The emergence of moral foundations in child language development

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

Emmy Liu* **(mengyan3@andrew.cmu.edu)**
Language Technologies Institute, Carnegie Mellon University

Renato Ferreira Pinto, Jr. (renato@cs.toronto.edu)
Department of Computer Science, University of Toronto

Spike W. S. Lee (spike.lee@rotman.utoronto.ca)
Rotman School of Management, Department of Psychology, University of Toronto

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

Abstract

One of the most influential modern theories of morality, Moral Foundations Theory, proposes that morality is formed on innate and shared modular foundations. Psychologists have studied the conceptual development of these moral foundations in childhood, but there exists no comprehensive effort on characterizing the early emergence of moral foundations in naturalistic settings. We explore the emerging order of moral foundations through child and caretaker speech. Using computational methods, we contribute an annotated dataset of moral utterances and find that the individualizing foundations emerge earlier than the binding foundations. Furthermore, caretakers tend to talk more about fairness and degradation, while children talk more about cheating. These results are robust across child gender, family's social class, and race.

Keywords: moral lexicon; moral foundations; child language development

Introduction

One of the most influential modern theories of human morality, Moral Foundations Theory (Graham et al., 2013), suggests that moral judgments are intuitive and driven by five core modular foundations. Each foundation involves two polarities representing the positive and negative aspects of morality: Care/Harm, Fairness/Cheating, Authority/Subversion, Loyalty/Betrayal, and Purity/Degradation. When do different moral foundations emerge in child development? We address this question by conducting a computational analysis of moral language in childhood.

Our study concerns two lines of research that have not been connected previously: computational work on textual inference of moral sentiment, and moral developmental psychology. Recent research has shown that language, or language use preserved in text corpora, can inform people's perception toward right or wrong. In particular, existing work has studied how text can be used to infer the sentiment of moral foundations that people express in different contexts such as political concerns, social movements, and opinions toward political figures (Graham, Haidt, & Nosek, 2009; Garten, Boghrati, Hoover, Johnson, & Dehghani, 2016; Mooijman, Hoover, Lin, Ji, & Dehghani, 2018; Hoover et al., 2020; Ramezani, Zhu, Rudzicz, & Xu, 2021; Roy, Pacheco, & Goldwasser, 2021). This line of work demonstrates that language is an effective medium for reflecting people's moral concerns. However, how moral language itself (particularly language for expressing the moral foundations) emerges in early childhood remains an open question.

A different strand of research from moral psychology has explored the development of moral sense in children (Bloom, 2013; Hamlin, 2013; Kohlberg, 1969). Work in this area has examined moral development typically under experimental settings and focused on understanding children's sensitivity to a particular moral foundation, which we summarize below. Our current study seeks to extend this line of research using a text-based approach to characterize the emergence of moral foundations comprehensively (beyond studying individual foundations in isolation) through the lens of language and in naturalistic settings.

An estimated developmental timeline of moral foundations

We begin with a foundation-centric review of previous studies to construct a rough estimated timeline for the emergence of moral foundations in child development. In doing so, we also identify gaps in the existing literature. We summarize the state of the literature in Figure 1 and described the details as follows.

Care. The preference for pro-social behaviors (e.g., helping an activity) over anti-social ones (e.g., hindering an activity) is an example of the importance of the Care moral foundation, and it has been shown that children as young as 5-6 months old prefer helping agents over hindering ones (Hamlin, Wynn, Bloom, & Mahajan, 2011; Hamlin & Wynn, 2011; Hamlin, Wynn, & Bloom, 2007).

Fairness. Previous work claims that children tend to respect and prioritize equal distribution of resources (Olson & Spelke, 2008; Shaw & Olson, 2012; LoBue, Nishida, Chiong, DeLoache, & Haidt, 2011). This sensitivity to fairness is observed in infancy: children with an average age of 15 months old show an expectation of a fair distribution rather than an unfair one (Schmidt & Sommerville, 2011), and prefer

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

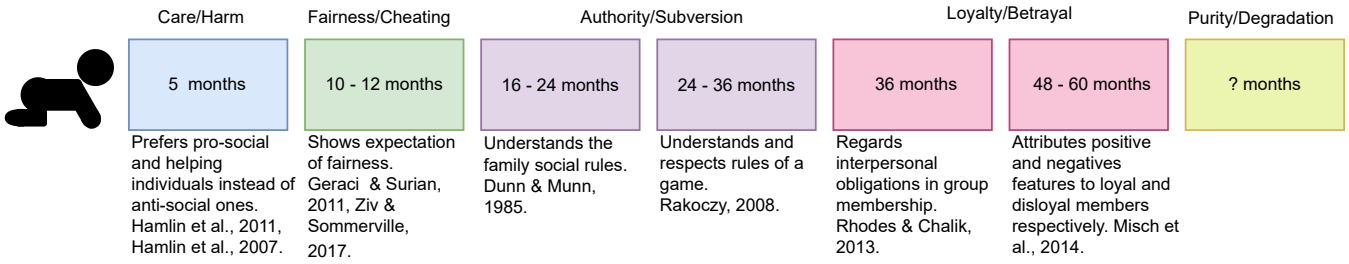

Figure 1: An estimated timeline for child development of moral foundations from the literature.

fair agents over unfair agents (Burns & Sommerville, 2014; Geraci & Surian, 2011; Ziv, Whiteman, & Sommerville, 2021). By repeating these experiments with infants in different age groups (6 months, 9 months, 12 months, and 15 months), studies have suggested that 9 months of age is the transitional period when infants start to develop expectations for fairness, and the first signs of this moral foundation are observed after this stage (Ziv & Sommerville, 2017).

Authority and Loyalty. Authority and Loyalty are observed to emerge later in childhood. For example, it was found that children in their second year of life, from 16 to 24 months old, understand the prohibition of acts by their mother, and have concerns about helping their mother after doing transgressions as a matter of respecting authority (Dunn & Munn, 1985). A related study supports the existence of perceptiveness to authority in children aged 2 to 3 years old, as they respect the rules when playing a game and protest against the transgressors (Rakoczy, 2008). The development of loyalty in children has led to several findings, such as the ability of 4 year olds and above to keep secrets of the group (Misch, Over, & Carpenter, 2016), the tendency to attribute positive and negative adjectives to loyal and disloyal members respectively (Misch, Over, & Carpenter, 2014), and expectation of out-group hostility (Chalik & Rhodes, 2014). Moreover, it is found that 3 year olds view harm within-group as violating intrinsic group obligations (Rhodes & Chalik, 2013).

The reviewed literature on moral development has offered an estimated timeline for the development of moral foundations, but it is limited in several respects. First, existing studies typically take place in controlled settings and emphasize a particular foundation of morality instead of providing a comprehensive account of how different moral foundations emerge through child development. Second, to the best of our knowledge, no study has examined the emergence of Purity, though some suggest that the disgust reaction emerges much later in development (Fallon, Rozin, & Pliner, 1984; Bloom, 2013). Third, how the distinction between positive and negative aspects of moral foundations manifests through the developmental time course is not clear. Studies have explored punishing unfairness and rewarding fairness (Ziv et al., 2021), helping versus hindering (Hamlin et al., 2007; Hamlin & Wynn, 2011), and group loyalty or betrayal (Misch et al., 2014; Rhodes, 2012; Abrams, Rutland, Pelletier, & Ferrell, 2009), but these approaches differ in many methodological details, making it difficult for a coherent analysis on the emergence of different aspects of moral foundations.

To address these limitations, we use a combination of computational and experimental means to identify morally relevant utterances in child and child-directed speech in order to quantify at scale the emergence of moral foundations through language. Our approach is based on the view that language is an effective tool for inferring moral perception (Garten et al., 2016), and we consider the emergence of moral foundations in child speech as a window into the developmental origins of moral foundations.

We contribute a dataset of morally relevant utterances in children's and caretakers' speech, spanning the first 6 years of childhood. This dataset includes 701 utterances from child speech and 670 from caretakers' speech annotated in accordance to the moral foundations.[1] Our work also offers a quantitative analysis on this dataset to reveal the emergence of childhood language expressing different moral foundations.

Hypotheses

We consider two hypotheses about the emergence of moral foundations in child language. Our first hypothesis is based on the empirical evidence of moral development reported in the psychological literature and shown in Figure 1. Specifically, we postulate that the emerging order of moral language in children mirrors the order of moral conceptual development, which follows as Care $\rightarrow$ Fairness $\rightarrow$ Authority $\rightarrow$ Loyalty (with the order of Purity being under-specified in the literature). This hypothesis is rooted in the view that children's conceptual development and linguistic development are closely related (Li, Ogura, Barner, Yang, & Carey, 2009). Although it is conceivable that moral conceptual development might precede moral language development, our current hypothesis does not presume that the timelines of moral conceptual development and linguistic development should be aligned. Rather, we postulate that the rank order should be preserved in moral language development. We predict that

[1]Data and code for replicating our analyses are available at https://github.com/nightingal3/moral-development.

language expressing what have been called the individualizing moral foundations (Care and Fairness) should emerge earlier than language expressing what have been called the binding moral foundations (Authority, Loyalty, and Purity).

Our second hypothesis is that the emerging order of moral language in children follows the order in which caretakers communicate moral foundations in the linguistic environment, which might or might not follow the same order found in the conceptual development of morality. This hypothesis is not orthogonal to our first hypothesis, but it can be different. For instance, parents might choose to emphasize Purity in the linguistic input more than Fairness, or more negative aspects of morality (don'ts) than positive aspects (do's) to prevent children from wrongdoing. Therefore the emerging order of moral language in children based on this example could be Care $\rightarrow$ Purity $\rightarrow$ Fairness $\rightarrow$ Authority $\rightarrow$ Loyalty, influenced by caretakers' emphasis on Purity. This hypothesis is motivated by existing evidence that shows child-directed speech exerts a substantial influence on children's language acquisition (Piper, 1998; Matychuk, 2005), and it is also consistent with Kohlberg's view on the pre-conventional stage of moral development whereby children's morality is shaped largely by adults (Kohlberg, 1969).

In addition to these primary hypotheses, we explore whether the emergence of moral foundations in child language depends on factors including child's gender (e.g., caretakers might emphasize one specific moral foundation more when talking to girls than to boys), and sociodemographic background of the family, particularly social class and race.

Data

We collected 44 text corpora from the CHILDES database (MacWhinney, 2014)—one of the largest public databases of childhood speech in naturalistic settings. Our collected corpus contains text transcripts of interactions between a child and caretaker reflecting linguistic communication for children ranging from 1 to 6 years, as it approximates the period of the conceptual development of morality in childhood suggested by the previous literature. In total, we collected $854,631$ unique transcripts, from which we extracted $356,081$ sentences of child speech (CS) and $524,396$ sentences of child-directed speech (CDS). We tagged each utterance with the age of the child at the time of recording. Other than age, CHILDES includes the child and caretaker gender information. The Hall corpus (Hall, Nagy, & Hillsdale, 1984) in this database also includes the family race and social class, which we use for the analysis of the influence of sociodemographic background on the order of moral language emergence. Overall we gathered $43,452$ CS and $32,952$ CDS utterances from the Hall corpus.

To collect morally relevant utterances from the transcripts in CHILDES, we use the Moral Foundations Dictionary (abbreviated as MFD) (Graham et al., 2009) version 2.0 (Frimer, Haidt, Graham, Dehghani, & Boghrati, 2017) as the base lexicon, which is comprehensive and includes around $2,000$ English moral words that signify different moral foundations. We exclude utterances without any mention of the MFD seed words because we want to study when children begin to use moral words in moral context.

Methodology

One way to estimate the order of emergence of moral foundations in child language is to track the normalized frequencies of utterances that include a moral word in child and child-directed speech. Specifically, the frequencies of the MFD seed words in childhood speech at different age groups can inform the emergence of moral language. However, there are limitations with this raw count-based method. In particular, one might use moral words like *fair* in polysemous ways to refer to an exhibition, or *hurt* to refer to a stomachache, which have minimal moral relevance in context. To alleviate this issue of polysemy (i.e., a word expressing different meanings in different context), we first describe a simple clustering technique that helps to group sentences containing MFD words into morally relevant or morally irrelevant clusters. We then describe how we use human annotation to further disambiguate the morally relevant clusters from the irrelevant ones and hence ensure the quality of moral language extraction.

Automatic clustering of moral utterances. In our approach, language expressing each moral foundation consists of a set of utterances that include at least one moral seed word from that foundation in MFD. We assign the utterances of a moral foundation to different clusters based on a clustering algorithm and then identify clusters that have morally relevant sentiments (or moral clusters) based on human annotations.

To do so, we first lower-cased the transcripts, split them into sentences, removed punctuation, and lemmatized the remaining tokens.[2] We then used SBERT (Reimers & Gurevych, 2019)—a state-of-the-art technique from natural language processing—to represent the utterances in a high-dimensional, contextually informed semantic space, and reduced the dimension with principal components analysis to keep 95% of variance. We next used a Gaussian mixture model (GMM) to assign the utterances to k clusters, whereby a cluster is specified as a Gaussian distribution. The number of clusters k ranged from 2 to 10 and is chosen by grid-search to maximize the Silhouette score (Rousseeuw, 1987) of the clustering. All implementation was done using the `scikit-learn` package. We followed this procedure for each of the ten moral foundations[3] in CS and CDS utterances separately, as children's and adults' speech can be structurally different, so overall we trained 20 GMM models[4].

Human annotation of morally relevant clusters. The clusters we have obtained are not guaranteed to be morally relevant due to the polysemy issue described earlier. To determine which clusters are morally relevant versus not, we conducted a survey to obtain human annotation. In this sur-

[2] All the pre-processing is done using the `NLTK` toolkit.

[3] Positive and negative poles are different moral foundations.

[4] We obtained 103 clusters in total.

Table 1: Sample utterances expressing each of the 10 moral foundations extracted from child and child-directed speech.

Moral foundation	Child-directed speech (CDS)	Child speech (CS)
Care	*we must rescue him*	*help Carrie wash dish*
Harm	*you wouldn't hurt Adam would you*	*and they always fight*
Fairness	*is that fair enough*	*next time I'm gonna make it fair*
Cheating	*they did steal the honey*	*I'm not cheating*
Authority	*go out and tell your father you're sorry*	*you mean she gave you permission*
Subversion	*why do you choose to be disobedient*	*except the dragon can't even kill the knight*
Loyalty	*are you being honest with me*	*you and me do this together*
Betrayal	*do you think it was one of his enemies*	*if they weren't my enemy either*
Purity	*there was a new punishment which is tell [Name] that he was gay*	*okay and the bishop told him*
Degradation	*he's filthy*	*he's he's dirty*

vey, the annotators were asked to 1) determine if a given sentence was spoken in a moral context, and 2) if so, identify the moral foundation(s) expressed in the sentence. We represent each cluster in the survey by deriving 10 prototypical and 10 peripheral sentences (at most) from the cluster, since it is infeasible to get annotation data for all the utterances. The prototypical sentences are the ones with the highest proximity to the cluster center (i.e., the average of all the utterances in the cluster) and the peripherals are the furthest to the center, with respect to the cosine similarity of their contextual distributed representations. Equation 1 specifies how the proximity of a sentence s to its cluster C is measured, where V_s is the semantic representation of sentence s.

$$proximity(s,C) = cosine(V_s, \frac{\sum_{s_i \in C} V_{s_i}}{|C|}) \quad (1)$$

In total, we gathered 670 utterances from CDS, and 701 utterances from CS. We recruited 300 participants. Each participant annotated 40 utterances, drawn randomly from the data. We used the Prolific recruitment platform, and the data collection was done through the Qualtrics platform.[5] We removed participants who failed the attention check ($N = 50$), resulting in an average number of 7.32 annotators per utterance. Table 1 provides sample utterances from each moral foundation that show a high degree of annotator agreement on moral relevance. The Krippendorff's α (Krippendorff, 2004) agreement among annotators is 0.25, where $\alpha = 1$ is perfect agreement and $\alpha = 0$ represents chance.

To determine if an utterance in the survey was morally relevant and to prevent ourselves from overestimating the moral language, we took the majority vote of the participants' responses: if more than 50% of the participants annotated an utterance as irrelevant, the utterance was considered non-moral. If at least 70% of the participants annotated an utterance as morally relevant, it was considered moral. The moral foundation of the utterance is then determined by taking the majority vote of the annotations. For example, an utterance like *he's really not to be trusted very much* was initially regarded as an example of Fairness (because the word *trusted* is a seed word for the Fairness moral foundation). However, the majority vote from the participants is Loyalty, thus would be counted as an example of Loyalty. If between 50% to 70% of the annotators believe an utterance is moral, we count the utterance as moral only if its initial moral foundation matches that from the majority vote of the annotators. The utterance would be excluded from the analysis otherwise, as there is not sufficient agreement between the annotators about its moral relevance.

Frequency estimation of moral foundations. Once we had obtained the moral (ir)relevance labels from the survey, we annotated the rest of the utterances based on these labels for their respective clusters. Each cluster is annotated based on the majority vote of the labels for its sentences (both prototypical and peripheral) that were present in the survey, and uncertain sentences are excluded. We define the agreement ratio metric as the number of survey utterances in a cluster whose moral annotation agrees with their cluster's moral label. For example, if a cluster has 20 utterances in the survey and 15 of which are annotated as moral, then the cluster is labeled as moral, and its agreement ratio would be 0.75. Among all the clusters, we obtain high average agreement ratios of 0.8 for the moral relevance label and an average agreement ratio of 0.81 for the moral foundations label. We discard all the utterances that are identified as morally irrelevant and estimate the fine-grained frequencies of the moral utterances for each moral foundation from age 1 to age 6.

Results

The emerging order of moral foundations in childhood language

To evaluate our main hypotheses, we track the frequencies of how often different moral foundations are talked about by caretakers and children. Figure 2a summarizes the normalized frequencies per age group. For better visual clarity, we re-display Fairness/Cheating and Purity/Degradation founda-

[5]We have obtained research ethics approval but omit the information here for anonymous review.

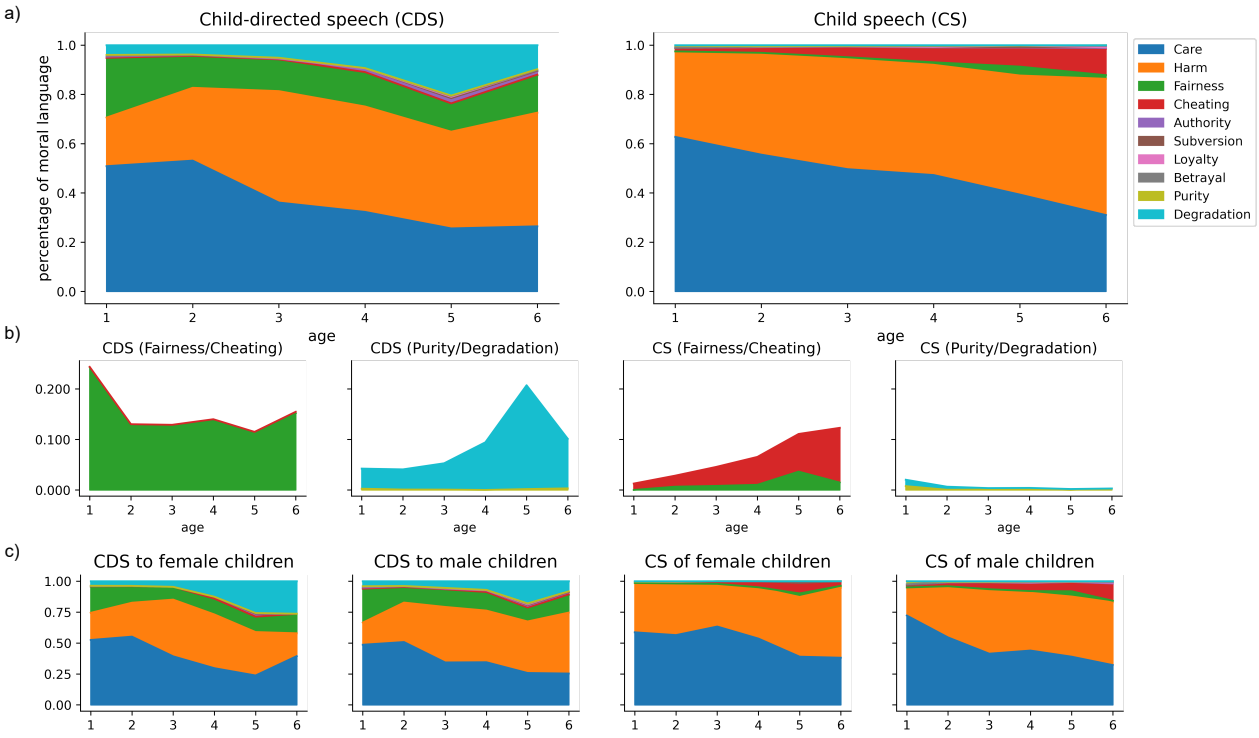

Figure 2: Stack area charts summarizing the frequencies of moral foundational language within different age and gender groups.

tions in Figure 2b. A first glance at the moral foundation frequencies indicates a clear dominance from the individualizing moral foundations (Care/Harm, Fairness/Cheating) over the binding ones (Authority/Subversion, Loyalty/Betrayal, Purity/Degradation). This initial finding provides evidence for our first hypothesis, namely that the development of moral language mirrors moral conceptual development.

More specifically for child speech, the Care/Harm foundation is already represented in moral language from the age of 1. Fairness/Cheating foundation becomes more frequently expressed after age 2 and shows a gradual rise throughout child growth. Precisely, Fairness/Cheating captures only 2% of child moral language at the age of 2 but exceeds over 12% by the age of 6. Purity/Degradation is scarcely represented through the course of development with no more than 2% expression in moral language. The Authority/Subversion moral foundation reaches to 0.5% representation only after the age of 3 and is almost non-existent before that. Loyalty/Betrayal also achieves a 0.5% language representation after the age of 4. These results indicate that the order of emergence of moral foundations in language development is as follows: Care/Harm $\rightarrow$ Fairness/Cheating $\rightarrow$ Purity/Degradation $\rightarrow$ Authority/Subversion $\rightarrow$ Loyalty/Betrayal. Our finding aligns with the conceptual order of development of morality in children, as stated in our first hypothesis, and extends the previous findings to locate the development of the Purity/Degradation foundation.

The emerging order reflected in caretaker speech is similar but not identical to child speech. One notable difference is the percentage of individual moral foundations expressed. Specifically, Purity/Degradation composes only 2% of child moral language, but in caretaker speech, this foundation is expressed up to 20%. This is also the case for Fairness/Cheating and Authority/Degradation, suggesting that although the order of moral foundations emergence in the language is similar between CS and CDS, the child-directed speech from adults consists of more intricate dimensions of morality while the child speech mainly focuses on Care, Harm, and Cheating foundations. Furthermore, as opposed to CS, the percentage of the Fairness/Cheating foundation stays relatively constant over time in CDS. Another difference between CDS and CS is that Cheating is more talked about by children, while Fairness is more predominant in caretaker speech. This asymmetry is presumably due to the caretakers' effort on teaching the quality of being fair to children. Another asymmetry is that Degradation is much more accentuated than Purity in CDS, which can be a result of caretakers preventing children from disgust-related matters that are not highly perceivable by children (Fallon et al., 1984; Bloom, 2013). All of our results were consistent across child gender (see Figure 2c).

Effects of social class and race

To examine the effects of sociodemographic factors, we ran additional analyses on the Hall corpus—independent of the main results we described—to understand whether our findings on the emergence of moral language are robust to fam-

ily's social class and race. The Hall corpus is a large subcorpus from CHILDES which consists of utterances from CS and CDS at age 4, and it was designed specifically for understanding differences in language development due to race (i.e., black or white) and social class (i.e., upper class or working class) (Hall et al., 1984). Since the utterances of this corpus were neither present in the survey nor used to train the clustering models, they can be considered an independent evaluation dataset for our hypotheses.

To determine the moral relevance and foundations of the utterances in the Hall corpus, we applied the same data preprocessing and used the clustering models obtained from the previous analyses to identify the cluster affiliations. The labels of the utterances are likewise determined by the moral labels of their assigned clusters.

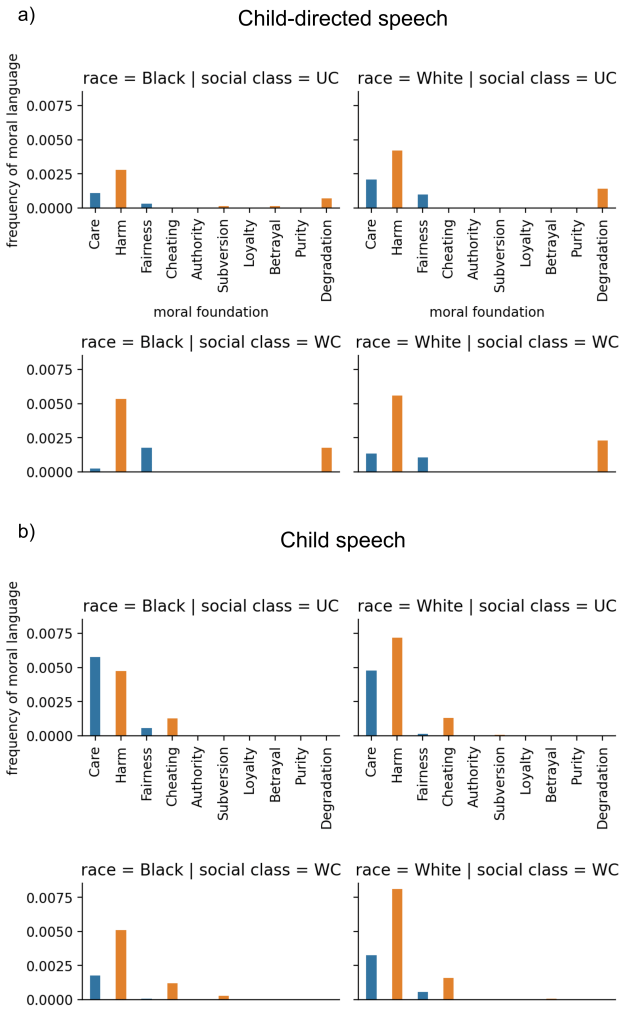

Figure 3: Comparisons of moral language frequency in families with age-4 children of different social classes and races.

The absolute frequencies of the moral foundations in language of different races and social classes are shown in Figure 3. We denote upper class families by UC, and working class by WC. A common pattern observed in all the groups is that the Care/Harm is the most frequently expressed foundation, followed by Fairness/Cheating (i.e., individualizing foundations), and next comes Degradation. Loyalty/Betrayal and Authority/Subversion (i.e., binding foundations) are almost never represented in this corpus. Beyond these overall patterns, we observed some nuances: Cheating appears more frequently in CS than in CDS, whereas Fairness appears more frequently in CDS than in CS. Degradation is conspicuously mentioned in CDS, but rarely used in CS. Although this corpus is an independent dataset, the observations here are similar to the results reported in the previous section which shows the robustness of our findings. These findings suggest that the development of moral foundations in child language is similar in families of different social classes and races.

As a final analysis, we ran permutation tests to assess statistically meaningful differences between the social groups. In each test, there was a control variable, which remained stable, and a changing variable, which was permutated between groups of race or social class for $1,000$ times. For example, we would control for social class by analyzing the difference between black UC and white UC families in how much a moral foundation, e.g., Degradation, was expressed by CDS. For each permutation, we randomly assigned a race (either black or white) to each Degradation utterance in UC CDS partition of the data while keeping the number of utterances for each race constant and equal to the original number in the Hall corpus. We repeated this procedure for all the possible queries based on our observations in Figure 3. In total we tested for 36 queries and used Bonferroni correction with the α of 0.05. In CS (Figure 3b), controlling for race, UC and WC differed significantly in how often the Care foundation was present in the language. Specifically, Care was more talked about in black UC families than black WC families ($p < 0.001$). However, this difference did not hold after p-value correction in white families ($p = 0.756$). Controlling for social class, black and white families differed significantly in how often Harm concerns were expressed. Specifically, Harm was more talked about in white families (for both social classes) than in black families ($p < 0.05$).

Conclusion

We examined the emergence of moral foundations in childhood through child and caretaker speech. Our work extends prior research on the emergence of moral foundations in experimental settings toward a comprehensive account of the developmental origins of all moral foundations in naturalistic settings. We show that the individualizing foundations emerge earlier than the binding foundations and that child speech and caretaker speech differ in what aspects of Fairness and Purity are emphasized. Our results are robust across child gender and family's social class and race. Future work could apply this approach to explore how children expand their moral circles.

Acknowledgments

AR is funded partly by a 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, a SSHRC Insight Development Grant 430202100060, a SSHRC Insight Grant 43520170127, and an Ontario ERA Award to SL.

References

Abrams, D., Rutland, A., Pelletier, J., & Ferrell, J. M. (2009). Children's group nous: Understanding and applying peer exclusion within and between groups. *Child Development*, *80*(1), 224–243.

Bloom, P. (2013). *Just babies: The origins of good and evil*. Broadway Books.

Burns, M. P., & Sommerville, J. (2014). "I pick you": The impact of fairness and race on infants' selection of social partners. *Frontiers in Psychology*, *5*, 93.

Chalik, L., & Rhodes, M. (2014). Preschoolers use social allegiances to predict behavior. *Journal of Cognition and Development*, *15*(1), 136–160.

Dunn, J., & Munn, P. (1985). Becoming a family member: Family conflict and the development of social understanding in the second year. *Child Development*, 480–492.

Fallon, A. E., Rozin, P., & Pliner, P. (1984). The child's conception of food: The development of food rejections with special reference to disgust and contamination sensitivity. *Child Development*, 566–575.

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.*

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science*, *14*(5), 1012–1020.

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.

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.

Hall, W. S., Nagy, W. E., & Hillsdale, R. L. (1984). *Spoken words, Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Erlbaum.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186–193.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30–39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences*, *108*(50), 19931–19936.

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.

Kohlberg, L. (1969). *Stage and Sequence; The Cognitive-developmental Approach to Socialization*. Rand McNally.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, *38*, 787–800.

Li, P., Ogura, T., Barner, D., Yang, S.-J., & Carey, S. (2009). Does the conceptual distinction between singular and plural sets depend on language? *Developmental Psychology*, *45*(6), 1644.

LoBue, V., Nishida, T., Chiong, C., DeLoache, J. S., & Haidt, J. (2011). When getting something good is bad: Even three-year-olds react to inequality. *Social Development*, *20*(1), 154–170.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Matychuk, P. (2005). The role of child-directed speech in language acquisition: a case study. *Language Sciences*, *27*(3), 301–379.

Misch, A., Over, H., & Carpenter, M. (2014). Stick with your group: Young children's attitudes about group loyalty. *Journal of Experimental Child Psychology*, *126*, 19–36.

Misch, A., Over, H., & Carpenter, M. (2016). I won't tell: Young children show loyalty to their group by keeping group secrets. *Journal of Experimental Child Psychology*, *142*, 96–106.

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.

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*(1), 222–231.

Piper, T. (1998). *Language and learning: The home and school years*. ERIC.

Rakoczy, H. (2008). Taking fiction seriously: Young chil-

dren understand the normative structure of joint pretence games. *Developmental Psychology*, *44*(4), 1195.

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* (pp. 1215–1228).

Reimers, N., & Gurevych, I. (2019, 11). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing.* Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084

Rhodes, M. (2012). Naïve theories of social groups. *Child Development*, *83*(6), 1900–1916.

Rhodes, M., & Chalik, L. (2013). Social categories as markers of intrinsic interpersonal obligations. *Psychological Science*, *24*(6), 999–1006.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Roy, S., Pacheco, M. L., & Goldwasser, D. (2021). Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.*

Schmidt, M. F., & Sommerville, J. A. (2011). Fairness expectations and altruistic sharing in 15-month-old human infants. *PloS one*, *6*(10), e23223.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*(2), 382.

Ziv, T., & Sommerville, J. A. (2017). Developmental differences in infants' fairness expectations from 6 to 15 months of age. *Child Development*, *88*(6), 1930–1951.

Ziv, T., Whiteman, J. D., & Sommerville, J. A. (2021). Toddlers' interventions toward fair and unfair individuals. *Cognition*, *214*, 104781.