Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm

Anonymous ACL submission

Abstract

While NLP models often seek to capture cognitive states via language, the validity of predicted states is determined by comparing them to annotations created without access the cognitive states of the authors. In behavioral sciences, cognitive states are instead measured via experiments. Here, we introduce an experiment-based framework for evaluating language-based cognitive style models against human behavior. We explore the phenomenon of decision making, and its relationship to the linguistic style of an individual talking about a recent decision they made. The participants then follow a classical decision-making experiment that captures their cognitive style, determined by how preferences change during a decision exercise. We find that language features, intended to capture cognitive style, can predict participants' decision style with moderate-to-high accuracy (AUC $\sim$ 0.8), demonstrating that cognitive style can be partly captured and revealed by discourse patterns.

1 Introduction

While language models grow in sophistication, NLP tasks increasingly focus on understanding the people behind the language (Choi et al., 2023; Dey et al., 2024). Such social and psychological NLP studies still rely primarily on **annotations** for evaluation. For example, recent social tasks have depended on annotated datasets for, e.g., emotions (Rosenthal et al., 2019; Mohammad et al., 2018), empathy (Sharma et al., 2020), politeness (Hayati et al., 2021), humor (Meaney et al., 2021), dissonance (Varadarajan et al., 2023), and reasoning abilities (Alhamzeh et al., 2022). However, while annotation-based work has pushed NLP towards capturing cognitive states of the language generators (i.e. people), it falls short of offering *ground truth* of psychological processes because annotations reflect *perception* of another person's state. (Sandri et al., 2023; Sap et al., 2021). For example, annotations of empathy point to linguistic cues that *appear* empathetic to observers but do not always reflect the actual human experience of empathy (Lahnala et al., 2022). Behavioral sciences, on the other hand, often emphasize the importance of *direct* assessment through experimental paradigms for the purpose of understanding constructs of interest.

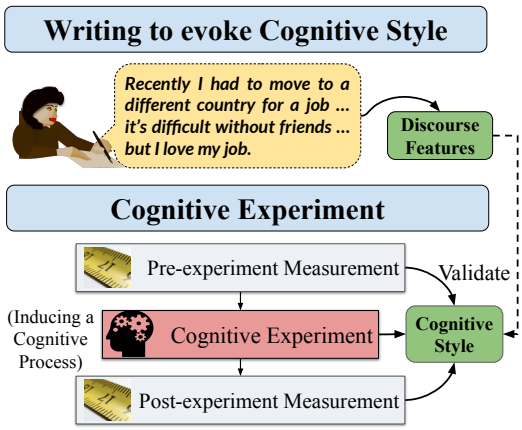

Figure 1: An alternate evaluation framework for validation of cognitive processes with language: The participants are first prompted to write about their experiences, eliciting their thought process. Then they are subjected to an experiment that would measure their behavior. The behavior is a *ground truth* measure of their cognitive style that can be tied to the expressed language.

We introduce an experimental framework that collects linguistic data alongside induced cognitive phenomena to evaluate the feasibility of discourse modeling approaches for capturing cognitive styles in decision making. By associating linguistic patterns with specific cognitive phenomena, we aim to understand individuals' unique cognitive styles, which are largely unseen and often only observable through the final decision (Campitelli and Gobet, 2010). Our study follows modern psychology–experimental designs to quantify how language use signals Cognitive Styles, or habitual patterns of thought related to various cognitive phenomena.

Our key contributions include: (1) An experiment-based evaluation framework to validate cognitive styles in language; (2) Exploring features for modeling decision-making cognitive styles; (3) Finding that language can be indicative of a person's cognitive styles even in the more stringent evaluation framework; (4) A corpus (*Decisions*) for the language of decision-making cognitive styles.

2 Related Work

While NLP for social science often relies on labels from annotators or questionnaires, behavioral science theory suggests carefully designed experiments can more objectively elicit and capture cognitive states. For instance, Saxbe et al. (2013) investigated emotional responses using an experimental design in which participants' brain activity was imaged as they listened to narratives eliciting different emotions. These methods can offer a more objective foundation for understanding the psychological processes at play (Brook O'Donnell and Falk, 2015). In our study, we focus on cognitive styles in decision making – reflecting one's tendency to maintain consistency, and resolve dissonance (Harmon-Jones and Harmon-Jones, 2007; McGrath, 2017). Since people show little awareness of their decision-making processes (Nisbett and Wilson, 1977), cognitive styles are measured experimentally by observing shifts in preferences after decision (Simon et al., 2004; Aguilar et al., 2022).

Our study draws from previous NLP research that validates author state measurements through annotations or self-report questionnaires. For example, past work has compared affective states with self-reported mental health by analyzing self-disclosures (Zirikly et al., 2019; Valizadeh et al., 2021), while others have examined cognitive styles through discourse frameworks (Juhng et al., 2023; Sharma et al., 2023) and causal reasoning patterns (Son et al., 2017, 2018a). However, annotations and self-reports are subject to perceptual biases, such as those observed in dialogue evaluations (Liang et al., 2020) or when assessing constructs such as humor, empathy, or offensiveness (Yang et al., 2021; Paulhus et al., 2007; Buechel et al., 2018). To address these limitations, we adopt an experimental approach that aims to objectively capture cognitive states, focusing on how individuals manage dissonance and consistency in their decision-making processes.

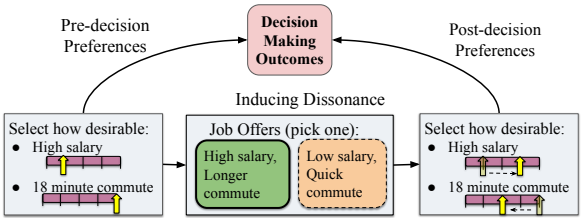

Figure 2: Participants wrote about recent decisions that they had made. The same participants separately completed a decision-making experiment wherein they encountered a simulated a job offer setting (See §3).

3 Experiment

A total of 514 participants were recruited in person for the study; 12 were excluded due to incomplete or invalid responses, resulting in a final dataset of 502 participants. Data collection was performed in 2 stages (see Figure 1). The questionnaire has been described in detail in Appendix B.

Writing Task Participants received 2 writing prompts to elicit language relevant to their decision-making cognitive style: 1) "Please describe a recent important and difficult decision that you have made," and 2) "What were the considerations that you thought about while making the decision? When answering, please consider all of the circumstances and details that went into the difficult decision." These questions were chosen to elicit detailed descriptions of a recent decision-making process, encouraging participants to discuss options and explain their reasoning. Please check Appendix A for a semantic description of the corpus collected, which we call *Decisions* corpus.

Constraint Satisfaction Experiment We replicated the experiment from Simon et al. (2004), modifying the preference score calculation to quantify overall preference changes rather than single attribute fluctuations.

1. Pre-Decision Preferences Participants answered questions assessing preferences on a 6-point scale (-5 to 5, interval of 2) for four attributes: Commute (com), Vacation (vac), Office space (off) and Salary (sal).

Each attribute had positive (+) and negative (-) questions for preference bounds. For example: com_+ (commute): "Please select how desirable the 18 minute commute is to you." com_-: "Please select how desirable the 40 minute commute is to you.". Participants rated each attribute's relative weight (W) on a 1-8 scale. Final preference(ρ_{com})

for each attribute was calculated as:

$$\rho_{\texttt{com}} = (com^{+} - com^{-}) \times \texttt{W}_{com}$$

2. Job Offers Two choices were offered to the participants, such that in choosing either of the jobs, they would likely make compromises on at least two attributes. The two options were:
Company A: com_{+}, vac_{+}, off_{-}, sal_{-}; and
Company B: com_{-}, vac_{-}, off_{+}, sal_{+}
where $_{+,-}$ (in subscript) refer to the favorable and unfavorable conditions for each of the four attributes. Therefore, the pre-decision preference score ψ for company A and B can be calculated as:
$\psi_{\texttt{A}} = +\rho_{com} + \rho_{vac} - \rho_{off} - \rho_{sal}$
$\psi_{\texttt{B}} = -\rho_{com} - \rho_{vac} + \rho_{off} + \rho_{sal}$
Participants choose between two options, typically aligning with their initial preferences. We randomly introduce an influencing factor, location (loc), describing the job as either near a fun mall or in a dull construction site. This aims to induce dissonance, compelling participants to compromise and potentially make contrarian decisions inconsistent with their initial preferences.

3. Post-Decision Preferences After selecting a job, participants answer the same questions from the pre-decision questionnaire again.

4 Methods

We explored common baseline models to capture the decision-making cognitive styles: a random baseline, zero- and four-shot prompting on both Llama3.1-8B-chat and Gemma-7B-Instruct, and finally, a predictive model from averaged embeddings of the text from L23 of RoBERTa-large. Further, we explored the following theoretically relevant **Discourse Relations**:

1. Causal explanations We extract individual reasoning behind decision-making behaviors using a causal explanation model trained on social media posts (Son et al., 2018b). We infer the proportion of messages containing causal explanations provided by the individual.

2. Counterfactuals These are statements of alternate reality; of what could have happened instead of actual events. We used the counterfactual relation recognition model based on a social media dataset (Son et al., 2017) to calculate the proportion of the messages from each individual that contains counterfactual statements.

3. Dissonance and Consonance We extracted linguistic dissonance and consonance using a model trained on social media posts introduced in Varadarajan et al. (2023), which captures signals of cognitive dissonance exhibited through language. We then calculated the average probability of dissonance for consecutive phrases predicted as dissonant or consonant.

4. Discourse Relation Embeddings To capture any discourse relation-level information, we use discourse relation embeddings (Son et al., 2022) aggregated at a message level to capture discourse relation.

Predictive Models for Decision Making outcomes We model 2 outcomes together: Choice-Induced Shift (CIS) and Influence (Inf); we define each construct and their measurement below:

1. Choice-Induced Shift (CIS) The change in preference is captured by subtracting the pre-experiment scores from post-experiment scores.
$CIS = \psi^{post}_{\texttt{A}} - \psi^{pre}_{\texttt{A}}$; `choice = A`
$CIS = \psi^{post}_{\texttt{B}} - \psi^{pre}_{\texttt{B}}$; `choice = B`
Here, we model binarized CIS which captures the direction of the preference change towards the job choice.

2. Influenced or Not (Inf) The job offer is further influenced by introducing a confounding attribute loc (location). Many participants choose the job influenced by the description of the location attribute, however not all of them change their minds. The change is a cognitive signal that measures if someone's choice was influenced by confounding attribute. This indicates that their initial preferences were not strong enough to begin with. This is captured as a binary variable: making the choice in the direction of the influenced variable or not.

Together, **CIS** and **Inf** variables capture the magnitude and direction of the tendency of a person to vacillate when exposed to conflict-inducing information. We combine them into a single variable **CIS_Inf** for modeling four distinct cognitive styles for decision making: (a) Negative CIS, Not Influenced (↓CIS↓Inf, 6%), (b) Negative CIS, Influenced (↓CIS↑Inf, 17%), (c) Positive CIS, Not Influenced (↑CIS↓Inf, 11%) and (d) Positive CIS, Influenced (↑CIS↑Inf, 66%).

We use a logistic regression model for 4-way classification with the features listed in Table 1, where we calculate stratified 5-fold cross-validation

accuracies using DLATK (Schwartz et al., 2017).

Baselines	AUC	Discourse feats	AUC	k
Random	0.50	Causal	**0.81**	1
Llama3.1 (0-sh)	0.56	Counterfactual	0.80	1
Gemma (0-sh)	0.56	Consonance	**0.81**	1
Llama3.1 (4-sh)	0.64	Dissonance	0.80	1
Gemma (4-sh)	**0.79**	DiscRE (full)	0.76	845
RoBERTa-L23	0.69	DiscRE (16-D)	0.79	16

Table 1: Accuracy of feature sets over the CIS_Inf outcome (**AUC**: mean Area Under the ROC Curve; **k**: number of input features). Linguistic measures from the participants' pre-experiment writing can predict CIS_Inf with moderate-high, non-trivial accuracy.

5 Results

We explore results for our primary application of the *experimental validation framework*: do discourse relation models, which capture explanatory styles and coherence in language of individuals, predict the cognitive style of a decision that an individual makes? Table 1 shows that cognitive styles, represented by CIS_Inf, have predictive correlates in language. CIS_Inf captures two different variables (Fig 3) – how much a person's preference shifts before and after the experiment and whether they were influenced in making the decision. While discourse relation embeddings themselves seem to have low predictive power, specific relevant relations such as **Causal** and **Consonance** have high predictive power towards the cognitive styles of individuals pertaining to actual decision-making. With discourse relation features achieving an AUC of ∼0.8, language shows promise in capturing cognitive styles of individuals that are exhibited through their behavior.

Theoretical Features	↓CIS↓Inf	↓CIS↑Inf	↑CIS↓Inf	↑CIS↑Inf
OCEAN				
Openness	-0.14	0.15	-0.18	0.03
Conscientiousness	-0.14	-0.18	0.20	0.05
Extraversion	-0.03	-0.02	0.03	-0.01
Agreeableness	-0.05	0.14	0.05	-0.09
Emotional Stability	0.00	0.00	-0.05	0.02
Anxiety	0.07	-0.10	0.24	-0.06
Stress	-0.17	0.06	0.08	-0.04
Loneliness	-0.19	-0.17	0.08	0.12
Empathic Concern	0.16	0.10	-0.15	-0.03
Discourse relations				
Causal	0.29	-0.13	-0.15	0.06
Counterfactual	-0.11	0.01	0.13	-0.04
Consonance	0.04	-0.15	-0.22	0.18
Dissonance	0.18	-0.16	0.04	0.03

Table 2: Cohen's *d* for theoretical features against cognitive style outcomes of CIS_Inf.

To explore language-specific patterns that relate to each type of cognitive style, we also extracted correlations for theoretically-relevant lexical and discourse relation features in predicting each class of CIS_Inf. Results are presented in Table 2, where we find that the four classes are highly differentiable along lexical-based measures for personality (Park et al., 2015), anxiety (Mangalik et al., 2024), stress (Guntuku et al., 2019a), loneliness (Guntuku et al., 2019b) and empathic concern (Giorgi et al., 2023). Discourse relations, especially the Causal relation has a Cohen's *d* of 0.29 with the class ↓CIS↓Inf. We find that individuals who use more causal explanations and dissonant statements in their description of a recent past decision are less likely to change their minds about a decision due to external influence, and are less likely to change their preferences after making a decision, whereas, individuals who use less consonant statements in describing their decisions are more likely to switch their preferences after making a decision in the experiment.

Interestingly, higher linguistic dissonance is associated with less change in preferences / tendency to be influenced, which may signal difficulty in resolving dissonance surrounding one's decision. Higher change in preferences with low tendency to be influenced also seems to be signaled by linguistic anxiety, and each of the cognitive styles have a distinct signature across personality and well-being dimensions. This indicates that individual decision-making cognitive styles derived from simulated real-life experiments can be gleaned from personal discourse and the explanatory style of the person.

6 Conclusion

We demonstrated that experimentally-evoked cognitive styles can indeed be captured by language, offering a more solid "ground truth" compared to annotations of perceived behavior, which often fail to reflect a person's true state. Experiment-based validation is a way to move NLP evaluations beyond the limitations of annotation- or questionnaire-based labels. Here, we found that discourse-based cognitive language features were indeed able to predict the cognitive style measured from a classic experiment. Methods for capturing cognitive style can enhance our understanding of human cognition and improve applications in mental health therapy and agent engagement by tailoring approaches to individual thought processes.

Limitations

While our experiment aims to capture cognitive dissonance through language in tandem with the replication of Simon et al. (2004), our study does not include direct questions in the writing prompts that explicitly prompt participants to discuss their decision-making process within the experiment itself. Despite the indirect writing prompt, we were able to capture promising cognitive style of individuals irrespective of the experimental outcome. However, this indeed created limitations in directly predicting participants' actual decision-making behaviors.

While discourse relations were originally intended to capture cognitive states through coherence and rhetorical structures, our annotative method for inferring these relations offers only a small boost to the correlations when compared to lexical measures and contextual representations. This suggests that regular contextual embeddings might contain enough information to pick up cognitive styles and human behavior from language.

Our study population introduces several limitations that should be noted. The experiment uses undergraduate students at a public university which may limit the generalizability of the findings to other populations or age groups. While the study's focus on job decisions was particularly relevant to undergraduate students, who are often navigating a transitional phase focused career personal development, their decision-making processes may vary considerably from those of individuals in diverse life stages or professional environments. Furthermore, the linguistic outcomes were constrained by the small number of participants limited to the university. Therefore, the effect size was influenced by the restricted diversity in the population and the size of the participants.

Ethics Statement

This study included an experiment with human subjects. The experiment followed closely to what that has been well replicated with no known risks in the past. The experiments were approved by ethical Institutional Review Board (IRB) who conducted a full review granting their approval.

All participants provided informed consent prior to their participation. Participants were informed that they have the right to withdraw from the study at any time without any repercussions. Participants were also informed about how their data would be used and the measures taken to protect their privacy. Additionally participants confidentiality and privacy have been maintained throughout the research and analysis process. Any identifiable information collected during the study has been securely stored on a password-protected server, ensuring that only authorized personnel could access the information. All data were anonymized, any identifying details were removed or coded so that individuals could not be readily identified from the dataset. These steps ensured that the study upheld the highest ethical standards, prioritizing the privacy and well-being of all participants. The participants were paid USD 25 for completing the questionnaire after being recruited through the university.

We run all of our experiments on an NVIDIA-RTX-A6000 with 50 GB of memory in an internal server, on open-sourced models. The LLMs were used for inferences rather than training for zero- and few-shot settings, with resource usage of about 15-20 hours on a single GPU.

This work is part of a growing initiative to improve NLP for the human context. The models produced are not intended for any clinical or industrial application, and in particular not for targeted marketing or in use case where one's language is assessed for individual targeted information without individual awareness. The primary aim is to enhance the way cognitive processes are understood, ensuring that technology serves to augment psychological processes and measures.

References

Pilar Aguilar, Isabel Correia, Jan de Vries, and Leda Tortora. 2022. Cognitive dissonance induction as an "inoculator" against negative attitudes towards victims. *Social and Personality Psychology Compass*, 16(12):e12715.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's time to reason: Annotating argumentation structures in financial earnings calls: The finarg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169.

Matthew Brook O'Donnell and Emily B Falk. 2015. Linking neuroimaging with functional linguistic analysis to understand processes of successful communication. *Communication Methods and Measures*, 9(1-2):55–77.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Guillermo Campitelli and Fernand Gobet. 2010. Herbert simon's decision-making approach: Investigation of cognitive processes in experts. *Review of general psychology*, 14(4):354–364.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H Andrew Schwartz. 2024. Socialite-llama: An instruction-tuned model for social scientific tasks. *arXiv preprint arXiv:2402.01980*.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H Ungar, H Andrew Schwartz, and Joao Sedoc. 2023. Human-centered metrics for dialog system evaluation. *arXiv preprint arXiv:2305.14757*.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. 2019a. Understanding and measuring psychological stress using social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 214–225.

Sharath Chandra Guntuku, Rachelle Schneider, Arthur Pelullo, Jami Young, Vivien Wong, Lyle Ungar, Daniel Polsky, Kevin G Volpp, and Raina Merchant. 2019b. Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ open*, 9(11):e030355.

Eddie Harmon-Jones and Cindy Harmon-Jones. 2007. Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, 38(1):7–16.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond user self-reported likert scale ratings: A comparison model for automatic dialog evaluation. *arXiv preprint arXiv:2005.10716*.

Siddharth Mangalik, Johannes C Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V. Ganesan, Shashanka Subrahmanya, Nikita Soni, Sean AP Clouston, et al. 2024. Robust language-based mental health assessments in time and space through social media. *NPJ Digital Medicine*, 7(1):109.

April McGrath. 2017. Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*, 11(12):e12362.

JA Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Richard E Nisbett and Timothy D Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Delroy L Paulhus, Simine Vazire, et al. 2007. The self-report method. *Handbook of research methods in personality psychology*, 1(2007):224–239.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Darby E Saxbe, Xiao-Fei Yang, Larissa A Borofsky, and Mary Helen Immordino-Yang. 2013. The embodiment of emotion: language use during the feeling of social emotions predicts cortical somatosensory activity. *Social cognitive and affective neuroscience*, 8(7):806–812.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations*, pages 55–60.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Dan Simon, Daniel C Krawczyk, and Keith J Holyoak. 2004. Construction of preferences by constraint satisfaction. *Psychological Science*, 15(5):331–336.

Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. 2018a. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3350–3359, Brussels, Belgium. Association for Computational Linguistics.

Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. 2018b. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658, Vancouver, Canada. Association for Computational Linguistics.

Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. 2022. Discourse relation embeddings: Representing the relations between discourse segments in social media. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 45–55, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Vasudha Varadarajan, Swanie Juhng, Syeda Mahwish, Xiaoran Liu, Jonah Luby, Christian C. Luhmann, and H. Andrew Schwartz. 2023. Transfer and active learning for dissonance detection: Addressing the rare-class challenge. In *Proceedings of The 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Zixiaofan Yang, Shayan Hooshmand, and Julia Hirschberg. 2021. Choral: Collecting humor reaction labels from millions of social media users. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4429–4435.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendix

A Semantic Description of the *Decisions* Corpus

In order to describe the main themes of the corpus while preserving privacy of our study participants, we employed topic modeling. The results show that respondents writing about their difficult decisions frequently mention topics related to college education, career goals, finances, mental health, friendships, family relationships, and vacation plans. These subjects are largely connected to common decision-making aspects of student life.

Figure A.1: Randomly selected topics emerging from LDA on the participant writing describing a recent decision, depicting the types of content evoked.

B Job Offer Questions

The detailed questionnaire administered to the participants is shown in Table B.1.

C Prompts

The zero-shot and few-shot prompts for eliciting the CIS_Inf scores are shown in Table C.1.

Questions	Response type
Writing about a recent difficult decision	
1. Please describe a recent important and difficult decision that you have made (20-150 words)	text
2. What were the considerations that you thought about while making the decision? When answering, please consider all of the circumstances and details that went into the difficult decision (100-300 words)	text
Background: Imagine that you have just graduated from college and have decided to look for a job. You have had interviews with a few companies, and are hoping to receive some job offers. In this experiment you will be asked to state how you feel about an assortment of aspects that might be included in job offers. Specifically, you will be asked to state how desirable or undesirable you find each aspect. There are no right or wrong answers to these questions. Please state how you personally feel about these aspects as if you were evaluating them in the context of making a real decision about your future career. You are not expected to have any special knowledge. You might find that the information given to you is less complete than you would like to have; nonetheless, respond as best as you can given the available information. The issues are unrelated, so simply consider each one independently.	
1. A company maintains a national training center in Jackstown, Tennessee. Every employee must spend 3 weeks of training at that center every year. Most employees describe the training as boring and the life in Jackstown as gloomy. - Please select how desirable participating in the training sessions at Jackstown is to you.	-5 to 5
2. The commute to work will take you about 18 minutes each way. - Please select how desirable the 18 minute commute is to you.	-5 to 5
3. The average annual salary for the position you are considering is $60,000. The salary you are being offered is $61,200. - Please select how desirable it is to you to receive $1200 above the average salary.	-5 to 5
4. You will be given a cubicle, which is located in a pretty noisy area. - Please select how desirable it is to work in a cubicle.	-5 to 5
5. Given your credentials, you should be considered for promotion within a year or two. Being promoted will mean that you will have more independence, but it also means that you will have many more responsibilities. Some veterans maintain that in this type of profession, it is best to gain more experience before being promoted. - Please select how desirable a promotion is to you.	-5 to 5
6. All companies give their employees at least two weeks of vacation a year. Some companies give additional vacation benefits. A company offers you only the minimum two-week vacation. - Please select how desirable it is to receive only the minimum two-week vacation.	-5 to 5
7. The commute to work will take you about 40 minutes each way. - Please select how desirable the 40 minute commute is to you.	-5 to 5
8. You are offered an office to yourself. The office is pretty small, though adequate. - Please select how desirable the private office is to you.	-5 to 5
9. The average annual salary for the position you are considering is $60,000. A company offers you $59,100. - Please select how desirable it is to you to receive $900 below the average salary.	-5 to 5
10. In addition to the standard two-week annual vacation, a company takes its employees and their families to a week-long retreat in San Diego. The retreat consists of work-related lectures and workshops, but it is usually quite a lot of fun. - Please select how desirable the retreat in San Diego is to you.	-5 to 5
11. A company has a policy of encouraging personnel mobility among its numerous branches located throughout the country and across Europe. Every employee is entitled to spend up to 3 months every 2 years working at any one of the company's branches. - Please select how desirable this mobility is to you.	-5 to 5
1. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 1. The office	1 to 8
2. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 2. The commute	1 to 8
3. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 3. The salary	1 to 8
4. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 4. The vacation package	1 to 8

[DISTRACTION] Synonyms task: Match the synonyms for 20 moderately difficult English words

Background:
In this experiment you will be asked to play the role of a person who has just graduated from college. You are currently looking for a job in the field of marketing.You have just received interesting job offers from two large department store chains, Splendor and Bonnie's Best. The two companies are similar in terms of their size, reputation and stability, and your prospects for promotion seem the same with both companies. You have already spent a couple of days at each of their offices, and have been interviewed by the key personnel. You found both companies to be stimulating and pleasant. After receiving more information about the two job offers, you will be asked to decide which one to accept.

Participants randomly get one of the two configurations (one with Splendor in a positive loc condition and the other with Bonnie's Best in a positive loc condition):

Option A: Splendor (positive loc condition)
Splendor is located in a fun part of town, next door to a new mall. There are many food joints, clothing stores, and cinemas close by. Most of the employees there go out to lunch in groups and eat at different places every day. They also do some convenient shopping on their way home from work. The average annual salary of a person at your position is $60,000. The salary you are being offered by Splendor is $59,100. At Splendor, you are offered an office to yourself. The office is pretty small, though adequate.The commute to the offices of Splendor takes about 18 minutes each way. Splendor offers its employees two weeks of vacation a year.

Option B: Bonnie's Best
Bonnie's Best is located in a dull, sparsely populated industrial area. There is only one mediocre cafeteria nearby. Most employees bring their own sandwiches and eat on their own, or spend much of their lunch break driving to eateries that are a fair distance away. The average annual salary of a person at your position is $60,000. The salary you are being offered by Bonnie's Best is $61,200. At Bonnie's Best, you will be given a cubicle, which is located in a pretty noisy area. The commute to the offices of Bonnie's Best takes about 40 minutes each way. In addition to the standard two-week annual vacation, every summer Bonnie's Best takes its employees and their families to a retreat in San Diego. The retreat consists of work-related lectures and workshops, but it is usually quite a lot of fun.

Option A: Bonnie's Best (positive loc condition)
Bonnie's Best is located in a fun part of town, next door to a new mall. There are many food joints, clothing stores, and cinemas close by. Most of the employees there go out to lunch in groups and eat at different places every day. They also do some convenient shopping on their way home from work. The average annual salary of a person at your position is $60,000. The salary you are being offered by Bonnie's Best is $61,200. At Bonnie's Best, you will be given a cubicle, which is located in a pretty noisy area. The commute to the offices of Bonnie's Best takes about 40 minutes each way. In addition to the standard two-week annual vacation, every summer Bonnie's Best takes its employees and their families to a retreat in San Diego. The retreat consists of work-related lectures and workshops, but it is usually quite a lot of fun.

Option B: Splendor
Splendor is located in a dull, sparsely populated industrial area. There is only one mediocre cafeteria nearby. Most employees bring their own sandwiches and eat on their own, or spend much of their lunch break driving to eateries that are a fair distance away. The average annual salary of a person at your position is $60,000. The salary you are being offered by Splendor is $59,100. At Splendor, you are offered an office to yourself. The office is pretty small, though adequate. The commute to the offices of Splendor takes about 18 minutes each way. Splendor offers its employees two weeks of vacation a year.

Questions	**Response type**
At this point you have all the available information, and you are now asked to make your decision. Take your time and feel free to look back at the information provided. Please consider all pros and cons of both job offers carefully. Try to make this decision as if you were really in the described situation, and were facing a choice that will strongly influence your future career. When you have made your decision, please choose one of the two options. I accept the job offer of:	Bonnie's Best / Splendor
You will now be requested to state your preferences towards the aspects of the job offers made by Splendor and Bonnie's Best. Specifically, you are requested to state how desirable or undesirable you find each of these aspects. There are no right or wrong answers to these questions. Please state your subjective preferences. You are requested to answer the following questions using the provided scales. You are encouraged to use the full range of the scale:	
1.The commute to the offices of Splendor takes about 18 minutes each way. - Please select how desirable the 18 minute commute is to you.	-5 to 5
2. Splendor does not offer any vacation benefits above the minimum two-week vacation a year. - Please select how desirable it is to receive only the minimum two-week vacation.	-5 to 5
3.The salary you are being offered by Bonnie's Best is $1,200 above the average salary in the field. - Please select how desirable it is to you to receive $1200 above the average salary.	-5 to 5
4. At Splendor, you are offered an office to yourself. The office is pretty small, though adequate. - Please select how desirable the private office is to you.	-5 to 5
5. At Bonnie's Best, you will be given a cubicle, which is located in a pretty noisy area. - Please select how desirable it is to work in a cubicle.	-5 to 5
6. In addition to the standard two-week annual vacation, every summer Bonnie's Best takes its employees and their families to a retreat in San Diego. The retreat consists of work-related lectures and workshops, but it is usually quite a lot of fun. - Please select how desirable the San Diego retreat is to you.	-5 to 5
7. The commute to the offices of Bonnie's Best takes about 40 minutes each way. - Please select how desirable the 40 minute commute is to you.	-5 to 5
8. The salary you are being offered by Splendor is $900 below the average salary in the field. - Please select how desirable it is to you to receive $900 below the average salary.	-5 to 5
1. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 1. The office	1 to 8
2. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 2. The commute	1 to 8
3. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 3. The salary	1 to 8
4. Please state the relative weight you would assign each of the aspects in the overall context of choosing a job (using the slider). You are encouraged to use the full range of the scale: - 4. The vacation package	1 to 8

Table B.1: Detailed description of the job offer questionnaire that the participants were administered.

Shot	Prompt
0-shot	You are an expert social and cognitive psychologist analyzing decision-making patterns from the 2004 study "Construction of Preferences by Constraint Satisfaction". You are tasked with evaluating how preferences change when participants choose between two job offers with multiple attributes, in a simulated setting. This experiment measured preferences before and after making a decision, revealing "coherence shifts" where preferences aligned more closely with the chosen job offer, occurring both with and without influencing attributes in the job description. Your goal is to estimate two scores: (1) the score of a coherence shift towards preferring the chosen job offer, expressed as a value between 0 and 1, where 0 indicates an increased preference for the rejected offer and 1 indicates a strong preference for the chosen offer; and (2) the score that the decision is influenced by the job descriptions, also on a scale from 0 to 1, where 0 signifies no influence and a rigid preference, and 1 signifies being easily swayed by minor incentives. Base your assessment on text provided by the user about a recent personal decision that need not be related to the job offer scenario. Consider the cognitive styles and patterns of decision making evident in their narrative. Present your findings in this format: "The score of a coherence shift towards the chosen job offer is: <score>and the score of being influenced by minor incentives is: <score>," with each score ranging between 0 and 1.
4-shot	You are an expert social and cognitive psychologist analyzing decision-making patterns from the 2004 study "Construction of Preferences by Constraint Satisfaction". You are tasked with evaluating how preferences change when participants choose between two job offers with multiple attributes, in a simulated setting. This experiment measured preferences before and after making a decision, revealing "coherence shifts" where preferences aligned more closely with the chosen job offer, occurring both with and without influencing attributes in the job description. Your goal is to estimate two scores based on user-provided text: (1) the score of a coherence shift towards preferring the chosen job offer, expressed as a value between 0 and 1, where 0 indicates an increased preference for the rejected offer and 1 indicates a strong preference for the chosen offer; and (2) the score that the decision is influenced by the job descriptions, also on a scale from 0 to 1, where 0 signifies no influence and a rigid preference, and 1 signifies being easily swayed by minor incentives. Here are four different examples of participants' narratives about recent personal decisions and with a score towards 1 if they had a coherence shift towards the chosen job offer, 0 if coherence shift is towards the rejected offer. Similarly, there is also a score for if being influenced by minor incentives (1 if influenced, 0 if not influenced): Example 1: User's Narrative: "I recently had to decide whether to buy a new car or keep my old one. The new car had better fuel efficiency and more features, but I was attached to my old car due to sentimental reasons. After considering the costs and benefits, I decided to go with the new car." Output: "The score of a coherence shift towards the chosen job offer is: 0.8 and the score of being influenced by minor incentives is: 0.6." Example 2: User's Narrative: "I was choosing between two vacation destinations: a beach resort and a mountain cabin. I love both settings, but ultimately chose the beach resort because it was more affordable and had better amenities." Output: "The score of a coherence shift towards the chosen job offer is: 0.7 and the score of being influenced by minor incentives is: 0.5." Example 3: User's Narrative: "I had to decide whether to take an online course or attend in-person classes for my professional development. The online course was more flexible, but I prefer face-to-face interaction. I chose the online course because it fit better with my schedule." Output: "The score of a coherence shift towards the chosen job offer is: 0.9 and the score of being influenced by minor incentives is: 0.4." Similarly, for the following user input text, estimate the scores. Base your assessment on text provided by the user about a recent personal decision that need not be related to the job offer scenario. Consider the cognitive styles and patterns of decision making evident in their narrative. Present your findings in this format: "The score of a coherence shift towards the chosen job offer is: <score>and the score of being influenced by minor incentives is: <score>," with each score ranging between 0 and 1 as a continuous value.

Table C.1: Zero- and 4-shot prompts for both Llama3.1 and Gemma models.