Word embeddings reveal growing moral concern for people, animals, and the environment

Stefan Leach[1], Andrew P. Kitchin, and Robbie M. Sutton[1]

[1]School of Psychology, University of Kent, UK

In Press, June 2023

British Journal of Social Psychology

Abstract

The Enlightenment idea of historical moral progress asserts that civil societies become more moral over time. This is often understood as an expanding moral circle and is argued to be tightly linked with language use, with some suggesting that shifts in how we express concern for others can be considered an important indicator of moral progress. Our research explores these notions by examining historical trends in natural language use during the 19th and 20th centuries. We found that the associations between words denoting moral concern and words referring to people, animals, and the environment grew stronger over time. The findings support widely-held views about the nature of moral progress by showing that language has changed in a way that reflects greater concern for others.

Keywords: *moral progress, moral expansiveness, moral circle, natural language processing, word embeddings*

Word embeddings reveal growing moral concern for people, animals, and the environment

The Enlightenment idea of historical progress asserts that civil societies become more moral over time (Kant, 1991; Pinker, 2018; Turgot, 1973). Although there is disagreement about what constitutes moral progress (Macklin, 1977; Moody-Adams, 1999), many scholars believe it involves increasing concern for the welfare of others, often referred to as an expanding moral circle (Bloom, 2010; Lecky, 1869; Nussbaum, 2007; Pinker, 2018; Singer, 1981). Moreover, it has been argued that language has played a central role in expanding the moral circle (Bloom, 2010; Pinker, 2018) and that shifts in how we express concern for others can be considered an important indicator of moral progress (Kant, 1991). Despite this, we know relatively little about how language indicative of concern for others has changed over time. To explore this, we used machine-learning methods (word embeddings) applied to natural language to test for historical shifts in the distribution of words denoting moral concern and words denoting people, animals, and the environment.

Moral progress and the expanding moral circle

Headlines and news cycles are often dominated by negative trends, such as declines in environmental quality and democratic norms (Pinker, 2011; Roser & Nagdy, 2014). Though pessimism is frequently warranted, positive historical trends are also apparent, including the abolition of slavery in many countries (Drescher, 2009), the rise of universal human rights (Buchanan, 2013), downward trends in prejudice and interpersonal violence (Charlesworth & Banaji, 2019; Pinker, 2011; Rosling et al., 2018), and increasing legal protections for animals and the environment (Ruby, 2012; Stoknes, 2015). These trends are consistent with the Enlightenment idea of historical progress, according to which civil societies become more moral as they develop over time (Kant, 1981; Pinker, 2011; Turgot, 1973).

These trends are indicative of a historical shift towards greater moral concern for people, animals, and the environment–a shift that is often referred to as an 'expanding moral circle' (Bloom, 2010; Pinker, 2011; Singer, 1981). The moral circle denotes the varying levels of concern people have for the welfare of others (Crimston et al., 2016; Laham, 2009; Lecky, 1869; Singer, 1981). Within a moral circle, concern is typically greatest for those entities that are closest to the centre, such as family members and friends, and lesser or non-existent for entities closer to the periphery, such as wrongdoers and invertebrate animals. A more expansive moral circle is one that has come to include more entities than it previously did, and is therefore one that expresses greater concern for others.

It has been suggested that our moral circles can be identified in how we speak and write about others. Kant argued that the public expression of universalized sympathy for the struggles of the French people against tyranny in the 1700s constituted a form of moral progress (Kant, 1991). This is consistent with contemporary philosophical thought about the level at which moral progress can occur (Macklin, 1977; Musschenga & Meynen, 2017). It is also consistent with the thinking of those who point to historical publications as markers of moral progress. Despite its contentious legacy, the publication of *Uncle Tom's Cabin* (1852) is looked upon as a moral landmark by virtue of its portrayal of the social injustices inflicted upon black Americans as slaves. Another example is found in Peter Singer's *Animal Liberation* (1975), which marks a shift towards viewing animals as having moral status and being worthy of concern. This understanding is predicated on broader ideas about the function of rational discourse (Haidt, 2001; Mercier & Sperber, 2011; Rhee et al., 2019) and how moral issues are expressed in language (Boyd & Schwartz, 2021; Garten et al., 2016; Graham et al., 2009; Pennebaker et al., 2003; Sagi & Dehghani, 2014). We take this view of moral progress and the expanding moral

circle as providing a strong impetus to explore how language reflecting moral concern for others has grown over time.

Historical trends in language

How we speak and write about others has undoubtedly changed over the last century. Large-scale trends in word usage are evident in the Google Books Corpus (comprising ~6% of all books published in English; Lin et al., 2012; Michel et al., 2011). Words such as *abuse, defend, kill, care, suffering,* and *peace* have increased in frequency over time (Wheeler et al., 2019), suggesting a greater focus on suffering and well-being. On the other hand, words such as *individual, self, get*, and *choose* have also increased, indicating a greater preoccupation with the self compared to others (Greenfield, 2013). At the same time, political figures have increasingly drawn on words associated with togetherness (Buttrick et al., 2020). There is also evidence that harm-related concepts, such as *bullying*, have become more expansive in their meaning over time, which is thought to be driven by an increasing sensitivity to harm (Haslam, 2016; Haslam et al., 2020; Vylomova et al., 2019). These findings provide important insights into how moral language has changed over time. However, they do not provide an answer to the question of whether language reflecting moral concern for others has changed over time.

Advances in machine learning and natural language processing, known as word embeddings, provide a novel means to investigate trends in language (Joulin et al., 2016; Mikolov, Chen, et al., 2013; Mikolov et al., 2017; Mikolov, Sutskever, et al., 2013; Mikolov, Yih, et al., 2013; Pennington et al., 2014). These methods analyse the distribution of words and quantify how they tend to co-occur, which can indicate how language is used in relation to others (Caliskan et al., 2017; Charlesworth et al., 2021; Garg et al., 2018; Grand et al., 2022; Leach et al., 2022; Lewis & Lupyan, 2020; Richie et al., 2019). Charlesworth et al. (2021) provide an illustrative example: take the statements "*dad is still at the office*" and "*get mom from the*

kitchen". The proximity of the words *dad–office* and *mom–kitchen* in these statements encodes the presence of stereotypical attributions to men and women–that men are more strongly associated with the workplace and women with the home. This kind of co-occurrence between words has been used to quantify how we communicate about entities with regard to, among other things, their competence (e.g., *man-strong;* Charlesworth et al., 2021), positivity (e.g., *flower-good*; Caliskan et al., 2014) and intelligence (e.g., *owl-smart*; Richie et al., 2022). Moreover, Leach et al. (2022) recently used this approach to quantify the lower moral concern afforded to animals compared to humans in language by examining how words such as *human-concern* and *animal-indifference* tend to co-occur. This work demonstrates that word co-occurrences, captured by word embeddings, can provide insight into how moral language is used in relation to others.

The relative abundance of historical records of text makes it possible to conduct quantitative analyses of linguistic change. This can be achieved by specifying multiple word embedding models on text from specific time periods (e.g., 1900-1910, 1920-1930, 1930-1940). The resulting embeddings can then be aligned so as to permit direct comparisons across models and therefore across time (Hamilton et al., 2016). Such models have provided insights into how language has shifted over time. For example, examining such shifts has provided insight into how linguistic biases associated with genders, occupations, and personality traits have changed over time (Garg et al., 2018). Taken together, this suggests that word embeddings can be used to test for historical shifts in how language reflecting moral concern has been used in relation to people, animals, and the environment.

Present work

Prominent perspectives argue that the expansion of our moral circles is intimately related to language (Bloom, 2010; Pinker, 2018) and that moral progress can be identified in

changes in how we speak and write about the welfare of others (Kant, 1991). We explored this by using word embeddings to examine shifts in how words denoting moral concern and words denoting others have tended to co-occur from the 1830s to the 1990s. We expected to find stronger co-occurrence relationships between these words over time, reflecting a change in language towards greater moral concern for others.

We also test two competing accounts about the nature of these changes. First, we test whether language has changed in a way that is consistent with a classic understanding of the expanding moral circle (Bloom, 2010; Lecky, 1869; Nussbaum, 2007; Pinker, 2018; Singer, 1981), which conceives moral progress in terms of improvements in the welfare of distant others and reductions in parochialism, such as might be brought about by targeted struggles to improve their rights and the reasoned extension of concern beyond close kin (Dixon et al., 2012; Singer, 1981). This view implies that language reflecting moral concern has increased more for entities on the margins of the moral circle (e.g., strangers, animals, and the environment) than it has for those found at the centre (e.g., family members). Second, we test if language has changed in a way that is consistent with a view of moral progress in terms of *a rising tide raises all boats*, which conceives moral progress as a general upward trend in the welfare of all, such as might be the product of global improvements in material welfare, societal conditions, and the adoption of egalitarian and universalist discourse (Evans, 2017). This view implies that language reflecting moral concern has increased in largely the same way for entities on the margins of the moral circle as well as those found at the centre.

Methods

The methods, data, and analysis script are available via the *Open Science Framework* (https://osf.io/z5vbt/).

Distributed semantics and word embeddings

Word embeddings refer to a set of techniques that model natural language by analysing the distribution of words and how they co-occur (Jurafsky & Martin, 2019). These models do this by representing each word in a corpus as a vector in a multidimensional space. The aim is to represent words that have similar co-occurrence relationships (e.g., fork-knife) with vectors that have similar directions; and words that have dissimilar co-occurrence relationships (e.g., fork-sky) with vectors that have dissimilar directions. Conceptually, we can think about similar words as occupying positions that are close together in this space, and dissimilar words as occupying positions that are further apart. Geometric measures, such as cosine similarity, can be calculated to quantify how similar words are with regard to the direction of their corresponding vectors.

The assignment of vectors can be achieved by training a model (a neural network) to predict target words from context words by using data from the underlying corpus (Mikolov, Chen, et al., 2013; Mikolov et al., 2017; Mikolov, Sutskever, et al., 2013; Mikolov, Yih, et al., 2013; cf. Pennington et al., 2014). To do this, the corpus is broken into samples, each of which is an individual context and target word. The context is the n words occurring before and after a target word. The goal of training is to increase the probability that the model outputs the target word given the context words as an input. To begin, each word in the training corpus is represented by a random vector in some multi-dimensional vector space. By iterating over the training samples, the positions of the vectors are shifted to better predict target words from context words. The error in these predictions is minimised by checking whether the cosine similarity between vectors better corresponds to how often and closely words co-occur in the corpus of text. Through this process, word embedding models arrive at a set of vectors that are ‘embedded’ in a space and represent the co-occurrence relationships of words in a corpus of text.

Word embeddings in psychology

By examining the relative positions of words in an embedding space, it is possible to explore how language is used in relation to entities (Caliskan et al., 2017; Caliskan & Lewis, 2020; Charlesworth et al., 2021; Garg et al., 2018; Leach et al., 2022; Lewis & Lupyan, 2020). For example, animals that are typically thought of as intelligent (e.g., *owl*) are generally found closer to words reflecting intelligence (e.g., *smart*) compared to animals that are not (e.g., *ant*). This seems to be a general property of language captured via word embeddings, and has been shown for animate (e.g., people) and inanimate (e.g., cities) entities and a range of features, including moral worth (Caliskan et al., 2017; Grand et al., 2022; Leach et al., 2022; Richie et al., 2019).

This property allows one to use word embeddings to answer social-psychological questions about how we use language in relation to others. For example, the relative positions of clusters of words in the embeddings space reveal racial biases in language, where words denoting European American names (e.g., Brad) are typically found to be relatively closer to pleasant (vs. unpleasant) words than African American names (e.g., Kareem; Caliskan et al., 2017). Similar approaches have revealed biases in how we communicate about the stereotypical features of men and women; for example, their associations with maths, arts, competence, and warmth (Charlesworth et al., 2021; Garg et al., 2018; Lewis & Lupyan, 2020). These results demonstrate that word embeddings can effectively quantify how we use language in relation to morally-relevant entities and therefore that they lend themselves to testing how language reflecting moral concern is used in relation to people, animals, and the environment.

Moral concern in word embeddings

We have argued that by examining the distribution of language, we can learn something about how language reflecting moral concern is used in relation to others. To pursue this idea,

we need a measure of moral concern for entities. Following prior work (Leach et al., 2022), we operationalize this as the relative association between clusters of words denoting the constructs of interest (Caliskan et al., 2017; Charlesworth et al., 2021; Garg et al., 2018; Grand et al., 2022; Lewis & Lupyan, 2020; Richie et al., 2019). That is, the relative association between words denoting entities (e.g., sister), concern (e.g., care), and indifference (e.g., apathy). We reason that if a given entity word is more strongly associated with words denoting concern than it is with words denoting indifference, this suggests that it is more likely to be the object of language reflecting moral concern.

To compute such an index, we defined a set of words that reflect relevant entities, concern, and indifference. We selected 84 words representing entities spanning six categories: family (e.g., husband), ingroup (e.g., ally), marginalised (e.g., unemployed), outgroup (e.g., enemy), animals (e.g., dog), and the environment (e.g., river). Where possible we selected these words from prior research (Crimston et al., 2016). The list of words denoting people, animals, and the environment is presented in Table 1.

Table 1

Words denoting people, animals, and the environment.

Category	Words
Family	husband, wife, father, mother, son, daughter, brother, sister, uncle, aunt, niece, nephew, grandmother, grandfather
Ingroup	acquaintance, ally, associate, colleague, counterpart, fellow, neighbor, patriot, confidant, companion, partner, supporter, member, follower
Marginalised	arab, beggar, blacks, crippled, disabled, jew, mexican, unemployed, native, elderly, indian, woman, chinese, pauper
Outgroup	emigrant, foreigner, intruder, settler, stranger, visitor, vagrant, opposition, rival, opponent, adversary, competitor, invader, occupier

Animal	animal, ape, bird, elephant, chicken, cow, dog, fish, pig, shark, bear, snake, monkey, lion
Environment	nature, forest, lake, mountain, ocean, reef, river, tree, sea, beach, island, coast, earth, planet

We drew on Leach et al. (2022) to compile a list of words denoting concern and indifference.[1] These words were originally compiled by consulting the wider literature on how moral concern is defined and measured. Moral concern can be defined as the level of consideration or precedence given to an entity's rights and welfare (Crimston et al., 2016). The empirical literature captures this in a number of ways, including whether entities are perceived to deserve care (Opotow, 1993; Piazza et al., 2014) and whether they evoke feelings of concern and sympathy (Crimston et al., 2016; Piazza et al., 2014; Waytz et al., 2019). We understand moral concern in similar terms: as being concerned for an entity and feeling sympathy for it. This conception captures the historical process of interest: expanding circles of moral concern and sympathy (Bloom, 2010; Pinker, 2018; Singer, 2011). We take indifference to be the opposite: being unconcerned for an entity and feeling indifferent towards it. On this basis, an initial list of 'seed' words was compiled (concern = *care, concern, compassion, sympathy*; indifference = *apathy, disregard, indifference, uncaring, unconcerned*; Leach et al., 2022). Following prior work (Liu, 2012), these lists were then expanded using WordNet (Miller, 1998). We only retained those words that had available vectors in all historic models. The final lists of words denoting concern and indifference are presented in Table 2.

[1] We conducted an additional analysis using words associated with the moral foundation of care (Graham et al., 2009, 2013; Frimer, 2019). This analysis conceptually replicated the findings reported in the Main Body (for further details see the Supporting Information).

Table 2

Words denoting concern and indifference.

Category	Words
Concern	care, cares, caring, cared, concern, concerns, concerned, concerning, compassion, compassionate, help, helps, helping, helped, sympathy, sympathize, sympathized, sympathetic
Indifference	apathy, unconcerned, disregard, disregarded, disregarding, indifference, indifferent, neglect, neglects, neglected, neglecting

Word Embeddings Association Test

We followed the logic of the Word Embeddings Association Test (WEAT; Caliskan et al., 2017) to provide a standardised estimate of the relative association between words representing entities (e.g., son), words representing concern (e.g., caring), and words representing indifference (e.g., uncaring). The WEAT is a widely-used method of estimating the relative association between groups of words. For example, Caliskan et al. (2017) used the WEAT to quantify the associations between words representing social groups (e.g., European-American names vs. African-American names) and words representing positivity and negativity (e.g., flower vs. hate). In a similar manner, Leach et al. (2022) used the WEAT to estimate the associations between words representing humans and animals (e.g., person vs. dog) and moral concern (e.g., care vs. apathy).

Following convention, we estimate the association between words by computing the cosine between two vectors (Caliskan et al., 2017; Charlesworth et al., 2021; Garg et al., 2018; Leach et al., 2022). The resulting cosine is bounded from negative one to positive one. A cosine of one indicates the embeddings are maximally similar, whereas a cosine of negative one indicates the embeddings are maximally dissimilar. We begin by computing the cosine between

an individual entity vector (e.g., son) and all the individual concern vectors (caring, concern, etc.). The cosine similarities (son-caring, son-concern, etc.) are then averaged to provide a mean association. We then do the same for the words vectors representing indifference (son-uncaring, son-unconcerned, etc.). Finally, we compute the difference between the two averages to provide a relative index of the strength of the association for the entity word (e.g. son) towards concern versus indifference. We repeat this process for each entity word. These scores are then standardised within models as a function of their standard deviation. A power analysis suggested these data (n_{entity} = 84) afforded 80% power (α = .050, two-tailed) to detect an effect of the following magnitude: r = .30. This seemed reasonable given the typical effect sizes observed in analogous work utilising word embeddings (rs > .30; e.g., Garg et al., 2018).

Models and data

We rely on the outputs of a set of pre-existing embedding models that are trained via a skip-gram procedure on decade-binned text from the Google Books Corpus and the Corpus of Historical American English spanning 1800-2000 (Hamilton et al., 2016). The Google Books Corpus is estimated to contain about eight million books totalling half a trillion English words (~6% of all books ever published), whilst the Corpus of Historical American English is a structured database containing about half a billion English words balanced by genre decade by decade. These corpora have been shown to contain traces of historical cultural trends in language (Lin et al., 2012; Michel et al., 2011; Pechenick et al., 2015). For example, the Google Books Corpus has shown that the frequency of words associated with slavery rose before the American Civil War and the civil rights movement (Michel et al., 2011).

We examined three models trained on the above corpora released via the HistWords project (Hamilton et al., 2016): the All English (1800s - 1990s) model trained on the full Google Books Corpus, the English Fiction (1800s - 1990s) model trained on a subset of the Google Books

Corpus, and the Genre-Balanced American English (1830s - 2000s) model trained on the Corpus of Historical American English. Because these models reflect the specific word co-occurrences in a given decade, they can be compared to examine differences in language over time. For example, the distance between *care* and *sister* in the 1940s embeddings might be larger than for the same words in the 1950s embeddings, suggesting that these words have a greater co-occurrence association in the 1950s compared to in the 1940s. All the HistWords models we analysed use orthogonal procrustes to align the learned embeddings so as to permit comparisons across models (Hamilton et al., 2016).

We focused our main analyses on the All English (1800s - 1990s) models because they were trained on the most comprehensive corpus. They encode about 100x more data than do the Genre-Balanced American English. Doing so allowed us to extend the historical analyses as far back in time as possible whilst maintaining the entire list of entity word vectors (1830s - 1990s, $n_{Entities}$ = 84), and therefore to present the most generalizable and broad picture of how language has changed over time. This can be contrasted with what is afforded by the English Fiction 1830s-1990s; $n_{Entities}$ = 38) and Genre-Balanced American English models (1830s-2000; $n_{Entities}$ = 54).

Results

Main analyses

We tested how moral concern WEAT scores have changed over time. To do this, we calculated WEAT scores for the same entities in each of the HistWords All English models 1830s-1990s (Hamilton et al., 2016). To account for the clustering of the data, we fit a mixed effects model predicting moral concern WEAT scores over time, where intercepts and slopes were allowed to vary randomly across entities. If language has shifted towards greater moral concern for others, we would expect the WEAT scores to have increased over time. Figure 1

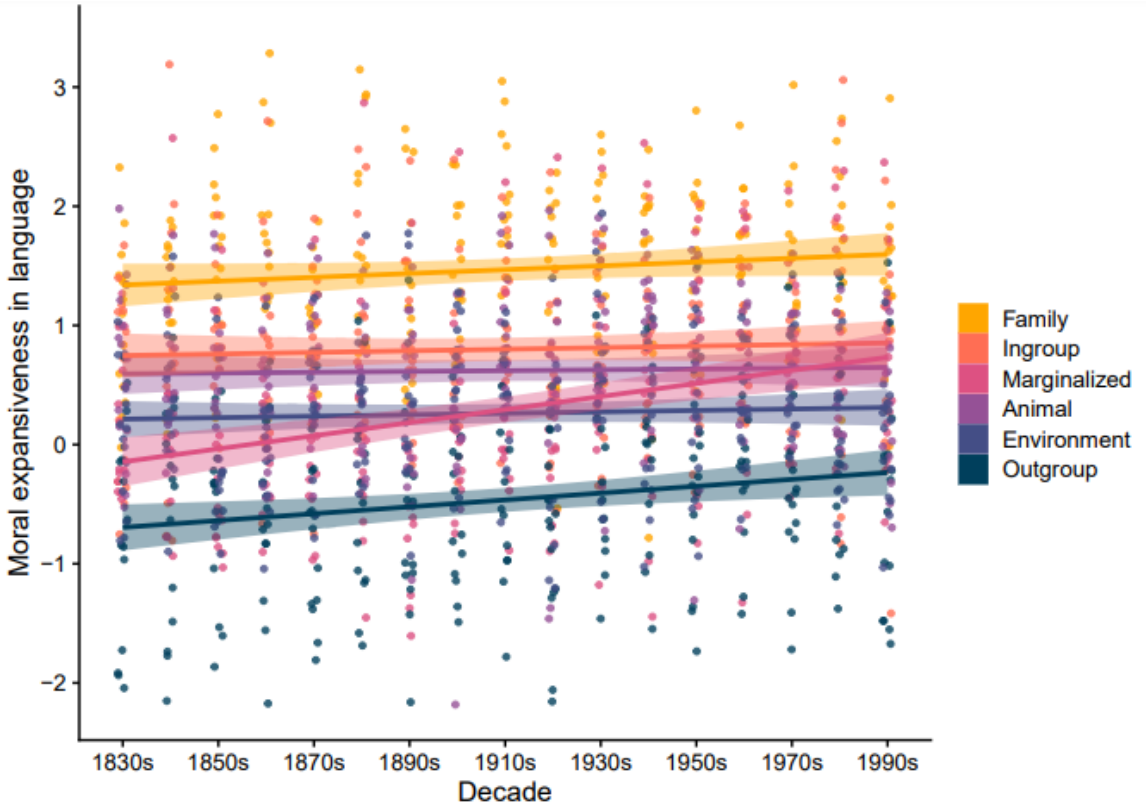
shows the trends over time. Looking across all groups, we found an overall upward trend, β = 0.12, SE = 0.03, 95% CI [0.05, 0.18], p = .001. These results suggest that language reflecting moral concern for others has increased over time.

Next, we examined the nature of these changes by testing whether WEAT scores have increased more, or less, for those entities that are typically afforded the most moral concern (i.e., those that are at the centre of the moral circle) compared to those that are afforded the least moral concern (i.e., those that are at the periphery of the moral circle). We first tested this in the most general way by ignoring the groups entities belonged to and simply testing if those entities with the lowest WEAT scores showed greater increases over time than those with the highest WEAT scores. A mixed effects quantile regression examining shifts in the 90^{th} quantile of WEAT scores compared to the 10^{th} quantile found no evidence that increases were any larger, or smaller, for entities at the centre of the moral circle (90^{th} quantile) compared to on the periphery (10^{th} quantile), β_{diff} = 0.04, 95% CI_{diff} [-0.14, 0.19]. This result points to a general *rising tide* rather than a specific or targeted increase in language reflecting moral concern for marginalised entities.

We also tested this by taking into account the entity groups. Figure 1 shows that expressions of moral concern have increased more towards at least one group at the periphery of the moral circle (marginalised others) compared to those closer to the centre (family members, ingroups), β = 0.25, SE = 0.11, 95% CI [0.02, 0.47], p = .038. Moreover, the particularly strong upward trend in moral concern for marginalised others does not entirely account for the overall trends, which remained evident even when excluding this group from the analyses, β = 0.08, SE = 0.03, 95% CI [0.01, 0.14], p = .026. Taken as a whole, these results support both views of moral progress: as one which reflects a general *rising tide* of improvements for all and as one which reflects more targeted and substantial improvements for those on the margins.

Figure 1

Moral expansiveness in language over time.



Note. Scores are derived from the Word Embeddings Association Test and reflect the average similarity between words denoting entities and words denoting concern (vs. indifference) in each decade. Similarity scores are extracted from the HistWords All English models 1830s-1990s (Hamilton et al., 2016).

Additional analyses

We conducted additional analyses to validate our approach and test alternative explanations. As a basic validity test, we examined which entities had the highest and lowest WEAT scores, irrespective of the time period. If our measure is valid, we would expect it to be

greatest for those entities that are typically found close to the centre of the moral circle (e.g., family members) compared to those that are on the periphery (e.g., outgroups; Crimston et al., 2016). Confirming this, WEAT scores were highest for family members (M = 1.48, SD = 0.75), followed by ingroups (M = 0.80, SD = 0.76), animals (M = 0.62, SD = 0.69), marginalised individuals (M = 0.29, SD = 0.88), the environment (M = 0.26, SD = 0.62), and finally, outgroups (M = -0.49, SD = 0.83), $F(5, 78) = 23.53$, $p < .001$, $\eta_p^2 = .60$. This pattern increases our confidence in the validity of the measure of language reflecting moral concern for others.

We also tested if there was any evidence of non-linear trends in the data, perhaps reflecting greater increases in concern for others in recent times. Adding a quadratic term did not improve model fit significantly compared to the simpler linear model, $\chi^2 = 3.62$, $\Delta df = 1$, $p = .057$. This suggests that the increases we observed in language are largely uniform across time.

The general upward trend in expressions of moral concern over time could be due to shifts in the sheer positivity of language referring to each entity, rather than moral concern. To explore this, we computed a new WEAT score for each entity reflecting its relative association with positive compared to negative words (positive = *caress*, *freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, gift, honor, miracle, sunrise, family, happy, laughter, paradise*; negative = *abuse, crash, filth, murder, sickness, accident, death, grief, poison, assault, disaster, hatred, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison*). We selected these on the basis of prior research, with the constraint that the corresponding vectors were available in all models (Caliskan et al., 2017). We found evidence of increasing positivity towards entities across time, $\beta = 0.10$, SE = 0.02, 95% CI [0.06, 0.15], $p < .001$. To test if our central findings could be accounted for by this upward trend in positivity, we re-analyzed shifts in moral concern whilst controlling for positivity. The upward trend in moral concern held when controlling for positivity, $\beta = 0.09$, SE = 0.04, 95% CI

[0.02, 0.14], *p* = .010. There was also no evidence to suggest that positivity in prior decades reliably predicted moral concern in subsequent decades, β = 0.01, SE = 0.03, 95% CI [-0.06, 0.08], p = .748. There was, however, evidence of the opposite. Moral concern predicted subsequent positivity, β = 0.06, SE = 0.02, 95% CI [0.02, 0.10], *p* = .005. These analyses suggest that the upward tick in language reflecting moral concern for others cannot be accounted for merely by shifts in the positivity of language. They also point to an interesting and unexpected insight about the potential effects of moral language on expressions of positivity–an idea we return to in the General Discussion.

Our results could be driven by an increase in the frequency with which entity words are mentioned. Numerous psychological accounts, including those that explain basic attitudinal processes (Zojonc, 1968) as well as prejudice and social cohesion (Allport, 1954), would predict that greater exposure to others can foster positive outcomes. In addition, it is plausible that a greater frequency of entity words in the underlying corpus could lead to stronger co-occurrence statistics. As such, we tested if the frequency of the entity words in each decade could account for the general upward trend in moral concern across time. The central effect persisted when including frequency as a covariate in the model, β = 0.14, SE = 0.04, 95% CI [0.07, 0.21], p < .001. This analysis suggests that the upward trend in language reflecting moral concern for others cannot be accounted for by a general increase in the frequency with which entity words are mentioned.

We examined the robustness of our findings by computing an alternative index of the association between words denoting entities and those denoting concern and indifference. The MAC is defined as the mean average cosine similarity (Manazini et al., 2019). This differs from the WEAT in that it is an unstandardized metric of the association between words. We found the same upward trend when computing this index, β = 0.005, SE = 0.001, 95% CI [0.00, 0.01], *p* =

.001. This result increases our confidence in the findings by showing they are not contingent on one specific metric of linguistic association.

Finally, we tested if similar patterns emerged in other corpora. Due to the limited availability of word vectors in the Google Books Fiction Corpus and the Corpus of Historical American English, we were only able to test for an upward trend in the association between a subset of the word vectors representing entities (*mother, father, husband, wife, son, daughter, sister, brother, niece, nephew, aunt, uncle, grandfather, acquaintance, member, fellow, companion, native, stranger, opposition, jew, woman, animal, dog, lion, fish, bear, bird, earth, forest, island, lake, mountain, coast , nature, river, sea, tree*) and moral concern (*care, compassion, concern, concerned, concerning, help, sympathy*). We found evidence of an overall upward trend in expressions of moral concern in the the Google Books Fiction Corpus, $\beta = 0.13$, SE = 0.03, 95% CI [0.08, 0.18], $p < .001$, and in the Genre-Balanced American English Corpus, $\beta = 0.13$, SE = 0.02, 95% CI [0.08, 0.18], $p < .001$. Importantly, these effects were indistinguishable from that of the Google Books Corpus, $\beta = 0.00$, SE = 0.03, 95% CI [-0.06, 0.06], $p = .981$; $\beta = -0.00$, SE = 0.03, 95% CI [-0.06, 0.06], $p = .966$. These results increase our confidence in the generalizability of the findings by suggesting that they are not constrained to a specific corpus of text.

General Discussion

Psychologists, philosophers, and other social scientists have argued for an intimate link between how we communicate about the welfare of others and the expansion of our moral circles (Bloom, 2010; Lecky, 1869; Nussbaum, 2007; Pinker, 2018; Singer, 2011), with some suggesting shifts in how we express concern for others can be considered an important indicator of moral progress (Kant, 1991). We tested this by analysing historical trends in the distributions

of words in natural language from the 1830s to the 1990s. Specifically, we examined whether words denoting concern (vs. indifference) have become more closely associated with words denoting people, animals, and the environment.

Growing expressions of concern for people, animals, and the environment in language

We found that words reflecting moral concern became more strongly associated with people, animals, and the environment over the 19th and 20th centuries. This finding adds to a body of work demonstrating historical changes in the usage of words denoting care and peace (Wheeler et al., 2019) and in the meaning of harm-related concepts (Haslam, 2016; Haslam et al., 2020; Vylomova et al., 2019). Shifts of this sort may reflect a greater preoccupation with harm and are likely spurred on by historically salient movements (e.g., towards gender equality, animal welfare, and environmental protection), including the rise of liberalism and civil rights discourses, which have served to bring the welfare of marginalised entities into focus. Taken as a whole, they corroborate long-standing perspectives about the nature of moral progress and how it presents itself in language (Kant, 1991). They are also in line with the broader hypothesis that moral concern is on an upward trajectory (Bloom, 2010; Pinker, 2011; Singer, 2011).

Our findings support multiple views of moral progress. By showing that language has changed in largely the same way for many entities on the margins of the moral circle (e.g., outgroups) as it has for those found at the centre (e.g., family), the results are consistent with the proverb *a rising tide raises all boats*. This corroborates perspectives which view moral progress in terms of general rises in the welfare of all driven by global improvements in material welfare, societal conditions, and the adoption of egalitarian and universalist discourse (Evans, 2017). On the other hand, the findings also show that language has changed more markedly for at least one group on the margin of the moral circle. Language directed towards marginalised others has shifted to such a degree that it is now indistinguishable from language directed

towards ingroups. This result is consistent with perspectives which view moral progress in terms of reductions in parochialism and improvements in the rights and welfare of those on the margins, such as might be brought about by collective social action and the reasoned extension of concern beyond close kin (Dixon et al., 2012; Singer, 1981).

Limitations, generalizability, and future directions

The results may indicate complimentary shifts in moral beliefs over time. This is because word embeddings derived from large corpora show strikingly similarities to beliefs and attitudes (Caliskan et al., 2017; Charlesworth et al., 2021). For example, the associations between words that represent genders and traits (e.g., he-assertive, she-assertive) correspond to people's stereotypes about men and women (Garg et al., 2018). Indeed, we found that moral concern was more strongly associated with family members than it was with outgroups, just as is the case when measured via self-reports (Crimston et al., 2016). Crucially, there is evidence that these linguistic associations change over time in a similar way as do beliefs captured by longitudinal survey data (Garg et al., 2018). This suggests that the strengthening linguistic associations between moral concern and people, animals, and the environment may reflect stronger convictions about their moral worth. This conclusion is tentative though, as the correspondence between word embeddings and self-reported beliefs over time is not strong and exactly why such a correspondence exists in the first place is not fully understood (Garg et al., 2018; Caliskan et al., 2020).

Our approach captures broad shifts in how words expressing concern are used in relation to others. The upward trends were generally quite modest and were typically overshadowed by differences between groups. They may also miss some important details of the historical process of interest. While word embeddings appear to capture important aspects of language and our collective psychologies (Caliskan et al., 2017; Caliskan et al., 2020), they

may not capture more nuanced and contextualised meanings. Language can reflect an expanding moral circle in many ways that do not require the use of words that explicitly express concern. For example, the novel Uncle Tom's Cabin (1852) is thought to have influenced attitudes towards slavery by vividly describing the reality of the institution (Bloom, 2012). Additionally, our analysis focused on large-scale shifts in language use. However, moral progress is often attributed to the publication of specific texts, such as Animal Liberation (Singer, 1978). It could be that the publication of such texts precede the more diffuse changes in language that we observed.

It is worth considering limits on the generalizability of our findings. We focused primarily on the Google Book Corpus which is estimated to contain about half a trillion words from 6% of all books ever published (Lin et al., 2012). We also found similar effects in a subset of the Google Book Corpus focused on fictional texts and in the Corpus of Historical American English. That similar patterns emerged in multiple corpora suggests that our findings are not a reflection of any idiosyncratic corpus-specific feature, such as the greater inclusion of scientific texts in the Google Book Corpus (Pechenick et al., 2015). It also suggests that the upward trend in expressions of moral concern for people, animals, and the environment reflect prevailing cultural representations. Finally, it is important to note that our conclusions are constrained to a single language, English. Future research may examine if similar trends are evident in different languages.

Our additional analyses revealed that expressions of moral concern predicted broader shifts in positivity, but not the other way around. This result promises an important insight into the nature of moral progress and how changes in moral values affect how we speak and write about others. It does so by suggesting that increases in moral concern for others may restrict the expression of negative attitudes and discrimination. This aligns with the perspective that

prejudice is constrained by beliefs, values, and societal norms (Crandall & Eshleman, 2003). Rhetoric which emphasises universal human rights and compassion for marginalised others may therefore drive moral progress by making public spaces inhospitable to those who explicitly express prejudice (Kant, 1981). This finding is of great potential interest and warrants replication and elaboration in future research

Conclusion

Our findings provide a new way to evaluate a widely-held belief about the nature of moral progress (Bloom, 2010; Lecky, 1869; Nussbaum, 2007; Pinker, 2018; Singer, 2011). By showing that language use has changed in a way that reflects greater concern for people, animals, and the environment, our results support optimistic views of history.

References

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Perseus Books

Bloom, P. (2010). How do morals change? *Nature*, *464*(7288), 490–490. https://doi.org/10.1038/464490a

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Buchanan, A. (2013). Moral progress and human rights. In C. Holder & D. Reidy (Eds.), *Human Rights: The Hard Questions* (pp. 399–417). Cambridge University Press. https://doi.org/10.1017/CBO9780511758553.029

Buttrick, N., Moulder, R., & Oishi, S. (2020). Historical change in the moral foundations of political persuasion. *Personality and Social Psychology Bulletin*, *46*(11), 1523–1537. https://doi.org/10.1177/0146167220907467

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Catalano, S. (2008). *Intimate partner violence in the United States*. U.S. Department of Justice.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit Attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science*, *30*(2), 174–192. https://doi.org/10.1177/0956797618813087

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240. https://doi.org/10.1177/0956797620963619

Crimston, C. R., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology*, *111*(4), 636–653. https://doi.org/10.1037/pspp0000086

Dixon, J., Levine, M., Reicher, S., & Durrheim, K. (2012). Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution? *Behavioral and Brain Sciences*, *35*(6), 411–425. https://doi.org/10.1017/S0140525X11002214

Drescher, S. (2009). *Abolition: A History of Slavery and Antislavery*. Cambridge University Press.

Evans, J. (2017). A working definition of moral progress. *Ethical Theory and Moral Practice*, *20*(1), 75–92. https://doi.org/10.1007/s10677-016-9753-2

Finkelhor, D., Jones, L., & Shattuck, A. (2010). Updated Trends in Child Maltreatment, 2008. *Crimes Against Children Research Center*. https://scholars.unh.edu/ccrc/57

Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehghani, M. (2019). *Moral Foundations Dictionary 2.0*. https://doi.org/10.17605/OSF.IO/EZN37

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/PNAS.1720347115

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. *Proceedings of IJCAI Workshop on Computational Modeling of Attitudes*. http://morteza-dehghani.net/wp-content/uploads/morality-lines-detecting.pdf

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01316-8

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science*, *24*(9), 1722–1731. https://doi.org/10.1177/0956797613479387

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, May). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of ACL*. http://arxiv.org/abs/1605.09096

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, *27*(1), 1–17. https://doi.org/10.1080/1047840X.2016.1082418

Haslam, N., Dakin, B. C., Fabiano, F., McGrath, M. J., Rhee, J., Vylomova, E., Weaving, M., & Wheeler, M. A. (2020). Harm inflation: Making sense of concept creep. *European Review of Social Psychology*, *31*(1), 254–286. https://doi.org/10.1080/10463283.2020.1796080

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. https://arxiv.org/pdf/1607.01759.pdf

Jurafsky, D., & Martin, J. H. (2019). Vector Semantics and Embeddings. In *Speech and Language*

Processing (3rd ed. Draft). https://web.stanford.edu/ jurafsky/slp3/6.pdf

Kant, I. (1981). *Lectures on ethics*. Hackett Publishing Co.

Kant, I. (1991). Kant: Political writings. In H. S. Reiss (Ed.), *Idea for a universal history with a cosmopolitan purpose* (pp. 41–53). Cambridge University Press.

Laham, S. M. (2009). Expanding the moral circle: Inclusion and exclusion mindsets and the circle of moral regard. *Journal of Experimental Social Psychology*, *45*(1), 250–253. https://doi.org/10.1016/J.JESP.2008.08.012

Leach, S., Kitchin, A., Sutton, R., & Dhont, K. (2021). Speciesism in everyday language. British Journal of Social Psychology, 00, 1–17. https://doi.org/10.1111/bjso.12561

Lecky, W. E. H. (1869). *History of European morals from Augustus to Charlemagne*. Longmans, Green and Co.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028. https://doi.org/10.1038/s41562-020-0918-6

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., & Petrov, S. (2012). *Syntactic annotations for the google books ngram corpus*. 6.

Macklin, R. (1977). Moral progress. *Ethics*, *87*(4), 370–382. https://doi.org/10.1086/292049

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57. https://doi.org/10.1017/s0140525x10000968

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P.,

Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January). *Efficient estimation of word representations in vector space*. http://arxiv.org/abs/1301.3781

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *ArXiv:1712.09405 [Cs]*. http://arxiv.org/abs/1712.09405

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of NIPS*. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT*. https://www.aclweb.org/anthology/N13-1090.pdf

Moody-Adams, M. M. (1999). The idea of moral progress. *Metaphilosophy*, *30*(3), 168–185. https://doi.org/10.1111/1467-9973.00120

Musschenga, A. W., & Meynen, G. (2017). Moral progress: An introduction. *Ethical Theory and Moral Practice*, *20*(1), 3–15.

Nussbaum, M. (2007). On moral progress: A response to Richard Rorty. *The University of Chicago Law Review*, *74*(3), 939–960.

Opotow, S. (1993). Animals and the Scope of Justice. *Journal of Social Issues*, *49*(1), 71–85. https://doi.org/10.1111/j.1540-4560.1993.tb00909.x

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus:

Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, *10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the EMNLP*. https://www.aclweb.org/anthology/D14-1162.pdf

Piazza, J., Landy, J. F., & Goodwin, G. P. (2014). Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition*, *131*(1), 108–124. https://doi.org/10.1016/J.COGNITION.2013.12.013

Pinker, S. (2011). *The Better Angels of Our Nature: A History of Violence and Humanity*. Penguin.

Pinker, S. (2018). *Enlightenment now: The case for reason, science, humanism, and progress*. Penguin.

Rhee, J. J., Schein, C., & Bastian, B. (2019). The what, how, and why of moralization: A review of current definitions, methods, and evidence in moralization research. *Social and Personality Psychology Compass*, *13*(12). https://doi.org/10.1111/spc3.12511

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), 50. https://doi.org/10.1525/collabra.282

Roser, M., & Nagdy, M. (2014). Optimism and Pessimism. *Our World in Data*. https://ourworldindata.org/optimism-pessimism

Rosling, H., Rosling, O., & Rosling, R., A. (2018). *Factfulness: Ten Reasons We're Wrong About the World-and Why Things Are Better Than You Think*. Flatiron Books.

Ruby, M. B. (2012). Vegetarianism. A blossoming field of study. *Appetite*, *58*(1), 141–150. https://doi.org/10.1016/j.appet.2011.09.019

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, *32*(2), 132–144. https://doi.org/10.1177/0894439313506837

Singer, P. (1975). *Animal liberation*. Harper Perennial.

Singer, P. (1981). *The Expanding Circle: Ethics, Evolution, and Moral Progress*. Princeton University Press.

Stoknes, P. E. (2015). *What we think about when we try not to think about global warming: Toward a new psychology of climate action*. Chelsea Green Publishing Co. https://doi.org/10.5860/choice.192627

Stowe, H. B. (1852). *Uncle Tom's Cabin*. Ingram, Cooke, & CO.

Turgot, A.-R.-J. (1973). *A philosophical review of the successive advances of the human mind*. Cambridge University Press.

Vylomova, E., Murphy, S., & Haslam, N. (2019). Evaluation of semantic change of harm-related concepts in psychology. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 29–34. https://doi.org/10.18653/v1/W19-4704

Waytz, A., Iyer, R., Young, L., Haidt, J., & Graham, J. (2019). Ideological differences in the expanse of the moral circle. *Nature Communications*, *10*(1), 4389. https://doi.org/10.1038/s41467-019-12227-0

Wheeler, M. A., McGrath, M. J., & Haslam, N. (2019). Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. *PLOS ONE*, *14*(2), e0212267. https://doi.org/10.1371/journal.pone.0212267

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social*

Psychology, 9(2), 1–27. https://doi.org/10.1037/h0025848