UNIVERSITY OF NOTRE DAME

Modeling Neo-Kohlbergian Moral Judgment Schemas Using Latent Semantic Analysis and Machine Learning

Mark Graves, Christian Llantero, Darcia Narvaez

ABSTRACT

Models created using artificial intelligence (AI) methods can provide novel perspectives on theories of moral development and yield new insights into human moral judgment. Using latent semantic analysis (LSA) and the machine learning method k-nearest neighbors (k-NN), we investigate moral judgment as characterized by the Standard Issue Moral Judgment Interview and Scoring System (Colby & Kohlberg, 1987). The resulting model consists of a collection of textual representations for moral schemas generated using LSA and k-NN from prototypical criterion judgment responses in the scoring system. Preliminary results demonstrate promise for the approach and suggest at least one way the computational approach could augment a previously identified challenge to human scoring.

BACKGROUND

As artificial intelligence (AI) systems become more autonomous, sophisticated, and pervasive within society, they begin to make decisions that benefit from what would be analogous to human moral judgment. Additionally, as AI becomes involved more closely in human social interaction and discourse, then their models of humans used in such interactions can benefit from a better “understanding” of human goals, values, and moral frameworks. Meanwhile, building these models can provide novel perspectives on theories of moral development and yield new insights into human moral judgment. Analyzing the semantics used in Kohlbergian moral judgment lays a foundation for AI models of moral judgment and can identify new perspectives on that significant theory.

METHOD

Source Material. The Standard Issue Moral Judgment Interview and Scoring System (MJI; Colby & Kohlberg, 1987) provides a semi-standardized method to identify the developmental level of moral judgments as well as operational definitions for those developmental stages. As a refinement, schema theory provides an alternative to Kohlberg’s hard stages and has been adapted both by artificial intelligence researchers to develop computational knowledge representation systems and moral psychologists developing a neo-Kohlbergian approach to moral judgment (Rest et al., 1999). Although Rest and others developed simpler and easier methods of evaluating moral judgment, e.g., DIT-2, the carefully constructed and organized moral judgment interview scoring system has sophisticated conceptual structures and numerous prototypical responses for criterion judgment that current AI techniques and text analysis methods can more effectively use for semantic analysis, schema development, and model building. Analysis was performed on Scoring System Form A, which has 1038 criterion judgments for three dilemmas.

Procedure. Dilemma-specific words and phrases are replaced with schema “slot” placeholders, then the resulting prototypical criterion judgment responses are mapped to a semantic space using Latent Semantic Analysis (LSA). The semantically-mapped criteria are then filtered for their ability to predict moral stage (or element) using the machine learning method k-nearest neighbors (k-NN). The filtering process is then repeated a second time.

RESULTS

For the 1038 criterion judgment response texts in Form A, k-NN classification correctly re-predicted stages for 65% of the texts (80% allowing for overlap between developmental and transitional stages). After removing all criterion judgments whose stage was not re-predicted exactly, the remaining 666 criterion judgments were accurately re-predicted 85% of the time (also 85% allowing for stage overlap). The elimination step based upon re-prediction was repeated, leaving 569 criterion judgments were accurately re-predicted 93% of the time (also 93% allowing for stage overlap).

Example Criterion Judgment Text: <PERSON> [should not steal] because it is a crime, wrong, or against the law.

Table 1: *Machine learning metrics for k-NN classification to predict moral stage from LSA-transformed criterion judgments (k=5, LSA 100-dim)*

Stage		Initial k-NN			k-NN After Filtering			k-NN After 2nd Filtering		
Developmental	Transition	F1 (10-fold CV)	Re-prediction Accuracy	Mean Cosine	F1 (10-fold CV)	Re-prediction Accuracy	Mean Cosine	F1 (8-fold CV)	Re-prediction Accuracy	Mean Cosine
1	1/2	27%	73%	.15	48%	89%	.17	55%	94%	.17
2	2/3	59%	76%	.13	63%	93%	.14	68%	94%	.14
3	3/4	53%	80%	.10	70%	87%	.10	78%	96%	.09
4	4/5	71%	77%	.09	78%	89%	.10	82%	95%	.09
5		57%	63%	.13	70%	76%	.16	67%	90%	.18

Tf-idf-transformed, 100-dimension LSA semantic space of criterion judgments by moral stage, transformed to two dimensions using MDS.

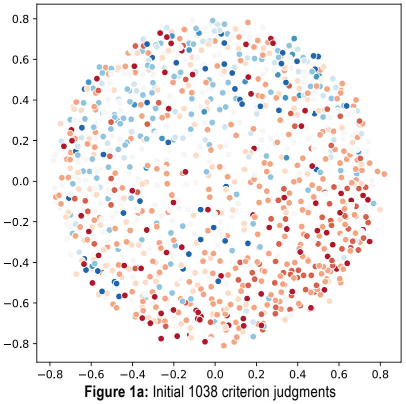

Figure 1a: Initial 1038 criterion judgments

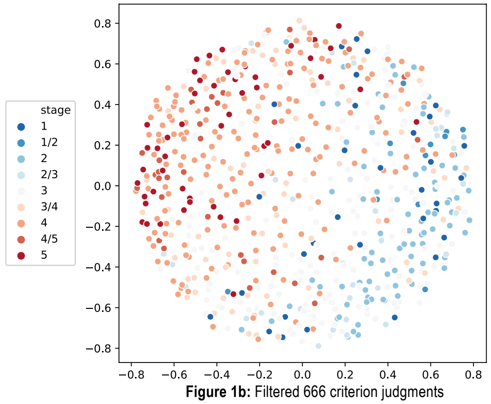

Figure 1b: Filtered 666 criterion judgments

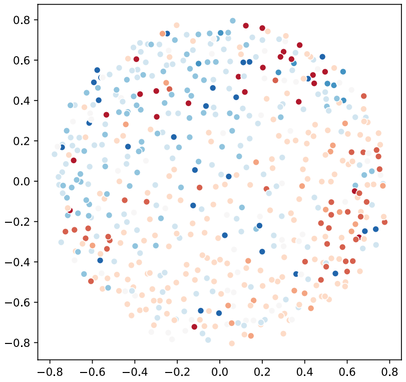

Figure 1c: Twice-filtered 569 criterion judgments

DISCUSSION

Machine learning creates more coherent subspaces for stages. Using machine learning (k-NN) to filter criterion judgments for LSA improves predictive performance on k-NN and appears to reorganize the semantic space to create more coherent subspaces for each moral stage. (See LSA semantic space visualization Figures 1a-1c).

1. Levels (preconventional, conventional, postconventional) are clearly outlined. In the first iteration, the projected space appears to split more clearly into pre-conventional, conventional, and post-conventional subspaces as shown in the more coherent grouping of stages in Figure 1b compared to Figure 1a.

2. Stage spaces are made more coherent. In the second filtering iteration, the splits for the intermediate stages appear to remain, but there appears to be more clustering occurring for each stage (though that may be an artifact of multi-dimensional scaling (MDS) as the clustering is not immediately apparent in the mean cosines scores). See Figure 1c compared to Figures 1a and 1b.

Binary Classification of Stage Miscalculations

Miscomprehension of higher stages into lower stages is a phenomenon well known in the literature (e.g., Rest, 1979) where, for example, respondents transform reasoning from stage 4 to stage 1, stage 5 to stage 2. To investigate whether the computational system suffered from the same problem, binary k-NN classifiers were created to compare stages 1-vs-4 and 2-vs-5, which demonstrated 91% and 84% accuracy, respectively (using 10-fold cross-validation). Thus, the computational system does not appear to have the same disadvantage as human scorers.

Latent Semantic Analysis (Details)

Latent Semantic Analysis (LSA) (Landauer et al., 2007) computes semantic similarity between texts. Two criterion judgment texts are translated into mathematical representations (bag-of-words vectors) and transformed using tf-idf and LSA into a 100-dimension semantic space. The semantic space is constructed so that words close in meaning are mapped to locations near each other. Thus each transformed vector represents the overall meaning of that criterion judgment, and the closeness (cosine of the angle) between vectors measures semantic similarity between the two texts.

k-Nearest Neighbors (Details)

Methodologically, a “grid walk” selected k=5 nearest neighbors and 100 dimensions for tf-idf-transformed LSA semantic space using F_1 scores. Cosine distance determined nearest neighbors; and the described metrics used hold out, full dataset, or 10-fold cross validation for grid walk, text elimination, and binary k-NN classification, respectively. Except for text elimination (filtering), correctness between predicted and actual stages allowed for overlapping stages among the five developmental stages and four intermediate stages in the scoring guide, e.g., stage “2” or “3/4” was considered correct if “1/2” or “4” were predicted, respectively. (Metric F_1 is the harmonic mean of precision and recall.)

CONCLUSION & FUTURE DIRECTIONS

Using statements from the MJI coding book, we showed that AI analyses could sort statements by level, by stage, and could deal with misattributed statements. Preliminary results show promise for using AI systems to identify moral reasoning, though further investigation is needed to distinguish among artifacts of the classifier, limitations of the scoring system, and results that can inform understanding of moral judgment. Future studies should apply these methods to everyday discourse to identify levels and stages.

ACKNOWLEDGMENTS

Mark Graves was funded by a fellowship from the John Templeton Foundation. Thank you!

Contact Information*: Mark Graves, mgraves@nd.edu*

REFERENCES

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment, Vol. 1. Theoretical foundations and research validation; Vol. 2. Standard issue scoring manua*l. New York, NY: Cambridge University Press.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007)*. Handbook of latent semantic analysis*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Rest, J. R. (1979). *Development in judging moral issue*s. Minneapolis, MN: University of Minnesota Press.

Rest, J. R., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: a Neo-Kohlbergian approach*. Mahwah, N.J.: L. Erlbaum Associates.