Erroneous Theory of Mind in the Context of Speed Dating

Andrew Scutt, Amitav Johri, Zijian Sun.

scuttand, johriami, sunziji1

andrew.scutt@mail.utoronto.ca, amitav.johri@mail.utoronto.ca, zijiansun.sun@mail.utoronto.ca

Department of Economics, University of Toronto

ECO482H1: Machine Learning Applications in Macroeconomic Finance

Dr. Marlène Koffi

April 6th, 2025

1 Introduction

Once a matter of fate or family, dating has evolved into a multi-billion dollar global industry, valued at $9.3 billion in 2023 (Research and Markets Ltd, 2024). Despite technological advances and sleek algorithms, most consumers still yearn for a compatible, long-term partner (Chavda, 2024). However, successfully forming a match requires that individuals develop a sufficient understanding of others—facilitated by empathic insight into others' thoughts and feelings, compatibility in attachment styles, and effective perspective-taking (Ainsworth et al., 1978; Dodell-Feder et al., 2015; Ickes and Hodges, 2013). Interestingly, theory of mind, an individual's capacity to accurately impute the mental state of themselves and others, underlies these factors (Premack and Woodruff, 1978). Research shows that brain activity related to thoughts about romantic partners is positively associated with that partner's well-being (Dodell-Feder et al., 2015), which in turn plays a crucial role in creating romantic relationships. Despite the growing academic interest in romantic match-making, one critical component remains largely unexplored: the ability to accurately predict whether someone wants to date you, reflecting theory of mind. Acquiring this knowledge would allow us to enhance decision-making and promote greater matching efficiency in romantic markets from the perspective of the individual decision-maker.

In this study, we ask: can we predict an individual's accuracy of their partner's desire to commit romantically? We define accuracy as a theory-of-mind-driven ability to correctly assess whether one's date is interested in a romantic relationship. Moreover, we know that theory of mind can be inferred through cognitive biases (Royzman et al., 2003). Therefore, we also investigate whether cognitive biases are important in predicting this accuracy.

Regarding these biases, previous psychological literature has identified crucial factors that obscure an individual's judgment of the perspectives of others. Firstly, overconfidence bias, defined as an individual's tendency to overestimate their abilities, or how

positively others view them, causes individuals to inflate their perceived matchmaking success, ultimately fostering misguided beliefs that their romantic commitment is reciprocated (Kruger and Dunning, 1999). Secondly, social projection bias — the tendency for individuals to assume that others share their current thoughts, feelings, or preferences — can lead to an overestimation of similarity between one's own mental state and that of others. As a result, individuals may overestimate how much their partner wants to commit to them depending on similarities (Robbins and Krueger, 2003; Bushong and Gagnon-Bartsch, 2024).

From our extensive literature review, we believe that we can predict a person's accuracy; in addition, we think that our cognitive biases will be crucial factors when predicting this accuracy. This research contributes to the broader field of marriage and family economics, which explores how individuals form partnerships. More precisely, it would aid in boosting matching efficiency in dating markets and avoid wasting resources on people who are not interested in romantically committing to you. In addition, it may provide unique insights about factors that render a person's accuracy erroneous.

In relation to previous economic literature, recent marriage economics research has highlighted younger generations' growing preferences for partners with similar character traits (Cheremukhin et al., 2024). Interestingly, this finding converges with social projection bias. Previous economic speed-dating research has primarily focused on identifying patterns in mate selection and examining how traits such as physical attractiveness, intelligence, and race influence romantic decision-making, often through the use of econometric models (Fisman et al., 2006). Unfortunately, due to the novel focus of our study, there is limited prior work directly comparable to ours, invalidating direct methodological comparisons. Crucially, Fisman et al.'s (2006) influential speed-dating dataset is especially well-suited to our question, as it includes participants' self- and partner-assessments, predicted match success, and actual romantic outcomes. From these, we construct measurable proxies for cognitive biases and romantic accuracy.

2 Data and Methodology

2.1 Data

Concerning our experimental data, Fisman et al. (2006) led speed dating sessions in which participants engaged in four-minute conversations. If both individuals expressed interest by selecting "yes," their contact information was shared afterward. These participants were drawn from Columbia University's graduate and professional schools. Recruitment was conducted via mass emails and flyers distributed on campus, with interested students registering online by providing their contact information and completing a pre-event survey. These speed dating events took place in a private room at a bar/restaurant near campus on weekday evenings between 2002 and 2004.

Upon arrival, participants were given a clipboard, pen, and a nametag displaying only their ID number. This clipboard recorded each participant's ID number and circled "yes" or "no" to indicate whether they wished to meet that person again. In addition to their decision, participants rated themselves and each partner on six attributes: attractiveness, sincerity, intelligence, funniness, and ambition. They also included demographic information, such as age, race, gender, and field of study as well as how much they thought their partner liked them (*guess_prob_like*). Finally, participants reported their personal interests, from which the researchers computed an interest correlation score to quantify similarity between individuals. Afterwards, all match data were compiled into a file where each observation represents a match.

After data cleaning, we had 4913 observations. In terms of descriptive statistics, the correlation between participants' perceived likelihood of being liked and their partner's actual decision (*decison_o*) was relatively weak (r = 0.14), suggesting that individuals are generally poor at accurately judging how much others like them. In contrast, the correlation between how they liked their partner and their own decision to match was substantially stronger (r = 0.52), indicating that personal attraction is a significant driver of commitment.

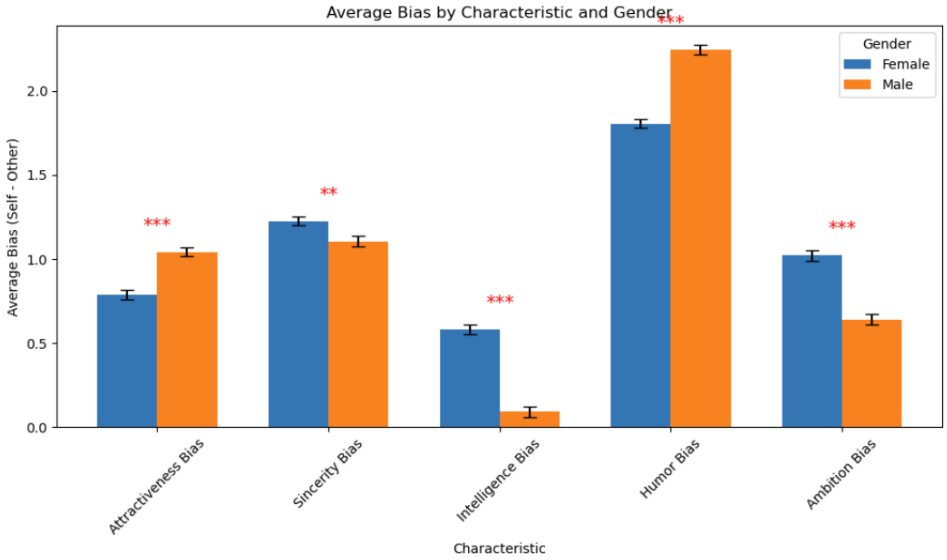

Figure 1: Characteristic Biases for Each Gender

Interestingly, men and women exhibit similar levels of characteristic rating bias (own rating - partner's rating of you) in the aggregate. However, women tend to overestimate their intelligence and ambition relative to how others perceive them, while men significantly overestimate their sense of humor. These results suggest that characteristic-specific biases may align with gendered self-perceptions and stereotypes. Finally, there were no economically or statistically differences in commitment accuracy across genders.

2.2 Methodology

To commence, we operationalize commitment accuracy as follows:

$$f(\text{guess\_prob\_like}, \text{decision\_o}) = \begin{cases} 1, & \text{if } (\text{guess\_prob\_like} > 5 \wedge \text{decision\_o} = 1) \\ & \quad \text{or } (\text{guess\_prob\_like} < 5 \wedge \text{decision\_o} = 0) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Albeit an indirect measure, *guess_prob_like* is the only variable we have to infer our participant's guess of how much their partner wants to commit. In addition, previous research on romantic decision-making shows that individuals who romantically like certain individuals tend to commit to them (Arriaga and Agnew, 2001); this notion is also

supported by our strong correlation between how much you like somebody and whether you choose to match with them (r = 0.52). In terms of predictors, we developed models to estimate each type of bias based on their respective definitions and properties.

Regarding overconfidence bias, we know that its domain-specific (West and Stanovich, 1997); therefore, we operationalize overconfidence bias as follows: *OverconfidenceBias* = {*Rating Bias*, where *RatingBias* = *OwnRating* - *PartnerRating*. Importantly, rating bias can be understood as a manifestation of overconfidence, as they stem from discrepancies between individuals' self-perceptions and how others perceive them. The first regression model assesses overconfidence bias by analyzing discrepancies between participants' self-ratings and their potential partner's evaluations across key personal characteristics where $C = \{$*Attrativeness*, *Sincerity*, *Intelligence*, *Humour*, *Amibition*$\}$. Positive discrepancies reflect overconfidence. Thus, we derive overconfidence bias for each characteristic for each individual.

$$RatingBias_{i,c} = \beta_c + \epsilon_{i,c} \tag{2}$$

To account for social projection bias, we employ age difference, interest correlate, and similarity scores for race and field of study, where $similarityscore_x = 1$ indicates a match and 0 otherwise. In addition, previous literature has documented gender-based differences in sexual and romantic interest, supporting the usage of gender as a stand-alone predictor (Haselton, 2003). Since variables like *diff_age*, *sim_race*, etc are transformations or interactions based on core variables such as *age*, *age_o*, *race*, and *race_o*, we retain the original variables in the model. This process follows the general principle in predictive modeling that discourages discarding original information, as it may still hold unique predictive value beyond its derived forms. The same logic is applied to characteristic ratings as well. All in all, we have 27 predictors X = {*diff_age*, *same_race*, *same_field*, *interests_correlate*,*bias_attractive*, *bias_sincere*, *bias_intelligence*, *bias_funny*, *bias_ambition*,*gender*, *race*, *age*, *field*, *race_o*, *age_o*, *gender_o*, *field_o*, *attractive_partner*, *sincere_partner*, *funny_partner*, *intelligence_partner*, *ambition_partner*, *attractive*, *sincere*, *intelligence*, *funny*, *ambition* }

In our analysis, we employed a diverse set of models to capture a range of behavioral patterns in romantic prediction data. We began with a logistic regression model, which is well-suited for binary classification problems and offers clear interpretability. Logistic regression assumes a linear relationship between the log-odds of the outcome and the predictor variables, making it a strong baseline for comparison and inference. On the other hand, this strong assumption may be problematic if it does not reflect our true underlying functional form. In addition, we incorporated ridge and lasso regularization to potentially improve model performance. Ridge penalization shrinks all coefficient estimates toward zero without eliminating any predictors, allowing us to retain all variables while reducing the influence of less important ones. In contrast, lasso performs both shrinkage and variable selection by setting some coefficients exactly to zero, effectively removing irrelevant predictors and reducing the dimensionality of the model.

Next, we implemented a K-Nearest Neighbors (KNN) model. This model assumes that individuals with similar observed characteristics on our X variables tend to behave similarly, warranting identical classification. However, this assumption may be problematic in our context, as being similar in terms of certain biases may not warrant identical predictive accuracy. Individuals with comparable demographics or preferences can exhibit markedly different behaviors, potentially weakening KNN's predictive power in this domain.

To model non-linear interactions and capture more nuanced relationships, we included random forests and XGBoost. Random forests takes many training sets from the population, builds a separate prediction model using each training set, and average the resulting predictions, reducing variance. It typically performs well when behavioral patterns are stable across individuals. In contrast, XGBoost (Extreme Gradient Boosting) builds trees sequentially, with each new tree correcting the residuals of the previous one. This model excels in edge cases where prediction is especially difficult—situations common in noisy, inconsistent behavioral data. However, its tendency to overfit noisy patterns if not properly tuned can be a concern.

Together, these four models allow us to account for a spectrum of psychological and

behavioral variability: from consistent patterns (best captured by random forests and logistic regression) to inconsistent, harder-to-predict behavior (better captured by boosting methods). This multi-model approach ensures robust comparisons across varying assumptions about human behavior in romantic contexts. Due to the class imbalance in our data, stemming from many more people being accurate than inaccurate, we balanced the classes by assigning a higher weight to the cases in which individuals were incorrect. If this were not the case, the models will learn to predict accuracy more often by default, which would naturally achieve a higher accuracy. By assigning a higher weight to the under-represented class, we balance the discrepancy in the data. In addition, we used 5-fold cross validation and 20-80 test-training split. Finally, all models used hyper-parameter tuning.

3 Results

Model Performance Comparison

Model	Accuracy	False Positive Rate	Auc Score	Recall Rate	F1 Score (weighted average)	Hyperparameter Tunning
KNN	0.63	0.89	0.58	0.99	0.57	K = 40
Lasso Logistic	0.36	0.05	0.58	0.04	0.23	Regularization Strength: 0.0886
Ridge Logistic	0.55	0.43	0.58	0.54	0.56	Regularization Strength: 0.0886
Logistic Regression	0.54	0.44	0.58	0.53	0.56	Regularization Strength: 0.0007
Random Forest	0.68	0.7	0.65	0.88	0.64	Max depth: 10 Min split: 5 Estimators: 200
Gradient Boosting	0.62	0.53	0.64	0.7	0.62	Learning rate: 0.1 Max depth: 3 Estimator: 50

Figure 2: Results Table

Using varying metrics for the accuracy of our models, we conclude that our random forest model was the most accurate. The random forest model has the highest accuracy rate, AUC score and F1 score, as well as the second-highest recall rate. The false positive

rate is relatively higher than the other models, but this is compensated for by the model's performance in other metrics. Note that many of our models have a high false positive rate, meaning that our models indicate that individuals accurately determine their partner's desire to commit when, in reality, they are not. In addition, high recall rates imply that the model was precise when it came to identifying accurate individuals. This suggests a class imbalance in our model, and the fact that the algorithm is assigning too many data points as "accurate"; hence warranting the usage of a class balance in our models. The high F1 scores imply that our models correctly identify true positives, while high AUC scores mean that our model correctly ranks those who are "accurate" above those who are "inaccurate" in their estimate of whether their partner likes them.

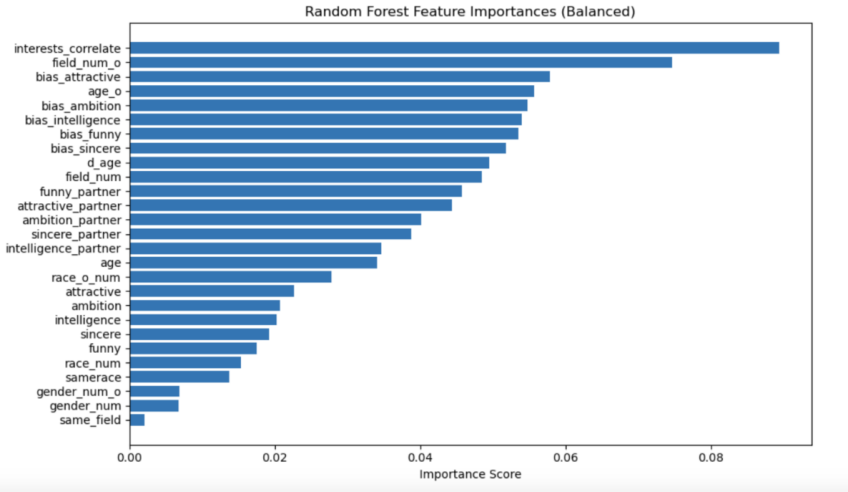

Figure 3: Importance Matrix

We find that all overconfidence bias variables—and some social projection bias variables—are significant, suggesting that overconfidence plays a larger role than social projection in this context. Additionally, the field of the other person significantly predicts accuracy, which may also reflect how perceived similarity or expertise influences judgment.

4 Conclusion

Our study demonstrates that individuals' ability to accurately predict their partner's romantic interest—an applied form of theory of mind—can be meaningfully predicted using machine learning models. Among the models tested, the random forest algorithm

performed best across most evaluation metrics, including accuracy, AUC, and F1 score. However, the high false positive rates observed across models indicate that people often overestimate their ability to judge others' intentions, aligning with psychological literature on theory of mind limitations.

Importantly, overconfidence bias consistently emerged as a key predictor of inaccurate judgment, while some social projection bias indicators also contributed, albeit to a lesser extent. This highlights the dominance of self-perception distortions—particularly inflated self-evaluations—in shaping erroneous social inference during dating interactions.

These findings have several implications. From a behavioral economics standpoint, improving match efficiency in romantic markets may require addressing these systematic cognitive biases. Applications could include nudging individuals toward more accurate self-perceptions or designing dating platforms that account for common misjudgment patterns. In the context of marriage and family economics, enhancing accuracy in partner evaluation may reduce mismatched investments of time and emotional energy, leading to more optimal relationship formation.

Future Directions

Future research could extend this work by incorporating richer psychological and behavioral measures, such as empathy, attachment style, or communication quality, to refine predictions of judgment accuracy. Additionally, longitudinal data would allow us to explore how theory of mind accuracy evolves over repeated interactions and whether it predicts long-term relationship success. Lastly, deeper examination of how different types of biases interact—particularly in high-stakes or emotionally charged contexts—would offer a more comprehensive understanding of decision-making in romantic settings.

5 References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. N. (1978). *Patterns of attachment: A psychological study of the strange situation.* http://ci.nii.ac.jp/

ncid/BB19374390

Arriaga, X. B., & Agnew, C. R. (2001). Being committed: Affective, cognitive, and conative components of relationship commitment. *Personality and Social Psychology Bulletin, 27*(9), 1190–1203. https://doi.org/10.1177/0146167201279011

Bushong, B., & Gagnon-Bartsch, T. (2024). Failures in forecasting: An experiment on interpersonal projection bias. *Management Science, 70*(12), 8735–8752. https://doi.org/10.1287/mnsc.2022.00655

Chavda, J. (2024, April 14). *Key findings about online dating in the U.S.* Pew Research Center. https://www.pewresearch.org/short-reads/2023/02/02/key-findings-about-online-dating-in-the-u-s/

Dodell-Feder, D., Felix, S., Yung, M. G., & Hooker, C. I. (2015). Theory-of-mind-related neural activity for one's romantic partner predicts partner well-being. *Social Cognitive and Affective Neuroscience, 11*(4), 593–603. https://doi.org/10.1093/scan/nsv144

Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics, 121*(2), 673–697. https://doi.org/10.1162/qjec.2006.121.2.673

Haselton, M. G. (2003). The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events. *Journal of Research in Personality, 37*(1), 34–47. https://doi.org/10.1016/s0092-6566(02)00529-9

Ickes, W., & Hodges, S. D. (2013). Empathic accuracy in close relationships. In *Oxford University Press eBooks*. https://doi.org/10.1093/oxfordhb/9780195398694.013.0016

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lee, A. J., Sidari, M. J., Murphy, S. C., Sherlock, J. M., & Zietsch, B. P. (2020). Sex differences in misperceptions of sexual interest can be explained by sociosexual ori-

entation and men projecting their own interest onto women. *Psychological Science, 31*(2), 184–192. https://doi.org/10.1177/0956797619900315

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526. https://doi.org/10.1017/s0140525x00076512

Research and Markets Ltd. (2024). *Dating services market size, competitors & forecast to 2032*. Research and Markets Ltd. https://www.researchandmarkets.com/report/dating-service?utm_source=GNE&utm_medium=PressRelease&utm_code=9m7xrt&utm_campaign=2016313

Robbins, J., & Krueger, J. (2003). Social projection to ingroups and outgroups: A meta-analytic integration [Dataset]. *PsycEXTRA Dataset*. https://doi.org/10.1037/e633872013-608

Restrepo-Echavarria, P., Tutino, A., & Cheremukhin, A. (2023c). *Marriage market sorting in the U.S.* https://doi.org/10.20955/wp.2023.023

Royzman, E. B., Cassidy, K. W., & Baron, J. (2003). "I know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology, 7*(1), 38–65. https://doi.org/10.1037/1089-2680.7.1.38

Sim, S. Y., Saperia, J., Brown, J. A., & Bernieri, F. J. (2015). Judging attractiveness: Biases due to raters' own attractiveness and intelligence. *Cogent Psychology, 2*(1), 996316. https://doi.org/10.1080/23311908.2014.996316

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review, 4*(3), 387–392. https://doi.org/10.3758/bf03210798