Evolution of moral semantics through metaphorization

Aida Ramezani (armzn@cs.toronto.edu)
Department of Computer Science, University of Toronto

Jennifer E. Stellar (jennifer.stellar@utoronto.ca)
Department of Psychology, University of Toronto

Matthew Feinberg (matthew.feinberg@rotman.utoronto.ca)
Rotman School of Management, University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

Although language is critical to supporting morality within society, it is not clear how moral language itself evolved. We investigate the evolution of moral semantics, hypothesizing that words evolved to take on moral meanings from concrete experiences through metaphorization. We test this hypothesis by analyzing moral semantic change in words from the Moral Foundations Dictionary and the Historical Thesaurus of English over the past hundreds of years. In contrast with the observation that words become concrete over time, we demonstrate that moral words in the English lexicon undergo concrete-to-abstract shifts, reflecting systematic metaphorical mappings to the moral domain. Our results provide large-scale evidence for the role of metaphor in the historical development of the English moral lexicon.

Keywords: lexical evolution; moral semantics; moral foundations; semantic change; metaphor

Introduction

Morality is a fundamental aspect of human society. From an evolutionary perspective, the emergence of morality helped early humans survive by allowing them to effectively take advantage of the benefits of group living (Haidt, 2007), and create mutual and shared expectations of how to treat one another (Tomasello, 2016; Tomasello, Melis, Tennie, Wyman, & Herrmann, 2012). As a result, humans have been able to exploit collective opportunities (e.g., big game hunting) and defend against collective threats (e.g., invasion by other groups). How exactly humans evolved these complex moral systems is not well understood, but many believe that it stemmed, in part, from language. Through communication, language may have facilitated the emergence of shared systems of moral norms and helped uphold them through rewarding moral behaviour and punishing immoral behaviour (Li & Tomasello, 2021; Poulshock, 2006). Despite the fundamental role of language in supporting morality, how moral language itself has developed over time is not clear.

We study the historical development of moral language by asking how words acquire moral meanings over time. We propose that the evolution of moral word meanings might critically depend on metaphorization, or our cognitive capacity to ground abstract moral thoughts in concrete experiences. Our theorizing builds on insights into the relationship between morality and metaphor from Conceptual Metaphor Theory (Lakoff & Johnson, 1980). This theory holds that people interpret abstract domains through concrete and perceptual experiences, which is enabled by our ability to identify structural similarities between domains (Gentner, 1983). Building on this notion, Moral Politics Theory argues that people develop abstract moral-political beliefs based on their understanding of family dynamics (i.e., family models) (Lakoff, 1996). Recent work in social psychology has proposed similar views suggesting that moral concerns (e.g., purity) are associated with concrete and physical experiences (e.g., physical dirtiness) (Lee & Schwarz, 2010b, 2021, 2010a).

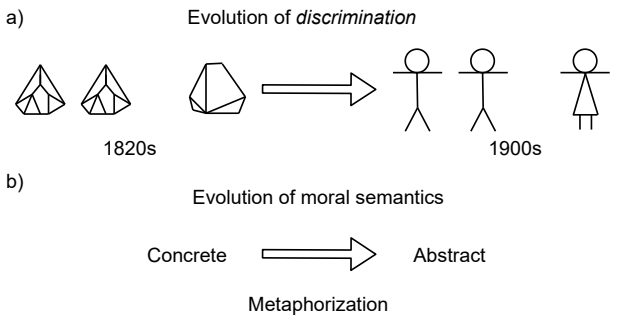

Figure 1: An illustration and overview of our hypothesis. a) Illustration of moral semantic change in the English word *discrimination*. In the 1820s, *discrimination* was used to indicate physical differences between objects (e.g., size of rocks). In the 1900s, *discrimination* evolved toward prejudices in society (e.g., gender discrimination). b) Overview of our hypothesis.

Although empirical work has shed light on the cognitive role of metaphor in moral judgment, to our knowledge there has been no comprehensive study on examining the role of metaphor in the evolution of moral word meanings. We hypothesize that the evolution of moral semantics has been made possible through metaphorical mappings. Figure 1 illustrates our overall hypothesis and provides an example of metaphorical semantic change in the English moral word *discrimination*. While expressing the concrete notion of dis-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

parity among physical objects 200 years ago, *discrimination* took on the more abstract moral meaning of social disparity about a century later presumably because people metaphorically mapped its meaning from the physical (source) domain to the moral (target) domain. This concrete-to-abstract meaning shift has been shown to be a primary force in the historical metaphorical mappings in English (Xu, Malt, & Srinivasan, 2017), and if words acquire moral meanings through metaphorization, we expect their historical trajectories to exhibit similar directionality. Our hypothesis is at odds with the observation that the English lexicon has a general tendency of gaining concreteness over time (Hills & Adelman, 2015; Snefjella, Généreux, & Kuperman, 2019), and here we investigate the possibility that moral words tend to follow the opposite trend by shifting from concrete to abstract meanings. We also examine valence (i.e., degree of pleasantness to capture the sentiment of a word) as a secondary factor to concreteness in metaphorical mapping (Xu et al., 2017), based also on the existing literature on Concept Creep suggesting that harm-related concepts tend to undergo negative sentiment change due to semantic expansion (Haslam, 2016). We therefore expect the semantic environment of a moral word to show a decrease in valence over time, although it is yet to be determined whether concreteness or valence might play the more dominant role in shaping moral semantic change.

Data

To test our hypothesis, we consider one of the largest lexical resources for contemporary moral terminologies in English ($n = 1,354$), The Moral Foundations Dictionary (MFD), which was developed to test the larger Moral Foundations Theory (MFT) (Graham et al., 2013). Moral Foundations Theory proposes five conceptual foundations to morality (care, fairness, authority, loyalty, and sanctity) each comprised of unique virtues and vices. The MFD provides a large set of English moral words for each moral foundation. For example, *compassion* is a virtue in the care category, while *brutality* is a vice in this same category. The MFD has been used widely and effectively in formal analyses of morality and language (Garten, Boghrati, Hoover, Johnson, & Dehghani, 2016; Mooijman, Hoover, Lin, Ji, & Dehghani, 2018; Hoover et al., 2020; Xie, Ferreira Pinto Junior, Hirst, & Xu, 2019; Mendelsohn, Tsvetkov, & Jurafsky, 2020; Ramezani, Zhu, Rudzicz, & Xu, 2021). We chose the MFD version 2 as the basis of the modern English moral lexicon because of its theoretical and empirical validity (Frimer, Haidt, Graham, Dehghani, & Boghrati, 2017), and that it covers a wider range of vocabulary. We also consider a second large resource of the moral lexicon drawing words from the Historical Thesaurus of English (HTE) (Alexander, Kay, Roberts, Samuels, & Wotherspoon, 2015). This resource provides a collection of moral words ($n = 1,722$) from the history of the English language. To collect these words from HTE, we extracted all the terms in HTE under the 'Society.Morality' category that were non-overlapping with MFD. For simplicity, we excluded all the compound words and phrases.

Additionally, we used the Metaphor Map of English (MME) (Alexander et al., 2015) database to study historical metaphorical mappings of the moral domain. The MME is one of the largest resources documenting metaphorical mappings between 415 semantic domains in the historical development of English words over the past millennium. We focused on the mappings that involve a sub-category of the 'Morality' semantic domain as either the source or the target. For example, the MME database identifies a metaphorical mapping from 'Direction' (source domain) to 'Virtue' (target domain), and lists the word *direct* as an exemplar for this metaphorization. Independently to the database, we extracted the human-annotated concreteness ratings of the semantic domains in MME from existing work (Xu et al., 2017).

Methods

To quantify word meanings from text through historical periods, we apply established methodologies from natural language processing to construct diachronic word embeddings (Hamilton, Leskovec, & Jurafsky, 2016). At each decade in between 1800 and 1990, a word embedding provides a latent high-dimensional representation of that word's meaning from its usages (or contexts), which allow us to track meaning changes over historical times.[1]

To analyze moral semantic change, we use recent methods from natural language processing to construct moral relevance time courses based on the diachronic embeddings and the Centroid model (Xie et al., 2019). The Centroid model is a classifier that consists of two centroids, 'moral' and 'neutral'. The 'moral' centroid is the average word embedding of moral words from the MFD, and the 'neutral' centroid is the average word embedding of a set of neutral words taken from large-scale empirical ratings of valence (Warriner, Kuperman, & Brysbaert, 2013). The word embeddings for the centroids were taken from the most contemporary decade in the diachronic word embeddings model (i.e., 1990s) (Hamilton et al., 2016). The degree of proximity of a word to the moral centroid determines its degree of moral relevance. At time point t, the proximity of word w to the centroid c is determined by the probability $p(c|w,t) \propto exp(-\|V_t(w) - E[S_c]\|)$, where $V_t(w)$ is the word embeddings for w at time t taken from the diachronic embeddings, and $E[S_c]$ is the word embeddings for the centroid c. The proximities are then converted to probabilities using a softmax function and used for classification. To estimate the time course of moral relevance for a word similar to the examples presented in Figure 2a, we applied this model incrementally at each decade from the 1800s to the 1990s.

Since the Google Ngrams historical text corpora (Lin et al., 2012) used to derive the diachronic embeddings spans only the period 1800-2000, we focused on analyzing words that underwent moral semantic change during this period. We identify these words by calculating the gross change

[1] Our code is available at `https://osf.io/mnsjk/`.

in moral relevance of a word between the flanking decades 1800s and 1990s. Specifically, we defined the gross change for each word w by subtracting its moral relevance at the initial time point (the 1800s, i.e., $p(moral|w,t=1800)$) from its moral relevance at the terminal time point (the 1990s, i.e., $p(moral|w,t=1990)$), and divided that quantity by the moral relevance at the initial time point. Formally, gross change is defined as $M(w) = \frac{p(moral|w,t=1990)-p(moral|w,t=1800)}{p(moral|w,t=1800)}$. We used bootstrapping to construct the distribution of moral relevance change for the word populations and partitioned the words into 'Stable' and 'Changing' groups. To do so, we resampled the $M(w)$ distribution of the words 100,000 times. Denoting μ and σ as the original mean and the standard deviation of the mean of the new samples respectively, the $M(w)$ of 'Changing' words is above $\mu+2\sigma$ (i.e., exhibiting meaning change significantly above the population average), and for the 'Stable' words it is between $\mu-2\sigma$ and $\mu+2\sigma$. Overall we identified 375 'Changing' words from MFD and 429 words from HTE. We focused our analyses on the 'Changing' group, as these are the words that have faced a significant increase in their moral semantics during 1800-2000.

To detect the time point (i.e., a decade) of moral semantic change, for each changing word, we applied an automatic change point detection algorithm (Kulkarni, Al-Rfou, Perozzi, & Skiena, 2015). This algorithm identifies the most significant change point given a time series, such that the mean of the time series after the change point is significantly different from the mean of the time series before the change point.

Historical evidence for concrete-to-abstract shift in moral words

To evaluate our hypothesis, we analyzed the degree of concreteness change before and after the change point and compared these paired concreteness values over the population of words we analyzed. We estimated concreteness of a word w at time point t by first finding its nearest semantic neighbours using the diachronic embeddings at t. We then took an average of its neighbours' concreteness ratings from an existing empirical study (Brysbaert, Warriner, & Kuperman, 2014) as an approximation of concreteness for that target word. We set $k = 100$ nearest neighbours for all the experiments reported. If none of a word's nearest neighbours' concreteness ratings are available, we would remove this word from the analyses.

We then calculated the concreteness of each word in the changing group before and after the change point and compared these paired concreteness values over the population of changing words we analyzed. Paired t-tests revealed that moral words are significantly more concrete before the change points as the degree of concreteness drops when they undergo moral semantic change ($t(375) = 4.38,\ p < 0.001$ for MFD; $t(429) = 7.76,\ p < 0.001$ for HTE).

Figure 2a illustrates our analysis using two example words. In each case, we observed that concreteness dropped in the semantic environment of a word as it acquired moral meanings after the historical change point was detected. For instance, the word *molestation* neighboured words such as *annoy* and *detain* prior to the 1860s, but neighboured words such as *sociability* and *intimidation* by the 1890s. We also repeated this analysis in the stable word groups and found that the degrees of concreteness in these words tend to remain relatively unchanged ($t(42) = 0.53,\ p = 0.3$ for MFD; $t(50) = 0.72,\ p = 0.24$ for HTE). These results provide support to the idea that modern English moral words originally had concrete meanings and gained moral semantics by abstraction.

To examine whether the concrete-to-abstract changes were specifically tied to the change points, we performed a randomized test on the degree of concreteness change. We defined the change in concreteness (denoted as $\Delta concreteness$) by taking the difference in concreteness of a word's semantic neighbours before and after the change point. Figure 2b shows the distributions of $\Delta concreteness$ for the word populations in MFD and HTE centering around negative values, which indicates an overall trend for concreteness to drop around the change point. For the randomized test, we repeated the t-test analysis from the previous experiment at random time points (instead of the change points) for 1,000 trials. We used these randomized statistics (i.e., t-values) to construct a null distribution for concreteness change. Figure 2c compares the null distribution from the random trials with the attested t-value obtained by anchoring at the change point. In both MFD and HTE, our results show that the concreteness change at the point of change, denoted by the t-values obtained initially ($t(375) = 4.38$, for MFD; $t(429) = 7.76$), is significantly greater than that at random time points ($Z = 1.95,\ p = 0.025$ for MFD; $Z = 6.47,\ P < 0.001$ for HTE). This analysis confirms that there is a significant difference between the concrete-to-abstract change at the moral change point versus at random time points. The results further support that moral semantic change follows the concrete-to-abstract pattern, (rather than abstract-to-abstract), implying that moral semantics are grounded in concrete experiences. We hypothesize that the semantic change at the detected change points is associated with metaphorization, which is generally recognized by a concrete-to-abstract shift.

Moreover, we examined whether valence also plays a role in moral semantic change inspired by the theory of Concept Creep (Haslam, 2016). We found that valence and concreteness are orthogonal dimensions, confirmed in the insignificant results from χ^2 tests of independence between human ratings of concreteness and valence ($\chi^2 = 0.61,\ P = 0.43$ for MFD; $\chi^2 = 0.06,\ P = 0.8$ for HTE), where we took both kinds of ratings from existing large-scale behavioral experiments (Brysbaert et al., 2014; Warriner et al., 2013).

Similar to our analysis on concreteness change, we measured the valence change of a word by considering the average valence ratings of its 100 semantic neighbours before and after the change point. For this analysis, we also partitioned the words into morally positive and negative groups based on the available valence ratings or the MFD catego-

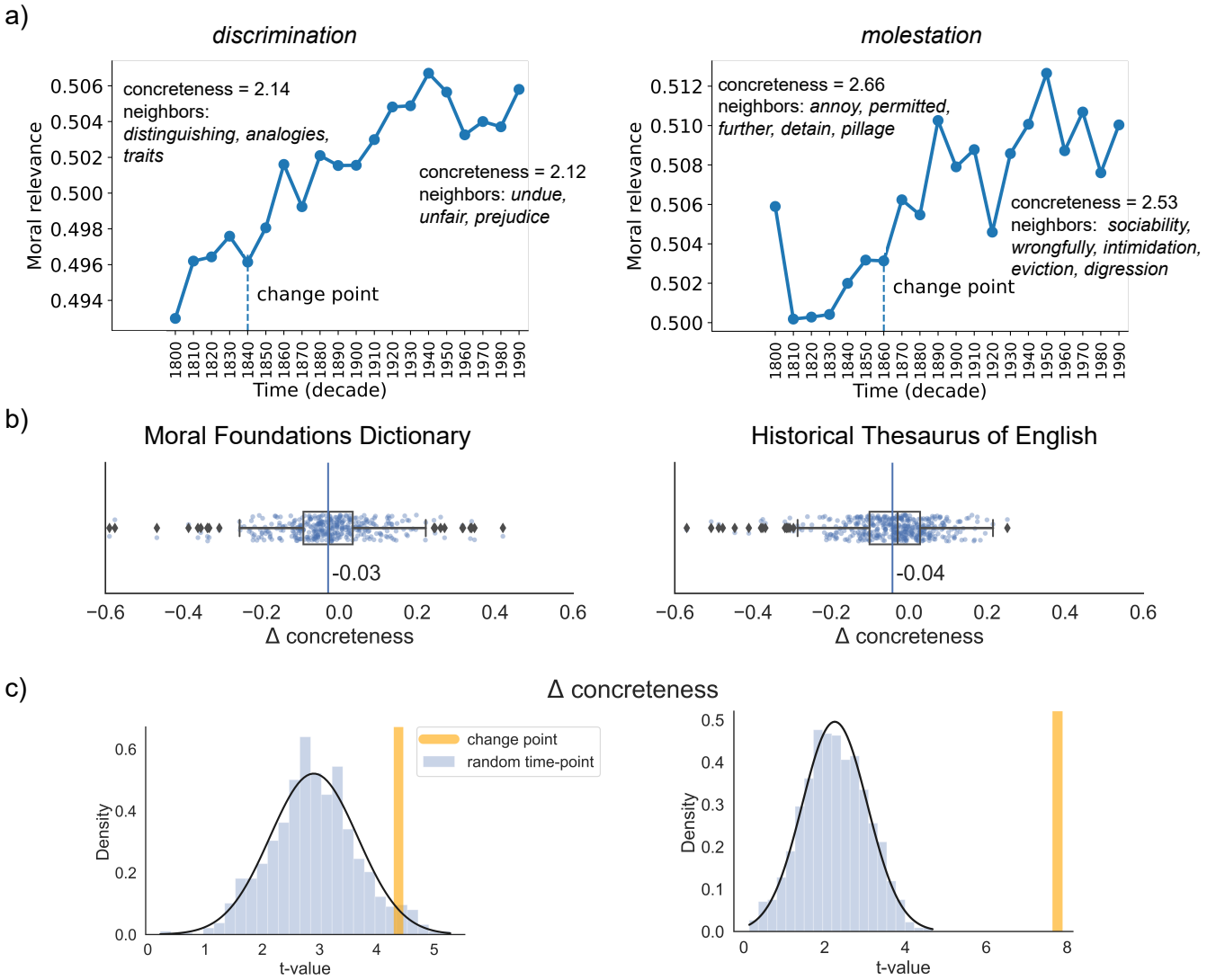

Figure 2: Summary of evidence for concrete-to-abstract shift in moral words. a) Illustration of concreteness shift in moral semantic change of two example English words *discrimination* and *molestation*. Each panel shows the time course of moral relevance of a word's meaning over the past two centuries. The vertical dash line marks the change point detected in the time course. b) Boxplots of concreteness change in the population of words that exhibit moral semantic change. Each dot shows the change in concreteness of a word's semantic neighbours before and after the change point in its time course of moral relevance. The vertical line marks the mean indicating a statistical tendency of concrete-to-abstract shifts around the change points. c) Randomized test on concreteness change in the moral words. The yellow vertical bar shows the attested t-value from a paired t-test on concreteness differences before and after the change points in the moral time courses of words taken from the two resources. The histogram shows a null distribution of similar t-values derived from using random time points instead of the actual change points.

rization. Table 1 summarizes the results from the paired t-tests. We observed significant changes in both concreteness and valence for the morally negative words (e.g., *spoiled* shows a decrease in both concreteness and valence), but only significant changes in concreteness for the morally positive words (e.g., *impartial* shows a decrease in concreteness but not in valence). The valence change in positive moral words was statistically insignificant ($t(240) = 0.374$, $p = 0.354$ for MFD; $t(184) = -1.284$, $p = 0.9$ for HTE). These results indicate that concreteness change is central to the overall evolution of moral semantics, whereas valence change is mostly relevant to negative change.

Source of moral words	Moral polarity	$\Delta Concreteness$	$\Delta Valence$
Moral Foundations Dictionary	Virtue ($n = 240$)	$t = 3.366$ ($p < 0.001$)	$t = 0.374$ ($p = 0.354$, n.s.)
	Vice ($n = 134$)	$t = 2.795$ ($p < 0.01$)	$t = 5.709$ ($p < 0.001$)
Historical Thesaurus of English	Virtue ($n = 184$)	$t = 4.849$ ($p < 0.001$)	$t = -1.284$ ($p = 0.9$, n.s.)
	Vice ($n = 129$)	$t = 4.165$ ($p < 0.001$)	$t = 2.465$ ($p < 0.01$)

Table 1: Comparative statistics of historical changes in concreteness versus valence of moral words.

Asymmetries in the historical metaphorical mappings of the moral domain

Our results so far were informed by the analyses of historical text corpora that span the period 1800-2000. We now extend our period of investigation to the past millennium by examining a database that records metaphorical mappings of the moral domain through the historical development of English.

We queried the Metaphor Map of English (MME) database, which is a dictionary-based resource of more than 14,000 metaphorical mappings from Old English to the present day recorded by lexicographers that have expertise in different periods of English (Alexander et al., 2015). Each recorded metaphorical mapping in MME is annotated with a source domain and a target domain, along with the directionality of mapping (e.g., source→target) and example English words that were identified to exemplify that metaphorical semantic change (e.g., the word *unhuman* exemplifies the metaphorical mapping from 'Humankind' to 'Moral Evil'). For our analysis, we considered all 273 recorded cases of metaphorical mapping that include the domain of 'Morality' as either target domain or source domain. In addition, we took the human ratings of concreteness for all the domains concerning these recorded cases from existing work on MME (Xu et al., 2017).

To evaluate our hypothesis, we focused on examining the asymmetries in the metaphorical mappings of morality in two respects: 1) asymmetry in the directionality of metaphorical mapping, namely whether domains of morality predominantly serve as the target domain as opposed to the source domain in the recorded cases of metaphorical mapping; 2) asymmetry in concreteness, namely whether in cases of metaphor where morality is the target domain, the source domain tends to be more concrete than the target domain.

Figure 3 summarizes the results. Regarding asymmetry in directionality, we found that 242 out of 273 cases recorded the metaphorical mapping direction to be $X \rightarrow$ moral domain, where X is a non-moral source domain, and only 31 cases in the opposite direction where a moral domain serves as the source. This result shows a significant asymmetry (binomial test $p < 0.0001$, $n = 273$) in the directionality of metaphorical mappings into the moral domain, which confirms that morality is predominantly the target domain in the historical process of metaphorization for English moral words. Regarding asymmetry in concreteness, we found that when morality is the target domain (i.e., a word gains a morally relevant meaning through metaphor), the degree of concreteness is significantly higher in the non-moral source domains than in the moral target domains ($t(242) = -33.37$, $p < 0.001$). This result confirms our corpus-based evidence and supports the view that word semantic change in the moral lexicon undergoes a concrete-to-abstract shift through the process of metaphorization. We also found that when morality is the source domain, the target domain is significantly more concrete ($t(31) = 6.54$, $p < 0.001$), suggesting that it is possible for moral words to extend toward concrete meanings (e.g., the word *virgin* has undergone a metaphorical semantic change from the 'Virtue' domain to 'Time').

This set of results provides direct evidence for the role of metaphor in the evolution of moral word meanings in the English lexicon, particularly how moral semantic change has happened by metaphorization.

Discussion

Previous work demonstrated that there is an increasing tendency toward using concrete words over abstract words in English (Hills & Adelman, 2015; Snefjella et al., 2019) and that word meanings typically move from abstract to concrete over time (Snefjella et al., 2019). However, when it comes to moral words, we found the opposite pattern suggesting that our findings cannot be explained by the general trend in the evolution of English language. The present research goes beyond past work (Wheeler, McGrath, & Haslam, 2019) to reveal how words in the MFD and HTE have changed meaning over time. It is important to note the exact causal relationships among these entities may be quite complex and potentially bidirectional. For instance, in our study and elsewhere (Xu et al., 2017), it has been found that word meanings can transfer from the moral domain to more concrete domains (e.g., the case of *virgin* after the semantic change period illustrated in Figure 3). Unraveling these dynamic interactions can be a fruitful direction for future work.

We found that the processes of moral semantic change involve changes in both the concreteness and valence (or pleasantness) of words, which is consistent with the claims for harm-related concepts from the Concept Creep theory (Haslam, 2016) (though the process of abstraction is notably more prevalent than valence). Furthermore, our findings resonate with work on semantic change suggesting metaphor as a key mechanism in historical meaning change (Sweetser, 1990), whereby semantic changes in the moral domain can be a manifestation of that general cognitive process.

Generally, our results provide evidence for determining the significant role of metaphorical mappings through grounding moral meanings in concrete experiences. We argue that

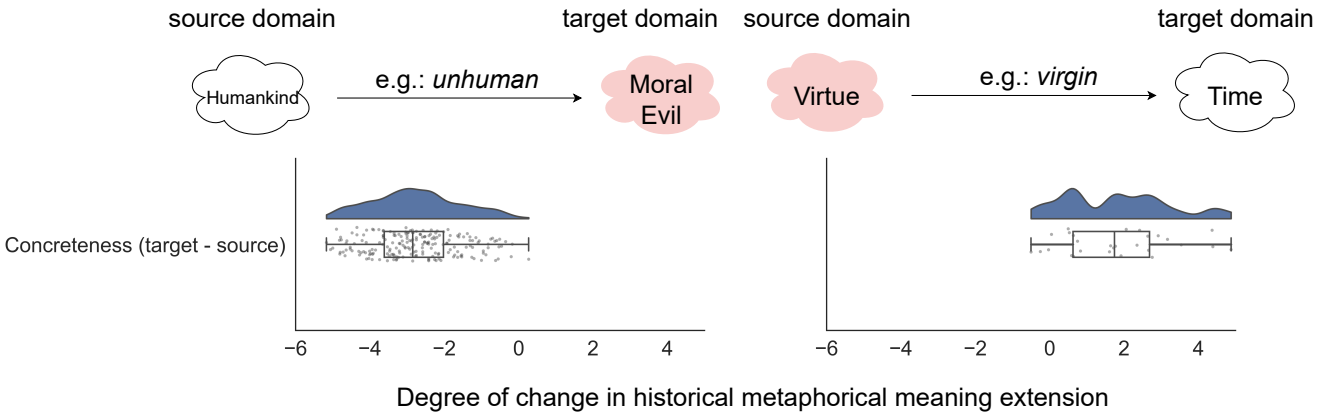

Figure 3: Evidence for asymmetries in the historical metaphorical mappings of the moral domain. The left panel shows that in a metaphorical mapping with morality as the target domain (e.g., 'Moral Evil'), the source domain tends to be more concrete (e.g., 'Human'). The right panel shows that in a metaphorical mapping with morality as the source domain (e.g., 'Virtue'), the target domain tends to be more concrete (e.g., 'Time'). The number of metaphorical mappings in this case is substantially lower than the case where morality is the target domain.

the observed patterns throughout moral semantic change, are associated with metaphorization more than other semantic change alternatives. For example, semantic restriction or semantic expansion cannot explain the concrete-to-abstract shift. Moreover, in semantic extension, the scope of a word's meaning expands to be used in a new broad context. Based on this notion, if moral semantic change is a semantic extension, then the new meaning (moral meaning) should be a more general version of the old meaning. This assumption is similar to the theory of Concept Creep, but does not explain the semantic change observed in positive cases of morality (Haslam, 2016). For moral semantic change to be a case of semantic restriction, the new moral meaning should be a specific case of the old meaning. Therefore, the frequency of the word with the specific meaning should be less than the broad meaning. However, the result of our analysis implies a significant increase in the frequency after moral semantic change ($t(375) = 3.85$, $p < 0.001$ for MFD; $t(429) = 2.00$, $p < 0.05$ for HTE). Another type of semantic change, which is applicable to morality, is semantic pejoration. During semantic pejoration, the sense of a word takes on more negative properties. As the result of our valence analysis implies, this can be a case for negative moral words. However, the concrete-to-abstract phenomenon is still captured in these words, suggesting a co-evolution of metaphorical mapping and semantic pejoration at the same time. We did not find any substantial evidence for word meanings to gain valence during moralization, therefore, semantic amelioration (i.e., words getting positive evaluations during the semantic change) is unlikely to explain the general trend of moral semantic change.

Regardless, we acknowledge that the patterns identified reflect strong statistical tendencies but not rigid laws, and propose a general association between moralization and metaphorization which has been possible through grounding moral meanings of words in their old concrete meanings.

Finally, the change points we focused on are not the only historical times that words undergo semantic change. A word might moralize at some point in history yet experience other kinds of semantic change at different times. For instance, we did observe a tendency for concreteness to rise in later decades after the point of moralization, which aligns with existing studies suggesting a general increase in concreteness in the English lexicon. This concreteness change may be due partly to the fact that some moral words become more frequently used and applied to describe everyday scenarios.

In sum, our work connects and extends two existing programs of research. Within philosophy, it connects to work examining the relationship between language and morality (Li & Tomasello, 2021; Poulshock, 2006), which points to language as an impetus for moral development. Within cognitive and social psychology it ties to concepts of grounding in morality (Lakoff, 1996; Lee & Schwarz, 2010a, 2010b, 2021), which proposes that metaphor influences moral judgment and behaviour. While language may have helped morality evolve, our work suggests that cognitive processes such as metaphorization may have helped the language for morality evolve.

Conclusion

Our work offers the first comprehensive quantitative study on moral semantic change in English. Our analyses of two large resources of moral vocabulary reveal that moral words tend to originate from concrete meanings and undergo metaphorical semantic change. Future work can assess the generality of our findings to languages other than English and explore questions such as why certain words became moralized over time (e.g., *discrimination*), whereas others did not (e.g., *difference*). Such lexicalization strategies may be similar or different across languages, which opens up exciting avenues for studying the evolution of moral lexicons across cultures.

Acknowledgments

AR is funded partly by Schwartz Reisman Institute for Technology and Society Graduate Fellowship. This work was supported by a NSERC Discovery Grant RGPIN-2018-05872, a SSHRC Insight Grant 435190272, and an Ontario ERA Award to YX.

References

Alexander, M., Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (2015). Metaphor map of english. In *Mapping metaphor with the historical thesaurus.* Glasgow: University of Glasgow. (http://mappingmetaphor.arts.gla.ac.uk)

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904–911.

Frimer, J., Haidt, J., Graham, J., Dehghani, M., & Boghrati, R. (2017). Moral foundations dictionaries for linguistic analyses, 2.0. *Unpublished Manuscript*.

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In *Proceedings of ijcai 2016 workshop on computational modeling of attitudes.*

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

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.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*, 1002 - 998.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, August). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers).* Berlin, Germany: Association for Computational Linguistics.

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, *27*(1), 1–17.

Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in american english from 1800 to 2000. *Cognition*, *143*, 87–92.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., ... others (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*, 625–635.

Lakoff, G. (1996). *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). Conceptual metaphor in everyday language. *The journal of Philosophy*, *77*(8), 453–486.

Lee, S. W., & Schwarz, N. (2010a). Dirty hands and dirty mouths: Embodiment of the moral-purity metaphor is specific to the motor modality involved in moral transgression. *Psychological science*, *21*(10), 1423–1425.

Lee, S. W., & Schwarz, N. (2010b). Washing away postdecisional dissonance. *Science*, *328*(5979), 709–709.

Lee, S. W., & Schwarz, N. (2021). Grounded procedures: A proximate mechanism for the psychology of cleansing and other physical actions. *Behavioral and Brain Sciences*, *44*.

Li, L., & Tomasello, M. (2021, 02). On the moral functions of language. *Social Cognition*, *39*, 99-116. (https://doi.org/10.1521/soco.2021.39.1.99)

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books Ngram corpus..

Mendelsohn, J., Tsvetkov, Y., & Jurafsky, D. (2020). A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, *3*, 55.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396.

Poulshock, J. (2006). *Language and morality: evolution, altruism, and linguistic moral mechanisms*. Unpublished doctoral dissertation, University of Edinburgh.

Ramezani, A., Zhu, Z., Rudzicz, F., & Xu, Y. (2021). An unsupervised framework for tracing textual sources of moral change. In *Findings of the association for computational linguistics: EMNLP 2021*.

Snefjella, B., Généreux, M., & Kuperman, V. (2019). Historical evolution of concrete and abstract language revisited. *Behavior research methods*, *51*(4), 1693–1705.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* (Vol. 54). Cambridge University Press.

Tomasello, M. (2016). *A natural history of human morality*. Harvard University Press.

Tomasello, M., Melis, A., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation. *Current Anthropology*, *53*, 673 - 692.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191–1207.

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*.

Xie, J. Y., Ferreira Pinto Junior, R., Hirst, G., & Xu, Y. (2019, November). Text-based inference of moral sentiment change. In *Proceedings of the 2019 confer-*

ence on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp) (pp. 4654–4663). Hong Kong, China: Association for Computational Linguistics. (https://www.aclweb.org/anthology/D19-1472)

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive psychology*, *96*, 41–53.