions needed to construe a context prior to making a fairness assessment.

One of the challenges of Machine Learning (ML) and Deep Learning (DL) in detecting patterns in data for classification is the need to correctly identify which properties to use. This can be straightforward when the data is easily characterizable using clear markers, such as colour or shape. However, when the data is highly dimensioned – in an abstract sense – identifying the appropriate dimensions presents a challenge. Language is no exception, with a sentence holding many possible dimensions: emotional, moral, power relations and aesthetic, to name a few. Thus, to elicit the appropriate dimensions for a universally acceptable fairness classification it becomes necessary to address this point.

As a starting point this paper considers the aforementioned primary perceptions that are typically elicited in humans when confronted with a situation in which they must make an ethical qualification: To do, or not to do.

To separate these out, we propose using an established technique, vector addition, subtraction and comparison.

1.7 Developing a fairness vector to assess words

While it may be possible to use the process of labelling to mark each sentence under investigation in terms of these abstractions—along with their causal properties, e.g., 'The boy kicked the baby':

(**Boy**): Agent, Irresponsible. (**Baby**): Patient, Pain, Loss. (**Kicked**) Causal-relation, Unfair. Then train a ML algorithm based on such abstractions, it is suggested by this paper that such a step in unnecessary.

This is based in the assumption that the process of word co-occurrence inherently captures these relational properties. For example: An agent acts on a patient (e.g., 'The boy kicked the ball, and it went far'), the causal outcome is contained ('it went far'). Yet, an alternative sentence, such as ('The ball was green, and it was large), one which has no agent acting on the patient, results in frame in which there is no outcome. The first sentence inherently holds the abstractions: agent – patient – outcome. Whereas the second does not. This dimension, if detected by a ML algorithm implicitly allows it to learn the concept of causality: A causal outcome is only found in texts in which there is a power interaction, that is, with two or more actors.

In the paper we consider that this information is inherent in Word Embeddings, even though such sentences are not labelled with such abstractions. Furthermore, as power interactions have their qualifications, that is, they are describable as either acts that one would wish for themselves, or not, i.e., fair or unfair, it can be argued that when embedding very large text documents, this fairness qualification will also present. Since words like 'slur', for example, are more likely to co-occur with words relating to sanction, irresponsibility and pain, than to responsibility, reward, and joy. Reflecting the aforementioned social propensity, a positive social bias in society, as previously detailed.

The Word Embedding of such a corpus would allow for each word vector to be partly representative of how it relates to the social ontological abstractions of all other words. As the corpus grows, the reflection of the human social condition, becomes more persuasive – unless the corpus is one of science fiction reflecting alternative realities, for example. As a vectorised corpus is characterizable based on Euclidean distances. Words can then be measured as to their closeness or distance to others.

The paper hypothesises that in making a single vector which captures the required dimensions of fairness, it will become possible to measure how similar such a vector is to any word act in the corpus, without the need of any training data.

Verbs reflect acts, typically between two or more agents. They are also ethically qualifiable: would I wish this 'verb' for myself? Whereby a fair act is one that I would, and an unfair one that I would not. Verbs also have certain grammatical expectations associated with them, such as an association with abstract units such as objects or complement clauses (Fortescue 2017). Thus, they inherently offer themselves up as contenders for agent-act-outcome-assessment co-occurrences.

To test this hypothesis, the paper presents the construal of what a Fairness Vector consists of. This is completed through adopting the terms that describe the abstract dimensions listed above from the social psychology literature. The dimensions that humans typically engage when making a fairness assessment. A test of the validity of using this vector to differentiate between fair and unfair acts is conducted. To do so a cosine similarity is calculated for the Fairness Vector against a collection of verbs. Where each verb is qualifiable as fair or unfair according to the golden rule. The verb list presented by a paper on this theme by (Jentzsch et al. 2019) was used. However, instead of using training data as they do, our paper presents a method to qualify acts with the power afforded by Word Embeddings using the appropriate psychological dimensions to elicit a fairness judgment.

Prior to the methods section, we present next a collection of hypothetical scenarios to describe how the fairness rule manifest itself in a manner that attracts universal appeal.

1.8 Scenario 1

Tom sees Jeff walking by. Tom has an urge to punch him, but he asks himself 'would I wish to be punched?' As he answers himself in the negative, he decides to desist. In turn not acting in an unfair manner towards Jeff.

1.9 Scenario 2

Tom does not mind people calling him 'four-eyed' for wearing glasses. In fact, he finds it amusing. One day he sees Jeff, also a wearer of glasses. Tom feels like calling Jeff 'four-eyed'. In the first instance, it appears that the fairness consideration 'would I wish the same on myself' will not help Tom to be fair. Yet, thinking it over, Tom concludes that the reason he does not mind people calling him 'four-eyed' is because he finds it amusing. Jeff, however, would not find it amusing, in fact he is sure that Jeff would find it insulting. Since Tom would wish that others do not insult him, and that calling Jeff 'four-eyed' would not amuse Jeff, rather, it would be insult Jeff, Tom thus uses the fairness consideration to treat him as he would wish to be treated, i.e., not to insult him, rather, to say something that would amuse him.

1.10 Scenario 3

Tom is travelling in a part of the world, where hosts welcome their guests with a large hot meal. Jeff is also a guest, but in another region of the world, one that welcomes guests with only a cup of tea. Two cultures, each valuing hospitality differently. Yet, despite the cultural differences, the fairness rule can also be applied: In the first culture it would be unfair to offer all but one guest, a meal, and to that singled-out guest, only a cup of tea. This is because no one wants to be given less than what they are due, in either culture. A host in one part of the world would wish to be offered a hot meal had they been the guest, whereas a host in another part of the world would feel no pain or indignation if they were not served more than a cup of tea. Each would consider fair what they would wish for themselves in their respective context.

1.11 Scenario 4

What if Jeff was about to get a ticket for speeding? Tom, an officer of the law, may not wish to get a ticket himself. Would his issuance of a ticket mean he is being unfair?

To unpack this, we can consider the following. If Jeff lived on a busy street, he would not wish his children or himself to be harmed by speeding cars. Thus, he supports a means to stop cars speeding. Let us say, through the use of speeding tickets.

If Jeff is then caught speeding, then to be consistent he will have to accept that being punished for speeding is a fair act, even if he gets annoyed. This can be considered a case in which the perpetuator admits that they 'deserve the punishment'. They may not enjoy it, or indeed emotionally wish it, but they believe it justifiable. However, if the punishment involved decapitation, for example, then Jeff would object, since Jeff would not wish the same on himself.

A basis for all these is the common factor that humans are typically harm averse. They recognize this in themselves and in others. Thus, humans recognize that all people typically do not want to be injured, irrespective of their culture. This characteristic gives strength to using the qualification of 'not treating others as one would not wish to be treated' as a basis for the fairness vector.

The use of the terms responsibility and irresponsibility to describe this heuristic is somewhat limited, in that the full question as given by its sentence form 'would I wish this act for myself' or similarly 'for my loved ones' is not fully captured. With this paper being focused on singular words, we consider using such sentences in our discussion on further work.

As such, and for this paper, we have selected the GloVe algorithm (Pennington et al. 2014) to make our embeddings due to its focus being on singular words. After preprocessing, the algorithm constructs a co-occurrence matrix which encodes the probability of two words appearing in the same context. It then employs various strategies (e.g., matrix factorization) to produce an embedding that preserves co-occurrence information (Liu et al. 2019).

1.12 Building a fairness vector

To use GloVe embeddings to make an assessment on singular words, it will be necessary to develop a method by which words, such as 'murder', 'theft' and 'help' are categorizable. This paper thus makes its contribution to the literature by suggesting that:

i Words in Glove embeddings (Pennington et al., 2014) carry social relations that are extractable.
ii By virtue of being a social species, these social relations are reflective of a propensity to be social.

iii Using vectors, it is possible to use this propensity as a classifier, through cosine similarity comparisons between a test word (e.g., 'murder') and a Fairness Vector.

iv A Fairness Vector is constructable when it is based on the appropriate social dimensions that are typically elicited when making a fairness evaluation.

2 Method

We use the Glove (Pennington et al. 2014) Common Crawl 840B tokens, 2.2 M vocab, cased, 300 dimensioned vectors as our corpus. Then using the abstract terms identified earlier, we perform a vector addition and subtraction to reflect the range of these concepts going from positive to negative. The process of adding and subtracting vectors allows one to consider a range of a dimension, with words closer to one scale reflecting a similarity more strongly than those on the opposite scale.

These will form our 'litmus' vector (FairVec) which will be used to test verbs, then to test singular words added to form sentences, as will be described. The test uses a cosine similarity between the verbs and the litmus vector. A cosine similarity is able to measure the similarity of two vectors. It does so using the dot product of each vector divided by the product of the two vectors' lengths. Its value ranges from -1 to 1, with -1 reflecting perfectly dissimilar vectors and 1 perfectly similar. Thus, a vector representing fairness, ought to be closer to a vector representing words such as 'thank' and 'appreciate' than to words such as 'slur' and 'insult'.

To build this 'litmus' vector, a combination that represents the ranges under examination is used though addition and subtraction:

$$\text{Fairness Vector (FairVec)} = \text{Vectors for [responsibility} - \text{irresponsibility} + \text{joy} - \text{pain} + \text{beneficial} - \text{harmful} + \text{reward} - \text{sanction].}$$

The use of addition and subtraction facilitates the range to be compared against. One change is made, however, for the Outcome dimension. Here the terms 'liberty and prison' were used based on the assumption that these are more commonly used in everyday language as descriptors of accountability compared to the more legal terms of 'reward and sanction'. As well as due to the double meaning afforded by the term 'sanction'.

It is also possible to add a list of similar words in place of a single word, for example: one could add to the word responsibility a list of similar words: [responsibility + responsible + duty + dependability + dutiful] to be part of the Fairness Vector. We have avoided this in this instance. As a concept, in this case fairness, can be described using its orthogonal dimensions through addition and subtraction. Such arithmetic narrows the possibilities of evoking other unrelated concepts (e.g., loyalty). The current addition and subtraction of FairVec may be considered as a form of narrowing of the intersection point of a Venn Diagram. Effectively narrowing the possible valid choices that lie close to the concept in the vector space.

To test how using only some of the dimensions of FairVec at the behest of others will affect the outcome of the Fairness Vector, an iteration of using only two, then four, then six dimensions of FairVec is made.

To consider how changing one word in FairVec, swapping it out for another, affects the outcome, we have taken the term 'joy' and replaced it with 'joyful' in one iteration. Similarly, we have swapped out 'responsibility' for 'accountable', then 'dutiful'.

For the initial test verbs, the shortlist of Do and Don't verbs presented by (Jentzsch et al. 2019) was used, removing words that were not present in the GloVe corpus, with the addition of 'rape' in the Don't verbs category for comparison with 'murder'.

To test whether the Fairness Vector is simply providing a false positive, a plot is made of the cosine similarity for each verb vector against only one dimension of FairVec, e.g., beneficial– harmful. This is repeated for each dimension of FairVec independent of the other dimensions.

A test of using the terms fair-unfair instead of using any of the Fairness Vector dimensions is also plot.

A further longer list of 200 verbs provided by (Jentzsch et al. 2019) is used with the eight dimensioned Fairness Vector, for which 12 verbs are removed (as they are not included in the Glove corpus)—8 Do Verbs and 4 Don't Verbs. To further test the accuracy of the results, they are correlated with the Python NLTK Vader sentiment package (Hutto 2020) outcome when it is applied to the verbs.

In terms of tuning the Fairness Vector to improve results, we consider how a change in a dimensional term, such as in going from 'joy' to the adjective 'joyous', allows for a change to be fed-back to the machine learning system through an optimization routine. This test is expanded on in the discussion section in terms of developing a fine-tuning mechanism for the Fairness Vector.

We then use the Google News 300 Word2Vec corpus (Google Code Archive—Long-Term Storage for Google Code Project Hosting. 2020) to replace the Glove corpus. An evaluation of the original Fairness Vector against the same list of 188 verbs was performed. A correlational Vader sentiment score was also made. The use of Word2Vec with the Google News corpus was in order test the ecological validity of FairVec by implementing cosine similarities in alternative documents using an alternative method of embedding, which in this case is a measure of co-occurrence at a local context as opposed to a global context (Pennington et al. 2014).

Lastly, FairVec is tested against a list of sentences. Each sentence being represented by an orthogonal iteration of its meaning, done for comprehensiveness. The paper limits this to simple sentences, removing stop words, i.e., 'boy kick baby'. Then its opposite sense 'boy help baby'. The length, agents and patients of each sentence are adjusted in each iteration as given in the results section below. Although sentences can be encoded into vectors using a variety of methods (Cer et al. 2018; Reimers and Gurevych 2019), the approach of adding each's representative Word Embedding vector was used (White et al. 2015, 2019).

All of the code and data files, including high resolution figures, are available on Github: https://github.com/AhmedIzzidien/FairnessVector/blob/master/FairnessVector%20v4.ipynb.

3 Results

The verbs that remained in the shortlist after removing those not found in Glove (Pennington et al. 2014) are presented below.

Do Verbs: smile, sightsee, cheer, picnic, snuggles, hug, brunch, gift, serenade, welcome, appreciate, acclaim, enjoy, thank, celebrate, delight, glorious, pleasure.

Don't Verbs: damage, harm, slander, slur, rot, contaminate, brutalise, poison, murder, disarticulate, demonise, negative, sicken, disorganise, miscount, rape.

For the first step, the cosine similarity is found for each dimension of the Fairness Vector against each of the verbs listed, scoring each through a subtraction from the number 1 (e.g., Liberty-Prison) as given in the top left panel of Fig. 1a. Then followed by using another independent dimension (e.g., Joy-Pain) as given in the top right panel (Fig. 1b). Allowing for a consideration of how each dimension is inadequate in and of itself to give a typical fairness assessment, as will be discussed.

To consider how the dimensions affect the outcome of FairVec, we plot an example of an iteration in which FairVec is initially represented by only the two dimensions of:

Liberty – Prison (Fig. 2, top left panel).

Then by four: Liberty – Prison + Responsibility – Irresponsibility (Fig. 2, top right panel).

Then by six: Liberty – Prison + Responsibility – Irresponsibility + Beneficial – Harmful (Fig. 2, bottom left panel).

Then by eight: Liberty – Prison + Responsibility – Irresponsibility + Beneficial – Harmful + Joy – Pain (Fig. 2, bottom right panel).

What can be observed in Fig. 2, are incorrect classifications appearing in all panels that use less than the eight dimensions. These incorrect classifications decrease in number as more dimensions are added, dimensions which allow for a narrowing of options to capture the overlapping concept that is represented by the addition and subtraction of the meanings inherent in the vectors. As more dimensions are added, the classifications improve in number and quality. Especially noted is the tempering of the final results which use all the dimensions (Fig. 2, bottom right panel) in a manner that generally reflects typical fairness evaluations, though not absolute. A point we shall return to.

The addition of the final dimensions has correctly classed 'murder' as one of a lower score than that of 'slander', and 'contaminate'. This distinction could not be made without the final two vectorised dimensions of pain and joy, as seen in comparing both bottom left and bottom right panels (Fig. 2).

What is of note from these results, is that any single pair would not have been sufficient to allow for an accurate fairness assessment. Using a combination of a few of the dimensions, at the behest of others did not proportionally capture a measure of whether an act is one that one would wish for themselves. For example, 'damage' and 'disarticulate' (the separation of two bones at their joint) are miscalculated as fair based on the first two dimensions (Fig. 2, top-left panel), then 'damage' is miscalculated as fair based on four dimensions (Fig. 2, top-right panel). This is improved with the addition of a further two dimensions (Fig. 2, bottom-left panel), whereby both are classed as unfair, though their relative score is still arguably inconsistent, with the 'damage' being close to the cut-off axis point.

The increased accuracy in correctly classifying acts as fair or unfair, with the addition of dimensions, appears to tally with the literature, whereby humans are typically guided by emotional responses, material gain, as well as a consideration of the consequences to their actions, and not just by one of these. As such a measure that captures these cognitions allows for a greater accuracy.

To determine the outcome of the using the terms 'fair-unfair' instead of the eight above dimensions, the cosine similarity of the vector 'fair-unfair' against the list of verbs is plot (Fig. 3). A correlation of cosine similarity score for the verbs in Fig. 3 with the Vader Sentiment compound score is found to = 0.047.

Using the Fairness Vector with its eight dimensions (FairVec) is plot in Fig. 4. The correlation between the cosine similarity score and the Vader Sentiment compound score = 0.85.

To test FairVec against the full list of 188 verbs presented by (Jentzsch et al. 2019), a cosine similarity of the Fairness Vector with each is presented in Fig. 5. The results are correlated with the Vader Sentiment Intensity Analyzer and found to = 0.71.

The misclassed verbs are given in Table 1.

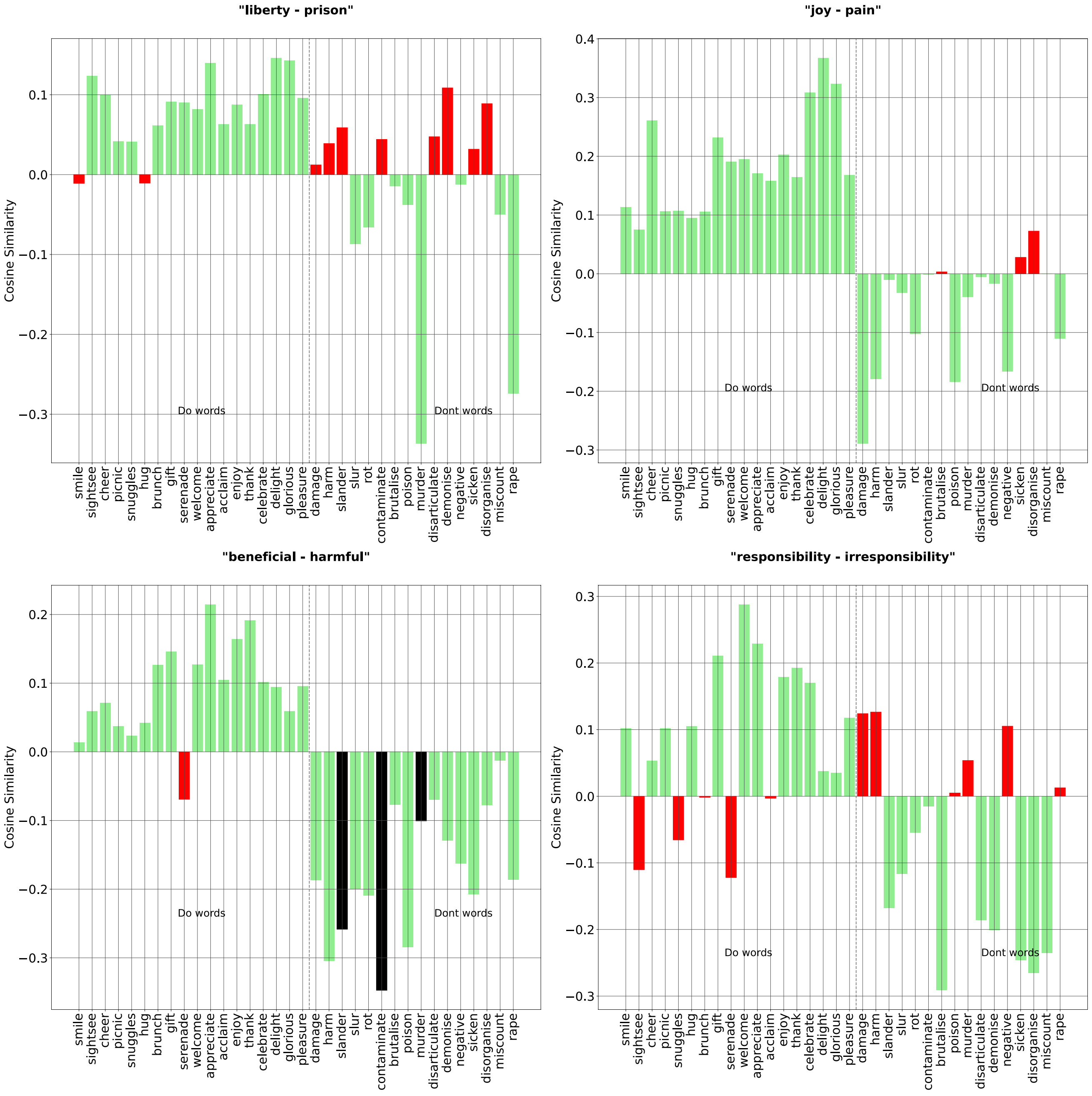

Fig. 1 Cosine similarity of verbs with four different word vector pairs. Green indicates correctly classed, red indicates incorrectly classed. All bars on the left of dotted line ought to be positive, while those on the right ought to be negative. Black is used to highlight the incongruence of relative scoring for 'Don't verbs' e.g., 'murder' is classed less than 'slander' and less than 'contaminate'

In Table 1, the misclassed verbs can be seen to be close to being correctly classed, with the defining line at y = 0. The mean of all of the Do Verbs = 0.1344 with a standard deviation of 0.1058. While the mean of the Don't Verbs = –0.1335 with a standard deviation of 0.0563.

The confusion matrix for Fig. 5 results is given in Table 2.

The top 15 and bottom 15 verbs are found to be as given in Table 3.

While it is arguably quite subjective to compare Do Verbs, the salience of Don't Verbs are more readily rankable based on common perceptions, for example, 'murder' is typically considered worse than 'steal'. It is at this stage a question is asked on how a slight adjustment to the Fairness Vector will play out. For this, the wording of one of the dimensions of the Fairness Vector 'joy' is altered to the adjective 'joyous', in the expectation that an adjective

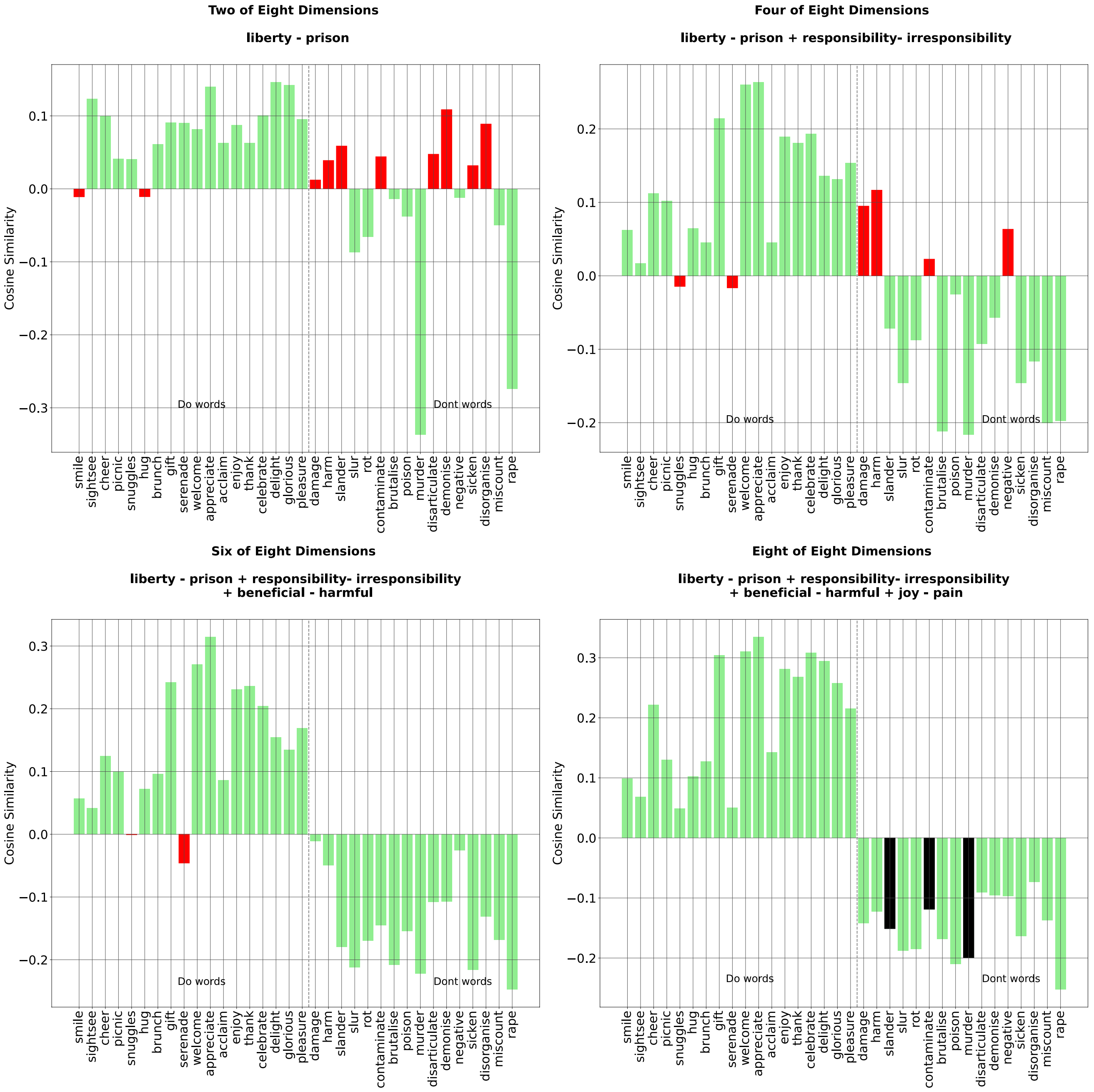

Fig. 2 Cosine similarity of verbs with increasingly dimensioned Fairness Vector (FairVec). Green indicates correctly identified classed; red indicates incorrectly classed verb. Black is used to highlight the congruence of relative scoring for 'don't verbs', whereby 'murder' is classed higher than 'slander', and 'contaminate', which is lower than both verbs

is more common when describing a noun or pronoun. In which case, the salience of the verbs may be better reflected.

Carrying this out, a more intuitive ranking of Don't verbs appears to present (Table 4). This finding is commented on in the discussion. The resultant overall accuracy for the ranking of the 188 verbs also improves with all Don't verbs classed correctly, and five Do verbs classed as Don't verbs (Table 5), producing an F1 score of: 0.97 and a Vader compound sentiment and cosine distance correlation of: 0.72. The misclassed words appearing in Table 6.

The misclassed verbs in Table 6 are close to being correctly classed, with the defining line at y = 0.

To test how swapping out words from FairVec affect the outcome, we replaced the word 'responsibility' with 'dutiful', which produced an F1 score of 0.97 (Table 7). Replacing 'dutiful' with 'accountable' produced an F1 score of

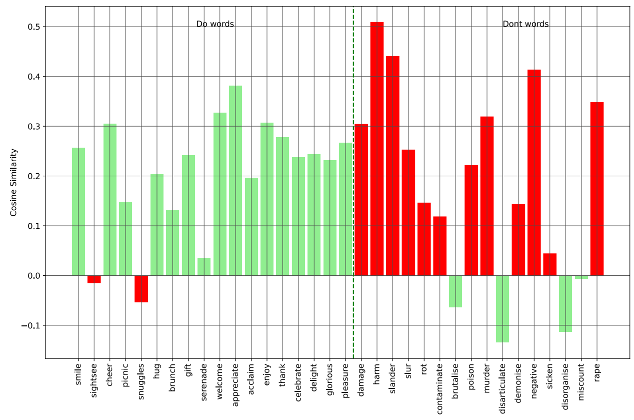

Fig. 3 Using “fair-unfair” as the only dimension for the fairness vector. All bars on the left of dotted line ought to be positive, while those on the right ought to be negative. Red determines incorrectly classed, green determines correctly classed ‘Do verbs’ and ‘Don’t verbs’

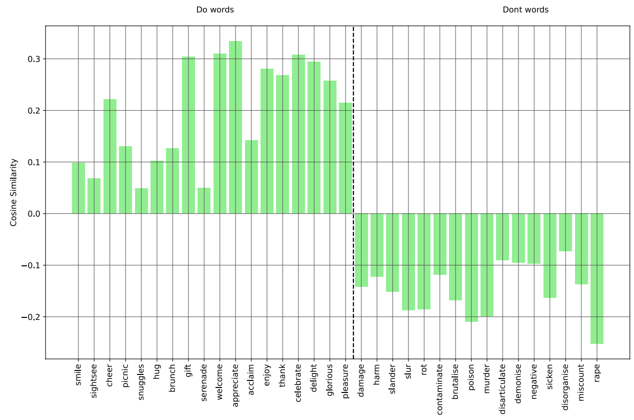

Fig. 4 Cosine similarity of the eight dimensioned Fairness Vector (FairVec) with ‘Do verbs’ and ‘Don’t verbs’, with all correctly classed

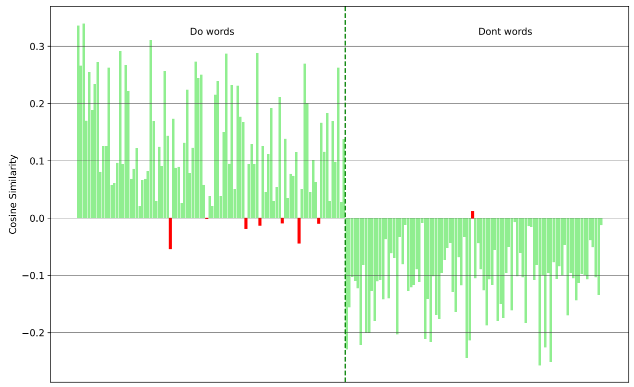

Fig. 5 Cosine similarity of 188 verbs. Seven ‘Do words’ were classed as a ‘Don’t word’. One ‘Don’t word’ was classed as a ‘Do verb’. All bars on the left of the dotted line ought to be positive, while those on the right ought to be negative. Red determines incorrectly classed, green correctly classed ‘Do verbs’ and ‘Don’t verbs’

Table 1 Misclassed verbs. Score rounded to three decimal places

Verb	Score
Schmooze	– 0.0098
Preconcert	– 0.0623
Nuzzle	– 0.0101
Unbend	– 0.0384
Effuse	– 0.0480
Sparer	– 0.0763
Spellbind	– 0.0641
Destroy	0.0147

Table 2 Confusion matrix, F1 = 95.7

N = 188		Actual class	
		Do verbs	Don’t verbs
Predicted class	Do verbs	89	1
	Don’t verbs	7	91

0.97 (Table 8). This minimal change tallies with the literature, as words with similar meanings lie close to each other in the vector space, and word embeddings using GloVe form clusters of conceptually similar words in the embedding space (Hu and Tsujii 2016). Furthermore, in using single words for our dimensions, we are not only comparing the literal sense of the word but the location of the word in vector space. A location that represents its meaning in relation to the whole of the corpus. A word’s broader neighborhood in the embedding space being typically populated by a multitude of terms with related meanings. Thus arithmetically producing similar results (Kozlowski et al. 2019).

To consider how changing the corpus affects outcome, we replaced Glove (Pennington et al. 2014) with the Google News 300 corpus vectorised with Word2Vec (Google Code Archive—Long-Term Storage for Google Code Project Hosting. 2020), repeating the original tests for the 188 Do and Don’t Verbs (Fig. 6) produced an F1 score of 0.97 (Table 9) and a Vader correlation of 0.66. With the misclassed verbs given in Table 10, all of which were close to the cut-off boundary line.

The mean of all of the Do verbs = 0.1267 with a standard deviation of 0.0876. While the mean of all of the Don’t verbs = –0.1250 with a standard deviation of 0.0664. The changing of the corpus appears to have little effect on the F1 score, a result we consider reflective of a universal social bias, a point we expand on in the discussion below.

Table 3 Cosine similarity of top and bottom 15 verbs to the Fairness Vector (FairVec). Scores rounded to four decimal places

Do verb	Score	Don't verb	Score
Joy	0.3834	Gangrene	– 0.2644
Appreciate	0.3346	Callous	– 0.2582
Cherish	0.3284	Torture	– 0.2531
Welcome	0.3103	Rape	– 0.2524
Celebrate	0.3083	Necrotising	– 0.2254
Spirit	0.3041	Exacerbate	– 0.2221
Gift	0.3040	Rearrest	– 0.2214
Delight	0.2948	Aggravate	– 0.2129
Endeavor	0.2846	Traumatize	– 0.2125
Congratulate	0.2830	Poison	– 0.2100
Enjoy	0.2812	Perjury	– 0.2083
Glory	0.2805	Incapacitate	– 0.2013
Adore	0.2714	Plague	– 0.2006
Thank	0.2679	Murder	– 0.1993
Joy	0.3834	Outgas	– 0.1971

3.1 Sentence level results

Table 11 presents the iterations of the stop-words removed sentence 'boy kick baby'. For each sentence, the Fairness Vector is applied, and scoring is displayed with a check on correctness given alongside it.

All the sentences were correctly classified with the exception of: 'boy kick baby happily' (0.054), 'boy kick baby away fire' (−0.0095), 'footballer help footballer' (−0.004). With the last two being close to the cut-off line.

Although the values are small, ordering the correctly classified sentences appears to offer a range (Table 12). We discuss the objectivity of judging a range as accurate and consistent in the discussion section.

4 Discussion

According to the good regulator theorem in cybernetics (Conant and Ross Ashby 1970) 'every good regulator of a system must be a model of that system'. One of the advantages of the use of Word Embeddings to tease out fairness assessments is its ability to represent an ethical dimension of the corpus it is vectorising, without the need for training. In this paper, it was demonstrated that by eliciting the appropriate dimensions of a fairness assessment it is possible to correctly class verbs as fair (Do verbs) or unfair (Don't verbs) with over 95% accuracy.

We accept that the totality of human judgment cannot be represented with such a simple approach; however, an approach is required. Huffington (2018) suggests an inherent danger of 'disentangling wisdom from intelligence', and from being drowned 'in data and starved for wisdom' (Gill 2020b). At the very least, having a friendly AI could potentially allow it to beat a malicious AI to the finish line (Davies 2016).

In using Word Embeddings, we have attempted to address an increasingly registered gap between AI system design and ethics (Gill 2020a), particularly when it comes to implanting such technology around humans. As merely programming in advance the vast systems of human norms is close to impossible, new computational learning algorithms are needed that allow AI to acquire and update, in a context-specific manner, norms that are relevant to their domain of deployment (Malle and Scheutz 2018).

H. L. A. Hart held that whence a legal system operated, people would not necessarily have to internalize the norms associated, only follow the law. In this respect, while an AI may be considered too sub-optimal a species to be able to internalize concepts of fairness, a secondary process by which it can functionally manifest these norms becomes possible (Burr and Keeling 2018). It is also the case with respect to the approach used for making ethical assessments. A computer assess language via calculations, in contrast to humans, who engage epistemically unique knowledge, making ethical judgment a unique human capacity (Weizenbaum 1976).

(Howard and Muntean 2017) have taken the view that artificial moral cognition is a process of developing moral dispositions, instead of learning moral rules. This is philosophically grounded in virtue theory as developed by Aristotle in Nichomachean Ethics. They argue that artificial morality is possible within the framework of a moral dispositional functionalism. It is premised on the theory that moral agents should not be constructed on rule-following methods, but on learning patterns from data. Such an approach incorporates moral functionalism and ethical particularism: principles are not impossible or useless to express, but they take less of a central role in the design. This contrasts with moral generalism, which embraces moral rules or principles (Bauer 2020a). For our system we sought to avoid specifying either approach using the rubric of 'would I wish it for myself'. It entails, firstly, an expectancy that humans only commit acts that they believe will bring them some form of gain, in one manner or another -even the unfortunate situation of self-harm is one that is engaged in due the perceived sense of relief that the perpetrator expects to gain.

Secondly, that humans recognize this same expectancy in others. Then thirdly, that the social outcome of these two premises is held in Word Embeddings.

This approach, whereby choices are qualified as ethical based on a human quality that appears to be inescapable, appears to satisfy the need of a cross-cultural operationalization of fairness.

The observation that humans are unable to commit absolutely gainless acts, has been recorded as far back as

Table 4 Replacing the noun dimension 'joy' (**a**), with the adjective (**b**) 'joyous' in the eight dimensioned Fairness Vector alters the order of ranking of the Don't Verbs

Don't Verb	Fairness Vector Score Using 'joy' (descending)
gangrene	-0.2644
callous	-0.2582
torture	-0.2530
rape	-0.2524
necrotising	-0.2254
exacerbate	-0.2221
rearrest	-0.2214
aggravate	-0.2128
traumatize	-0.2125
poison	-0.2100
perjury	-0.2083
incapacitate	-0.2013
plague	-0.2005
murder	-0.1993
outgas	-0.1971

(a)

Don't Verb	Fairness Vector Score Using 'joyous' (descending)
rape	-0.2929
torture	-0.2881
gangrene	-0.2548
poison	-0.2517
murder	-0.2317
rot	-0.2279
plague	-0.2268
callous	-0.2127
necrotising	-0.2105
perjury	-0.2083
aggravate	-0.2059
exacerbate	-0.2056
tar	-0.1964
assault	-0.1935
scum	-0.1920

(b)

Highlighted: 'murder' and 'rape' both jump up the list, while 'callous' is brought down to a less negative value

Table 5 Confusion Matrix using ‘joyous’ in FairVec. F1 = 0.97

N = 188		Actual class	
		Do verbs	Don’t verbs
Predicted class	Do verbs	91	0
	Don’t verbs	5	92

Table 6 Misclassed verbs

Verb	Score
Preconcert	– 0.0029
Effuse	– 0.0135
Sparer	– 0.0186
Spellbind	– 0.0288
Care	– 0.0136

All were Do Verbs incorrectly misclassed as Don’t Verbs. Scores rounded to four decimal places

Table 7 Confusion Matrix using ‘dutiful’ instead of ‘responsibility’ in FairVec. F1 = 0.97

N = 188		Actual class	
		Do verbs	Don’t verbs
Predicted class	Do verbs	95	4
	Don’t verbs	1	88

Table 8 Confusion Matrix using ‘using ‘accountable’ instead of ‘responsibility’ in FairVec. F1 = 0.97

N = 188		Do verbs	Don’t verbs
Predicted class	Do Verbs	89	1
	Don’t Verbs	7	91

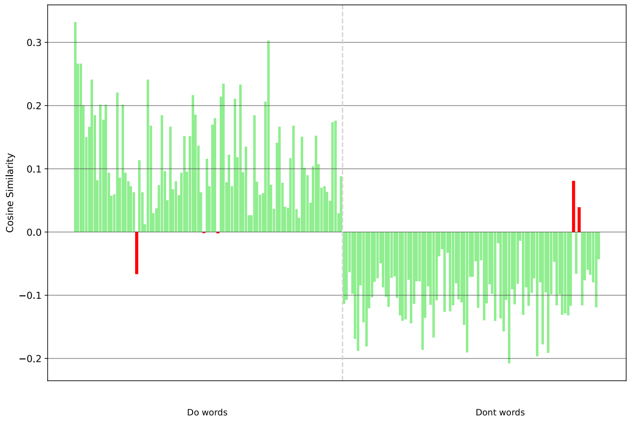

Fig. 6 Cosine similarity of 188 verbs using the Word2Vec Google News Corpus. Three ‘Do words’ were classed as ‘Don’t words’. Two ‘Don’t words’ were classed as ‘Do words’. All bars on the left of dotted line ought to be positive, while those on the right ought to be negative. Red determines incorrectly classed, green correctly classed ‘Do verbs’ and ‘Don’t verbs’

Table 9 Confusion Matrix using Google News corpus and FairVec. F1 = 0.97

N = 188		Actual class	
		Do verbs	Don’t verbs
Predicted class	Do verbs	93	2
	Don’t verbs	3	90

Table 10 Misclassed verbs of the Google News corpus. Three Do verbs incorrectly misclassed, and two Don’t verbs misclassed. Scores rounded to four decimal places

Verb	Score
Purl	– 0.0659
Nuzzle	– 0.0010
Friend	– 0.0014
Cause	– 0.0802
Blame	– 0.0386

Socrates (Morrison 2010) and exhibits a modal prohibition, as opposed to a deontic prohibition.

In effect, the approach of combining the salient abstract features of an act; emotional, material and consequential, without specifying them materially, offers a perspective of ethical variantism while tethering this qualification with ‘would I wish it on myself’ offers a perspective of ethical invariantism based on the fixed human aversion to absolutely gainless activity.

In our attempt to approximate this, we used the finding that human traits bias word embeddings. That these biases are accurate reflections of the culture of society the corpus details. Numerous studies have tested Word Embedding outcomes with independent measures, and found them to tally, giving them epistemological validity, as detailed in the epistemological introduction of this paper. As such, we believed we would be able to tap into a well-documented human social bias by applying a fairness measure, one that articulated the salient features of a social propensity to be social. The main attribute of comprehensive Word Embeddings, that is, those that are built from a representative corpus of everyday life and not based on fiction or fantasy, is their ability to represent multifactorial considerations. Whereby individual points within this vector space do not merely represent singular words, but the depth and forces of interaction of human laden concepts. We propose that it is from this, that it gains epistemological authority to represent the human condition. The vectorized word, is not representative of the word itself, but of a location, a barycenter that sits within the gravitational negotiation of social memes and social reasoning manifested in common human discourse.

Table 11 Classifications of sentences without stop words

Change in	Sentence Words Vectorised	Cosine similarity score	Correctly classed
Verb	**boy kick baby**	-0.0057	Yes
	boy help baby	0.0472	Yes
Context	**boy kick baby toy**	0.0219	Yes
	boy kick baby head	-0.0137	Yes
Description	**boy kick baby toy ball**	0.0327	Yes
	boy kick baby head side	-0.0261	Yes
Adjective	**boy kick baby happily**	0.0537	No
	boy help baby happily	0.0977	Yes
Inferred Intention	**boy kick baby away himself**	-0.0072	Yes
	boy kick baby away fire	-0.0095	No
Stated Intention	**boy kick baby offence**	-0.0332	Yes
	boy kick baby defence	0.0222	Yes
Agent	**footballer kick baby**	-0.0168	Yes
	footballer help baby	0.0425	Yes
Agent and patient	**footballer kick footballer**	-0.0501	Yes
	footballer help footballer	-0.0040	No
Agent-related-object	**footballer kick baby ball**	0.0062	Yes
	footballer kick baby head	-0.0237	Yes

Red words signify change in sentence

Furthermore, the method, in attempting to engage relevant dimensions explicitly, has the added advantage of explainability. Explainability as to why things are fair and what makes them intrinsically fair. Indeed, a genuinely wise agent, it has been posited, must be able to realize what makes good things for well-being good (Tsai 2020).

Imbuing an AI with a framework to make decisions that are socially relevant also requires the agent to have a language in which to represent the structure of the actions being judged (Mikhail 2007). For humans, the most natural way to describe a moral dilemma is to use natural language, hence the emphasis of using this space in the paper. This goes beyond deontic abstractions that have been used in the past as a basis for social negotiation, such as with game-theoretic representation of interactions between individuals (Conitzer et al. 2017).

A further use of developing a fairness metric for texts, beyond fair-AI, is its potential ability to be used to qualitatively assess policy and legal documents. ML algorithms are often trained on examples, with the assumption that it is able to identify the correct dimensions by which to judge new documents (Medvedeva et al. 2020). However, if it is possible to identify the most pertinent dimensions of a text, then such a process becomes even more homed. Indeed, it is widely accepted that a qualification fairness as a balance of rights and responsibilities within a social power interaction provides a comprehensive measure of the said interaction, as demonstrated by Hohfeld (Wenar 2005). It presents a means to adjudicate on the fairness of documents such as contracts as articulated in both EU and UK legislation (Unfair Contract Terms Directive; Consumer Rights Act 2015, c. 15, Part 2, 64:2). Thus, instead of relying on the algorithm to

Table 12 Sorted list of sentences using FairVec. Scores to four decimal places

Sentence	Cosine similarity score sorted in descending order
Boy help baby happily	0.0977
Boy help baby	0.0472
Footballer help baby	0.0425
Boy kick baby toy ball	0.0327
Boy kick baby defence	0.0222
Boy kick baby toy	0.0219
Footballer kick baby ball	0.0062
Boy kick baby	−0.0057
Boy kick baby away himself	−0.0072
Boy kick baby head	−0.0137
Footballer kick baby	−0.0168
Footballer kick baby head	−0.0237
Boy kick baby head side	−0.0261
Boy kick baby offence	−0.0332
Footballer kick footballer	−0.0501

identify the underlying construct that is being sought, it becomes possible to use the correct dimensions in ensemble.

By identifying the pertinent markers in such documents, it also becomes possible to use the wording of the legal and policy documents in a causation analysis with the said legal-policies' outcome. Here, changes in policy wording could be tracked more closely as to their role in the outcome the policy is attempting to seek. Centrally, with interactions between agents being framed as power interactions, it becomes necessary to accurately qualify such power interactions. We put forward the argument that in order for AI to be fully harnessed for its power, it ought to be able to perceive such documents in terms of rights and responsibilities due. One that incorporates the dimensions of fairness. In using Word Embeddings an AI is given epistemic access to society, instead of being closed in by a set of rules beyond which is it unable to learn.

4.1 Improvements

In our results, it was found that the accuracy of the Fairness Vector could be improved if one of the terms was adjusted. This presents both an opportunity and a challenge. Since a small change can have a noticeable effect on the scoring, it may be asked, who determines what the exact, correct, wording ought to be. One could replace 'pain' with 'painful' in line with adjusting 'joy' with 'joyous'. A second challenge also arises when it is attempted to objectively measure the scoring. Most individuals would rate 'murder' as worse than 'rot', however, with lesser scored verbs, and especially positive verbs (Do Verbs) it becomes difficult to measure the validity and accuracy of the scale. One way this paper attempted to do this was to use a sentiment analyser. The assumption being that verbs reflect a relative sentiment and that a plausible ranking ought to correlate positively—though not fully. A positive correlation was assumed as sentiment captures a degree of emotion attached to the verbs. An expectation that it would not fully correlate was made, since a such a score would indicate that the dimensions used in the Fairness Vector act as a sentiment analyser.

Table 13 Comparing a test sentence against 'The (subject) would wish it' vs. 'The (subject) would not wish it'

Test sentence	Perform a cosine-similarity test with
'The man killed the child'	'The child would wish it' 'The child would not wish it'

Two outstanding issues thus present themselves, the heuristic (vector wording) needed, and the expected result by which one can ascertain that the correct wording has been used. To address this, it is proposed for further research to approach this problem using the same method employed in Word2Vec (Mikolov et al. 2013). Word2vec uses a feed-forward fully connected architecture (Le and Mikolov 2014). It estimates the probability of the occurrence of a word given the input of other words. It then tests this result against the correct -expected- result. Finding the discrepancy between its own result and the correct result, it feeds back a loss to the neural network which then re-adjusts. Minimising the loss function until the best words are found. Only in our case, instead of using words as our inputs, it is proposed to use the permutations of possible alternative vector words (i.e., pain, paining, painful, pains) as the inputs, with the correct -expected- result being the consistency of output with the corpus itself.

This consistency could be measured by employing sentence level comparisons. For example, if the output were to rank two verbs in descending order (a then b), then we would input these words into sentences: it is worse to..a.. than to..b.. (e.g., it is worse to kill than to kick). This sentence would be vectorised using the Universal Sentence Encoder (USE) (Cer et al. 2018) or Sentence-BERT (Reimers and Gurevych 2019), for example, and a juxtaposition of a and b in the sentence would be compared with part of the corpus. In such a case, the first sentence 'it is worse to kill than to kick', will be more closely matched than 'it is worse to kick than to kill'. Echoing earlier work on this area using SBERT (Schramowski et al. 2019, 2020).

This may be further improved in using a sentence level fairness assessment in which the wording of the question is explicitly analyzed 'Would (Agent 1) wish (Verb 1) on themselves in (Context 1)?' In this case, a cosine similarity could be calculated comparing a reformulation of the original sentence against two opposite senses (Table 13).

Such symmetrical analysis would potentially allow for a universal assessment of the act, one that took into consideration the cultural nuances of the individuals. The context can also be further incorporated using follow on sentences, such as: 'Then for the (verb) the (subject) was applauded/chastised'.

Thus, instead of focusing on individual verbs, it is suggested that vectorising whole sentences, then using the same non-training methodology employed in this paper could be a better way forward. In this regards the approach would take the form of an error minimisation function that employs variations of sentences which elicit fairness dimensions to be tested against the corpus for both ethical congruence and qualification. For example, on prison-liberty, a test sentence would read: The criminal was jailed by the court vs. The court was jailed by the criminal. Other sentences that test for each of the dimensions identified in the paper could also be included to allow for a feedback loop that fine-tuned the system. Furthermore, the use of SBERT or the USE would solve the issue of word order in Bag-of-Words and vector embedding addition methods, which do not preserve sentence word order or context (Cer et al. 2018; Reimers and Gurevych 2019).

The use of a phrase FairVec in this paper is thus only to qualify the type of measure being used, and not to be considered in absolute terms. This is seen in the necessary addition of further details to the selected dimensions when using sentences, improving validity and reliability.

5 Limitations

Word embeddings in GloVe can be initialized randomly – as starting point to the process of learning. Different initial starting points have been shown to maintain a high level of accuracy when comparing tasks in each. However, divergences have been found in the relationships they learnt. One manner of monitoring and enhancing this has been to use a metric to improve the performance of NLP tasks downstream (Tian et al. 2016). The use of small corpora is also to be avoided, as fine-grained distinctions between cosine similarities become less reliable. Long documents that use small corpora have been found to be more susceptible to variation in the cosine similarities between embeddings (Antoniak and Mimno 2018).

It has also been shown that vector-spaces contain hubs made of vectors that are in close proximity to a large number of other vectors (Radovanovi´c et al. 2010). This manifests when words have a high cosine similarity with many other words (Dinu et al. 2014). While different distance normalization schemes have been proposed to ameliorate this phenomenon (Dinu et al. 2014; Tomašev et al. 2011, Wilson and Schakel 2015), it would be worth considering words that share less commonality when building comparative vectors, as well as implementing subtractions and additions to minimize the noise introduced through the use of spurious word locations in the vector space. Computationally, one may also exploit the feature that words which only appear in similar contexts tend to have longer vectors than words of the same frequency that appear in a wide variety of contexts (Wilson and Schakel 2015).

A further limitation lies in homonyms words, for which Huang et al. (2012) introduced the Stanford Contextual Word Similarity dataset (SCWS) to compute similarity between two words given the contexts they occur in, e.g., money vs. bank: 'along the east bank of the river', and 'the basis of all money laundering'. Further work has suggested the use of multiple vectors per word-type to account for different word-senses (Neelakantan et al. 2014).

A general limitation that is often mentioned in the literature is that the choice of corpora can affect outcome (De Vine et al. 2014a, b). We attempted to test this using the Google News corpus, which we found produced similar results. We temper the caution of the literature on the choice of corpora by suggesting that for FairVec to work, it inherently relies on a large corpus, one that captures the range of human activity. It would inherently not work with fantasy and sci-fi documents that represent activity that runs contrary to the natural order of our world, for example, where causal effects are suspended, where anarchy is the sought-after norm in families and societies. However, given the nature of human society, such a fairness vector is potentially replicable in various languages and even using corpora from different time periods.

6 Conclusion

This paper demonstrated the plausibility of using Word Embedding vectors to make fairness assessments. Its premise being that the human propensity for both society and an inherent aversion to harmful gainless activity introduces a pro-social bias into Word Embeddings. Whereby acts that meet this propensity are qualified as being closer in the vector space to the latent concept of fairness. We demonstrated that this latent concept can be elicited by building a vector that specified as its dimensions the principal perceptions engaged by humans when making a fairness assessment. The dimensions were found based on the social psychology literature covering the perception of social interaction. The recognition of loss, pain and punishment are seen as blameworthy. Whereas gain, joy and liberty as praiseworthy, but only when filtered according to their associated score of being responsible, or irresponsible, respectively. The use of vector embedding for [responsible -irresponsible] acted to conceptually moderate the other fairness dimensions to organize them into an ethical vector-space. The limitation of this study was its focus on singular verbs. A number of suggestions were made to improve the performance of FairVec through sentence embeddings and dimension optimization routines using neural feedback loss minimization. The approach used in this paper also demonstrates a method to make an ethical assessment that forgoes the need to program deontic rules into an AI algorithm, or to use training data, relying instead on is the efficacy of Word Embeddings.

7 Availability of data and material

The data used in this paper is available at the Github link below.

Acknowledgements This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Next Generation Internet TRUST grant agreement no. 825618. The author wishes to thank David Stillwell for kindly commenting on the paper, Bao Sheng Loe, Doradistria Rota, Sander van der Linden, Lee De-Wit, Patrick McClanahan, and Joseph Watson for their kind feedback, the Department of Psychology Social Decision-Making Laboratory, the Cambridge Forum for Legal and Political Philosophy, and the Psychometrics Centre, all at the University of Cambridge. I would also like to thank the two anonymous reviewers for their comments.

Funding This research was funded by the European Union's Horizon 2020 research and innovation programme under the Next Generation Internet TRUST grant agreement no. 825618.

Code availability https://github.com/AhmedIzzidien/FairnessVector/blob/master/FairnessVector%20v4.ipynb.

Declarations

Conflict of interest No known conflicts of interest or competing interests to disclose.

References

Almeida F, Xexéo G (2019) Word embeddings: A survey

Andrews M, Frank S, Vigliocco G (2014) Reconciling embodied and distributional accounts of meaning in language. Topics Cogn Sci 6(3):359–370

Antoniak M, Mimno D (2018) Evaluating the stability of embedding-based word similarities. Trans Assoc Comput Linguist 6:107–119. https://doi.org/10.1162/tacl_a_00008

Bauer WA (2020a) Virtuous vs. Utilitarian artificial moral agents. Ai Soc 35(1):263–271. https://doi.org/10.1007/s00146-018-0871-3

Bauer WA (2020b) Expanding nallur's landscape of machine implemented ethics. Sci Eng Ethics. https://doi.org/10.1007/s11948-020-00237-x

Berkowitz L (1972) Social Norms, Feelings, and Other Factors Affecting Helping and Altruism11The author's research reported in this paper was carried out under grants from the National Science Foundation. In Berkowitz L (Ed.), Adv Experim Soc Psychol 6:63–108. Academic Press. https://doi.org/10.1016/S0065-2601(08)60025-8

Berkowitz L, Daniels LR (1963) Responsibility and dependency. J Abnormal Soc Psychol 66(5):429

Bicchieri C (2006) The grammar of society: The nature and dynamics of social norms (pp. xvi, 260). Cambridge University Press

Borghi AM, Barca L, Binkofski F, Tummolini L (2018) Varieties of abstract concepts: Development, use and representation in the brain. Philosoph Trans Royal Soc B 373(1752):20170121. https://doi.org/10.1098/rstb.2017.0121

Boyd R, Richerson PJ (2009) Culture and the evolution of human cooperation. Philosoph Trans Royal Soc B 364(1533):3281–3288. https://doi.org/10.1098/rstb.2009.0134

Boyd RL, Wilson SR, Pennebaker JW, Kosinski M, Stillwell DJ, Mihalcea R (2015) Values in words: Using language to evaluate and understand personal values. Ninth International AAAI Conference on Web and Social Media

Brañas-Garza P, Rodríguez-Lara I, Sánchez A (2017) Humans expect generosity. Sci Rep 7(1):1–9. https://doi.org/10.1038/srep42446

Brewer MB (2004) Taking the social origins of human nature seriously: Toward a more imperialist social psychology. Pers Soc Psychol Rev 8(2):107–113

Brosnan SF, Bshary R (2016) On potential links between inequity aversion and the structure of interactions for the evolution of cooperation. Behaviour 153(9–11):1267–1292. https://doi.org/10.1163/1568539X-00003355

Brunet M-E, Alkalay-Houlihan C, Anderson A, Zemel R (2019) Understanding the origins of bias in word embeddings. International Conference on Machine Learning, 803–811

Brunnermeier MK (2001) Asset pricing under asymmetric information: Bubbles, crashes, technical analysis, and herding/Markus Brunnermeier K. Oxford University Press, Oxford

Burkart JM, Allon O, Amici F, Fichtel C, Finkenwirth C, Heschl A, Huber J, Isler K, Kosonen ZK, Martins E, Meulman EJ, Richiger R, Rueth K, Spillmann B, Wiesendanger S, van Schaik CP (2014) The evolutionary origin of human hyper-cooperation. Nat Commun 5(1):1–9. https://doi.org/10.1038/ncomms5747

Burr C, Keeling G (2018) Building machines that learn and think about morality. Proceedings of the convention of the society for the study of artificial intelligence and simulation of behaviour (AISB 2018) Society for the study of artificial intelligence and simulation of behaviour

Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356:6334. https://doi.org/10.1126/science.aal4230

Carey S (2011a) Précis of the origin of concepts. Behav Brain Sci 34(3):113–124. https://doi.org/10.1017/S0140525X10000919

Carey S (2011b) The Origin of Concepts. Oxford University Press, Oxford

Castelfranchi C, Giardini F, Lorini E, Tummolini L (2003) The prescriptive destiny of predictive attitudes: From expectations to norms via conventions. Proce Ann Meet Cogn Sci Soc 25(25)

Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C (2018) Universal sentence encoder

Cervantes J-A, López S, Rodríguez L-F, Cervantes S, Cervantes F, Ramos F (2020) Artificial moral agents: a survey of the current status. Sci Eng Ethics 26(2):501–532

Chen D, Peterson JC, Griffiths TL (2017) Evaluating vector-space models of analogy

Civai C (2013) Rejecting unfairness: Emotion-driven reaction or cognitive heuristic? Front Hum Neurosci. https://doi.org/10.3389/fnhum.2013.00126

Clark S, Pulman S (2007) Combining symbolic and distributional models of meaning

Conant RC, Ross Ashby W (1970) Every good regulator of a system must be a model of that system. Int J Syst Sci 1(2):89–97

Conitzer V, Sinnott-Armstrong W, Borg JS, Deng Y, Kramer M (2017) Moral decision making frameworks for artificial intelligence. In Thirty-First Aaai Conference on Artificial Intelligence (pp. 4831–4835)

Consumer rights act (2015) (n.d.). Queen's Printer of Acts of Parliament. Retrieved December 7, 2020, from https://www.legislation.gov.uk/ukpga/2015/15/part/2/enacted

Corradi-Dell'Acqua C, Civai C, Rumiati RI, Fink GR (2013) Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. Social Cogn Affect Neurosci 8(4):424–431. https://doi.org/10.1093/scan/nss014

Cremer DD, Lange PAMV (2001) Why prosocials exhibit greater cooperation than proselfs: The roles of social responsibility and reciprocity. Eur J Pers 15(S1):S5–S18. https://doi.org/10.1002/per.418

Davies J (2016) Program good ethics into artificial intelligence. Nature News. https://doi.org/10.1038/538291a

De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P (2014a) Medical semantic similarity with a neural language model. 1819–1822

De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P (2014b) Medical Semantic Similarity with a Neural Language Model. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 1819–1822. https://doi.org/10.1145/2661829.2661974

Decety J, Michalska KJ, Kinzler KD (2012) The Contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. Cereb Cortex 22(1):209–220. https://doi.org/10.1093/cercor/bhr111

Drozd A, Gladkova A, Matsuoka S (2016) Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. Proceedings of Coling 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 3519–3530

Duke, George RP (2017) The Cambridge companion to natural law jurisprudence/edited by George Duke, Deakin University, Robert P. George, Princeton University. Cambridge : Cambridge University Press

El-Amir H (2020) Deep learning pipeline: Building a deep learning model with TensorFlow / Hisham El-Amir, Mahmoud Hamdy. Apress LP, Berkeley, CA

Fehr E, Rockenbach B (2004) Human altruism: Economic, neural, and evolutionary perspectives. Curr Opin Neurobiol 14(6):784–790. https://doi.org/10.1016/j.conb.2004.10.007

Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation, and the enforcement of social norms. Hum Nat 13(1):1–25. https://doi.org/10.1007/s12110-002-1012-7

Fessler DM, Haley KJ (2003) The strategy of affect: Emotions in human cooperation 12. The Genetic and Cultural Evolution of Cooperation, P. Hammerstein, Ed, 7–36

Firth JR (1958) A Synopsis of Linguistic Theory, 1930–1955

Fortescue M (2017) The Abstraction Engine. In Aicr.94. John Benjamins Publishing Company

Garg N, Schiebinger L, Jurafsky D, Zou J (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci 115(16):E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gill KS (2020) Strange affair of man with the machine. AI Soc 35(4):777–782. https://doi.org/10.1007/s00146-020-01078-9

Gill KS (2020) Dance of the artificial alignment and ethics. AI Soc 35(1):1–4. https://doi.org/10.1007/s00146-019-00923-w

Google Code Archive—Long-term storage for Google Code Project Hosting. (n.d.). Retrieved October 11, 2020, from https://code.google.com/archive/p/word2vec/

Hai-Jew S (2017) Psychological text analysis in the digital humanities. In Data Analytics in Digital Humanities. Springer International Publishing. https://doi.org/10.1007/978-3-319-54499-1

Handgraaf MJJ, Van Dijk E, Vermunt RC, Wilke HAM, De Dreu CKW (2008) Less power or powerless? Egocentric empathy gaps and the irony of having little versus no power in social decision making. J Pers Soc Psychol 95(5):1136–1149. https://doi.org/10.1037/0022-3514.95.5.1136

Hewstone M, Stroebe W, Jonas K (2012) An Introduction to Social Psychology (4th ed.). John Wiley & Sons

Howard D, Muntean I (2017) Artificial moral cognition: moral functionalism and autonomous moral agency. In: Powers TM (Ed.), philosophy and computing: essays in epistemology, philosophy of mind, logic, and ethics (Vol. 128, pp. 121–159). https://doi.org/10.1007/978-3-319-61043-6_7

Hu W, Tsujii J (2016) A latent concept topic model for robust topic inference using word embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 380–386

Huffington A (2018) "We're Drowning in Data But Starved for Wisdom." Medium. March 27, 2018. https://medium.com/thrive-global/were-drowning-in-data-but-starved-for-wisdom-bd2375baca5

Hutto CJ (n.d.). vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. (3.3.2) [Computer software]. Retrieved September 25, 2020, from https://github.com/cjhutto/vaderSentiment

Izzidien A, Chennu S (2018) A Neuroscience Study on the Implicit Perceptions of Fairness and Islamic Law in Muslims Using the EEG N400 Event Related Potential. J Cogn Neuroeth 5(2):21–50

Jentzsch S, Schramowski P, Rothkopf C, Kersting, K (2019a) Semantics derived automatically from language corpora contain human-like moral choices. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 37–44. https://doi.org/10.1145/3306618.3314267

Kahane G (2016) Moral Brains: The Neuroscience of Morality. In Is, Ought, and the Brain. Oxford University Press. https://oxforduniversitypressscholarship-com.ezp.lib.cam.ac.uk/view/https://doi.org/10.1093/acprof:oso/9780199357666.001.0001/acprof-9780199357666-chapter-13

Köchling A, Wehner MC (2020) Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 1–54

Kozlowski AC, Taddy M, Evans JA (2019) The geometry of culture: analyzing the meanings of class through word embeddings. Am Sociol Rev 84(5):905–949. https://doi.org/10.1177/0003122419877135

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. International Conference on Machine Learning, 1188–1196

Liu Y, Jun E, Li Q, Heer J (2019) Latent space cartography: visual analysis of vector space embeddings. Computer Graphics Forum 38(3):67–78. https://doi.org/10.1111/cgf.13672

Malle BF, Scheutz M (2018) Learning how to behave. Moral competence for social robots. Springer, Wiesbaden, Germany, pp 1–24

Medvedeva M, Vols M, Wieling M (2020) Using machine learning to predict decisions of the European Court of Human Rights. Artific Intellig Law 28(2):237–266

Mikhail J (2007) Universal moral grammar: Theory, evidence and the future. Trends Cogn Sci 11(4):143–152

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. https://openreview.net/forum?id=idpCdOWtqXd60&noteId=mmlAm0ZawBraS

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inform Process Syst 3111–3119

Milinski M, Semmann D, Krambeck H-J (2002) Reputation helps solve the 'tragedy of the commons.' Nature 415(6870):424–426. https://doi.org/10.1038/415424a

Morrison DR (2010) The Cambridge companion to socrates. Cambridge University Press. https://doi.org/10.1017/CCOL9780521833424

Nerbonne J, Hinrichs E (2006) Linguistic Distances. Linguistic Distances, 1–6

Nowak MA (2006) Five rules for the evolution of cooperation. Science 314(5805):1560–1563

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543

Peysakhovich A, Nowak MA, Rand DG (2014) Humans display a 'cooperative phenotype' that is domain general and temporally stable. Nat Commun 5(1):1–8. https://doi.org/10.1038/ncomms5939

Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks

Rheault L, Cochrane C (2020) Word embeddings for the analysis of ideological placement in parliamentary corpora. Polit Analy 28(1):112–133. https://doi.org/10.1017/pan.2019.26

Sadler-Smith E (2012) Before virtue: biology, brain, behavior, and the "Moral sense." Bus Ethics Q 22(2):351–376. https://doi.org/10.5840/beq201222223

Schramowski P, Turan C, Jentzsch S, Rothkopf C, Kersting K (2019) BERT has a Moral Compass: Improvements of ethical and moral values of machines

Schramowski P, Turan C, Jentzsch S, Rothkopf C, Kersting K (2020) The Moral Choice Machine. Front Artific Intellig. https://doi.org/10.3389/frai.2020.00036

Schwartz SH, Howard JA (1982) Helping and cooperation: A self-based motivational model. In: Grzelak J, Derlega VJ (Eds.), Cooperation and Helping Behavior—1st Edition. New York: Academic Press

Simon HA (1990) A mechanism for social selection and successful altruism. Science 250(4988):1665–1668. https://doi.org/10.1126/science.2270480

Smith EA (2010) Communication and collective action: Language and the evolution of human cooperation. Evol Hum Behav 31(4):231–245. https://doi.org/10.1016/j.evolhumbehav.2010.03.001

Tabibnia G, Satpute AB, Lieberman MD (2008) The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol Sci 19(4):339–347. https://doi.org/10.1111/j.1467-9280.2008.02091.x

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: statistics, structure, and abstraction. Science 331(6022):1279–1285. https://doi.org/10.1126/science.1192788

Tian Y, Kulkarni V, Perozzi B, Skiena S (2016) On the convergent properties of word embedding methods

Tomasello M (2014) The ultra-social animal. Eur J Soc Psychol 44(3):187–194

Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46(1):35–57

Tsai C (2020) Artificial wisdom: A philosophical framework. AI Soc 35(4):937–944. https://doi.org/10.1007/s00146-020-00949-5

Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson KA, Ceder G, Jain A (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571(7763):95–98. https://doi.org/10.1038/s41586-019-1335-8

Unfair contract terms directive. (n.d.). [Text]. European Commission - European Commission. Retrieved December 7, 2020, from https://ec.europa.eu/info/law/law-topic/consumers/consumer-contract-law/unfair-contract-terms-directive_en

van Dijk E, Vermunt R (2000) Strategy and fairness in social decision making: sometimes it pays to be powerless. J Exp Soc Psychol 36(1):1–25. https://doi.org/10.1006/jesp.1999.1392

Wallach W, Allen C, Smit I (2008) Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. AI Soc 22(4):565–582

Weizenbaum J (1976) Computer power and human reason: from judgment to calculation. Computer power and human reason: from judgment to calculation. Oxford, England: W. H. Freeman & Co.

Wenar L (2005) The Nature of Rights. Philos Public Aff 33(3):223–252

White L, Togneri R, Liu W, Bennamoun M (2015) How well sentence embeddings capture meaning. Proceedings of the 20th Australasian Document Computing Symposium, 1–8. https://doi.org/10.1145/2838931.2838932

White L, Togneri R, Liu W, Bennamoun M (2019) Sentence Representations and Beyond. In: White L, Togneri R, Liu W, & Bennamoun M (Eds.), Neural representations of natural language (pp. 93–114). Springer. https://doi.org/10.1007/978-981-13-0062-2_5

Wilson BJ, Schakel AMJ (2015) Controlled Experiments for Word Embeddings

Youyou W, Kosinski M, Stillwell D (2015) Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci 112(4):1036–1040