Essays in Experimental Economics

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften der Universität Mannheim

vorgelegt von

Can Çelebi

im Herbst-/Wintersemester 2024

Abteilungssprecher:	Prof. Dr. Thomas Tröger
Referent:	Prof. Stefan P. Penczynski, Ph.D.
Korreferent:	Dr. Stefano Balietti, Ph.D.
Tag der mündlichen Prüfung:	16. September 2024

Acknowledgments

A big thank you to Prof. Dr. Stefan Penczynski. His support and guidance have been essential throughout my PhD journey. Our conversations, which often extend beyond academic topics, are both enjoyable and insightful. We've achieved a good amount of work together, and I look forward to continuing our collaboration and exploring new research questions.

And another big thank you to Dr. Stefano Balietti. His guidance was crucial in transitioning my experiments to an online format and venturing into cryptocurrencies, leading to the world's first global cryptocurrency investment survey. I've learned so much about coding from him and thankful to him for giving me an excuse to master generative AI image tools for research purposes. His support during COVID-19, connecting me with colleagues at the Max Planck Institute, was invaluable and helped me navigate those tough times.

This page is intentionally left blank

Anneme

This page is intentionally left blank

Contents

List of Tables

List of Figures

This page is intentionally left blank

Preface

The dissertation consists of three distinct chapters, all related to experimental economics. Each chapter incorporates artificial intelligence (AI) in a unique way: as a complementary tool to ensure robustness (Chapter 1), as a novel methodological tool under investigation (Chapter 2), or as the central concept around which research questions revolve (Chapter 3).

A primary motivation for my research is to challenge traditional methods and the established academics who defend them. These researchers, comfortable with long-established methodologies, are often skeptical of new approaches and tend to dismiss them. They have dominated the field for decades, claimed the most intriguing questions, accumulated extensive experience, and have relatively easy access to resources. Consequently, competing with them on their terms is an uphill battle I cannot win.

Given the accumulated wisdom in the literature and the established protocols using traditional methods, one approach to gaining a stake in the field is to gradually deviate from traditional methodologies by incorporating new technologies. To surpass established researchers, one must rewrite some of the rules or add new ones. AI offers this possibility, introducing new research questions and novel methodological practices to experimental economics.

Older individuals and those with high self-esteem are less inclined to adopt new technologies (Mahmud et al., 2022). Therefore, it is likely that established researchers will continue to overlook the power of AI and its potential to revolutionize the field. Even if they eventually recognize its significance, they may struggle to adapt and devise novel research questions due to their day-to-day lack of experience with it. Even if they do not struggle, I already have the first mover's advantage in this game of accumulated knowledge and experience.

By leveraging AI, I aim to open new research avenues, create a distinct niche that builds upon yet transcends traditional methodologies, and gain a competitive advantage against established researchers in the field.

Chapter 1

Chatper 1 is based on joint work with Dr. Stefan Penczynski. This chapter investigates strategic voting under the unanimity rule in jury settings, challenging the traditional view that unanimity protects against wrongful convictions. Previous studies by Feddersen and Pesendorfer (1998) theoretically demonstrated that strategic voting under unanimity should increase as the jury size increases, which, in turn, can lead to higher error rates (false conviction) in larger juries. Guarnaschelli et al. (2000) confirmed strategic voting behavior experimentally but found no significant difference in the percentage of jurors who vote strategically between jurors of difference sizes. To address these discrepancy between the theory and empirical results, we propose a non-equilibrium level-k model to explore the cognitive processes behind strategic voting. Our model suggests a cyclical pattern where players at different levels alternate between informative and strategic voting. Furthermore, unlike the traditional Bayesian Nash equilibrium predictions of Feddersen and Pesendorfer (1998), our level-k model does not predict a change in the jurors' likelihood to vote strategically as the jury size increases. We tested our theory through a two by two experimental design, varying jury sizes (3 and 6) and information signal sampling methods (with and without replacement).

Our findings indicate that strategic complexity, influenced by these variations, affects the distribution of strategic sophistication among jurors. Our results align with Guarnaschelli et al. (2000), showing no significant difference in strategic voting frequency between jury sizes but highlighting an unexpected decrease in convicting the innocent and an increase in acquitting the guilty. This is explained by the prevalence of less sophisticated level-1 players in more complex settings. Sampling without replacement, perceived as less complex, resulted in higher strategic sophistication. Our analysis suggests that the level-k model effectively captures the dynamics of strategic voting in juries and can predict performance across different settings. To ascertain subjects' level of thinking, we conduct the experiment with the intra-team communication protocol, introduced by Burchardi and Penczynski (2014), that yields incentivized written accounts of subjects' individual reasoning. The messages are classified independently by two research assistants (RA). For each individual decision's message, RAs classified the level of reasoning that the message corresponds to most closely. We have additionally utilized GPT-4 to check the robustness of the classification outcomes by RAs. The classifications by GPT-4 aligned with 91.3% of the RAs' classifications.

Chapter 2

Chatper 2 is based on joint work with Dr. Stefan Penczynski. It investigates the classification capabilities of GPT models in processing text data from economic experiments, particularly focusing on promises and strategic thinking. This chapter seeks to answer research questions regarding the performance of GPT models compared to human annotators and traditional machine learning methods, the adaptability of human classification instructions for large language model (LLM) prompts, and the efficacy of various prompting techniques. Specifically, we utilized GPT-3.5 and GPT-4 models to classify text transcripts. We varied the prompts between 0-shot and few-shot techniques and experimented with the use of the 0-shot chain-of-thought (CoT) prompting technique.

Additionally, the chapter aims to serve as a primer for researchers interested in integrating LLMs into their work by providing a comprehensive review of the computer science literature specific to prompt engineering and offers a critique of the current application of LLMs in social sciences.

Results indicate that GPT-4 consistently outperforms GPT-3.5, achieving accuracy levels near or above 90% for most tasks, and 73% for more complex tasks. Furthermore, GPT-4 demonstrates notable gains with n-shot and 0-shot-CoT prompting, particularly for relatively more complex tasks, while GPT-3.5 shows mixed results. Additionally, the chapter highlights the cost-effectiveness and efficiency of using LLMs over human annotators. LLMs provide immediate, consistent results with detailed justifications and are not susceptible to fatigue, making them a valuable tool for researchers working with text data. Moreover, results showed that GPT-4's performance is comparable to human annotators and surpass traditional natural language processing methods. Overall, the chapter underscores the potential of LLMs to revolutionize text data analysis in social sciences, offering a scalable, cost-effective, and highly accurate alternative to traditional methods and human annotation.

Chapter 3

Chapter 2 is based on joint work with researchers from Max Planck Institute for Intelligent Systems.[1] Chapter 3 investigates how offering algorithmic advice as an optional tool influences its utilization compared to mandatory advice. It also examines the effect of additional accuracy information on advice utilization, considering subjects' prior expectations and experience with AI systems. The research employs facial image pair classification tasks, where subjects determine whether pairs of facial images belong to the same person and face recognition system (FRS) as its AI system.

Results show that subjects are more likely to use algorithmic advice when they actively solicit it. However, because advice is not solicited frequently enough, the overall utilization of algorithmic advice is lower when it is optional compared to when it is mandatory. Additionally, the study finds that additional accuracy information generally

[1]See Chapter 3, section 3.8 for details.

has a negative effect on algorithmic advice utilization. This negative effect occurs because subjects with near-perfect performance expectations adjust their beliefs downward upon seeing information indicating FRS performance below near-perfect levels. Furthermore, the role of prior understanding and experience with FRS is found to be significant. Subjects with higher levels of understanding and experience have more accurate performance expectations and are less affected by additional accuracy information. Conversely, those with less understanding and experience are more prone to the perfect automation bias and adjust their beliefs downward, reducing advice utilization. Lastly, the chapter documents that when advice is solicited, subjects tend to ignore additional accuracy information, nullifying its debiasing effect. This behavior is consistent regardless of subjects' FRS understanding or experience levels. The findings suggest that providing further autonomy to human decision makers by making advice optional can lead to lower overall utilization of algorithmic advice and the disregarding of important accuracy information. Consequently an attempt to provide additional agency to the human decision maker by giving them the choice to get algorithmic advice is not recommended for in Algorithm-in-the-loop systems.

Chapter 1

Strategic thinking in jury decisions

Joint with Stefan P. Penczynski

1.1 Introduction

Predicated on the assumption that jurors vote non-strategically, the unanimity rule in juries has traditionally been seen as a safeguard against convicting the innocent. Building upon the work of Austen-Smith and Banks (1996), Feddersen and Pesendorfer (1998, FP) challenged this perspective and showed that, under the a priori assumption that a defendant is equally likely to be innocent or guilty, an equilibrium with the unanimity rule can encompass strategic voting. In contrast to informative voting, strategic voting might not simply reveal private information.

FP proved that the presence of strategic voting leads to higher error rates in convicting the innocent and acquitting the guilty compared to juries under simple majority or supermajority rules, where strategic voting is absent in equilibrium under the same initial assumptions. Furthermore, they highlighted that the issue of an increased rate of convicting the innocent under unanimity voting becomes notably exacerbated with larger jury sizes.

Guarnaschelli et al. (2000, GMP) experimentally confirmed the presence of strategic voting under unanimity voting, but also documented results that to some extent contradict FP's claims on error rates. They sought to reconcile this partial discrepancy between theory and data by utilizing the Quantal Response Equilibrium (QRE) model's statistical nature (McKelvey and Palfrey, 1995).

We introduce an alternative non-equilibrium framework, using level-k modeling, to address this discrepancy and offer a fresh perspective on jury voting under the unanimity rule, both theoretically and experimentally. Despite a large literature on strategic thinking and heterogeneous level-k reasoning that includes work on auctions, social learning and other related settings (Crawford et al., 2013), to the best of our knowledge the precise cognitive processes of strategic voting and their concrete influence on jury accuracy have not yet been studied.

We model the jury voting via iterative best responses. We begin with the assumption that all level-0 players vote uninformatively. A level-1 player best-responds to level-0 players by voting her signal (informative voting); and a level-2 player best-responds to level-1 players by voting "guilty" irrespective of her signal (strategic voting). The intuition is that informative voting with imprecise signals will otherwise never lead to convictions under the unanimity rule. With strategic voting being inherently uninformative, level-3 players revert to informative voting in response. This establishes a cyclical best-response pattern: higher odd-level players respond to strategic voting with informative voting and higher even-level players counteract informative voting with strategic voting.

We attained nuanced insights into individual reasoning beyond mere vote observation by employing an intra-team communication protocol that provides written accounts of subjects' decision justifications (Burchardi and Penczynski, 2014; Penczynski, 2017). These accounts substantiated the existence of the modelled level-k reasoning types within our sample and allowed us to differentiate among overlapping voting behaviors of various level-k types, that arises due to our model's cyclical best-response characteristic. Additionally, this approach enhanced the robustness of identifying strategic voters by screening out level-0 players who might otherwise be misidentified as strategic voters if categorizations were exclusively based on observed votes.

In a 2×2 within-subject experimental design, we investigate strategic voting across two jury sizes ($n = 3, 6$) and two sampling methods for the information signal (with and without replacement). While our theoretical model predicts that specific level-k types do not change their behaviour when jury sizes or signal sampling methods vary, both variations are predicted to affect the Nash equilibrium (NE) behavior and are expected to influence the complexity of the strategic task.

Given a specific level-classification, the observed voting behavior closely mirrors the theorized behavior and remains largely unaffected by treatments. Yet, these treatments influence the aggregate voting behavior by altering the distribution of strategic sophistication. This influence partly explains the discrepancies between observed jury accuracy and the optimal accuracy predicted by NE. Crucially, the level distributions are primarily affected by the strategic complexity introduced by the treatment variation, so that the aggregate voting behavior not necessarily aligns with the expected changes under optimal voting.

We find results in line with GMP, who found no significant difference in the frequency of strategic voting between jury sizes of 3 and 6. Both studies rather find a significant decrease in convicting an innocent, and an increase in acquitting a guilty, contrary to the NE predictions. We can explain this with the informative votes of less sophisticated level-1 types, which become more frequent in the more complex setting with $n = 6$ jurors. We similarly find that the sampling without replacement seems less complex and leads to higher strategic sophistication compared to sampling with replacement.

In light of these results, we propose that a model of heterogeneous types of voters such as the level-k model is well-suited to understand and predict jury performances across a number of different settings because it represents the underlying cognitive processes of strategic voting. Furthermore, our results suggest that more research is needed to understand how task complexity leads to changes in the distribution of sophistication.

Our results relate the setting of jury voting to other settings, in which the plurality of types has been identified as important for a good overall outcome. For example, in social learning, the heterogeneity of types improves upon the inefficiencies of fully rational behaviour – information cascades and herding – thanks to occasional and private-information-revealing level-1 decisions (Penczynski, 2017). Furthermore, the fact that the observed level distribution still leads to a good jury performance resembles the coordination "magic" in market-entry games, in which heterogeneous beliefs seem to provide a useful mechanism of sorting otherwise homogeneous players into market entrants and non-entrants (Rapoport et al., 1998; Camerer et al., 2004).

Due to the observed influence of the treatment variations on the level-k distribution, our work also relates to the idea of endogenous depth of reasoning (Agranov et al., 2012; Alaoui and Penta, 2016). If the strategic sophistication is the result of a cost-benefit analysis of additional steps of reasoning as modeled by Alaoui and Penta (2016), our analysis suggests that a larger number of possible signal realizations within the jury, be it due to a larger jury or to sampling without replacement, increases the complexity of the game and the cost of deliberation and hence lowers the average observed sophistication.

Our results suggest that conditioning on being pivotal – casting a decisive vote that determines the jury decision – in voting is much less inhibitive of strategic sophistication than conditioning on bidding highest in first-price auctions (Eyster and Rabin, 2005; Crawford and Iriberri, 2007; Li, 2017). In auctions, an intra-team communication analysis suggests that "subjects may actually have problems to form even a basic belief" as only 15% of subjects deliberate the other players' decision and thus qualify for anything other than level-0 (Koch and Penczynski, 2018, p. 79). In contrast, the level distribution in voting is very similar to commonly found distributions in other settings.

Incorporating sampling without replacement in our experiment is inspired by Rabin (2002), who models inference via a sampling process without replacement in order to reflect the common belief in the law of small numbers (Tversky and Kahneman, 1974). Similar to our motivation, the increased difficulty of dealing with independent signal and the simplification of draws that are more "representative" of the urn composition have led Grimm and Mengel (2020) to use sampling without replacement in their experiment. Our results support the intuition that this sampling is easier to understand.

1.2 Theory

1.2.1 Model Setup

Consider a game with n jurors. Nature determines the state of the world to be red or blue, $S \in \{R, B\}$, where each state is equally likely to occur. The realization of S is not observable by the jury members. After S is determined, each juror receives a private red or blue signal, $s \in \{r, b\}$. Signals are informative, as the colors for state and signal coincide with $p \in (\frac{1}{2}, 1)$, and differ with probability $1 - p$.

Assume the state of the world is represented by the color of an urn and the jury members' signals are balls drawn from the urn. Given $p > \frac{1}{2}$, relatively more blue (red) balls are in the blue (red) urn.

We distinguish between two types of sampling of the balls: without replacement (O) and with replacement (W). In O, the private signal observed by the juror is dependent on the private signals observed by the other $n - 1$ jury members. More specifically, p is a realization of a hypergeometric random variable. In contrast, in W, the private signal observed by the player is independent of the private signals observed by the other jury members, and p is a realization of a Bernoulli random variable.

After each juror receives her private signal, she votes as a part of the jury to correctly guess the true state of the world with vote $v \in \{R, B\}$. The votes of the n jurors are aggregated into a jury decision $\hat{v} \in \{R, B\}$ according to the unanimity voting rule: the jury decides for the red urn if and only if all the jury members vote red, and decides for the blue urn otherwise.

Given this notation, the probabilities of convicting an innocent defendant and acquitting a guilty defendant are respectively represented as $Pr(\hat{v} = R|S = B)$ and $Pr(\hat{v} = B|S = R)$. For the ease of notation, in the rest of the paper, we will refer to these error rates as $Pr(R|B)$ and $Pr(B|R)$.

Every juror has the same payoff function $U(\hat{v}, S)$ and is assumed to receive $\pi = 0$ if the jury correctly guesses the color of the urn; and bears a cost of $q \in (0, 1)$ for wrongly identifying a blue urn as red, and a cost of $(1 - q)$ for wrongly identifying a red urn as blue. In summary, we have:

$$\begin{aligned}
U(R, R) &= U(B, B) = \pi = 0 \\
U(R, B) &= -q \\
U(B, R) &= -(1 - q)
\end{aligned}$$

Given these payoffs, the parameter q defines a juror's boundary for reasonable doubt. A juror who believes the defendant to be guilty with probability higher than q will strictly prefer to convict the defendant. A greater value of q indicates that the juror is more tolerant of the risk of acquitting a guilty defendant compared to the potential harm of convicting an innocent.

1.2.2 Nash Equilibrium

Define $\sigma(s)$ as the probability to vote red given signal s. FP show the existence of a mixed strategy equilibrium in which every juror votes red with some positive probability, $\sigma(b)$, when her signal is blue; and always vote red, $\sigma(r) = 1$, when her signal is red. We adapt their relevant findings – for sampling with replacement – to our terminology and summarize them in the following proposition:

Proposition 1

Given the signals are independent from each other, the unique symmetric mixed strategy equilibrium is defined as

$$\sigma(r) = 1, \tag{1.1}$$

$$\sigma(b) = \frac{Kp - (1-p)}{p - K(1-p)}, \tag{1.2}$$

where

$$K = \left(\frac{(1-q)(1-p)}{qp}\right)^{\frac{1}{n-1}}. \tag{1.3}$$

Moreover, the probability of an incorrect jury decision to vote red when the true state is blue, $Pr(R|B)$, and the probability to vote blue when the true state is red, $Pr(B|R)$, are defined as

$$Pr(R|B) = (\rho_B)^n, \tag{1.4}$$

$$Pr(B|R) = 1 - (\rho_R)^n, \tag{1.5}$$

where

$$\rho_R = p\sigma(r) + (1-p)\sigma(b) \text{ and} \tag{1.6}$$

$$\rho_B = (1-p)\sigma(r) + p\sigma(b) \tag{1.7}$$

are the probabilities that a juror votes red for the respective states of the world, R and B.

Proof. See Feddersen and Pesendorfer (1998).[1] ∎

[1] More specifically, see pages 24-26 and Appendix A in FP for the proposition and its proof. For a brief overview of the explicit functional forms see pages 408-409 in GMP. For a brief discussion of the strategic voting and pivotality see pages 376-377 and 386-387 in Coughlan (2000) and pages 35-39 in Austen-Smith and Banks (1996).

Proposition 1 defines an equilibrium solution for the case where the private signals are drawn independently (W). We provide the mixed strategy equilibrium for the hypergeometric case (O) in the following corollary.

Corollary 1.1 *Given the signals are drawn from the hypergeometric distribution, the unique symmetric mixed strategy equilibrium is defined as*

$$\sigma(r) = 1, \tag{1.8}$$

$$\sigma(b) = \left(\frac{qp}{(1-p)(1-q)}\right)^{\frac{1}{n(1-2p)}}. \tag{1.9}$$

Moreover, the probability of an incorrect jury decision to vote red when the true state is blue, $Pr(R|B)$, and the probability to vote blue when the true state is red, $Pr(B|R)$, are defined as:

$$Pr(R|B) = (\rho_B)^n, \tag{1.10}$$

$$Pr(B|R) = 1 - (\rho_R)^n \tag{1.11}$$

where

$$\rho_R = \sigma(r)^{pn}\sigma(b)^{(1-p)n} \text{ and} \tag{1.12}$$

$$\rho_B = \sigma(r)^{(1-p)n}\sigma(b)^{pn}. \tag{1.13}$$

Proof. See Appendix 1.A.1. ■

1.2.3 Best responses

Define α^s as the juror's belief on being pivotal given the signal s and conditional on state R, and define β^s for this belief conditional on state B. In the following proposition, we identify the conditions for informative and strategic voting under any beliefs α^s and β^s.

Proposition 2
For every juror i, define $u_i(\cdot)$ as the utility given her vote. Then, we have:

$$\mathbb{E}(u_i(\sigma_i(r) = 1)) = ((1-q)\alpha^r p - q\beta^r(1-p)) + \mathbb{E}(u_i(\sigma_i(r) = 0)) \tag{1.14}$$

$$\mathbb{E}(u_i(\sigma_i(b) = 0)) = (q\beta^b p - (1-q)\alpha^b(1-p)) + \mathbb{E}(u_i(\sigma_i(b) = 1)) \tag{1.15}$$

Proof. See Appendix 1.A.2. ■

Using Proposition 2, we identify the conditions under which a juror votes informatively or strategically in the following corollary.

Corollary 2.1

Assume $\alpha^s + \beta^s > 0$ for $s \in \{r, b\}$, and define $w = \frac{1-q}{q}$ then a juror votes her signal (informative voting) if and only if $p > \frac{w\alpha^b}{w\alpha^b+\beta^b}$ and $p > \frac{\beta^r}{w\alpha^r+\beta^r}$; and a juror always votes red (strategic voting) if and only if $\frac{w\alpha^b}{w\alpha^b+\beta^b} > p > \frac{\beta^r}{w\alpha^r+\beta^r}$.

Proof. See Appendix 1.A.3. ■

The assumption $\alpha^s + \beta^s > 0$ assures that a juror is pivotal in at least one state of the world.

Notice that if we assume that either type of errors are equally costly as in our experimental setup, then we have $w = 1$; and the boundary conditions for informative and strategic voting become solely dependent on the pivotality probabilities. In the continuation of the paper, for the ease of notation and to be in line with our experimental setup, we will assume that $U(R, B) = U(B, R)$ and hence set q to $\frac{1}{2}$.

Moreover notice that if we further assume that $\sigma(r) \geq \sigma(b)$, then since $p > \frac{1}{2}$, it is trivial to show $\alpha^s \geq \beta^s$ for $s \in \{r, b\}$. Given $\alpha^s \geq \beta^s$ for $s \in \{r, b\}$, the inequality $\frac{\beta^r}{\alpha^r+\beta^r} \leq \frac{1}{2} < p$ is always satisfied, and the inequality $\frac{\beta^r}{\alpha^r+\beta^r} > p > \frac{1}{2}$ is never satisfied. As a result, informative and strategic voting conditions respectively simplifies to $\frac{\alpha^b}{\alpha^b+\beta^b} > p$ and $\frac{\alpha^b}{\alpha^b+\beta^b} < p$.

1.2.4 Level-k Modeling

Consider a model of heterogeneous types of strategic reasoning, with types $k \in \mathbb{N}^0$, who apply k iterated best responses to a level-0 belief (See Nagel, 1995; Stahl and Wilson, 1995; Crawford et al., 2013). The types are distributed in the population according to the level-k distribution $d(k)$, and each level-k juror believes all other jury members to be level-$(k-1)$. Furthermore, each juror chooses strategy $\sigma_k(s)$ and believes to be pivotal with probabilities α_k^s and β_k^s for states R and B respectively.

We assume a level-0 juror to vote uninformatively and hence independently of her signal, $\sigma_0(r) = \sigma_0(b) > 0$.[2] With $\sigma_0(s) > 0$ for $s \in \{r, b\}$, we avoid a trivial setup and assure the level-1 juror to be pivotal with positive probability.

Given the uninformative level-0 voting, a level-1 juror's pivotality is independent of the state. Hence, a level-1 juror's expected utility solely depends on her signal's informativeness. Consequently, given the received signal has some degree of informativeness (i.e. $p > \frac{1}{2}$), a level-1 juror votes the same color as her signal.[3]

[2] We can relax this assumption by introducing some degree of informativeness to level-0 voting, and maintain the same level-1 predictions (see Appendix 1.A.6 for details). In addition, notice that $\sigma_0(r) = \sigma_0(b) > 0$ entails the case for level-0 jurors to vote randomly, i.e. $\sigma_0(s) = \frac{1}{2}$ for $s \in \{r, b\}$.

[3] If we relax the assumption that the jury errors are equally costly and assume without loss of generality that convicting the innocent is more costly for the juror than acquitting the guilty ($q > \frac{1}{2}$), then the degree of the received signal's informativeness needs to be higher than the juror's threshold of reasonable doubt, $p > q$, in order for her to vote informatively.

A juror's expected utility for either choice of vote depends both on her received signal's strength, p, and her perceived pivotality for a given state, α^s or β^s. Given the asymmetric nature of unanimous voting, a level-2 juror believes to be less pivotal if the true state is B. Hence if her perceived pivotality under state R is relatively high enough to offset her received blue signal's strength ($\alpha^b(1-p) > p\beta^b$), she will vote red upon a blue signal.[4]

Because a level-2 juror votes strategically, i.e. always votes red, her action becomes as uninformative as a level-0 juror. Hence, a level-3 juror best responds in the same way a level-1 juror does and votes informatively.

Proposition 3

Consider the case where each juror receives an independent private signal. $\forall n > 2$, *jury members at levels* $\{1, 3, ...\}$ *vote informatively; and jury members at levels* $\{2, 4, ...\}$ *vote strategically.*

Proof. See Appendix 1.A.4. ■

For the case where the private signals are dependent (O), given $n \geq \frac{2}{1-p}$,[5] a level-2 juror never believes to be pivotal as she believes that there is always a level-1 juror in the group that receives the blue signal and votes blue. Therefore, for $n \geq \frac{2}{1-p}$, a level-2 juror is indifferent between voting blue and red.

To eliminate this indifference, we introduce a small error term, ϵ, into the juror's belief about other players' choices. Specifically, we assume that every juror believes with some small probability $\epsilon > 0$ that every other juror votes the color other than the color they are expected to vote, i.e. makes a mistake.[6] The introduced noise enables us to parallel the predictions we made in Proposition 3, for O in the following proposition.

Proposition 4

Consider the case where each juror receives a hypergeometrically dependent private signal. Assume that each juror believes other jury members to make a mistake with some probability ϵ *such that:*

$$\epsilon < \frac{1}{1 + \left(\frac{p}{1-p}\right)^{\frac{1}{(2p-1)n}}}$$

[4]If we relax the assumption that the jury errors are equally costly, then strategic voting will also be dependent on the relative costs of acquitting a guilty and convicting an innocent. For instance, if the jury members are primarily concerned with not convicting an innocent, then the threshold of reasonable doubt can be high enough to offset the relatively small probability of being pivotal under the blue state and a "level-2" juror will not vote strategically. On the other hand, given the juror is not pivotal in the blue state, $\beta^b = 0$, (as can be the case under without replacement sampling), then even if she has a very high threshold of reasonable doubt, she will still prefer to vote strategically.

[5]Given $p = \frac{2}{3}$, for $n \geq 6$.

[6]Alternatively, it can be assumed that every juror believes with some small probability $\epsilon > 0$ that every other juror votes randomly. This has no effect on the predictions of our model, but changes the upper bound for ϵ stated in Proposition 4.

Then $\forall n > 2$*, jury members at levels* $\{1, 3, ...\}$ *vote informatively, and jury members at levels* $\{2, 4, ...\}$ *vote strategically.*

Proof. See Appendix 1.A.5. ■

On the basis of Propositions 3 and 4, Table 1.1 spells out the level-k predictions for level-0 to level-3 by sampling method, jury size and signal. Notably, each level-k type behaves the same across these different aspects. Differences in $\sigma(s)$ across treatments can therefore only result from changes in the level-k distribution $d(k)$.

	O				W			
	$n = 3$		$n = 6$		$n = 3$		$n = 6$	
$d(k)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$
$k = 0$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$k = 1$	0	1	0	1	0	1	0	1
$k = 2$	1	1	1	1	1	1	1	1
$k = 3$	0	1	0	1	0	1	0	1

Table 1.1: Level-k predictions for $\sigma_k(s)$.

Given the importance of $d(k)$, asking for an ideal level-k distribution becomes insightful. Specifically, which ideal distribution $d^*(k)$ minimizes the aggregate probability of a jury's errors, $Pr(R|B) + Pr(B|R)$? We answer this question in an optimization problem that features three distinct types, level-0, level-1, and level-2. Higher levels shall be reflected by level-1 or level-2 because odd-level and even-level behaviors coincide. Table 1.2 indicates $d^*(k)$ by sampling method and jury size. Notably, the uninformative behavior of level-0 is not useful in minimizing errors and therefore $d^*(0) = 0$. Furthermore, the average level μ_d^* is relatively high, especially for $n = 6$, and translates in a relatively high ideal share of level-2.

	O		W	
$d^*(k)$	$n = 3$	$n = 6$	$n = 3$	$n = 6$
$k = 0$	.00	.00	.00	.00
$k = 1$	.50	.29	.68	.34
$k = 2$	.50	.71	.32	.66
μ_d^*	1.50	1.71	1.32	1.66

Table 1.2: Optimal Level-k distribution $d^*(k)$ by treatment.

Table 1.3 shows the aggregate predictions $\sigma(s)$ following $d^*(k)$. For both O and W, levels of $\sigma(b)$ increase with jury size. Interestingly, these predictions coincide with the NE predictions, which also minimize the probability of jury error.[7] Intuitively, only level-

[7]Our W treatment predictions slightly differ from the W predictions of GMP who implement $p = \frac{7}{10}$ and not $p = \frac{2}{3}$. They have not considered the case without replacement.

2 jurors vote strategically red after a blue signal, which is required with some probability in order to correctly decide for the true state S under the unanimity rule. Table 1.4 gives the minimized probabilities of jury errors.

	O		W	
	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$
$n = 3$	1	.50	1	.32
$n = 6$	1	.71	1	.66

Table 1.3: Nash equilibrium and d^* predictions for $\sigma(s)$ by treatment.

Another prediction relates to the jury's accuracy to predict the true state of the world. The two types of errors are quantified in Table 1.4. Under Nash and $d^*(k)$ behaviour, these do not change with the jury size in O.

	O		W	
	$Pr(R\|B)$	$Pr(B\|R)$	$Pr(R\|B)$	$Pr(B\|R)$
$n = 3$	.25	.50	.16	.54
$n = 6$	.25	.50	.21	.52

Table 1.4: Probabilities of incorrect jury decisions in Nash equilibrium and under $d^*(k)$.

1.3 Experimental design

In this study, we employ a 2×2 within-subject experimental design, varying the size of the jury, $n \in \{3, 6\}$, in one dimension and the sampling of the signal, without (O) or with replacement (W), in the other. This implies four treatments, 3O, 6O, 3W, 6W. Signals are balls drawn from a red or blue urn, $S \in \{R, B\}$, each of which is set to be equally likely to occur. Moreover, we assume the cost of either type of error to be equal ($q = \frac{1}{2}$) and the probability of drawing a ball of the same color as the urn to be twice more likely ($p = \frac{2}{3}$).

1.3.1 Team Communication

To ascertain subjects' level of thinking, we conduct the experiment with an intra-team communication protocol that yields incentivized written accounts of their individual reasoning (Burchardi and Penczynski, 2014).[8]

The communication protocol incentivizes the subjects' messages within their respective teams as follows. Subjects are randomly assigned into teams of two players. The two

[8]The experimental instructions are reprinted in section 1.B of the online appendix. The experiments were programmed and conducted with the software z-Tree (Fischbacher, 2007).

members are connected through the modified chat module of the experiment software. Once subjects know the decision problem, each team member can state a so-called "suggested decision" and justify it in a written message. After the suggested decision is made, the suggestions and messages get exchanged simultaneously. In a next step, both team members individually state their "final decision". For each decision, one of the two team members' final decisions is chosen randomly by the computer to count as the "team's action". This construction provides incentives to state the full reasoning underlying the suggested decision in a clear and convincing way. The message is entered in free form without explicit space or time limitations.

Note that each team takes the role of a single juror and members of the same team always observe one and the same signal. Juries are therefore composed of n teams and the jury decision derives from the n final decisions. However, our main analysis focuses on the suggested decisions and messages of the individual subjects.[9]

1.3.2 Experimental procedures

We conducted six experimental sessions in the Experimental Economics Laboratory at the University of Mannheim (mLab). A total of 96 subjects participated in the experiments. Subjects were recruited from the University of Mannheim's subject pool. The sessions respectively had 12, 12, 24, 24, 12 and 12 subjects. All of the subjects were either Bachelor (69), Master (20) or Doctoral (7) students of the University of Mannheim. Subjects' mean and median ages were 22. 52% of the subjects were female. Out of 96 subjects, 27 were studying Economics, 11 of them first year, 6 of them second year, 6 of them third year and 4 of them forth year or above. 45% of the subjects had prior training in game theory and 95% of them had previously participated in laboratory experiments.

Each session began with an initial test phase, during which subjects familiarized themselves with the team communication procedures. During this test phase, subjects answered two unrelated questions involving guessing the date of two different historical events. Subsequently, subjects played the jury voting game in four consecutive parts, with each part corresponding to a different treatment. Each part consisted of two periods, with subjects playing the treatment-specific variation of the game in each period. For every period, subjects were randomly reassigned to a new team and a jury.[10] Prior to each part, subjects were provided with instructions relevant only to that specific treatment. At the end of each period, subjects were provided with the aggregated decision of their jury and the resulting payoff for that period. After completing a part, subjects received a new set

[9] Arad et al. (2022) show that the team setup does not introduce systematic differences in strategic decisions compared to a setup with individuals.

[10] Upon investigation, we have found no significant difference between the consecutive periods for each treatment. Consequently, we combined the data from both periods to increase the number of observations for our analysis.

of instructions for the subsequent one. The experimenter read the instructions aloud and addressed any clarification questions publicly.

The sessions were organized such that half of them followed an order in which subjects first played O (without replacement) treatments for jury sizes of three and six, respectively, followed by W (with replacement) treatments. The other half of the sessions featured a reversed order for the sampling treatments while maintaining the same order for the jury sizes.[11]

	O	W	$\sum$
$n = 3$	192	192	384
$n = 6$	192	168	360
$\sum$	384	360	744

Table 1.5: Number of observations by treatments.

In sum, 744 observations are collected in 6 sessions. Table 1.5 indicates 24 fewer observations for 6W treatment because of an imposed end in one session, in which several subjects took significantly more time to make their decisions than expected.[12]

Remuneration for subjects was structured as follows: for each period, subjects earned €2 if their jury reached a correct decision, and €0.20 if the decision was incorrect. A show-up fee of €4 was provided, with additional earnings based on the accuracy of their jury's decisions averaging at €8.9, and ranging from a minimum of €5 to a maximum of €12.4. The subjects received their payment after the experiment in private and in cash.

1.3.3 Classification Process

The messages are classified independently by two research assistants (RA). For each individual decision's message they indicated the level of reasoning that the message corresponds to most closely. For this task, the RAs are introduced to the level-k model and received detailed instructions about characteristics of the individual types.[13]

The following features of the levels of reasoning are derived from the model and guide the classification process (similar to Burchardi and Penczynski, 2014; Penczynski, 2017). Level-0 play corresponds to choosing randomly, entirely without justification or with some justification completely unrelated to the task. Level-1 jurors always follow their own signal. They may argue in favor of playing their own signal through some probability argument. Level-2 reasoning assumes that all other jury members follow their signal and suggests a way to best respond to that. Level-3 reasoning is aware that people best respond to a belief that others follow their signal by voting red. Since level-3's

[11]Order effects are analyzed and discussed in section 1.6.1.

[12]GMP faced a similar issue with their 6W treatment.

[13]Classification instructions for the RAs are reprinted in section 1.C of the online appendix.

best response is to follow their signal, level-3 reasoning might have similarities to level-1 reasoning.

The classification procedure starts with both RAs providing independent sets of classifications. Then, both are anonymously informed about the classifications of the other RA and have the possibility to simultaneously revise their own classification. This revision process is repeated twice. These iterations allow them to reconsider diverging classifications and to screen errors or misperceptions.[14]

Table 1.6 indicates that out of 744 observations and hence message opportunities, 529 (71.1%) sent a message, which was classified in all except 4 cases. Out of the classified messages, 493 (93.2%) had a matching level classification by the RAs. Only observations with classified messages enter our analysis. Table 1.7 shows that the percentage of messages classified out of all observations is stable across treatments. Table 1.8 shows message examples that are classified for each type of level of thinking.

	Message Sent		No Message Sent
	Classified	Unclassified	
	525	4	215
Matched	493		

Table 1.6: Number of observations by messages and classification.

	O	W	$\sum$
$n = 3$	.69	.65	.67
$n = 6$	.65	.66	.66
$\sum$	.67	.66	.66

Table 1.7: Ratio of observations with classified messages by treatments.

GPT Classification

Additionally, we have utilized GPT-4 to benchmark the classification outcomes by RAs and to investigate the accuracy of classifications by Large Language Models (LLMs). The classifications by GPT-4 aligned with 91.3% of the RAs' classifications. The details of the GPT-4 classification procedure are provided in Chapter 2.

[14] The initial agreement rate was 77.6%. After the first revision it increased to 87.4%, and after the second revision, the final agreement rate was 93.2%. See Burchardi and Penczynski (2014), Eich and Penczynski (2016), and Penczynski (2019) for further evidence on the robustness and replicability of this kind of classification.

Level	Message
L0	I don't have a clue.
L0	Did not exactly understand this experiment seems to be just depending on luck.
L1	The probability of our urn being the color of the ball is 2/3 while the probability of the other color is 1/3.
L1	Hm.. the chance of it being the correct color is higher than it being the wrong one.
L2	I'd take red, because if any other is taking blue, it'll be blue anyways.
L2	I suggest we go for red because our decision won't be decisive in the committee's vote if the others go for blue. We don't hurt anyone with this decision.
L3	I think we should stay at blue because the probability of the urn to be blue is 50/50. So the others may decide to take red as they assume that one team will choose blue but if every team thinks in this way we would lose.
L3	Risky to vote blue but others may not vote blue even when draw blue. I say we vote blue.

Table 1.8: Examples of messages

1.4 Hypotheses

The literature has identified many classes of games, in which subjects apply level-k reasoning. We therefore expect that such reasoning is also used for jury voting (Crawford et al., 2013).

Hypothesis 1 *Jury member decisions are governed by level-k reasoning.*

Hypotheses 2 and 3 express some of the theoretical findings from section 1.2.4.

Hypothesis 2 *Given a level of reasoning, the behaviour will not differ by treatment. Hence, aggregate $\sigma(b)$ will not depend on the treatments, but on the level-k distribution $d(k)$ only.*

Hypothesis 3 *The jury error rates in terms of $Pr(R|B)$ and $Pr(B|R)$ are a function of the level-k distribution $d(k)$.*

What could influence the level-k distribution? At various points in the literature, the dependence of the level-k distribution on game and population characteristics has been discussed and empirically documented (Alaoui and Penta, 2016; Penczynski, 2016b, 2017; Koch and Penczynski, 2018).

Increasing task complexity raises the cognitive cost of strategic thinking, while reduced perceived pivotality curtails the motivation for such thought. In our experiment, we selected the sampling method and jury size as treatment dimensions, as each influences task complexity by changing signal realizations within the jury, which in turn affects a juror's perceived pivotality. A jury size of 3 presents fewer signal realizations than a jury of 6, irrespective of the sampling method. We hypothesize that such decreased complexity leaves more cognitive capacity for strategic deliberation and thus leads to a higher mean level of reasoning.

Hypothesis 4 *The smaller number of possible signal realizations in jury size 3 compared to jury size 6 frees cognitive capacity and leads to the observation of a higher average level of reasoning μ_d.*

In O, for a given state S, the distribution of red and blue signals within the jury is predetermined and hence known by the jury members. Rabin (2002) models the belief in the law of the small numbers by means of "without replacement" sampling, leading us to expect that O makes deliberation easier for subjects. In W, many more signal realizations are possible, especially with larger jury sizes n.

Hypothesis 5 *The smaller number of possible signal realizations in O compared to W frees cognitive capacity and leads to the observation of a higher average level of reasoning μ_d.*

1.5 Results

Section 1.3.3 has shown that the RAs agreed about the content of messages in 93.2% of the cases. Table 1.9 shows the aggregated level-k distribution according to these classifications. The distribution $d(k)$ is non-degenerate and features a heterogeneity of types, a hump-shape with mode behaviour at level-1, and hardly any level-3 behaviour. All of these are expected and common traits of level-k distributions as observed in other contexts (Crawford et al., 2013). An average level μ_d of 1.12 is well within the range between 1 and 1.5 that the literature commonly observes (Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Burchardi and Penczynski, 2014).[15]

Result 1 *According to the message classification, jury member decisions are governed by level-k reasoning in a similar fashion as other strategic decisions in the literature.*

Table 1.10 shows behaviour $\sigma(s)$ for both signals by treatment and levels. While levels are inferred from messages without explicit knowledge of the suggested action,

[15] In each treatment, the balance between the received signals and the state of the urn is checked. The proportions of the red signals received were as follows: 0.48 for 3O, 0.54 for 6O, 0.53 for 3W, and 0.51 for 6W. Similarly, the proportions of the red states were found to be: 0.47 for 3O, 0.66 for 6O, 0.53 for 3W, and 0.49 for 6W. In sum, barring the imbalance of red and blue states for 6O, the treatments are balanced in terms of red and blue signals and states.

	$d(k)$
$k=0$	.21
$k=1$	.49
$k=2$	.29
$k=3$	.02
μ_d	1.12

Table 1.9: Aggregate level-k distribution $d(k)$.

the behaviour of $\sigma(s)$ within levels is not statistically different across treatments (Fisher exact test, $p_{min} > 0.187$) except for level-0 behavior for $s = b$ between 6O and 6W (Fisher exact test, $p = 0.081$).

Some degree of informativeness in level-0 voting is identified for W treatments (Fisher exact test, 3W: $p = 0.134$, 6W: $p = 0.041$), while level-1 and level-2 voting closely align with predictions. Level-1 voting is primarily informative and significantly differs between signals in all treatments (Fisher exact test, $p_{max} < 0.001$). Level-2 voting is predominantly strategic and not significantly dependent on the signal (Fisher exact test, $p_{min} > 0.23$)

	O				W			
	$n=3$		$n=6$		$n=3$		$n=6$	
$d(k)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$
$k=0$	.75	.64	.71	.73	.40	.86	.38	.75
$k=1$	.03	.97	.09	.97	.09	.97	.04	.97
$k=2$	.92	1.00	.91	1.00	.92	1.00	1.00	1.00
$k=3$	.00	–	.00	–	.00	–	–	–

Table 1.10: $\sigma(s)$ per treatment and level.

Table 1.11 aggregates these numbers over levels. For $\sigma(r)$, Table 1.11a shows that our results (ÇP) are close to the NE predictions and $d^*(k)$ implication of $\sigma(r) = 1$. Yet, due to level-0 voting, we reject the null hypothesis that they are 0.99 or above for all treatments (one-sample binomial test: for 3O, 6O, 6W $p_{max} < 0.001$; for 3W $p = 0.071$). For $\sigma(b)$, Table 1.11b shows that the jury size has less influence than NE and $d^*(k)$ would predict, both in our data and in GMP's results. Specifically, in aggregate, $\sigma(b)$ is found not to be significantly different between jury sizes 3 and 6 for both O or W (Fisher exact test, O: $p = 0.371$; W: $p = 0.69$), and it is found to be significantly lower than predicted in 6O and 6W (one-sample proportion test , 6O: $p = 0.004$, 6W: $p < 0.001$).

In addition, note that in Table 1.11b, the aggregate strategic voting proportions, $\sigma(b)$, includes strategic votes by level-2 jurors but also non-strategic votes from level-0 jurors as well as mistakes of level-1 jurors. As a consequence, the $\sigma(b)$ proportions in Table 1.11b reflect more than purely strategic voting. When the non-strategic votes are accounted for, the true aggregate strategic voting rates for 3O, 6O, 3W, and 6W decrease to .35, .34, .20, and .22 respectively. Under these adjusted values the $\sigma(b)$ proportions for 3O and 3W are also significantly lower than the NE predictions (one-sample proportion test, 3O: $p = 0.011$; 3W: $p = 0.037$).

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	1	.92	1	.97	.95
$n = 6$	1	.93	1	.90	.90

(a) $\sigma(r)$.

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	.50	.46	.32	.32	.36
$n = 6$	.71	.55	.66	.36	.48

(b) $\sigma(b)$.

Table 1.11: Nash equilibrium and $d^*(k)$ predictions and empirical results for behavior.

The logit regression in Table 1.12 summarizes these findings and shows on the one hand the different and significant impact of different levels on the voting behavior compared to level-0 – especially for $\sigma(b)$ – and on the other hand the minor and insignificant impact of the treatments.

Result 2 *Given a level of reasoning, behavior does not differ across treatments. Aggregate behavior $\sigma(b)$ depends less on the treatments than predicted by NE and $d^*(k)$.*

The observed behavior in terms of $\sigma(s)$ implies error rates as presented in Table 1.13.[16] For each type of error, the observed error rate deviates in the same direction from the prediction in all treatments. In Table 1.13a, the error rate $Pr(B|R)$ is significantly higher in 6 than in 3 for both O and W and consequently higher than under NE or d^* (Wilcoxon ranksum test[17], $p < 0.001$). Conversely, in Table 1.13b, the error rate $Pr(R|B)$ is significantly lower (Wilcoxon ranksum test, $p < 0.001$). Both movements can be explained with the lower than optimal sophistication in the observed $d(k)$ and the resulting infrequent strategic voting with $\sigma(b) = 1$. With the increased blue votes, the probability $Pr(B|R)$ – acquitting a guilty – increases and $Pr(R|B)$ – convicting an innocent – decreases.

[16] Due to our intra-team methodology, half of the time, a decision made by a subject is not reflected in the final team decision. As a result, it is not possible to trivially sum the cases in which the jury's aggregated decision is correct using the experimental data of jury decisions. Instead, using the mean values for $\sigma(r)$ and $\sigma(b)$, for each treatment and state, we calculated the expected jury accuracies by plugging $\sigma(r)$ and $\sigma(b)$ into the formulas stated in Proposition 1 and Corollary 1.1 for ρ_R and ρ_B for W and O cases respectively. In GMP, accuracy values are derived directly from the jury decisions.

[17] Wilcoxon ranksum test is performed on the bootstrapped distributions for the error rates. These distributions are generated by calculating the error rates from the re-sampled with replacement vote variable.

Table 1.12: Logit regressions with average marginal effects on $\sigma(s)$.

	M_1		M_2	
	$\sigma(b)$	$\sigma(r)$	$\sigma(b)$	$\sigma(r)$
Level-1	-.508***	.002***	-.502***	.002***
	(.072)	(.001)	(.070)	(.001)
Level-2	.503***	.090***	.518***	.093***
	(.123)	(.027)	(.118)	(.025)
Level-3	-.407***		-.407***	
	(.056)		(.055)	
$6O$	.062	-.000		
	(.149)	(.000)		
$3W$	-.049	.000		
	(.119)	(.000)		
$6W$	-.096	.000		
	(.101)	(.000)		

Notes: Values in parenthesis represent the standard errors clustered by subjects. '***' represents $p < 0.001$ significance. There are no level-3 observations for $\sigma(r)$ cases. M2 only considers the level variable while M1 additionally includes treatment variable. We compared the fit of two logistic regression models using a likelihood-ratio test, we fail to reject M2 over M1 for each subsample ($p_{min} > 0.52$). Hence, we did not find evidence that including the treatment factors in the model significantly improved the fit to the data.

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	.50	.61	.54	.57	.53
$n = 6$	.50	.78	.52	.86	.73

(a) $Pr(B|R)$.

	O		W		
	NE/d^*	ÇP	NE/d^*	ÇP	GMP
$n = 3$	.25	.19	.16	.16	.19
$n = 6$	.25	.08	.21	.02	.03

(b) $Pr(R|B)$.

Table 1.13: Nash equilibrium and $d^*(k)$ predictions, and empirical results for the error rates.

Result 3 *In line with the observed level-k distribution, the jury accuracy deviates from the optimal accuracy in that convictions are less likely, independent of the state S.*

Table 1.14 shows the level distribution $d(k)$ by treatment. At first sight, the distributions and the mean levels μ_d are supportive of Hypotheses 5 and 4 that both treatment dimensions have an influence on the sophistication of strategic thinking.

	O		W	
$d(k)$	$n = 3$	$n = 6$	$n = 3$	$n = 6$
$k = 0$	.15	.24	.14	.32
$k = 1$	.45	.40	.58	.51
$k = 2$	.37	.35	.26	.16
$k = 3$	.03	.01	.02	.00
μ_d	1.28	1.13	1.14	0.84

Table 1.14: Level-k distribution $d(k)$ by treatment.

Pooling the level-k distribution across two jury sizes, Table 1.15a shows that the average level of reasoning is significantly higher under $n = 3$ compared to $n = 6$ (Wilcoxon ranksum test, $p < 0.001$).[18] Specifically, level-0 is higher while level-1, level-2, and level-3 are lower in $n = 6$ compared to $n = 3$.[19]

$d(k)$	$n = 3$	$n = 6$
$k = 0$	.14	.28
$k = 1$	.52	.45
$k = 2$	.32	.27
$k = 3$	.03	.00
μ_d	1.23	1.00

(a) Jury size $n = 3$ and $n = 6$.

$d(k)$	O	W
$k = 0$	.19	.23
$k = 1$	.43	.55
$k = 2$	.36	.22
$k = 3$	.02	.01
μ_d	1.22	1.01

(b) Sampling O and W.

Table 1.15: Level distributions $d(k)$ by sampling and jury size.

Result 4 *The treatments with jury size $n = 3$, which produce a smaller number of possible signal realizations in a jury, feature higher average levels of reasoning.*

Pooling the level-k distributions across different sampling methods, Table 1.15b shows that average level of reasoning is significantly higher under O compared to W (Wilcoxon

[18]Since the number of level-3 jurors are relatively low, we have additionally considered the same test on the subset that excludes the level-3 data, and still found a significant difference ($p = 0.003$)

[19]The observed differences for level-0 and level-3 are found to be significantly different (Fisher exact test, level-0: $p < 0.001$; level-3: $p = 0.02$), for level-1, it is found to be marginally different (Fisher exact test, $p = 0.149$), and for level-2, it is found not to be significantly different (Fisher exact test, $p = 0.315$)

ranksum test, $p < 0.001$).[20] Specifically, the level-1 fraction is higher and the level-2 fraction is lower in W compared to O.[21]

Result 5 *The sampling method O, which produces a smaller number of possible signal realizations in a jury, features higher average levels of reasoning.*

We have additionally compared the level distribution controlling for the signal. Although the average level of sophistication is observed to be higher for the blue signal, the difference between the two average levels of reasoning is found to only be marginally significant (Wilcoxon ranksum test, $p = 0.122$). As can be seen in Table 1.16, except for the level-3 ratios, the level of sophistication between the two distributions are quite close to each other [22]. Furthermore, when we exclude the few level-3 players from the data, the average strategic sophistication level for the blue signal subset becomes 1.11, and the marginal significance between the two subsets is lost (Wilcoxon ranksum test, $p = 0.318$). Upon receiving a red signal since there is no distinction in terms of voting behavior among levels, subjects potentially did not have the motivation to consider a higher level of reasoning. This might, in turn, have led to the lack of level-3 thinkers for the red signal cases, producing the observed marginal significant difference in the level of strategic sophistication between the two signals.

$d(k)$	$s = b$	$s = r$
$k = 0$	.20	.21
$k = 1$	.46	.51
$k = 2$	.31	.28
$k = 3$	.03	–
μ_d	1.17	1.08

Table 1.16: Distribution of levels controlling for signal

1.6 Further explorations

1.6.1 Order effects

Considering our experiment involved four consecutive treatments, and the fact that the order of W and O treatments alternated between sessions, we explored potential fatigue

[20] Since the number of level-3 jurors are relatively low, we have additionally considered the same test on the subset that excludes the level-3 data, and still found a significant difference ($p = 0.002$).

[21] For level-1 and level-2, the observed differences are found to be significantly different (Fisher exact test, level-1: $p = 0.009$; level-2: $p < 0.001$) while for level-0 and level-3, they are found not to be significantly different (Fisher exact test, level-0: $p = 0.32$; level-3: $p = 0.29$).

[22] The observed differences for level-0, level-1 and level-2 are found not to be significantly different (Fisher exact test, $p_{max} > 0.24$), while for level-3, it is found to be significantly different (Fisher exact test, $p = 0.003$).

and learning effects. These effects could be in action and might potentially confound our previously presented results.

Fatigue effect

As subjects advance through the treatments (rounds), they might get tired, leading them to exhibit reduced strategic sophistication in later parts of the experiment. Therefore, if there is a noticeable fatigue effect, regardless of the specific treatments involved, one would anticipate a diminished average sophistication level in the later rounds of the experiment.

$d(k)$	H_1	H_2
$k=0$	.25	.17
$k=1$	.44	.52
$k=2$	.3	.29
$k=3$	.01	.02
μ_d	1.07	1.17

Table 1.17: Fatigue effect

In Table 1.17, we categorized the data into two subsets based on treatments: those played during the first two rounds, labeled as H_1, and those from the last two rounds, labeled as H_2. We then analyzed the strategic level distribution across these subsets. Contrary to the expectation of a drop in sophistication due to potential fatigue, the results indicate a higher sophistication level in H_2 (Wilcoxon ranksum test, $p = 0.094$). As a result, we deduce that fatigue did not have a predominant influence on the subjects, underscoring the robustness of our previously presented findings.

Learning effect

The learning effect, in contrast to the fatigue effect, might not manifest uniformly across treatments. Specifically, while the fatigue effect might consistently impact all the later rounds, the influence of the learning effect could differ based on the altering ordering of treatments.

When subjects first play O and then transition to W (O2W), an easy-to-hard learning effect might occur. The relatively lower complexity of the strategy space in O may enable subjects to better grasp the strategic nature of the game. This enhanced initial understanding of the game's strategic aspects could then foster greater strategic sophistication when subjects tackle more complex W that follow.

Conversely, when subjects start with W followed by O (W2O), a potential hard-to-easy learning effect might emerge. Initially engaging with W exposes subjects to a broader signal and strategy space, demanding heightened cognitive effort. As they transition to O, characterized by a strategy space with fewer possible outcomes, their prior experience in the complex dynamics of W could enhance their grasp of the game, leading to increased strategic sophistication compared to scenarios where O precede W.

Empirical evidence from domains such as motor skills (Wulf and Shea, 2002) and auditory skills (Liu et al., 2008; Church et al., 2013; Wisniewski et al., 2017) suggests that an easy-to-hard ordering of tasks can improve learning and performance compared to random or hard-to-easy orderings. In the domain of test-taking, some studies find no overall effect of question ordering on performance (Plake et al., 1982; Klimko, 1984), while others demonstrate a positive impact with easy-to-hard ordering (Bassey et al., 2022; Hambleton and Traub, 1974). In contrast, hard-to-easy ordering either shows no significant benefit (Hauck et al., 2017) or a negative effect on performance (Hambleton and Traub, 1974; Newman et al., 1988).

$d(k)$	O2W	W2O
$k = 0$	.16	.24
$k = 1$	.49	.48
$k = 2$	.34	.26
$k = 3$	.01	.02
μ_d	1.19	1.06

Table 1.18: Learning effect

In Table 1.18, we do observe that jurors portray a significantly higher level of strategic sophistication in O2W relative to W2O (Wilcoxon ranksum test, $p = 0.016$). Given that O2W represents the learning effect via the easy-to-hard ordering while W2O represents the learning effect via the hard-to-easy ordering, this initial investigation hints at the possibility that only an easy-to-hard learning effect is present in our data.

Comparing the average level of strategic thinking between Tables 1.19a and 1.19b, we observe an easy-to-hard learning effect in W: the average level in O2W (1.13) is significantly higher than in W2O (0.89) (Wilcoxon ranksum test, $p = 0.003$). For O, however, the strategic thinking level in O2W (1.26) is higher than in W2O (1.19) (Wilcoxon ranksum test, $p = 0.1752$), which makes a hard-to-easy learning effect unplausible.

$d(k)$	O	W
$k = 0$	.19	.14
$k = 1$	.37	.59
$k = 2$	.43	.26
$k = 3$	.01	.01
μ_d	1.26	1.13

(a) O to W order, Ord_1

$d(k)$	O	W
$k = 0$	.19	.31
$k = 1$	.47	.5
$k = 2$	.31	.18
$k = 3$	.03	.01
μ_d	1.19	0.89

(b) W to O order, Ord_2

Table 1.19: Level distributions $d(k)$ by sampling and order.

Given the statistically significant influence of the easy-to-hard learning effect on W, we further examined whether our previously discussed results regarding the overall effect of sampling and group size remain robust when controlling for order. As illustrated in Table 1.19, the average sophistication level in O is consistently higher across both the O2W and W2O subsets. This disparity is statistically significant for both subsets (Wilcoxon ranksum test, O2W: $p = 0.053$; W2O: $p < 0.001$).

Similarly, Table 1.20 reveals that for both subsets, the average sophistication level is higher in the treatments with a smaller group size, and these differences are also significant (Wilcoxon ranksum test, O2W: $p = 0.102$; W2O: $p = 0.002$). Thus, our main results remain consistent after accounting for the ordering effect.

$d(k)$	$n = 3$	$n = 6$
$k = 0$	.12	.21
$k = 1$	.51	.46
$k = 2$	.35	.33
$k = 3$	.02	–
μ_d	1.26	1.13

(a) O to W order

$d(k)$	$n = 3$	$n = 6$
$k = 0$	.16	.34
$k = 1$	.52	.44
$k = 2$	.29	.22
$k = 3$	.04	.01
μ_d	1.20	0.89

(b) W to O order

Table 1.20: Level distributions $d(k)$ by group size and order.

Given the documented presence of the easy-to-hard learning effect in our data, a natural inquiry arises: could this effect also influence performance across jury sizes of 3 and 6? It is worth noting that in our experimental sessions, we alternated only the order of sampling, while always maintaining the sequence of $n = 6$ treatments following $n = 3$ treatments. This consistent order means we lack the variation to explore this effect empirically. However, if the easy-to-hard learning effect is inflating the strategic sophistication in jury size 6 treatments, this implies an even more pronounced innate difference between the two jury sizes than reported. Hence, the possibility of such a learning effect would only further emphasize our findings on the differences in strategic thinking across the two jury sizes.

1.6.2 Influence of team communication on votes

In light of our intra-team communication experimental design, we sought to understand how communication impacts juror members' decisions after their interactions (see Penczynski, 2016a; Arad et al., 2022) with their teammates (partners). This inquiry is explored in Table 1.21.

Of the 493 messages, RAs were able to classify the strategic thinking levels of both the juror and her partner for 336 instances (68.2%). Of these, only 85 cases (25.3%) had differing pre-communication votes between the partners.

For partners that start with the same suggested vote, Table 1.21a shows that their subsequent decisions remain largely unchanged and unaffected by the level of strategic thinking conveyed with their communication. In contrast, for diverging suggested votes, Table 1.21b shows that the communicated message's strategic depth has a discernible influence. Specifically, jurors tend to change their vote in more instances when a higher level of reasoning is exhibited by their partner.

	Vote Changed?	
Higher Lv.	$\times$	✓
Player	76	0
Neither	100	1
Partner	74	0

(a) Same Initial Vote.

	Vote Changed?	
Higher Lv.	$\times$	✓
Player	23	9
Neither	13	6
Partner	16	18

(b) Different Initial Vote.

Table 1.21: Comparison of vote changes.

Note: "Vote Changed?" indicates whether the *Player*'s final vote is different from their suggested vote.

1.7 Concluding remarks

Our study provides evidence that strategic thinking is relevant in jury voting and appropriately modeled by the level-k model of reasoning. The predicted types and their behaviour align with the evaluation of written accounts and decisions observed in the experimental games. The model correctly predicts that – given a level of sophistication – behaviour is unresponsive to changes in jury size and signal sampling method. The experimental text and decision data support this prediction and show that behaviour reacts to treatments primarily because strategic sophistication responds to the cognitive complexity of the task at hand. Specifically, we find evidence that larger juries and the more involved "with replacement" sampling of signals lead to a lower strategic sophistication.

The deviations of the observed jury error rates from the minimal error rates in equilibrium can be viewed as a result of the sub-optimal distribution of the levels of reasoning in the subject sample. Interestingly, these error rates would be much higher, and jury votes under unanimity in large juries would basically be uninformative if only one type of level-k reasoning with informative or strategic voting was present. From that perspective, the heterogeneity in types being so close to the optimal distribution is what leads to the observed low error rate in the jury voting. So while many arguments in favor of jury diversity already exist, the study adds and exhibits in detail a stark reason for jury diversity in terms of strategic sophistication.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work I used GPT-4 in order to improve readability and language of the text. After using this tool, I reviewed and edited the content as needed.

Appendix to Chapter 1

1.A Proofs

1.A.1 Proof of Corollary 1.1

Following Feddersen and Pesendorfer (1998), we are looking for a responsive symmetric equilibrium in mixed strategy profiles. First note that a necessary condition for a mixed strategy profile is for a juror who receives a blue signal to be indifferent between voting red and voting blue. This occurs when the probability that the urn is red (conditional on the juror i's vote being pivotal and on her private signal to be blue) is equal to the threshold of reasonable doubt, q. Let $Pr(S|s, piv_i)$ represent the probability of state S, conditional on the signal s and on juror i being pivotal. Then a juror who receives a blue signal is indifferent between voting red and blue when

$$Pr(R|b, piv_i) = q. \tag{1.16}$$

Using Bayes formula we expand on equation (1.16) as:

$$\frac{Pr(piv_i|R, b)Pr(b|R)Pr(R)}{Pr(piv_i|R, b)Pr(b|R)Pr(R) + Pr(piv_i|B, b)Pr(b|B)Pr(B)} \tag{1.17}$$

where due to the assumption that signals are drawn without replacement we have:

$$Pr(piv_i|R, b) = \sigma(r)^{pn}\sigma(b)^{(1-p)n-1} \tag{1.18}$$

$$Pr(piv_i|B, b) = \sigma(r)^{(1-p)n}\sigma(b)^{pn-1} \tag{1.19}$$

In words, equation (1.18) describes the probability of being pivotal given n many balls and jurors where pn many balls are red and received by pn many other jurors, and $(1-p)n-1$ balls are blue and received by $(1-p)n-1$ many other jurors. Similarly, equation (1.19) describes the probability of being pivotal given n many balls and jurors where $(1-p)n$ many balls are red and received by $(1-p)n$ many other jurors, and $pn-1$ many balls are blue and received by $pn-1$ many other jurors. Note that in FP $Pr(piv_i|R, b)$ and $Pr(piv_i|B, b)$ are additively defined as $(\sigma(r)p + \sigma(b)(1-p))^{(n-1)}$ and $(\sigma(r)(1-p) + \sigma(b)p)^{(n-1)}$ respectively, while in the without replacement case, we have a *simpler* form as in equations

(1.18) and (1.19). Moreover, under the assumption $\sigma(b) > 0$, we have $Pr(R|b, piv_i) = q$. Clearly, $Pr(R|r, piv_i) > Pr(R|b, piv_i) = q$. Thus, we have $\sigma(r) = 1$. Then, using (1.18) and (1.19) in (1.17), we get:

$$Pr(R|b, piv_i) = \frac{(1-p)\sigma(b)^{(1-p)n-1}}{(1-p)\sigma(b)^{(1-p)n-1} + p\sigma(b)^{pn-1}} = q \tag{1.20}$$

Isolating $\sigma(b)$ in equation (1.20), we get:

$$\sigma(b) = \left(\frac{qp}{(1-p)(1-q)}\right)^{\frac{1}{n(1-2p)}} \tag{1.21}$$

The probabilities for the jury's decision to be wrong given the true state of the world are defined as $Pr(B|R)$ and $Pr(R|B)$. In FP, for the with replacement case, they are defined as $(\sigma(r)p + \sigma(b)(1-p))^n$ and $(\sigma(r)(1-p) + \sigma(b)p)^n$ respectively. In the without replacement case, due to the hypergeometric nature of the signals, they are defined as $Pr(B|R) = \sigma(r)^{pn}\sigma(b)^{(1-p)n}$ and $Pr(R|B) = \sigma(r)^{(1-p)n}\sigma(b)^{pn}$ in a similar fashion defined in equations 1.18 and 1.19.

Lastly, note that given (1.20), it can easily be shown that $\lim_{n\to\infty}\sigma(b) = 1$.

1.A.2 Proof of Proposition 2

For every juror i, her expected payoff for voting red given she receives a red signal is defined as follows:

$$\begin{aligned}
\mathbb{E}(u_i(\sigma_i(r) = 1)) = \;& U(R,R)Pr(R|r)\alpha^r + U(R,B)Pr(B|r)\beta^r \\
& + U(B,B)Pr(B|r)(1-\beta^r) + U(B,R)Pr(R|r)(1-\alpha^r) \\
=\;& p(\pi\alpha^r - (1-q)(1-\alpha^r)) \\
& + (1-p)(\pi(1-\beta^r) - q\beta^r) && (1.22) \\
=\;& -p(1-q)(1-\alpha^r) - (1-p)q\beta^r && (1.23) \\
=\;& -(1-q)p + p(1-q)\alpha^r - (1-p)q\beta^r && (1.24)
\end{aligned}$$

Note that the step from equality (1.22) to (1.23) has been taken via the assumption $\pi = 0$ and the same step has also been taken in the rest of the derivations of this subsection.

Through similar steps, one can define the juror i's expected payoff for voting blue given she receives a red signal as:

$$\begin{aligned}
\mathbb{E}(u_i(\sigma_i(r) = 0)) &= U(B,B)Pr(B|r) + U(B,R)Pr(R|r) && (1.25) \\
&= \pi(1-p) - (1-q)p \\
&= -(1-q)p && (1.26)
\end{aligned}$$

Notice that in equation (1.25) we do have a relatively simplified right hand side equation without the payoffs $U(R,R)$ and $U(R,B)$, and the pivotality probabilities. This is because since the juror votes blue, $U(R,R)$ and $U(R,B)$ cases never occurs and the belief on what the other jurors will do becomes irrelevant. Using equalities (1.24) and (1.26), we get the desired equality in (1.14) of the Proposition 2.

Next, for every juror i, we calculate her expected payoff for voting red given she receives a blue signal as:

$$\begin{aligned}
\mathbb{E}(u_i(\sigma_i(b)=1)) = \; & U(R,R)Pr(R|b)\alpha^b + U(R,B)Pr(B|b)\beta^b \\
& + U(B,B)Pr(B|b)(1-\beta^b) + U(B,R)Pr(R|b)(1-\alpha^b) \\
= \; & \pi(1-p)\alpha^b - qp\beta^b + \pi p(1-\beta^b) - (1-q)(1-p)(1-\alpha^b) \\
= \; & -qp\beta^b - (1-q)(1-p)(1-\alpha^b) \\
= \; & -qp\beta^b + \alpha^b(1-q)(1-p) - (1-q)(1-p)
\end{aligned} \tag{1.27}$$

Through similar steps, one can define her expected payoff for voting blue given she receives a blue signal as:

$$\begin{aligned}
\mathbb{E}(u_i(\sigma_i(b)=0)) &= U(B,B)Pr(B|b) + U(B,R)Pr(R|b) \\
&= \pi p - (1-q)(1-p) \\
&= -(1-q)(1-p)
\end{aligned} \tag{1.28}$$

Using equalities (1.27) and (1.28), we get the desired equality in (1.15) of the Proposition 2.

1.A.3 Proof of Corollary 2.1

Using Proposition 2, under the additional simplifying assumption that $U(R,B)=U(B,R)$, we identify the strict inequality conditions for voting either red or blue (conditional on the signal) as in the Tables (1.22) and (1.23).

Note that for the inequalities in Tables (1.22) and (1.23) to be well defined, we assume, respectively, the conditions $\alpha^s + \beta^s > 0$ and $\alpha^s, \beta^s > 0$ to hold for $s \in \{r, b\}$. Either condition aims at avoiding the cases where a jury member is not pivotal in either state of the world. Table (1.22) offers a more general format for the conditions, and as a result, is presented in the main section of the paper. Table (1.23) provides a better starting point for various algebraic manipulations that takes places in the subsequent proofs that utilizes these boundary conditions.

In the with replacement case, since the signals are independent of each other, we have $\alpha^i = \alpha^j$ and $\beta^i = \beta^j$ for $i \neq j$ and $i, j \in \{r, b\}$. Hence, for ease of notation, we can drop the signal superscripts. Then, given p and n, we define the beliefs about pivotalities for

	$Vote = r$	$Vote = b$
Signal = b	$\frac{\alpha^b}{\alpha^b + \beta^b} > p$	$\frac{\alpha^b}{\alpha^b + \beta^b} < p$
Signal = r	$\frac{\beta^r}{\alpha^r + \beta^r} < p$	$\frac{\beta^r}{\alpha^r + \beta^r} > p$

Table 1.22: Conditions for informative and strategic voting

	$Vote = r$	$Vote = b$
Signal = b	$\frac{\alpha^b}{\beta^b} > \frac{p}{1-p}$	$\frac{\alpha^b}{\beta^b} < \frac{p}{1-p}$
Signal = r	$\frac{\alpha^r}{\beta^r} > \frac{1-p}{p}$	$\frac{\alpha^r}{\beta^r} < \frac{1-p}{p}$

Table 1.23: Alternative conditions for informative and strategic voting

each state as:

$$\alpha = (p\sigma(r) + (1-p)\sigma(b))^{n-1} \tag{1.29}$$
$$\beta = ((1-p)\sigma(r) + p\sigma(b))^{n-1} \tag{1.30}$$

In the without replacement case, given p and n, we define the beliefs on pivotalities for each state and signal received as:

Given signal is red

$$\alpha^r = \sigma(r)^{(pn-1)}\sigma(b)^{((1-p)n)} \tag{1.31}$$
$$\beta^r = \sigma(r)^{((1-p)n-1)}\sigma(b)^{(pn)} \tag{1.32}$$

Given signal is blue

$$\alpha^b = \sigma(r)^{(pn)}\sigma(b)^{((1-p)n-1)} \tag{1.33}$$
$$\beta^b = \sigma(r)^{((1-p)n)}\sigma(b)^{(pn-1)} \tag{1.34}$$

Using equations (1.31)-(1.34) and Table 1.23, we have that, in the without replacement case, a juror votes informatively if and only if $\left(\frac{1-p}{p}\right)^{\frac{1}{n(2p-1)}} < \frac{\sigma(r)}{\sigma(b)} < \left(\frac{p}{1-p}\right)^{\frac{1}{n(2p-1)}}$; and a juror votes strategically if and only if $\left(\frac{p}{1-p}\right)^{\frac{1}{n(2p-1)}} < \frac{\sigma(r)}{\sigma(b)}$.

1.A.4 Proof of Proposition 3

First of all, note that for the with replacement case, since the signals are independent of each other, we have $\alpha^r = \alpha^b$ and $\beta^r = \beta^b$. Henceforth, in the rest of the proof we will omit the signal subscripts for ease of notation.

Secondly, we denote the level of the juror in their pivotality probability as α_k and β_k for $k \in \mathbb{N}^0$.

Level-1 By assumption, a level-0 juror votes uninformatively. This translates to $\alpha_1 = \beta_1 > 0$ and $\frac{\alpha_1}{\beta_1} = 1$. By Table 1.23, a level-1 juror votes blue when the signal is blue, $\sigma_1(b) = 0$, if and only if $\dfrac{\alpha_1}{\beta_1} < \dfrac{p}{1-p}$, and votes red when the signal is red, $\sigma_1(r) = 1$, if and only if $\dfrac{\alpha_1}{\beta_1} > \dfrac{1-p}{p}$. Since $p > \frac{1}{2}$ and $\frac{\alpha_1}{\beta_1} = 1$ by assumption, the informative voting condition for both signals is satisfied and level-1 juror always votes informatively.

Level-2 By the above discussion, a level-1 juror always votes informatively. By definition, a level-1 juror receives the signal r with probability p in state R and with probability $1-p$ in state B. Hence, assuming the state is R and there are n jurors in total, a level-2 juror believes to be pivotal with probability $\alpha_2 = p^{n-1}$; and assuming the state is B, she believes to be pivotal with probability $\beta_2 = (1-p)^{n-1}$. Thus, we have $\frac{\alpha_2}{\beta_2} = \left(\frac{p}{1-p}\right)^{(n-1)}$. Given $p > \frac{1}{2}$, we have $\frac{p}{1-p} > 1 > \frac{1-p}{p}$. Moreover note that $\forall k > 1$, we have $\frac{p}{1-p} < \left(\frac{p}{1-p}\right)^k$. Setting $k = n-1$ and noting $k > 1$ is equivalent to $n > 2$, we have $\frac{1-p}{p} < \frac{p}{1-p} < \left(\frac{p}{1-p}\right)^{(n-1)}$. Hence both conditions for strategic voting in Table 1.23 are satisfied.

Level-3 Given a level-2 juror always votes red, a level-3 juror believes to be always pivotal, $\alpha_3 = \beta_3 = 1$. Given $\frac{\alpha_3}{\beta_3} = 1$, analogue to the level-1 behavior, every level-3 juror votes informatively.

Level-4 and above Due to the assumed degenerate population belief on the next lower level $k-1$, every even-leveled juror behaves the same way as a level-2 juror and always votes red. Furthermore, every odd-leveled juror behaves the same way as a level-3 juror behaves and always votes informatively.

1.A.5 Proof of Proposition 4

In the following proof, we denote the level of the juror and the received signal in their pivotality probability as α_k^s and β_k^s for $k \in \mathbb{N}^0$ and $s \in \{r, b\}$. If the signal subscript is omitted in the subsection of the proof, this indicates that we have $\alpha^r = \alpha^b$ and $\beta^r = \beta^b$.

Level-1 By assumption a level-0 juror votes uninformatively. Introducing an ϵ possibility to make a mistake in a symmetric manner does not change this fact. Hence, we have $\alpha_1 = \beta_1 > 0$ which, in turn, implies $\frac{\alpha_1}{\beta_1} = 1$. By Table (1.23), a level-1 juror votes blue when the signal is blue if and only if $\dfrac{\alpha_1^b}{\beta_1^b} < \dfrac{p}{1-p}$, and votes red when the signal is red if and only if $\dfrac{\alpha_1^r}{\beta_1^r} > \dfrac{1-p}{p}$. Since $p > \frac{1}{2}$ and $\frac{\alpha_1}{\beta_1} = 1$ by assumption, informative voting condition for either type of signal received is always satisfied and a level-1 juror always votes informatively.

Level-2 Based on the above discussion, a level-1 juror always intend to vote informatively. Also note that she is assumed to make a mistake with some probability $\epsilon > 0$ and votes against her signal. Given a juror receives a red signal, her probability of being pivotal under states R and B are respectively defined as:

$$\alpha_2^r = (1-\epsilon)^{(np-1)}\epsilon^{(n(1-p))} \tag{1.35}$$

$$\beta_2^r = (1-\epsilon)^{n(1-p)-1}\epsilon^{(np)} \tag{1.36}$$

Using equation (1.35) and (1.36), we have $\frac{\alpha_2^r}{\beta_2^r} = \left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)}$. Using Table (1.23), given the received signal is red, in order for a level-2 juror to vote strategically we need:

$$\left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)} > \frac{(1-p)}{p} \tag{1.37}$$

With a bit of algebra, inequality (1.37) becomes:

$$\epsilon < \frac{1}{1+\left(\frac{(1-p)}{p}\right)^{\frac{1}{(2p-1)n}}} \tag{1.38}$$

Hence, given ϵ satisfies inequality (1.38), a level-2 juror votes red when she receives a red signal.

Next, assume that the level-2 juror receives a blue signal. Then we have:

$$\alpha_2^b = (1-\epsilon)^{(np)}\epsilon^{(n(1-p)-1)} \tag{1.39}$$

$$\beta_2^b = (1-\epsilon)^{n(1-p)}\epsilon^{(np-1)} \tag{1.40}$$

Using equation (1.39) and (1.40), we have $\frac{\alpha_2^b}{\beta_2^b} = \left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)}$. Using Table (1.23), given the received signal is blue, in order for a level-2 juror to vote strategically we need:

$$\left(\frac{1-\epsilon}{\epsilon}\right)^{n(2p-1)} > \frac{(p)}{1-p} \tag{1.41}$$

With a bit of algebra, equation (1.41) translates to:

$$\epsilon < \frac{1}{1 + \left(\frac{p}{(1-p)}\right)^{\frac{1}{(2p-1)n}}} \tag{1.42}$$

Hence, given ϵ satisfies inequality (1.42), a level-2 juror votes red when she receives a blue signal.

Lastly, since $p > \frac{1}{2}$, we have:

$$\frac{1}{1 + \left(\frac{p}{(1-p)}\right)^{\frac{1}{(2p-1)n}}} < \frac{1}{1 + \left(\frac{(1-p)}{p}\right)^{\frac{1}{(2p-1)n}}}$$

Hence, inequality (1.42) is a sufficient condition for inequality (1.38) to be satisfied. Thus, given inequality (1.42) is satisfied, a level-2 juror always votes red.

Level-3 Given a level-2 juror always votes red, a level-3 juror believes to be pivotal with probability $\alpha_3 = \beta_3 = (1-\epsilon)^{(n-1)}$. Since $\frac{\alpha_3}{\beta_3} = 1$, just like a level-1 juror, every level-3 juror votes informatively.

Level-4 and above Given the cyclical nature of the behavior at even and odd levels of thinking, every even-leveled juror behaves the same way as a level-2 juror and always votes red.

Given the cyclical nature of the behavior at even and odd levels of thinking, every odd-leveled juror behaves the same way as a level-1 juror and always votes informatively.

1.A.6 Relaxing the level-0 jurors' uninformativeness assumption

Assume that a level-0 juror votes her signal with probability θ and votes the opposite of her signal with probability $1-\theta$. Given this assumption, a level-0 juror votes randomly given $\theta = \frac{1}{2}$ and as θ goes to 1 or 0, the informativeness of level-0 juror's vote increases. For $\theta = 1$, a level-0 juror is equivalent to a level-1 juror. Given this assumption, when the sampling is without replacement, we have the following pivotality values for the level-1 juror for the case where the signal is red as:

$$\alpha_1^r = \theta^{np-1}(1-\theta)^{n(1-p)} \tag{1.43}$$

$$\beta_1^r = \theta^{n(1-p)-1}(1-\theta)^{np} \tag{1.44}$$

Given equations (1.43) and (1.44), using the Table 1.23, we get the following inequality for level-1 juror to vote informatively when she receives a red signal:

$$\left(\frac{\theta}{1-\theta}\right)^{n(2p-1)} > \frac{1-p}{p} \tag{1.45}$$

Similarly for the case when the level-1 juror receives a blue signal, we have the following pivotality values:

$$\alpha_1^b = \theta^{np}(1-\theta)^{n(1-p)-1} \tag{1.46}$$

$$\beta_1^b = \theta^{n(1-p)}(1-\theta)^{np-1} \tag{1.47}$$

Given the equations (1.46) and (1.47), using the Table 1.23, we get the following inequality for level-1 juror to vote informatively when she receives a blue signal:

$$\left(\frac{\theta}{1-\theta}\right)^{n(2p-1)} > \frac{p}{1-p} \tag{1.48}$$

Define K as $\left(\frac{p}{1-p}\right)^{\frac{1}{n(2p-1)}}$. Then with a bit of algebra, we get the following condition for the level-1 juror to vote informatively:

$$\theta \in \left(\frac{1}{1+K}, \frac{K}{1+K}\right) \tag{1.49}$$

First note that given $p > \frac{1}{2}$ we have $\dfrac{1}{1+K} < \dfrac{K}{1+K}$. Then recall that θ represents the probability for a level-0 juror to vote informatively, and note that given the constraint for θ in (1.49), a level-1 juror votes informatively. Hence, given a level-0 juror votes her signal, i.e. votes informatively with probability θ such that it satisfies the constraint (1.49), a level-1 juror votes informatively.

Setting p to $\frac{2}{3}$, for $n = 3$, the inequality in (1.49) becomes $\left(\frac{1}{3}, \frac{2}{3}\right)$; and for $n = 6$, it becomes $\left(\frac{1}{1+\sqrt{2}}, \frac{\sqrt{2}}{1+\sqrt{2}}\right) \sim (0.41, 0.59)$.

For the with replacement case, first note that since the signals are independent, we have $\alpha_1^r = \alpha_1^b$ and $\beta_1^r = \beta_1^b$. Hence, we drop the superscript and define the following pivotality values:

$$\alpha_1 = (p\theta + (1-p)(1-\theta))^{n-1} \tag{1.50}$$

$$\beta_1 = ((1-p)\theta + p(1-\theta))^{n-1} \tag{1.51}$$

Given the equations in (1.50) and (1.51), using Table 1.23, we have the following inequality conditions for a level-1 juror to vote informatively:

$$\frac{p\theta + (1-p)(1-\theta)}{(1-p)\theta + p(1-\theta)} > \frac{1-p}{p} \tag{1.52}$$

$$\frac{p\theta + (1-p)(1-\theta)}{(1-p)\theta + p(1-\theta)} < \frac{p}{1-p} \tag{1.53}$$

Define $\widetilde{K}$ as $\left(\frac{1-p}{p}\right)^{\frac{1}{n-1}}$. Then with a bit of algebra, we get the following condition for a level-1 juror to vote informatively:

$$\theta \in \left(\frac{p(1+\widetilde{K}) - 1}{(2p-1)(1+\widetilde{K})}, \frac{p + \widetilde{K}(p-1)}{(2p-1)(1+\widetilde{K})} \right) \tag{1.54}$$

Setting p to $\frac{2}{3}$, for $n = 3$, the inequalities in (1.52) and (1.53) approximately translate to $\theta \in (0.24, 0.76)$ and for $n = 6$, they translate to $\theta \in (0.4, 0.6)$.

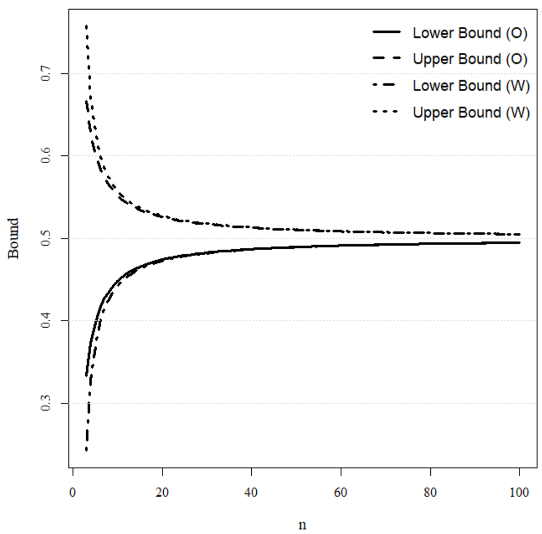


Figure 1.1: Bounds for θ given n and sampling method

As can be observed in Figure 1.1, for both sampling cases, as the group size increases, the range for θ, or in other words, the degree of a level-0 juror's informativeness in her vote decreases. For a level-1 juror to not vote informatively, i.e. to vote strategically, her belief on the informativeness of the aggregation of the other jurors' vote should overweight the informativeness of her signal. As the number of other jurors in the jury increases the necessary informativeness each of these jurors should provide with their vote for their aggregate informativeness to overweight the informativeness of the juror's signal decreases.

1.B Experiment instructions

Introduction

You are about to participate in an experiment in team decision making. Please follow the instructions carefully.

In the experiment you may earn a considerable amount of money. Your decisions and the decisions of the other participants determine the amount. You will be instructed in detail how your earnings depend on your and the others' decisions. All that you earn is yours to keep, and will be paid to you in private and in cash, after today's session.

It is important to us that you remain silent and do not look at other people's screens. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, exclaim out loud, etc., you will be asked to leave. Thank you.

Since this is a team experiment, you will at various times be matched randomly with another participant in this room in order to form a team that plays as a single entity. Your team's earnings will always be shared equally between you and your team partner.

The experiment consists of four parts (**Parts I, II, III and IV**). The parts are independent of each other but feature the same task in different settings. Each part consists of two rounds that require you to take a single decision. The way you interact as a team to take decisions will be the same throughout the experiment. Common features to the Parts will be given in the general instructions section.

Now, let us explain how your ***Team's Decision*** is determined. First of all, both you and your team partner will individually submit a ***Final Decision*** and the computer will randomly choose one of these two final decisions as your team's decision. The probability that your team partner's final decision is chosen is equal to the probability that your final decision will be chosen (i.e. your chances are 50:50). However, you have the possibility to influence your partner's final decision in the following way: Before you enter your final decision, you can propose to your partner a ***Suggested Decision*** and send her one and only one text ***Message***. *Note that this message is your only chance to convince your partner of the reasoning behind your suggested decision. Therefore, use the message to explain your suggested decision to your team partner.* After you finish entering your suggested decision and your message, these will be shown to your team partner. She will then make her final decision. Similarly, you will receive your partner's suggested decision and message. You will then also make your final decision. As indicated above, once you both enter your final decision, the computer chooses randomly one of your final decisions as your team's decision.

If you have any questions at this point, please raise your hand. In order for you to get familiar with the messaging system, you will now try it out in a **Test Period**. Please turn the page for further instructions.

Test period

A participant in this room is now randomly chosen to be your team partner. The **Test Period** has two rounds, with one question to answer in each round. Since this is only a test, your earnings will not depend on any decision taken now. In both test rounds, you will need to answer a question about the year of an historic event. The team that is closest to the correct year wins. Ties will be broken randomly by the computer.

As described, you will be able to send one ***Suggested Decision*** with your proposed year and an explaining ***Message***. After having read your partner's suggested decision and message, you will enter your ***Final Decision***. As described earlier, either your or your partner's final decision will be chosen randomly to be your ***Team's Decision***.

The messenger allows ***Messages*** of any size. However, you have to enter the message line by line since the input space is only one line. Within this line, you can delete text by using the usual "Backspace" button of your keyboard. By pressing "Enter" on the keyboard, you add the written sentence to the message. Please note that only added sentences will be sent and seen by your partner. *The words in the blue input line will* ***not*** *be sent.* You can always delete previously added sentences by clicking the "Clear Input" button. The number of lines you send is not limited. You can therefore send messages of any length. You finally send the message to your partner by clicking the "Send Message" button.

When you are ready, please click the "Ready" button to start the **Test Period**.

General Instructions

In every round of every part of the experiment you will be matched with a single, randomly chosen, different team partner. Together with other teams, which will also differ in every round, your team will face the following situation.

Your team along with other teams will constitute a ***committee*** in which each team has the right to a ***single vote***.

There will be *two* colored urns containing *some number* of colored balls. The urns will either be ***red or blue***, and either colored urn contains some number of red balls and some number of blue balls.

The color of the urn will determine the ratio of the number of red and blue balls in it. ***In the red urn***, there will be ***two times more red balls*** than blue balls and ***in the blue urn***, there will be ***two times more blue balls*** than red balls.

Every round, your newly formed committee will be assigned to a ***single urn*** whose ***color*** will be ***randomly determined*** by the computer as either blue or red ***with the equal probability***.

Your team – like all other teams in your committee – will observe the color of only one ball drawn from the urn assigned to your committee. Your task as a committee is to ***correctly guess the color of the urn***. The guess of the committee is determined by the votes of its teams.

The votes of the teams in the committee will be aggregated to a committee decision according to the ***unanimity voting rule***. The rules of the voting rule are as follows:

- If all the teams vote for the red color, then the color red will be the committee's guess.
- If at least one of the three teams votes for the blue color, then the color blue will be the committee's guess.

In other words, in order for the committee to select the red color as the guess, all the teams have to vote red. Thus, the teams need to *reach a consensus* in order to guess red; otherwise, the guess of the group will be blue.

If the committee correctly guesses the color of the urn they are assigned to, ***every team member*** in the committee will receive ***200 Eurocents***. If the committee does not correctly guess the color of the urn ,then ***every team member*** in the committee will receive ***20 Eurocents***.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

The instructions of the Parts will specify the number of teams in the committee, the number of balls in the urn and the exact procedures of drawing a ball from the urn. Are there any questions at this point?

Part I

You are about to start Part I of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of three teams (your team and two other teams).

In this part, the balls are drawn from an urn that contains **two balls** of the same color of the urn and **one ball** of the ”opposite” color. Your team will only observe the color of a single ball drawn from the urn. For all three teams, the balls will be drawn **without replacement**. That means that a drawn ball is not returned back to the urn for subsequent draws. There will therefore always be two teams observing the correct color and one team observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will* ***not*** *be sent. Press "Enter" to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the "Ready" button, you will start **Part I** of the experiment.

Part II

You are about to start Part II of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of six teams (your team and five other teams).

In this part, the balls are drawn from an urn that contains **four balls** of the same color of the urn and **two balls** of the ”opposite” color. Your team will only observe the color of a single ball drawn from the urn. For all six teams, the balls will be drawn **without replacement**. That means that a drawn ball is not returned back to the urn for subsequent draws. There will therefore always be four teams observing the correct color and two teams observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be sent. Press "Enter" to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the "Ready" button, you will start **Part II** of the experiment.

Part III

You are about to start Part III of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of three teams (your team and two other teams).

In this part, the balls are drawn from an urn that contains **two balls** of the same color of the urn and **one ball** of the "opposite" color. Your team will only observe the color of a single ball drawn from the urn. For all three teams, the balls will be drawn **with replacement**. That means that a drawn ball is returned to the urn for the subsequent draws. Independently of other teams' draws, each team will have a 2/3 chance of observing the correct color and a 1/3 chance of observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be sent. Press "Enter" to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team's Decision***.

When you click the "Ready" button, you will start **Part III** of the experiment.

Part IV

You are about to start Part IV of the experiment. In each of the two rounds you will be matched with a new team partner and a new committee. Your team belongs to a committee that consists of six teams (your team and five other teams).

In this part, the balls are drawn from an urn that contains **four balls** of the same color of the urn and **two balls** of the "opposite" color. Your team will only observe the color of a single ball drawn from the urn. For all three teams, the balls will be drawn **with replacement**. That means that a drawn ball is returned to the urn for the subsequent draws. Independently of other teams' draws, each team will have a 2/3 chance of observing the correct color and a 1/3 chance of observing the incorrect color (with respect to the color of the urn assigned the teams). Your team will ***not*** know or observe the colors of the balls given to the other teams in your committee.

Upon the observation of the color of your team's ball, you will send your team partner a ***Suggested Decision*** and a ***Message***. Remember to explain in the message your reasoning behind your suggested decision. *(And note again that the words in the blue input line will **not** be* *sent. Press “Enter” to add them to the message.)* After this information is exchanged, both of you enter your ***Final Decision***, from which the computer randomly chooses the ***Team’s Decision***.

When you click the “Ready” button, you will start **Part IV** of the experiment.

1.C Classification instructions

Welcome

Thank you for participating in this experiment. In this document you will find instructions as to how this experiment works. You will be asked to classify messages that have been collected in an experiment on voting games. You as well are in an experiment which allows us to give you particular incentives and makes it easier for us to pay you.

To take part in the experiment, we assume that you are familiar with the level-k model as it has been introduced by Nagel (1995). However, in order to clarify potential questions of terminology, we reproduce the main features of the level-k model. In addition we provide detailed instructions of the original experiment, which explain the voting game and also give you a short introduction to voting games. Please read all information carefully.

Classification Task

General Comments

Your task is to classify the messages sent by subjects to their team member into their according level of thinking. You will read each comment and classify it according to the guidelines below. You can enter your assessment into the excel sheet provided to you. The excel sheet will have 6 different columns: The first four columns (which are already provided/filled) will identify the message by indicating the experiment, the part, the subject, the period. The fifth column is the classification column that we want you to fill and the sixth column is to indicate your personal comments on your task to further clarify your classification choice if necessary. The order of these columns will follow the transcript we will provide you.

It is very important that you double check whether the first 3 columns are filled correctly, i.e. that you enter the data for the correct subject, period, part and experiment. Based upon the guidelines below your task will be to fill up the classification column with an integer between 0 and 3 ("0" for level-0, "1" for level-1, "2" for level-2 and "3" for level-3) or leave it empty. If you find interesting elements that occur frequently but that have not been picked up by us, feel free to add a new column and mark all messages that contain the element. You can then specify to us in an email what exactly this element is.

For each individual classification, your assessment will be benchmarked against another classifier's assessment. Your personal remuneration is based on the number of matches of the level classification. A match is a classification that is congruent with the classification of another independent classifier. Each match will be remunerated with 0.07 Euro.

Please read this document and the instructions for the experiment entirely in order to get an overview and only then start the classification based on the player's sent message and action proposed. If you have any questions please do not hesitate to contact us.

The Original Experiment

A single experimental session consists of 4 parts and every part consists of 2 periods. In every part, every subject is randomly paired with another one to form a team. Depending on the treatment, every team then is randomly matched with 2 or 5 other teams to form a voting group. Then every group will be assigned to a blue or red urn and every team in the group will draw a ball from their assigned urn. The blue (red) urn contains twice more blue (red) balls then red (blue) balls. One treatment variable is the group size, in every experiment, in parts 1 (2) and 3 (4) the groups will consists of 3 (6) teams and accordingly the urns will contain 3 (6) balls. The other treatment variable is the draw mechanism, depending on the session, either for the first 2 parts or the last 2 parts, the balls will be drawn from the urn without replacement (i.e. any ball picked will not be placed back to the urn). For the other 2 parts, the balls will be drawn with replacement (i.e. any ball picked will be placed back to the urn before the next draw). In addition, we have included classroom experiments to the data set (experiments 7 and 8). In both experiments, the balls are drawn from the urn without replacement. Both experiments have two parts where in the first part the group size is 3 while in the second part the group size 6.

The goal of every group is to correctly state the color of the urn they are assigned to. Every team submits a vote (either red or blue) and if every team in the group votes for red then the group's decision will be red, otherwise it will be blue. Beyond this brief description of the experimental setup, please read the instruction sheets given to the subjects to have a better understanding of the situation of subjects [23].

How do the messages get produced? Every subject in a team observes the team's drawn ball and then makes a decision on the color of the urn. Then, they send a message to their team member explaining why they should vote for the suggested color. Next, the team members receive the suggestions and the message of their teammates, and make their final decision on their own. Either of the team member's decision is equally likely to be chosen as the team's final voting decision. This experimental methodology is developed to elicit subjects' reasoning through their sent messaged (see Burchardi and Penczynski, 2014). In the classroom experiments (experiments 7 and 8), subjects were not paired into teams but instead they were asked to elaborate on their reasoning for taking the action they have taken.

Model

General Model

It is assumed that you are familiar with the level-k model as it has been introduced by Nagel (1995) or represented by Camerer (2004). In order to clarify potential questions of terminology and introduce the main features of the model we quickly reproduce the main features of the model in the terminology used in this document. The level-k model of bounded rationality assumes that players only think through a certain number (k) of best responses. The model has four main ingredients:

Population distribution: This distribution reflects the proportion of types with a certain level $k \in N_0 = \{0, 1, 2, 3, 4, 5, \ldots\}$.

[23] there are two versions of the instruction sheet where the versions vary in their of treatments. We are providing you with one copy. Please do not base the treatment ordering on the ordering of the instruction sheet.

Level-0 distribution: By definition, a level-0 player does not best respond. Hence, his actions are random to the game and distributed randomly over the action space. In our case, the action space is $\mathcal{A} = \{Red, Blue\}$ where Red and $Blue$ represent the voting choice of the player. Note that, our model does not incorporate salience by assuming higher probabilities in the level-0 distributions for the action that is salient due the signal received (i.e. if a blue ball is received and the player is playing random, the player, due to the availability of the blue signal, will chose to vote blue)

Level-0 belief: In the model, the best responses of players with level $k > 0$ are anchored in what they believe the level-0 players play. Their level-0 belief might not be consistent with the level-0 distribution. For best responding, all that matters is the expected payoff from choosing an action from the action space $\mathcal{A} = \{Red, Blue\}$.

Population belief: Players do not expect other players to be of the same or a higher level of reasoning. For a level-k player, the population belief is therefore defined on the set of levels strictly below k. It follows that level-0 players have no defined belief, level-1 players have a trivial belief with full probability mass on $\{0\}$ (i.e. the belief that everyone else is level-0), level-2 players have a well defined belief on $\{\{0\}, \{1\}\}$. From level 3 higher order beliefs are relevant as level-3 players have to form a belief about level-2's beliefs.

Specific Model

We consider a game with n players (jurors in the voting context). The game starts by nature choosing a state of the world S in $\Omega = \{Red, Blue\}$ with probability r and $1 - r$ respectively. The players do not observe the state, but each acquires a private signal s about the realized state of the world. If the true state is $Blue$, then each player observes an independent (or geometrically dependent[24]) Bernoulli random variable (the private signal) which is $blue$ with probability p and red with probability $1 - p$ (and conversely for when the true state is Red). After observing their private signals, players chose an action a (a vote) from the action space $\mathcal{A} = \{Red, Blue\}$. Given the votes of the players, $1 \leq k* \leq n$ represents the number of votes needed for Red to be chosen for the aggregate decision. In other words, if $k*$ many or more players vote for Red, then the group decision is Red otherwise it is $Blue$. The utility of jury j when she takes action a with certainty given her signal is s and given the state is S is defined as $u_j(\sigma_j(s) = a, S)$. Given any signal s, the utility $u : A$ x $\Omega \mapsto \mathcal{R}$ for jury j is further defined by $u_j(\sigma_j(s) = blue, Blue) = u_j(\sigma_j(s) = red, Red) = 0$, $u_j(\sigma_j(s) = red, Blue) = -q$ and $u_j(\sigma_j(s) = blue, Red) = -(1-q)$, where $0 < q < 1$.

In all our experiments, we used $k* = n$ (i.e., the unanimity voting rule), $r = 0.5$, $q = 0.5$ and $p = \frac{2}{3}$. As previously explained treatments vary in the number of players n between 3 and 6.

Under the experimental set-up, the model provides the prototypical behavior of the subjects given their level as follows:

$n = 3$ ***or*** $n = 6$ ***and with replacement:***

When the blue ball is observed, optimal strategy for:

- level-1 player: vote blue

[24]Independent case refers to the aforementioned with replacement draws case while the geometrically dependent case refers to the without replacement case

- level-2 player: vote red

When the red ball is observed, optimal strategy for:
- level-1 player: vote red
- level-2 player: vote red

$n = 3$ ***and without replacement:***

When the blue ball is observed, optimal strategy for:
- level-1 player: vote blue
- level-2 player: vote red

When the red ball is observed, optimal strategy for:
- level-1 player: vote red
- level-2 player: vote red or blue

(Voting red is strictly preferred to voting blue under additional assumption that the level-2 player assumes with some small probability ϵ that the other players are level-0

$n = 6$ ***and without replacement:***

When the blue ball is observed, optimal strategy for:
- level-1 player: vote blue
- level-2 player: vote red or blue[25]

(Voting blue is strictly preferred to voting blue under additional assumption that the level-2 player assumes with some small probability ϵ that the other players are level-0 and not level-1)

When the red ball is observed, optimal strategy for:
- level-1 player: vote red
- level-2 player: vote red or blue

(Voting red is strictly preferred to voting blue under additional assumption that the level-2 player assumes with some small probability ϵ that the other players are level-0)

Guidelines for classification

General Comments:

- Subjects do not necessarily describe every step of their thinking; therefore, it may not always be obvious to decide which level they are. In many comments, any indications of a level of thinking may be partial or implicit, you should then indicate the most likely level of reasoning of the player.
- If the message indicates to simply refer to a previous message (”same as before/above”),

[25]Note that for $n = 3$ voting red is strictly preferred to voting blue

then you can use the previous message's evaluation to determine the level of the current message. Please indicate this inference with a 1 in the column ”Other message inference”.

- If you are unsure of the level of the message, you should indicate the level you think is more likely.
- We have deliberately chosen not to disclose the action taken by the subject. You may still see in their comment which action they chose. We do not want you to base your classification on the action taken as it may be misleading.

Empty classification:

If no message has been formulated you should leave the classification empty. Also, you should leave the classification empty, if you are not sufficiently certain that any of the types below is capturing the strategic thinking in the message.

Level-0 Player:

Characteristics Chooses randomly, without justification or through some justification completely unrelated to the task. Might not have understood the game or shows no interest in the game or in thinking about it.

Examples ”50 50 chance to get red at least 50 50 could also be 100 percent.”
”I like blue, so I chose blue.”
”Think it will be red again.”
”definitely red this time”
”We have to go for red. No other way than that. I like turtles”

Note Comments such as ”It is obviously blue” or ”Play red, trust me!” should not be considered as level-0 thinking as these comments to some extent signal some level of understanding/interest of the task. Such comments are likely to be level-1 comments yet without any additional information, you should leave the specific cell empty.

Level-1 Player:

Characteristics Always follows his own signal. The subject may argue in favor of playing his own signal through some probability argument

Examples ”Our signal is blue. Let's play blue.”
”The probability that the red ball we observe is out of the red urn is twice the probability that it is out of the blue urn”
”1/3 of all teams is observing wrong color, so we would try to find out whether we have wrong or right ball, keep with red.”

Note The key idea in defining a level-1 player is to identify some thinking process that signals the subject's interest/understanding of the task and the private signal. Furthermore, it is important that the subject does **not** offer any argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e. adjusting the strategy given what others are expected to do).

Level-2 Player:

Characteristics Assume that all other players almost always follow their signal (i.e. she assumes almost all the other players are level-1 while an epsilon portion of them are level-0). Player does offer an argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e. a best response given others are most likely playing their signal). In other words, if you identify any comment that indicates that the subject assumes (or considers the case) where the other players in her group play their signal, you should consider the possibility that the subject is a level-2 player.

Examples ”Let’s take red because if the urn is red and we got the opposite color and we take blue, the decision will be blue.”

”We need to chose Red. If we are the only ones who picked blue, then the urn is red and we guess correct If the urn is blue, then the other guys will pick blue so there will be at least one blue vote and we win as well If the others guys (also blue) think the same way then we lose But this is too many ifs”

”I have a blue ball. If we have the blue urn, someone else also has a blue ball and as a result our group will chose blue regardless of my vote. If we have the red urn, I am the only one with the blue ball and if I vote blue, we will chose the wrong urn. So I should vote for red.”

”In case two teams choose red and one chooses blue, blue will be taken. That means that choosing red has a higher chance of being a good decision.”

”I guess this is more about luck because there is no way to know it for sure. I would say blue just because of the higher probability. Also I like turtles Also it is likely that one other team will pick blue and then it is that color anyways”

”There is no point for us to take blue I think the chances for us to get the right color are higher if we stick with red” [red ball is observed]

”I suggest red because we donat hurt anyone with this decision If the others go for blue because they have a blue ball, the committeeas decision will be blue regardless of our decision”

”We could be the deciding vote for blue if the other two choose red. Choosing blue isnt as helpful as choosing red, because: only one blue ball can overturn our whole decision but only a unanimous decision for red can help us the same way”

Note In order to discern the two types, you should look for more than any trivial arguments such as the ones given under level-1. There may be cases where the message starts as a level-1 argument and then as the subjects elaborates on her reasoning, she starts considering the strategy of the other teams and justify her decision accordingly (see the third example above). In such cases, this message should be considered as level-2. The acknowledgment of other teams’ voting strategy may not always be obvious or may be worded differently such as ”hurting the other’s decision” or ”not being helpful” (see the last three examples above)

Level-3 Player:

Characteristics Assumes that almost all other subjects are level-2 players (partially degenerate beliefs). The reasoning in a level-3 player message will have similarities with a level-1 player message but it will have additional arguments indicating that she assumes others are level-2 players.

Examples ”In my opinion, if there is another person with blue they may be afraid of voting blue so we should vote blue to make sure.”
”Let’s now pick the shown colour because the others now will probably enter their opposite colour.”
”Risky to vote blue but others may not vote blue even when they draw blue. I say we vote blue.”

Note As stated above, level-3 players are likely to follow their signal like a level-1 player yet they will argue to do so through a much more intricate argument (unlike a level-1 player merely stating probabilities to argue her action). Level-3 players are rare. Higher levels (level-4 etc.) are assumed to not occur; therefore, you should consider only the first 4 levels of thinking.

Thank you.

Chapter 2

Using Large Language Models for Text Classification in Experimental Economics

Joint with Stefan P. Penczynski

2.1 Introduction

Advances in information technology increase the performance of computerised Natural Language Processing (NLP), which in turn increases NLP's potential contribution to the scientific endeavour. The accessibility of Large Language Models (LLM) will further increase the use of "text as data" in the social sciences (Gentzkow et al., 2019) because, as we show here, text can easily and accurately be classified according to challenging social scientific concepts.

In this study, we employed GPT models to classify text transcripts from economic experiments focused on promise and the level of strategic thinking. We aim to answer the following research questions:

- How do GPT models compare to expert human annotators and traditional machine learning methods in classifying these concepts?
- Can classification instructions designed for human annotators be minimally modified to serve as prompts for LLMs, delivering comparable performance levels?
- How effective are established prompting techniques in classifying these concepts, and how does model size influence task performance and the efficacy of these techniques?

We vary the prompts along two dimensions in order to explore LLMs' performance and to understand which prompting techniques are most effective for the given classification tasks. In the first dimension, we alternate between zero-shot and few-shot prompting, which relates to the number of example messages and classifications that are included in the prompts (Dong et al., 2022). In the second dimension, we alternate between requiring responses to be with or without the chain-of-thought (CoT) feature by employing the 0-shot CoT prompting technique introduced by Kojima et al. (2022). With CoT, the LLM provides written justifications for each of its classification decisions. Furthermore, we conducted all these classifications using OpenAI's GPT-3.5 turbo as well as GPT-4 turbo to investigate performance gains achievable with larger LLMs.

We find that GPT-4 outperforms GPT-3.5 and achieves high levels of agreement, reaching near or above 90% and up to 73% in more complex tasks. Importantly, these high levels can at times be attained through zero-shot classification instructions originally designed for human annotators. Unlike traditional supervised machine learning approaches, which necessitate partitioning data for training and testing (Penczynski, 2019), use of LLMs obviates the need for data separation, making it particularly advantageous in low-resource environments, such as those commonly found in economic experiments. Moreover, we document that while n-shot prompting consistently improves the performance of both models, 0-shot CoT prompting's effect on models' performances is dependent on the model, the prompt and the task difficulty. It often enhanced GPT-4's performance but did not reliably improve GPT-3.5's performance across tasks and prompts, generated more consistent performance gains for relatively more difficult tasks, and when used in conjunction with n-shot prompting, always resulted in the best performing outcomes for GPT-4.

Additionally, by examining the classification of two concepts –promises and strategic thinking– which potentially vary in terms of their availability in the pre-training corpus of the models, we investigate whether the classification task primarily involves recognition-based or learning-based subtasks, depending on the concept (Pan et al., 2023). While we find no significant difference in performance gains switching from GPT-3.5 to GPT-4 among concepts, we do observe a higher degree of performance gain from incorporating demonstrations into the prompts for the classification of level of thinking compared to the classification of promises, and thereby, find some evidence that the classification of level of thinking involves more learning components than the classification of promises.

As a scientific tool, LLMs offer a number of advantages over RAs, making them an increasingly attractive option for researchers working with 'text as data' and enhancing the scientific insights derived from text. The costs associated with RAs arise from recruitment, instruction, and, most importantly, their work-time. In contrast, LLM services are paid by tokens and are comparatively inexpensive. For example, classifying 100 messages in the most complex dataset cost 6.60 USD (as of February 2024) and took only 43 minutes. LLMs offer attractive and distinctive performance features that are poised to improve further. They deliver immediate results, provide on-demand detailed justifications (CoT), and ensure consistent classifications that are not subject to fatigue. Additionally, testing and gradually refining prompts is straightforward and cost-effective with LLMs, whereas it is prohibitively expensive with RAs.

2.2 LLMs and related literatures

An LLM is a statistical language model trained on a large corpus to predict the next word for any given textual input. By inputting text instructions, one can strategically leverage this predictive capability to steer the model's output towards a desired outcome, a practice commonly referred to as prompting. The appeal of prompting stems from the ease with which natural language allows us to convey complex ideas. Yet, this very flexibility may introduce inaccuracies or ambiguities if concepts are not clearly defined or presented with insufficient context. The effectiveness of a prompt hinges both on the user's adeptness at crafting instructions with clarity and contextual relevance, and on the model's ability to accurately *interpret* these instructions within their context. While an LLM's capacity to process text and follow instructions are fundamentally based upon its pre and post training and its parameter size, for downstream tasks, the user can still attempt to refine her mode of interacting with the LLM by engineering her prompts to align more closely with model's operational framework, in order to effectively leverage its capabilities(Reynolds and McDonell, 2021).

2.2.1 GPTs in the computer science literature

FLAN (Wei et al., 2021), OPT (Zhang et al., 2022b), and PaLM (Chowdhery et al., 2023) are examples of LLMs that have showcased remarkable proficiency in natural language understanding (NLU) tasks (Ye et al., 2023). Particularly, the generative pre-trained transformer (GPT) series (Brown et al., 2020), more specifically GPT-3 and its subsequent iterations GPT-3.5, GPT-3.5-turbo, GPT-4 and GPT-4-turbo introduced by OpenAI have sparked considerable attention due to their exceptional performance in integrating various NLU tasks into generative ones (Ye et al., 2023).

Earlier GPT models, GPT-1 and GPT-2, are limited in their ability to recognise textual patterns across diverse tasks due to their relatively smaller training corpus and parameter size (Radford et al., 2019). Consequently, these models require substantial fine-tuning on task-specific datasets to achieve satisfactory performance. Yet fine-tuning poses several problems: first, it requires large volumes of task-specific data; second, there is a risk that these training datasets do not cover the full spectrum of task variations, which could lead to suboptimal performance on data not represented within the training set (lack of generalizability due to over-fitting) (Brown et al., 2020). Furthermore, fine-tuning an LLM on data that introduces new knowledge is documented to increase the model's likelihood to make up information (hallucinate) (Gekhman et al., 2024).

Building on its predecessor, GPT-3 has been trained on a significantly larger corpus, consisting of approximately 400 billion tokens, compared to GPT-2 which was trained on 1.5 billion tokens (Brown et al., 2020). This extensive training has markedly imrpoved its ability to detect diverse textual patterns (Brown et al., 2020) and has enabled reasoning-like emergent qualities (Wei et al., 2022b). Notably, GPT-3 can perform specialised tasks when provided with few examples demonstrating how to perform it (Brown et al., 2020). This capability, which significantly reduced the need for parameter adjustments through fine-tuning, catalysed the development of the In-Context Learning (ICL) paradigm for LLMs comparable in size to GPT-3 or larger (Dong et al., 2022).

In the ICL paradigm, the process of demonstrating task execution through a small number of input-output pairs, where the input serves as the question and the output as the answer, is referred to as n-shot prompting. In the specific case of classification tasks, the term "n" indicates the number of examples provided in the prompt. n-shot prompting quickly gained popularity as it requires only a few demonstrations to guide the model toward achieving performance comparable to that of fine-tuned models trained on extensive datasets. Notably, a single demonstration can be as effective as fine-tuning the model with approximately 300 to 3,500 input-output pairs, depending on the task (Scao and Rush, 2021). Additionally, by modifying the format of the demonstrations from <input, output> to <input, reasoning, output>, one can enable the model to demonstrate reasoning capabilities. This approach, referred to as n-shot-CoT (Chain of Thought), has proven to significantly enhance the models' performance, especially on more involved tasks that can potentially benefit from multi-step reasoning (Huang and Chang, 2022).

Although the ICL paradigm offers a flexible and data-efficient way to "teach" the model at inference, the efficacy of the method on improving the model's task performance, or in other words the model's ability to "learn" from these demonstrations, relies to a greater extent on the choice of examples, the sequence in which they are presented within the prompt, and the frequency with which examples for each category to be classified are provided (Lu et al., 2021; Kumar and Talukdar, 2021a; Zhao et al., 2023). Although various methods to select the optimal set of examples and their order have been proposed, and documented to improve the model's performance (Li and Qiu, 2023a; Su et al., 2022; Liu et al., 2021; Luo et al., 2023; Chang and Jia, 2022), their technically demanding procedures may be less accessible for social scientist to implement. This lack of accessibility juxtaposes with the appeal of GPT's out-of-the-box usability, which we believe is necessary for any prompting method to be widely adopted by social scientists.

Given n-shot-CoT prompting is an extension to n-shot prompting, its effectiveness in enhancing the model's task performance also relies on the choice and order of examples. In addition, however, its effectiveness is also dependent on the manner with which the rationales are provided for each demonstration (Huang and Chang, 2022). To improve the reliability of n-shot prompting, the most cited method is the "self-consistency" method proposed by Wang et al. (2022). In this method, the temperature hyperparameter is set to a strictly positive value (e.g. 0.5 or 0.7) in order to increase the variability in the reasoning sequence of the output, and multiple request (e.g. 20 or 40) to the model are made using the n-shot-CoT method for every single task instance. Then the most frequently outputted answer is picked as the decisive (consistent) output. Noting that requesting a reasoning prior the answer already puts additional token cost, by generating k many outputs with reasoning for each task instance, "self-consistency" method is increasing the computational cost approximately k times more. Making the classification task approximately k times more costly to improve the reliability of the model's output could make this approach prohibitively expensive for social scientists, and potentially deter them from viewing it as a viable alternative to human annotators.

To refine the reasoning provided with each examples, it has been suggested, among many, to use multiple human annotators to provide a diverse set of reasoning for each example (Li and Qiu, 2023b), to use the LLM itself to generate a diverse sets of reasoning for each examples and then select the reasoning with the most steps (Fu et al., 2022), or to generate multiple outputs using the generic n-shot-CoT prompting method and to choose the most frequently provided output as the final output of the model (Wang et al., 2022) (see Huang and Chang (2022) for details and other methods to refine the rationale). Arguably, although these proposed methods for refining the reasoning in demonstrations are not computationally demanding, they still require careful selection of the most suitable reasoning for each example, and may lead the researchers to doubt whether the explanations considered were adequate, particularly when the model's performance does not meet their expectations.

In addition to the unreliability concerns with n-shot and n-shot-CoT prompting, it is also unclear how the model "learns" at inference to perform a task via few demonstrations of the task, and whether it genuinely learns via demonstrations (Reynolds and McDonell, 2021). Min et al. (2022) demonstrate that even when the output labels of the input-output pairs in n-shot prompts are replaced with incorrect labels, the model's performance remains unaffected. They suggest that models do not learn from demonstrations in the same way humans do from examples; rather, these examples primarily serve to delineate the label space and the distribution of the input text, thereby aiding the model in task execution[1]. Reynolds and McDonell (2021) argue that demonstrations do not actually "teach" the model how to perform the task but simply enables the model to locate the tasks in the model's existing knowledge of tasks that it acquired during its pre-training[2]

[1]Yoo et al. (2022) revisited the assertions of Min et al. (2022) and found instances where employing incorrect output labels adversely affects model performance. Consequently, the question of whether and how models learn from demonstrations, and precisely what they learn, remains an open question.

[2]As a consequence, In-Context learning is also referred to as priming (Webson and Pavlick, 2021). Yet, the term priming encompasses a broader spectrum of prompting techniques. For instance, priming can also be done by pre-pending a prompt with a series of token (instead of or in addition to the input-output

(meta-knowledge). Similarly, with n-shot-CoT prompting, it is unclear whether the model genuinely engages in reasoning and, if so, how this reasoning improves the task performance (Madaan and Yazdanbakhsh, 2022). Furthermore, the ability of a model to reason effectively has been shown to correlate with the frequency of a task's presence in its pre-training corpus: the more frequently a task is represented in the training data, the more likely the model is to exhibit sound reasoning and produce accurate outputs (Razeghi et al., 2022).

Given the aforementioned challenges with n-shot-CoT method, a notable, demonstration-free alternative is the 0-shot-CoT prompting method (Kojima et al., 2022). This approach simply involves appending the phrase "Let's think step-by-step" to the instructions and thereby triggers a reasoning process before generating the output. 0-shot-CoT, devoid of the reliability concerns regarding the selection of examples or the quality of reasoning, is task-agnostic and can seamlessly be integrated into an existing prompt by adding the keyword "think step-by-step" into the specific sections of the instructions where the user wishes to invoke reasoning steps in the model (OpenAI, 2023b). While 0-shot-CoT has been shown to significantly enhance performance across a range of tasks, its efficacy diminishes with relatively more involved tasks that potentially requires an explicit outline of the reasoning steps to be followed by the model. In such scenarios, the method often yields suboptimal outcomes due to the model's failure to accurately execute or complete the necessary reasoning steps, either by omitting steps or by making errors within specific steps of the reasoning process (Zheng et al., 2023; Wang et al., 2023). Consequently, alternative 0-shot reasoning-invoking methods have been proposed to address these limitations (Huang and Chang, 2022). Furthermore, 0-shot-CoT's effectiveness is similarly influenced by the model's pre-training corpus. Tasks less represented within the training corpus is observed to provide diminished performance improvements when the method is integrated. This highlights, once again, the dependency of the model's performance on the attributes of its training corpus (Wu et al., 2023).

OpenAI has progressively enhanced GPT-3 by fine-tuning on a collection of instruction-answer pairs[3] (Ouyang et al., 2022). These improvements enabled the model to more closely follow instructions, and reduced to a considerable degree the need to instruct the model to perform a task via demonstrations (Chung et al., 2022). Consequently, recent literature has begun to emphasise the instruction learning paradigm which shifts focus from learning through demonstrations to learning via instructions (Lou and Yin, 2024). It's worth noting that this paradigm does not preclude the inclusion of demonstrations

demonstration pairs) that do not necessarily make intuitive sense (Kumar and Talukdar, 2021b).

[3]Fine-tuning GPT-3 to better follow instructions resulted in the development of GPT-3.5 (Ouyang et al., 2022). GPT-3.5 was then further fine-tuned using reinforcement learning from human feedback (RLHF) method to further enhance its capacity to understand instructions and to better engage in conversational interactions with its users (Ye et al., 2023). These advancements led to the creation of GPT-3.5-Turbo, the underlying model of the ChatGPT application.

within the instructions; rather, it puts more weight on structuring and designing prompts that combine instructions and examples to optimise the model's performance (Lou et al., 2023). There are only a few studies that aim to provide cross-task generalizable prompt design tips intended to enhance demonstration-free instructions to be considered in prompt engineering (Mishra et al., 2021a; Reynolds and McDonell, 2021; Mishra et al., 2021b; Gu et al., 2022; White et al., 2023; Peskine et al., 2023). Additionally, as of writing of this paper, only a single paper has systematically explored various instructional prompting techniques across a wide range of tasks and documented each instructional design component's contribution on improving the model's performance (for a more detailed discussion and the application of these design choices, see Section 2.4.1) (Mishra et al., 2021a).

The concept of instructing an LLM in a manner akin to how one might instruct a human is appealing because it renders the act of prompting both intuitive and flexible. Furthermore, prompting via detailed instructions, free of examples, circumvents various methodological issues inherent with the ICL paradigm. Yet, the availability and diversity of instructions in the training corpus of the model is also observed to be a determinant factor on the effectiveness of 0-shot instructions on improving the model's performance. When instructions tailored for human annotators (such as mTurkers on Amazon Mechanical Turk) are considered verbatim as prompts to assess the model's ability on following human-tailored instructions (turking test), the smaller GPT-2 model has demonstrated poor performance(Efrat and Levy, 2020), while the larger GPT-3 model, when fine-tuned on a large set of human-tailored instructions, has demonstrated the ability to effectively follow unseen human-tailored instructions (Mishra et al., 2021a; Ouyang et al., 2022). Moreover, human-tailored instructions are documented to outperform basic prompts that instruct the model with one or two sentences devoid of any additional descriptive context for categories (Mishra et al., 2021a). In brief, irrespective of how well instructions are constructed to improve the model's performance, the size of the model and whether the model was fined-tuned on instructions are observed to play a major role in the model's performance; and if the model is fine-tuned on instructions, then using instruction provides an improvement over the performance of the model.

Similar to the ICL paradigm, where it is uncertain whether the model genuinely learns from demonstrations, it is also unclear whether the model truly grasps the task's context and execution conditions from a set of 0-shot instructions (Webson and Pavlick, 2021). If the model truly learns from instructions, variations between two different sets of instructions that convey the same meaning should not affect its performance. However, it has been observed that, without sufficient fine-tuning on task-specific examples, changes in word choices that preserve semantic textual similarity in 0-shot instruction prompts can impact model performance as significantly as training it with an additional 200 task-specific examples (Puri et al., 2022). On the other hand, when the model is fine-tuned with a large collection of task-specific examples, its performance demonstrates robustness to variations in the wording of the instructions (Puri et al., 2022). These results, on the one hand, demonstrates the importance of providing carefully designed instructions, while, on the other hand, hints at the fact that the model does not only learn from a set of instructions but also leverages the provided descriptions to locate the task on its existing knowledge. Lastly, demonstrations and carefully provided descriptions for each classification category are observed to complement each other. Irrespective of whether the model is sufficiently trained on task-specific instructions, when instructions are supplemented with a few demonstrations, the model's performance is observed to remain stable despite variations in the choice of words and phrases (Gu et al., 2022).

A recurring theme in our discussion of various prompting techniques is that the effectiveness of any such technique in enhancing a model's performance largely depends on how well the task's contextual components are represented in the model's training corpus. Certain elements of any given task might already be familiar to the model, while others may be novel. This distinction categories any given task as either a recognition task, where the model identifies elements it has seen before, or as a learning task, where the model encounters new contextual elements. Through a series of carefully designed experiments, Pan et al. (2023) demonstrate that the marginal effect of additional task examples in a prompt diminishes for recognition tasks, since only a few examples are observed to be sufficient for the model to recognize the task, and any additional examples do not enable the model to "further recognize" it. Conversely, for learning tasks, the effect of additional demonstrations is observed to be somewhat linear, with each additional example helping the model grasp more of the task's contextual nuances a bit more. Another distinction between recognition and learning tasks identified by Pan et al. (2023) is the scale of the model[4]. The model size is observed not to significantly enhance its performance on recognition tasks, whereas it is observed to be a crucial factor for the model's ability to learn from demonstrations when faced with a novel task (Pan et al., 2023). This observation is supported by Wei et al. (2022b), who argue that as the model scales, it acquires an emergent capability to learn from the examples. Lastly, it is important to note that any given task may consist of subtasks that fall into two either recognition or learning task category. Furthermore, certain contextual elements of a task might be categorised as learning tasks, while others are more appropriately considered as recognition tasks. Thus, both the size of the pre-training corpus and the scale of the model are potentially crucial factors that impact the performance of the model for a given task. Yet this impact may vary depending on the proportion of recognition to learning components within the task.

[4]The scale or size of a model refers to the number of parameters in its neural network. Largest GPT-2 model has approximately 1.5 billion parameters (Radford et al., 2019) while GPT-3 has approximately 175 billion parameters (Brown et al., 2020), and although not exactly known, GPT-4 is estimated to have over 1 trillion parameters (Baktash and Dawodi, 2023).

Consequently, recognising that the impact of prompting techniques on model performance can vary significantly with each specific task is crucial, as this variability necessitates a case-by-case investigation of the effectiveness of different prompting techniques and model configurations. Therefore, it is unreasonable to universally generalise that one prompting technique or a model with a larger training set or more parameters will consistently perform better across all tasks.

2.2.2 GPTs in the social science literature

In Table 2.1, we compile a selection of studies that explore the annotation capabilities of various GPT models. Although this list is not exhaustive, it effectively showcases the diverse prompting techniques used by researchers in a range of annotation tasks within the social sciences.

Despite well-established guidelines from computer science literature, there is a noticeable oversight in the related social science literature with respect to the integration of various prompting strategies such as instruction learning, n-shot prompting, 0-shot-CoT, and the proper usage of the temperature hyperparameter. Our aim is to provide insights relevant to our current work and highlight methodological oversights such as the misuse of the temperature hyperparameter (Reiss, 2023; Törnberg, 2023; Gilardi et al., 2023; Pangakis et al., 2023; Matter et al., 2024; He et al., 2024; Li et al., 2024), misuse of 0-shot-CoT (Zhu et al., 2023; Kuzman et al., 2023; Li et al., 2024), classifying messages in batches (Zhang et al., 2022a; Amin et al., 2023; He et al., 2024; Heseltine and Clemm von Hohenberg, 2024; Matter et al., 2024), and using chatGPT rather than the underlying GPT model (Zhang et al., 2022a; Kuzman et al., 2023; Zhong et al., 2023; Amin et al., 2023; Bhat and Varma, 2023; Heseltine and Clemm von Hohenberg, 2024). If unaddressed, these oversights could compromise the perceived utility of LLMs in text annotation tasks and could misdirect the literature towards suboptimal prompting practices. Moreover, certain papers are often cited for their claims that GPT models are unreliable in text annotation tasks, yet their conclusions rest on methodologically questionable practices (Reiss, 2023; Savelka et al., 2023). Meanwhile, other papers advocate for the adoption of specific prompting methodologies, but these recommendations either lack clarity (Pangakis et al., 2023), or show inconsistencies between proposal and practice (Ziems et al., 2024). Consequently, it is crucial to examine these studies more closely to ensure that they do not mislead future research or get perpetuated uncritically in subsequent works.

In Table 2.1, under the "Prompt" column, the "Basic" tag is used for studies that employ a very basic prompt format such as "Classify X as $Y_1, Y_2, Y_3, ...$". We labeled slightly more involved basic prompts as "Basic$_+$". These prompts either provide additional context for the task, "Given context C, classify ...", invoke a specific persona from the model "Act as R and classify ...", or offer explanations for each classification category,

Table 2.1: Papers in social sciences

Paper	Field	GPT	Temp.	Prompt	Shot	CoT
Rytting et al. (2023)	Poli. Sci.	3	?	Structured*	$2, 3$	
Chae et al. (2023)	Poli. Sci.	3	0	Basic*	$0, 1, 2$	
Reiss (2023)	Poli. Sci.	3.5	$0.25, 1$	Basic^*_+	0	
Gilardi et al. (2023)	Poli. Sci.	3.5*	$0.2, 1$	Original	0	
Zhu et al. (2023)	Psychology Poli. Sci.	3.5*	?	Basic	0	
Li et al. (2024)	Poli. Sci.	3.5*	$0, 1$	Original Basic*	0	$\sim$
Zhang et al. (2022a)	Poli. Sci.	$\widetilde{3.5^*}$	~ 0.7	Basic	0	
Aiyappa et al. (2023)	Poli. Sci.	$\widetilde{3.5^*}$	~ 0.7	Basic	0	
Kuzman et al. (2023)	Linguistics	$\widetilde{3.5^*}$	~ 0.7	Basic*	0	$\sim$
Zhong et al. (2023)	Linguistics	$\widetilde{3.5^*}$	~ 0.7	Basic*	$0, 1, 5$	✓
Amin et al. (2023)	Psychology	$\widetilde{3.5^*}$	~ 0.7	Basic	0	
Bhat et al. (2023)	Psychology Linguistics Poli. Sci.	$\widetilde{3.5^*}$	~ 0.7	Basic	0	$\sim$
Heseltine et al. (2024)	Poli. Sci.	$\widetilde{4}^*$	~ 0.7	Basic*	0	
Törnberg (2023)	Poli. Sci.	4	$0.2, 1$	Basic	0	
Pangakis et al. (2023)	Psychology Linguistics Poli. Sci.	4	0.6	Original*	0	
Savelka et al. (2023)	Law	4	0	Original & Structured*	0	✓
He et al. (2024)	Linguistics	4	$0.2, 1$	Original	0	
Rathje et al. (2023)	Psychology	$3.5^*, 4^*$	0	Basic*	$0, 1$	
Matter et al. (2024)	Sociology	$3.5^*, 4^*$	0.1	Basic*	0	$\sim$
Ziems et al. (2024)	Psychology Linguistics Poli. Sci.	$3.5^*, 4$	0	Basic_+	$0, 1$	
Our Paper	Economics	3.5^*, 4^*	0	Basic_+, Original & Structured*	$0 - 19$	✓

Notes: The "Field" column represents the broad field under which the annotation tasks can be categorized. In the "Model" column, the asterisk indicates that the turbo version of the model is used, and the tilde indicates that the model is not leveraged via the API but through the ChatGPT platform. In the "Prompt" column, "Basic" indicates a basic instructions to classify a text, "Basic_+" indicates a basic instruction accompanied by a short definition of for each category, " Original" indicates that the original human instructions are used verbatim as the prompt, and "Structured' indicates that a prompt template is used to structure the prompt into distinct components such as instructions, context, definitions, examples and so on. Moreover, in the "Prompt" column, asterisk superscript indicates that the study investigated either to improve the model's performance via restructuring or augmenting the prompt through rephrasing, incorporating additional context or definitions, making the instructions more precise, etc. or to investigate the effect of a specific variation on the prompt such as considering the prompt in an other language, instructing the model to output a non-binary classification, etc. The "Shot" column indicates the number of demonstrations used in the prompt (n-shot prompting). The "Temp." column indicates the temperature parameter(s) used for the respective model(s), question mark indicates that this value is not provided in the respective paper. Moreover, the exact temperature value for ChatGPT is not known and 0.7 the unconfirmed yet commonly assumed value for it. The "CoT" column not only indicates whether the study used some form of chain-of-thought prompting technique (✓) but also points out studies that considered asking for an explanation after the classification is done ($\sim$) either as an attempt to improve the performance or to further investigate the outputs provided.

"Classify X as $Y_1, Y_2, ...$ where $Y_1 :< description >, Y_2 :< description >, ...$ ". The "Structured" tag is assigned to prompts that imposes a structured template that organises explanations, context, constraints, demonstrations and additional prompting techniques into modular components via textual cues such as titles or delimiters. The "Original" tag is assigned to prompts that verbatim use instructions tailored for human annotators. Consequently, using this tag in conjunction with the "Structured" tag indicates that the original instructions have been reframed and restructured for its use as a prompt, and a markup language is leveraged to impose this structure. All the papers with a "Structured" tag in Table 2.1 used the `Markdown` language to structure their prompts. Lastly, the superscript "*" is used to denote studies that explored variations on their initial prompts to enhance model performance either by rephrasing, adding further information, simplifying existing descriptions or using established prompt engineering techniques such as CoT.

A major determinant of a model's task performance is the nature of the task itself, which can be considered under two main dimensions: the representational depth of task-specific categories in the model's pre-training corpus and the complexity of the task. Representational depth reflects the frequency and variety with which the categories to be annotated are represented in the training corpus. A greater representational depth ensures that the model is exposed to a wider range of conceptual diversity for given a category which, in turn, impacts the model's recognition and learning capabilities from provided demonstrations or descriptions for a given annotation task (Reynolds and McDonell, 2021; Razeghi et al., 2022; Pan et al., 2023). Zhu et al. (2023) document that GPT-3.5 performs relatively poorly when tasked with classifying topics that occurred after its training, such as the Ukraine-Russia war. In a similar vein, for GPT-4, Ziems et al. (2024) report strong performance in tasks involving categories common in everyday conversations, such as "anger" in an emotion recognition task, while in tasks requiring expert knowledge and involving non-conventional categories, such as "white grievance" in hate speech classification, GPT-4's performance is notably weaker.

From an information perspective, a task that requires a more diverse set of information is considered more complex (Liu and Li, 2012). Complexity can also be defined by the level of abstraction necessary and the extent of inferential reasoning needed to effectively interpret and act on information (Yang et al., 2016). For example, while utterance-level classification involves analysing individual statements, conversational-level classification should be considered as more complex as it requires understanding the broader context and dynamics within entire dialogues (Arad et al., 2024). Similarly, analysing court opinions to interpret legal concepts (Savelka et al., 2023) or classifying nth level strategic thinking in jury voting (see Chapter 1) involves far more complex cognitive processes than identifying promises (Charness and Dufwenberg, 2006). This increased complexity necessitates a model that not only understands expert specific information but also integrates and reasons about it in a manner that emulates higher-order cognitive processes (Huang and Chang, 2022). For instance, Bhat and Varma (2023) investigate the annotation performance of GPT-3.5 across three tasks of varying complexity and find that the model performs poorly with linguistically more challenging task of news category classification (51% average accuracy) compared to sentiment analysis (84% accuracy).

Similarly, Savelka et al. (2023) observe poor GPT-4 performance in the task of analysing court opinions to interpret legal concepts (46% average accuracy) yet this performance is found to be still on par with expert level annotators. Lastly, Ziems et al. (2024) document that as the complexity of the text increases, moving from standalone messages to conversational texts, the performance of both GPT-3.5 and 4 models deteriorate in classification tasks. In sum, the results corroborate that the performance of GPT-3.5 and GPT-4 models are shaped by both the depth of representational coverage in their training data and the complexity of the tasks they are assigned.

An additional aspect of representational depth is the linguistic diversity within the model's pre-training corpus. While specific details are often undisclosed, it is widely inferred that the primary training data for GPT-3.5 and GPT-4 consist predominantly of English texts (Lai et al., 2023). The limited representation of multilingual data can in turn cause models' performance on annotation tasks to deteriorate primarily on under-represented language families with a syntactic order in the form of Subject-Object-Verb (SOV, e.g., Hindi, Turkish, Arabic or Amharic) compared to the family of languages with syntactic order in the form of Subject-Verb-Object (SVO, e.g., German, Italian, Spanish or Slovenian) that English is a part of (Bjerva et al., 2019). This is because variations in syntactic order alter the word inter-dependencies crucial for models' language comprehension (Bender, 2011; Nivre et al., 2016). These structural differences result in unique word co-occurrence patterns and grammatical dependencies that LLMs, primarily trained on English, rely on to infer the semantics of the text. Consequently, GPT models may face greater difficulties in effectively processing the semantics in language families with syntactic structures distinct from English, while languages with a similar syntactic order, such as German, may present fewer challenges (Conneau and Lample, 2019). In practice, Heseltine and Clemm von Hohenberg (2024) find that GPT-4 demonstrates consistent performance across tasks in German, Italian, and Chilean Spanish compared to English. Conversely, Bhat and Varma (2023) observe that GPT-3.5 struggles with Indic languages, although it was not tested against English texts to ascertain if the prompts would have fared better in English. Rathje et al. (2023) report that GPT-3.5-turbo and GPT-4-turbo show comparable performance in Turkish, various African languages[5], and Arabic. Kuzman et al. (2023) document that while Slovenian texts are classified as effectively as English texts by these models, prompts in Slovenian yield poorer results compared to their English counterparts. The empirical findings largely corroborate the theoretical expectations that various GPT models perform better with languages syntactically similar to English, as observed with German, Italian, and Spanish. However, the comparable

[5] African languages display diverse syntactic structures across several language families. The Niger-Congo family mainly uses SVO order, similar to English, but variations like SOV are found in some Bantu languages. Nilo-Saharan languages typically feature SOV order, with dialectic variations (Vossen and Dimmendaal, 2020).

performance in SOV languages, like Turkish or Arabic, suggests that the models have the capability to effectively adapt to diverse linguistic structures.

The 0-shot improvement technique of providing context and label descriptions on a "Basic" prompt is shown to improve the classification performance of the models (Peskine et al., 2023). Chae and Davidson (2023) initially consider a simpler prompt than the "Basic" prompt to instruct GPT-3 to classify political stance of Twitter messages with the prompt "Stance:". This method led the model to incorrectly output labels associated with sentiment classification ("positive", "negative") rather than the intended stance labels ("support", "oppose"). Consequently, they opted for the "Basic" prompt to ensure the correct classification labels, and then tested two incremental improvements on the "Basic" prompt: first, by adding a statement indicating the potential expression of a stance, and second, by providing a general definition of what "stance" entails. These modifications improved the prompt's performance by 11%.

Similarly, Heseltine and Clemm von Hohenberg (2024) started with a "Basic" prompt and enhanced it by adding a single-sentence description for each category, which on average boosted the model's performance by 8% across different tasks.

The 0-shot improvement technique of reframing instructions (Mishra et al., 2021b), either by making existing instructions more precise or by adding descriptions to clarify category definitions, has also been documented to enhance model performance. Savelka et al. (2023) initially employed an "Original & Structured" prompt. They observed that definitions provided for the categories were too broad and adversely affecting the classification performance. By reframing these definitions to be more precise, they managed to increase the model's performance by an average of 28% across multiple prompts. Matter et al. (2024) first consider a "Basic" prompt. Their prompt has three main labels: explicit violence, implicit violence and non-violence. For violence cases they have additional three subcategories: directed to a specific person, directed to self or undirected. They first test this on a randomly selected subset of the dataset. Then, to enhance model's performance, they provide a sentence description for each label. They, then, further refined the prompt through a process of iteratively testing the refined prompt on additional randomly selected subsets. This process involved identifying misclassifications, instructing GPT to generate additional instructional text to address these errors, and then integrating these GPT suggestions into the prompt. Through these iterations, they developed a final version of the prompt that provides additional descriptions for explicit and implicit violence, and directed and undirected violence categories and an example for explicit, direct category and another example for implicit undirected violence each category. They claim the final prompt significantly improves the model's performance. However, since neither the initial "Basic" prompt nor the prompts from each iteration were tested on the full dataset, they were not able to provide a specific percentage number for the significance of each modification and for the overall effectiveness of the process compared to the initial "Basic" prompt. Furthermore, it is unclear why only specific categories are provided with additional instructions or examples. A similar prompt refinement procedure is proposed by Pangakis et al. (2023) where a randomly selected subset of data was first classified by human annotators, and then GPT was prompted to classify the same subset using identical instructions. If the model's classifications substantially diverge from the human annotations (no specific match percentage is given to define "substantial"), they identify where the model misclassifies and adjust the prompts accordingly to improve the model's performance. This process was repeated until a "satisfactory" alignment was achieved between the model's and humans' classifications (yet again, no specific percentage is given to define a "satisfactory" alignment). This method was also found to significantly enhance the model's performance and promoted as a procedure to be adopted by other researches. However, neither the precise effect of this method relative to a "Basic" prompt nor the details of the adjustments made to the prompt were reported.

Although Matter et al. (2024) and Pangakis et al. (2023) provide case studies highlighting the importance of effective prompting on model performance, their methods risk overfitting on a subsample of the messages, which could not only fail to improve but also potentially degrade classification performance on the rest of the data. Furthermore, their prompt enhancement procedures lack clear documentation. In Matter et al. (2024), one must scrutinise the final prompt to discern the prompting techniques used, and it remains unclear why certain labels are given additional descriptions or examples. In Pangakis et al. (2023), the final prompts and details of the prompt refinements are not even disclosed. As a result, neither study offers valuable insights into prompt engineering that could be adopted by other researchers.

Attempts to enhance model performance using either 0-shot-CoT or n-shot-CoT prompting methods are observed in social science applications, albeit in very few instances, despite the accessibility of the 0-shot-CoT method and detailed guidelines on its use (OpenAI, 2023b). Depending on the complexity of the task, CoT is observed to have a varying improvement on the model's performance. Zhong et al. (2023) document that 0-shot-CoT increases the performance by 15%, 1-shot-CoT increases performance by 8%, and 5-shot-CoT by 21%. Yet, they note that generating the reasoning examples for n-shot-CoT prompts were challenging. They have first constructed a hand written reasoning example for an input then instructed GPT to provide similar reasoning demonstrations for other inputs which they used as the additional four reasoning examples in their 5-shot-CoT prompt. Conversely, in Savelka et al. (2023) where the task is argued to be more complex, 0-shot-CoT is documented to worsen the performance of the initial prompt by 13% and only to improve the performance of the improved prompt by 4%. In addition to the established CoT prompting techniques, we identified various studies where the model is prompted to provide reasoning after the classification. In Table 2.1, studies employing this post-classification reasoning approach are indicated in the "CoT" column with a tilde

($\sim$). Asking the model to reason after providing the response is documented to either not improve or provide a minor improvement to the model's performance compared to a baseline of no CoT (Wei et al., 2022a). Therefore, future studies should consider avoiding this methodological oversight to avoid suboptimal model performance.

Few studies have explored n-shot prompting. Chae and Davidson (2023) experimented with various examples for 1-shot and 2-shot prompting and documented that the choice of example significantly affects model performance, with F1 scores on average varying between 42% and 73% across tasks. This variance corroborates concerns voiced in the computer science literature about the high dependency of performance on the choice of examples (Zhao et al., 2023). Rathje et al. (2023) implemented 1-shot prompting across five tasks, with performance deteriorating in two tasks and slightly improving in three, resulting in an average performance improvement of 3.5%. These outcomes emphasise the majority label bias, where the model tends to favor labels that appear more frequently in the demonstrations used in the prompt. Consequently, this bias is particularly pronounced in 1-shot prompting, as the model often replicates the classification from the single example provided, leading to either reduced performance or minimal gains (Zhao et al., 2023). Furthermore, Rytting et al. (2023) experimented with up to 30 demonstrations in few-shot prompting and noted that although performance improvements were observed, these gains plateaued after two or three demonstrations. The diminishing marginal effect of additional demonstrations after at most three examples suggests that the various tasks considered by Rytting et al. (2023) may primarily be considered as recognition tasks (Pan et al., 2023).

Although only three studies have considered few-shot prompting, six other studies on our list have engaged in batch classification, where multiple distinct texts are classified within a single request sent to the model (Zhang et al., 2022a; Amin et al., 2023; Savelka et al., 2023; He et al., 2024; Matter et al., 2024; Heseltine and Clemm von Hohenberg, 2024). Given the autoregressive nature of GPT, any classified example in a batch classification process effectively acts as a demonstration for the subsequent examples within that batch. For instance, a batch classification of $2n + 1$ messages is, on average, equivalent to n-shot prompting, with the first message evaluated under 0-shot conditions and the last message under $2n$-shot conditions. Considering that the choice, order and number of examples used in n-shot prompting significantly affect the model's performance (Lu et al., 2021; Kumar and Talukdar, 2021a; Zhao et al., 2023), and given that in a batch classification each input is classified using a different number, order, and choice of examples, the performance for each classification can vary drastically when performing classification in batches. This method has been promoted by all the stated studies for reducing costs and time for classification. Matter et al. (2024) took batch classification a step further by experimenting with different batch sizes to identify the optimal batch size that maximises the model's performance. We find it important to highlight this oversight in these studies' prompting methodology with the hope that it will be avoided in future research.

In the computer science literature, various papers that investigate prompting techniques consistently set the temperature hyperparameter of the model to 0 to maximise model consistency (Brown et al., 2020; Kojima et al., 2022; Wei et al., 2022a). OpenAI's code examples also state that for classification tasks, they set the temperature value to 0 (OpenAI, 2023a). Technically, the temperature hyperparameter, T, adjusts the softmax function commonly used in machine learning. As depicted in Equation 2.1, the softmax function normalises the raw input scores from a neural network's final layer and transforms these scores into probability values. The outputted probability values are proportional to the input values. A lower temperature value generates a probability distribution of the input scores where the input score with the highest score is given more weight. As the temperature value approaches to 0, the softmax function effectively becomes the argmax function that maps the highest input value to 1 and the other values to 0. As the temperature increases, the distribution generated by the softmax function becomes more uniform, reducing the weight on the highest scored tokens and increasing the weight on lower scored tokens. This results in the LLM becoming more likely to pick lower-scored tokens, essentially adding randomness to the token selection process.

$$\text{Softmax}(z_i) = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}} \tag{2.1}$$

Pangakis et al. (2023) set the temperature value to 0.6 and repeated the annotation task at least three times, although the exact number of repetitions for each task was not specified. They observed a positive correlation between the consistency of the classification across repetitions and the accuracy of the classification for each message within each task. More specifically, they found that a classification was 19.4% more likely to be correct if the model classified it the same way three or more times. This suggests that the model's next token probability distribution is indicative of the difficulty of the task classification for the model. This observation aligns with OpenAI's recommendation to use the probability distribution of the next token predictions as a way to measure the confidence level of the model on its next token prediction (OpenAI, 2023c).

Gilardi et al. (2023); Törnberg (2023); Li et al. (2024) and He et al. (2024) investigated the impact of using two different temperature settings on classification performance, as detailed in Table 2.1. They conducted multiple classification iterations for the same message and reported the internal consistency of classifications at each temperature setting to demonstrate the robustness of their results under temperature variations. Although not explicitly argued in these studies, the high consistency of classification results at higher temperature settings suggests that the tasks were not too challenging for the model, or in other words, the model had high confidence in its next token predictions, which was reflected in its next token probability distribution being close to degenerate (given the high consistency of the classification results even at high temperature values). Furthermore, the correlation between high confidence and high accuracy, documented by Pangakis et al. (2023), is further supported in these studies by the high accuracy of their results in conjunction with the high consistency of their results at high temperatures. All these studies documented that the model's performance was either on par with or superior to that of either online or expert annotators. However, these insights were lacking in these studies as the goal of Gilardi et al. (2023); Törnberg (2023); Li et al. (2024); He et al. (2024) in comparing the consistency of two sets of temperature values was only to argue that at lower temperatures, the model is more consistent and recommended lower temperature to be considered in future studies.

A more cost-efficient way of investigating this is to utilise the "logprob" functionality of GPT models that became available via the API right before Christmas 2023. By ensuring that the model outputs a single token as a classification output, the "logprob" functionality can be used to obtain the probability distribution for each classification label (OpenAI, 2023c). This functionality not only provides probabilities associated with each class prediction but also allows users to set their own confidence thresholds for the classifications (OpenAI, 2023c). Although this functionality was not incorporated into our current study due to its irrelevance for our research questions, it is important to note that it presents a viable alternative for running multiple classifications to approximate the model's classification distribution and to assess the confidence level of the model's classification for each message.

In Table 2.1, studies that used ChatGPT instead of directly interacting with the models via the OpenAI API are denoted under the "GPT" column with a tilde. There are several issues associated with using ChatGPT for research purposes. First, the temperature setting of the underlying model is not disclosed by OpenAI, and it cannot be altered by users. The temperature of models used in ChatGPT is commonly assumed to be 0.7; however, as previously stated, for task classification, the recommended temperature value is 0 to achieve robust results. Second, ChatGPT employs a pre-defined system prompt that precedes every conversation on the platform (see Section 2.4.1 for details), which can confound any prompts considered and, in turn, undermine the robustness and replicability of results [6]. Third, there is a limit to the number of requests to ChatGPT. Although this varies, the typical limit set by OpenAI allows only about 40 requests every three hours which equates to a maximum of 320 classifications per day. In contrast, when using the OpenAI API, our experience shows that depending on how involved the task is, one can classify 100 messages in as little as 1.5 minutes and up to 43 minutes (see Section 2.9 for more details). Moreover, just like the temperature hyperparameter, it is unclear which specific GPT model is the underlying model used for ChatGPT or ChatGPTplus, which in

[6]The only way to get rid of the system prompt is to build a GPT agent where the platform allows the users to defined their own system prompts. Yet, none of the studies in Table 2.1 considered this option.

turn further undermines the robustness and replicability of results (Aiyappa et al., 2023). Lastly, any classification task conducted in the platform has the risk of data leakage, as a set of messages that are used for a classification via ChatGPT has the possibility of becoming part of the training data for the next iteration of the model used for the platform (Aiyappa et al., 2023). In sum, we strongly recommend the researchers to not use ChatGPT due to the closed nature of its underlying GPT models and due to the risk of data leakage.[7]

Reiss (2023), unlike other studies in our list, solely focuses on the consistency of GPT-3.5. It is frequently cited in social science literature for its recommendation against using GPT for text classification, highlighting the model's unreliability due to documented output inconsistencies. The author claims that the model's classification is inconsistent, and therefore, the model is unreliable in two dimensions. First, the author compares the classification results for temperature values of 0.25 and 1 and shows that the classification results are not consistent between these temperature values, as for a single repetition of each message, the Krippendorf's alpha is below 0.8 (0.71). Yet when he repeats the classification three and ten times, the classifications results becomes consistent with a Krippendor's alpha above 0.8, reaching 0.91 for ten repetitions. The author's fundamental misunderstanding lies in his assumption that the output distribution of tokens varies with changes in temperature. However, in reality, the model's generated token distribution is independent of the temperature hyperparameter (hence the prefix "hyper"). In other words, the model produces a similar token distribution across classifications of the same instance, regardless of the temperature value used. When the temperature is high, the model tends to select less likely tokens more frequently, which in turn results in variation in the outputted token (which corresponds to the classification provided by the model). Therefore, what is perceived as inconsistency is merely a characteristic of the model's functionality, which can be mitigated by setting the temperature to 0, as recommended by OpenAI (OpenAI, 2023a).

Second, the author compares the classification results for 10 different prompt variations where he describes the differences in instructions between these prompts as "minor". Firstly, it is unclear whether the author pools the classification results from both temperature settings for this analysis. Assuming that the results are not pooled by temperature and that only the lower temperature value of 0.25 is considered, he still misleads the reader with his characterisation of the variations in the 10 prompts he compares as "minor". A closer examination of the 10 prompts under consideration reveals that the first prompt is the original human instruction prompt, which is significantly longer than the others and

[7] Additionally, researchers should be cautious not to use "ChatGPT" to refer generically to any GPT model they use in their research, as this is analogous to calling an Intel processor a Dell computer simply because it is used within a Dell product. The classification tasks are performed by the underlying GPT model leveraged by the ChatGPT platform.

is written mostly in German with some parts in English. This original prompt provides considerably more information about the labels “news” and “not news”. Previous studies have documented that prompting in a language other than English significantly affects the performance of the model (Kuzman et al., 2023). Moreover, providing additional definitions in the prompt is expected to effect the classification results of the model (Chae and Davidson, 2023; Peskine et al., 2023). Therefore, given that his subsequent prompts are in English and do not provide additional descriptions of the categories, it is not surprising that the results from this first prompt differ form the classification results of the other prompts. His second and third prompts are “Basic” prompts that indeed involve only minor changes. His fourth and fifth prompts also exhibit minor alterations; they maintain the original semantics of the prompt while adding additional emphasis on how to label categories.

On the other hand, prompts 6 and 7 employ the prompt engineering technique of invoking a persona on the model. In prompt 6, the model is instructed to “take a human perspective”, and in prompt 7, it is instructed to act as “a research assistant in a scientific project”. It has been documented that invoking a persona on the model significantly changes the model’s performance and, consequently, the classification results Kong et al. (2023); Salewski et al. (2024). Therefore, the fact that these prompts generate classifications that differ from the other prompts should be expected. In prompt 8, the model is instructed to base its decision on the article “What is News? News values revisited (again)” by Tony Harcup and Deirdre O’Neill. This represents a significant deviation from the other prompts. Moreover, it is unclear how the model is influenced by being instructed to use information from an article, as this approach has not traditionally been recognised as a prompting technique. However, it is expected to significantly affect how the text is classified, and therefore, it should not be surprising that the classification results differ.

Lastly, in prompts 9 and 10, a weaker definition for the categories is used. For instance, instead of the direct instruction “if the text is news classify it as 1”, the prompts state “1 means all or most in the text is news”. Such a variation in the prompt can potentially alter the classification outcomes even for human annotators. Therefore, the fact that the model provides a different set of classifications when the classification category is presented in a weaker form suggests that the model can discern semantic nuances in the instructions and closely follows them; and this capability should not be considered as evidence of inconsistency in the model’s performance.

In brief, what the author describes as “minor” variations in the prompts are in fact significant changes, which naturally lead to different classification results. Therefore, his arguments concerning variations in temperature and prompt design do not substantiate the claim that GPT is unreliable, and his recommendation against using GPT for text classification is unwarranted. More importantly, it is imperative that researchers take the time to thoroughly investigate the claims of a study by examining the prompts used to ensure the validity of its claims.

A similar mistake is made by Savelka et al. (2023). They compare the classification results of a prompt with 0-shot-CoT and without 0-shot-CoT, in both single and batch classifications. Given that batch classification is effectively akin to n-shot prompting, the authors inappropriately compare results from established prompting methods like 0-shot-CoT and n-shot prompting to a prompt without these techniques, to argue that GPT classification is not robust to "minor" prompt changes. However, the modifications to the prompt are substantial enough to expect changes in the model's classification outcomes, and thus should not be cited as evidence of the model's prompt "brittleness" (Kaddour et al., 2023).

Lastly, we would like to address a major issue we observed with the prompts considered in Ziems et al. (2024). While their study offers valuable guidelines for effectively conducting 0-shot prompting, a closer examination of their various prompts[8] reveals several inconsistencies and issues. Despite their claims of using 0-shot prompts, we identified that two of their prompts inadvertently provide examples for each label, effectively making them 1-shot. Additionally, while some of their prompts are "Basic", others include additional descriptions for each label. We also discovered that three prompts employed the technique of invoking a persona. Moreover, although they arbitrarily used additional explanations in some prompts and additional demonstrations in others, in one prompt, they instructed the model to categorise labels "based on formal workplace social norms". "Social norm" is a term that is too broad and varies significantly across cultures. Consecutively, the models' performance would have benefited significantly from a more detailed description of what these social norms entailed, yet they arbitrarily decided not to provide any. These inconsistencies across tasks are noteworthy because they compare the model's performance across tasks without controlling for the prompt techniques used. Furthermore, while some prompts include descriptions, others lack any explanatory detail, and no efforts are made to standardise or improve these descriptions across different tasks. Yet, they boldly claim that based on their results, LLMs should not be used for annotation tasks. We believe that to make such bold claims, one must first ensure that their prompts are optimised to maximize the LLMs' performance to the fullest extent possible. Without such rigorous optimisation, their recommendation against using LLMs for text classification seems rather unwarranted.

[8]It was a challenge to access their prompts. They did not provide a supplementary online appendix where they clearly displayed the various prompts they have used. We took the effort to search through their code to find the prompts that they have used.

2.2.3 Text analysis in Economics

In economics, the applications of text analysis include the evaluation of policy platforms, understanding news impact on stock prices, central bank communication influence on financial markets, media slant and more (Gentzkow et al., 2019). In Table 2.2, we provide a representative list of the papers that used GPT, which are so far confined to the areas of central bank communication, financial markets (sentiment analysis of firm specific news) and corporate finance (analysis of conference call transcripts of firms).

Multiple studies (Hansen and Kazinnik, 2023; Alonso-Robisco and Carbó, 2023; Lopez-Lira and Tang, 2023; Jha et al., 2024) in Table 2.2 have documented that GPT models outperform existing text classification techniques such as BERT[9] (or its variants) or dictionary-based methods (Penczynski, 2019; Hüning et al., 2022a). On the other hand, in classifying Central Bank communication transcripts, GPT models are shown to perform poorly compared to expert annotators in classifying Central Bank communication transcripts in all (Hansen and Kazinnik, 2023; Smales, 2023; Alonso-Robisco and Carbó, 2023; Peskoff et al., 2023) but one study (Fanta and Horvath, 2024).

In finance, the effectiveness of these models is typically assessed based on their ability to predict investment returns or the value of companies over a set period, rather than by comparing the classification results to a ground truth established by human annotators. For instance, Lopez-Lira and Tang (2023) and Glasserman and Lin (2023) implemented basic investment strategies, where stocks are bought or sold based on the news sentiment classified by the model the day before the transaction. This approach resulted in cumulative returns of 550% (Lopez-Lira and Tang, 2023) and 350% (Glasserman and Lin, 2023) over a two-month period.

Almost all studies in Table 2.2, used either a "Basic" or "Basic$_+$" prompt. Majority of the studies related to finance have additionally leveraged invoking a persona of a "Financial Expert" (Jha et al., 2024; Lopez-Lira and Tang, 2023; Glasserman and Lin, 2023). Differently from any other studies we have reviewed, Obaid and Pukthuanthong (2024) used as a prompt a set of 14 survey type questions, each requiring a likert-scale response that is traditionally used for human subjects. Hence, although their prompt is not borrowed from an existing human instruction, we classified their prompts as "Original".

[9] BERT (Bidirectional Encoder Representations from Transformers) and its derivative models, such as finBERT, sBERT, roBERTa, are transfomer-based language models that are relatively "small" (Devlin et al., 2018; Liu et al., 2019; Huang et al., 2023) with a parameter size of 110 million for the base model and 355 million parameters for its variants (see Section 2.2.1 for comparison to GPT models). Unlike GPT models, or any other unidirectional LLMs such as Gemini or Claude, BERT cannot simply be inputted with instructions that are potentially accompanied with detailed descriptions of categories and annotation demonstrations, and be expected to either recognise or learn from these, nor can it provide reasoning in a similar fashion to GPT-3.5 and GPT-4. Furthermore, BERT and its variants have a relatively small token limit of 512 (Devlin et al., 2018; Liu et al., 2019; Huang et al., 2023), compared to token limits of 4096 for GPT-3.5-turbo, 8192 for GPT-4-turbo, and 32768 for GPT-4-32k. On the other hand, because BERT is a significantly smaller language model, it is feasible to run BERT in a local system or to fine-tune it using a training dataset at a comparatively lower cost.

Table 2.2: Papers in economics

Paper	Field	GPT	Temp.	Prompt	Shot	CoT
Alonso-Robisco and Carbó (2023)	Macroecon.	$\widetilde{3.5^*}$	~ 0.7	Basic	0	$\sim$
Smales (2023)	Macroecon.	$\widetilde{3.5^*}, \widetilde{4^*}$	~ 0.7	Basic	0	
Hansen and Kazinnik (2023)	Macroecon.	$3, 4$	?	Basic	0	
Peskoff et al. (2023)	Macroecon.	4	?	Basic_+ & Structured	$0, 10$	
Fanta and Horvath (2024)	Macroecon.	$\widetilde{3.5^*}, \widetilde{4^*}$	~ 0.7	Basic	$0, 1$	
Glasserman and Lin (2023)	Finance	3.5	0	Basic_+	0	
Kim et al. (2023)	Finance	3.5^*	0	Basic_+	0	
Lopez-Lira and Tang (2023)	Finance	$3.5^*, 4$	0	Basic_+	0	
Jha et al. (2024)	Finance	$\widetilde{3.5^*}$	~ 0.7	Basic	0	$\sim$
Obaid and Pukthuan-thong (2024)	Finance	4	?	Original	0	
Our Paper	Exp. Econ.	$3.5^*, 4^*$	0	Basic_+, Original & Structured*	$0 - 19$	✓

Notes: The "Field" column represents the subfield of economics under which the annotation tasks can be categorized. In the "Model" column, the asterisk indicates that the turbo version of the model is used, and the tilde indicates that the model is not leveraged via the API but through the ChatGPT platform. In the "Prompt" column, "Basic" indicates a basic instructions to classify a text, "Basic_+" indicates a basic instruction accompanied by a short definition of for each category, " Original" indicates that the original human instructions are used verbatim as the prompt, and "Structured' indicates that a prompt template is used to structure the prompt into distinct components such as instructions, context, definitions, examples and so on. Moreover, in the "Prompt" column, asterisk superscript indicates that the study investigated either to improve the model's performance via restructuring or augmenting the prompt through rephrasing, incorporating additional context or definitions, making the instructions more precise, etc. or to investigate the effect of a specific variation on the prompt such as considering the prompt in an other language, instructing the model to output a non-binary classification, etc. The "Shot" column indicates the number of demonstrations used in the prompt (n-shot prompting). The "Temp." column indicates the temperature parameter(s) used for the respective model(s), question mark indicates that this value is not provided in the respective paper. Moreover, the exact temperature value for ChatGPT is not known and 0.7 the unconfirmed yet commonly assumed value for it. The "CoT" column not only indicates whether the study used some form of chain-of-thought prompting technique (✓) but also points out studies that considered asking for an explanation after the classification is done ($\sim$) either as an attempt to improve the performance or to further investigate the outputs provided.

Differently from our model and previous studies discussed in Section 2.2.2, Peskoff et al. (2023) imposed a format structure upon their prompt through `XML` tags. Moreover, only two studies consider few-shot prompting technique and Fanta and Horvath (2024) document that 1-shot prompting technique did not provide any improvement on the model's performance which is most likely due to their attempt to do classifications in batches. Lastly, just as with almost all the studies reviewed in Section 2.2.2, none of the studies in Table 2.2 considered either 0-shot or n-shot CoT prompting technique. Yet, few considered to leverage models' reasoning capabilities for non-performance related inquiries (such as getting a more detailed understanding of the classification made by the model) by instructing the model to provide a reasoning after it provided its classification.

Similar issues with prompting methodologies, albeit minor, are also observed within these studies. Few studies have failed to disclose the temperature hyperparameter they have used for their models. One study have only provided a brief description of their prompt but did not disclose it (Kim et al., 2023). And few others have used ChatGPT platform for their classification rather than directly accessing the GPT models through the OpenAI API. In addition, one puzzling prompting technique we observed with two studies (Lopez-Lira and Tang, 2023; Glasserman and Lin, 2023) is to begin their prompts with the statement: "Forget all your previous instructions". This is the most basic prompt injecting method to "jailbreak" a model from its pre-defined system prompt which serves to prevent the user from leveraging the model to generate harmful content (Shen et al., 2023). However, neither of the studies that used this statement did their classification via the ChatGPT platform, hence there was no need to attempt to overwrite a system prompt. Moreover, this specific prompt injection phrase is commonly known and most likely already accounted for by the companies that provide the LLM services (Anthropic, 2023). Therefore, even if it was used as intended, it would not have worked, and would have potentially resulted in their accounts to be flagged.

In experimental economics, the analysis of text has increased with the augmentation of experimental action data with choice-process data (Cooper et al., 2019). Starting with a prominent exploration of strategicness in games by means of team chat (Cooper and Kagel 2005), further investigations have used intra-team communication (Burchardi and Penczynski, 2014), talk-aloud protocols (Capra, 2019) and written advice (Schotter, 2003a). Naturally-occurring language has also been analysed to understand, for example, cooperative behaviour in large stake game shows (Van den Assem et al., 2012).

In an earlier attempt to computerise text classifications, Penczynski (2019) describes the effectiveness of supervised machine learning techniques in classifying intra-team communication in various games according to the level of strategic sophistication. More recently, Hüning et al. (2022a) and Hüning et al. (2022b) consider both "traditional" dictionary-based methods and BERT for classification of "premises", and documented that BERT performs as good as dictionary-based methods. Howevever, while their automated classification results show very good performance (87% match with human classification), the model's performance heavily relies on the size of the training data, and deteriorates as the size of the training data decreases, or as the concepts to classified become more nuanced[10] (Hüning et al., 2022a). Notably, Hüning et al. (2022a) state that effec-

[10]Hüning et al. (2022a) demonstrate the difficulty of classifying nuanced text with the following pair of messages: "Rent control will lead to fixed and projectable prices for renters." and "Rent control will lead to fixed prices that cannot fluctuate anymore.". GPT-4-turbo successfully identifies the nuance between these two messages, and classifies them correctly using the following "Basic$_+$" prompt:

```
- Classify whether the following message is against or for rent control.
- Provide a step-by-step reasoning before providing your classification.
- Code 'for' as 1 and 'against' as 0.
- Refrain from providing any classification other than 'for' or 'against'.
```

tive performance with automated classification using BERT or dictionary-based methods requires "a few hundred training data per classification category". Unlike these methods, the use of GPT in this study obviates the need for "supervision" – the training of a model with substantial appropriate data.

2.3 Research questions

The following research questions guide our analysis of GPT classification performance.

RQ 1: **[Prompts]** Can classification instructions intended for human annotators be minimally modified into prompts that deliver GPT performance levels comparable to those of human annotators?

- **1.a:** **[Promises]** How does GPT's performance in classifying "promises" compare to that of expert-level human annotators and of the aggregated classifications of groups of human annotators?
- **1.b:** **[Level of reasoning]** How does GPT's performance in classifying levels of strategic thinking, as well as label and payoff salience, compare to that of expert-level human annotators and of traditional machine learning methods?

RQ 2: **[n-shot and CoT]** How effective are n-shot and 0-shot-CoT prompting techniques in classifying "promises" and various concepts related to strategic thinking?

RQ 3: **[GPT-3.5 vs. GPT-4]** How does the size of the model influence performance in classifying "promises" and various concepts associated with strategic thinking?

2.4 Procedures

2.4.1 General Prompt Structure

In order to investigate whether GPT can be considered as a viable alternative to human annotators, it is essential to ensure the observed performance is not compromised by suboptimal prompt design choices. Research has shown that using instructions tailored for human annotators directly as prompts leads to significantly poor GPT performance (Efrat and Levy, 2020). On the other hand, reframing these human-tailored instructions into cross-task generalizable prompt templates has been shown to substantially improve GPT's performance across a variety of tasks (Mishra et al., 2021b). While our objective is not to identify the ultimate prompt design, we are nevertheless dedicated to optimising our prompts. By doing so, we aim to ensure that if GPT's performance falls short, it

`- Follow the format: \n Reasoning: \n ... \n Classification: 0/1.`

is more likely a reflection of its own limitations rather than the result of our potentially suboptimal prompt design.

Recall that the tasks we examine are grouped into two distinct classification concepts: "promise" and "strategic thinking". Within each of these groups, the tasks exhibit differences in complexity and context, leading to natural variations in their instruction design and structure. Our interest lies in examining how the model's performance adjusts as the complexity within each task varies. However, since our prompts incorporate certain crucial parts of the human instructions verbatim, there is a significant variation in wording and, as a result, in the style of the prompts, especially noticeable between the two "strategic thinking" classification instructions and, to a much lesser extent, between the two "promise" classification instructions. Given the documented impact of word choice on the model's performance (Yuan et al., 2021; Haviv et al., 2021; Jiang et al., 2020), the inherent potential for variability in the effectiveness of our instructions that a prompt template cannot fully address remains. Nevertheless, it has also been established that even minor variations in a prompt, such as spacing between statements or the choice of separators among arguments, can affect an LLM's performance (Sclar et al., 2023). Therefore, employing a general prompt template allowed us to at least mitigate variations stemming from structural and formatting differences within the prompts of the two classification tasks. In brief, our choice to use a general prompt template was also driven by the goal of imposing a degree of control and consistency in the structure and format to the classification instructions. This choice, in turn, enabled a more robust investigation and comparison of the model's performance across tasks that vary in complexity and context.

In brief, to understand how task-specific context and complexity influence GPT's performance and to assess its potential as an alternative for human annotators, adopting a general prompt template was deemed essential. This strategy reduced the variability caused by differing instructions and enhanced our ability to isolate and evaluate GPT's true performance consistently across tasks. Hence, following the guidelines (Zhao et al., 2023; Ziems et al., 2024), recommendations (White et al., 2023), and investigations (Mishra et al., 2021b; Clavié et al., 2023; Yuan et al., 2023; Chae and Davidson, 2023; Savelka et al., 2023) for effective prompt design, we developed and utilised the prompt template depicted in Figure 2.1.

All four human-tailored instructions that we used as a basis to construct our prompts consisted of two consecutive parts: a first part providing the background information on the experiment, followed by a second part detailing how the human annotator should classify each message. The background information consisted of, to a varying degree, a detailed explanation of various components of the experiment: the decision process of the subjects, the payoff structure of the game, the communication protocol, and a theoretical background for the game played. The 'Context' section of our prompt template served to provide all these background information, in line with previous studies that have shown

Figure 2.1: General prompt structure

 # General Task
- Classify <X> in <E>
 # Role Persona
- Act as a behavioral economist specialized in text classification, concept <C> and game <G>
 # Context
- <Game mechanics>
- <Experimental design/Decision Process>
- <Communication protocol>
- <Theory>
- ...
 # Classification Task
- Classify <X> as <Y> given conditions <Y_1, Y_2, ...>
- ...
 # Classification Coding
- Code <X> as <Z> if it is classified as <Y>
- ...
 # Examples (only used in n-shot prompts)
- <Example text> <classification>
- ...
 # Classification Process (only used in CoT prompts)
- Provide a step-by-step reasoning before providing your classifications.
 # Constraint(s)
- Follow the below output format.
- ...
 # Output Format
<Desired output format>

incorporating additional background details into prompts positively impacts GPT's performance (Chae and Davidson, 2023; Savelka et al., 2023; Yuan et al., 2023). However, rather than incorporating these information into the 'Context' section verbatim, we opted to include only the most crucial information deemed necessary for GPT to properly infer the context of the message to be classified (further details will be provided for each game in Sections 2.5-2.8). Given that background information primarily serves to provide domain specific linguistic patterns to facilitate the model to better interpret the context of the message (White et al., 2023), we conjectured that an effective summary of the background information is sufficient enough as long as this summary manages to maintain and present the key words and phrases that encompasses these patterns. Although the literature presents mixed outcomes regarding the effectiveness of presenting information in a more succinct manner –as it has been documented to either improve a prompt's performance (Beltagy et al., 2020; Kuznia et al., 2022) or have no significant effect (Mishra et al., 2021b; Li and Qiu, 2023b) depending on the model deployed– it, at the very least, served to significantly reduce the cost of our classifications by minimising the number of inputted tokens.

In line with our objective to reduce the effort needed to utilise GPT for classification tasks, we opted to incorporate the second part of the human-tailored instructions, which specifically detail the classification task, verbatim into the "Classification Task" section of our prompt template. This approach not only allowed us to investigate the possibility of using GPT for classification tasks with minimal effort but also provided us with the opportunity to assess if instructions designed for human annotators are effective enough to elicit high-level performance from GPT. Peskine et al. (2023) compared the effectiveness of label descriptions to having no descriptions and also evaluated the performance impact of descriptions provided by experts versus those generated by GPT. They documented that both types of label descriptions significantly enhanced the model's performance, with descriptions from experts leading to even greater improvements. Moreover, research by Mishra et al. (2021a) indicates that task descriptions crafted by experts generally surpass the effectiveness of basic instructions commonly found in NLP literature, such as those presented by Bach et al. (2022) in PromptSource, e.g., "Classify whether the following message constitutes a promise or not." Additionally, Logan IV et al. (2021) have shown that expert-crafted prompts from Schick and Schütze (2020) typically outperform automatically generated (soft) prompts[11]. Hence, based on these findings, we argue that it is ideal to use existing classification instructions prepared by experts and to restructure them in a manner that is more easily processed by the model, following the guidelines established by Mishra et al. (2021b).

The "Example" section was designated to separately provide examples provided in the original instructions. However, in instances where the original instructions lacked examples, this section was omitted from the prompt. Additionally, there were scenarios where instructions on how to classify messages were interwoven with examples –forming a pattern of instructions followed by supporting examples, then more instructions, and so forth. In these situations, due to our commitment to use the classification instructions from the original instructions verbatim, and considering that extracting examples from their instructional context could compromise the coherence of the instructions, we chose not to isolate these examples into a distinct "Example" section. Consequently, in such cases, an independent "Example" section was also omitted. Moreover, there were instances in the original codebook where examples were provided separately from the classification instructions, yet each example or set of examples was accompanied by additional remarks. In these cases, we opted to create a separate "Example" section while preserving the structure of each example followed by its remark. This approach was taken with the aim of staying as close as possible to the original instructions to minimise the effort needed to restructure and reframe the codebooks into prompts.

[11] Soft prompts, in contrast to discrete, human-readable prompts, are vector-like, non-textual parameters fine-tuned to steer the outputs of language models. They constitute an optimised set of tokens (words or subwords) designed to influence a pre-trained language model's output for specific tasks, facilitating task-specific adjustments without modifying the core model (Refer to Li and Liang, 2021; Lester et al., 2021, for additional information). Though this comparison primarily involves very basic single-sentence prompts and soft prompts, it underscores the efficacy of expert-crafted prompts.

Apart from the “Context”, “Classification Task”, and “Examples” sections, the other sections in our prompt template were not directly derived from the original instruction text. These additional sections were included based on recommendations found in the literature regarding optimal prompt design (Reynolds and McDonell, 2021; Mishra et al., 2021b; White et al., 2023).

The first section, “General Task”, serves as a direct task specification (Reynolds and McDonell, 2021), that serves to summarise the task broadly by incorporating key terms like “classify”, “message”, “promise”, or “strategic thinking”, without specifying how to accomplish the task. The efficacy of this section in enhancing GPT’s performance is predicated on the assumption that the model has already acquired an understanding of these fundamental concepts during its pre-training phase (recognition task). Therefore, by offering a high-level task description that includes these keywords, it is posited that GPT is more aptly primed to produce the intended output (Mishra et al., 2021a; White et al., 2023).

The “Role Persona” section acts as an augmented task specification that employs memetic proxy concepts to deepen the task description (Reynolds and McDonell, 2021). This section seeks to subtly expand upon key task concepts like “classify”, “message”, “promise”, and “strategic thinking” by placing them within the context of “a behavioural economist”. This method enhances the model’s contextual understanding of the task and aligns its operation with the persona’s style of reasoning. Notably, research has shown that directing the model to emulate a specific persona can elevate its performance similarly to the impact observed with CoT prompting (Kong et al., 2023). Hence, differently from the “General Task” section, this section implicitly instructs the model on *how* to perform the task, by drawing on the model’s pre-existing knowledge of such roles. However, it’s worth noting that the specific traits of the persona adopted by the model are unclear. This is primarily because there is no detailed knowledge about the specific data on which GPT has been trained no. Hence, while the role persona technique has been effective in elevating GPT’s performance (Kong et al., 2023), there’s a risk it may highlight biases from its training dataset (Salewski et al., 2024).[12] However, in our context, we do not foresee any biases associated with assuming a behavioural economist persona negatively impacting GPT’s task performance.

The “Constraint”, “Output Format”, and “Classification Coding” sections collectively shape GPT’s output generation, each serving a complementary role in guiding the model towards producing outputs in a specified format. The “Constraint” section ensures GPT adheres to the particular format outlined in the “Output Format” section, which specifies the exact formatting requirements for the model’s outputs. Together, these sections are pivotal in achieving consistently formatted outputs and facilitate the extraction of

[12]For example, Salewski et al. (2024) observed that GPT-3.5’s ability to classify car models improves when prompted to assume a male persona over a female one.

classification outcomes using basic string pattern matching algorithms. Moreover, when specific output criteria are necessary beyond the conventional format, the "Constraint" section introduces additional directives to meet these tailored requirements to guarantee that outputs precisely match the classification task's needs[13]. Similarly, the "Classification Coding" section, akin to both "Constraint" and "Output Format", furthers this objective by instructing the model on how to encode various label categories in its output[14].

The "Classification Process" section was devised strictly to employ the 0-shot-CoT prompting method, and was incorporated into our prompts only when we explored this prompting technique's effect on the classification performance of GPT. Our use of the 0-shot-CoT process diverges from the methodology presented in the foundational paper by Kojima et al. (2022). In Kojima et al. (2022), the technique involves appending "Let's first think step-by-step" to the classification prompt, explicitly guiding the model to begin with reasoning before tackling the assigned task. We've chosen an alternative strategy that better fits our existing prompt template by instructing the model to "provide a step-by-step reasoning before providing a classification" under the "Classification Process" section. Furthermore, in order to ensure that this order is strictly followed by the model, when CoT is considered in the prompt, the "Output Format" explicitly outlines that GPT should structure its response by initially presenting a reasoning section, subsequently followed by the classification in a dedicated section as depicted in Figure 2.2.

Figure 2.2: Output format section for CoT

```
# Output Format
  ## Reasoning
  ...
  ## Classification
  <Desired output format>
```

Segmenting a lengthy set of instructions into a list format is argued to enhance the model's comprehension and response accuracy, and is documented to improve GPT's performance (Mishra et al., 2021b). Hence in order to optimise GPT's classification performance, we adopted this reframing technique by converting the original instructions into sequences of semantically coherent statements. Each itemised statement was no longer than two sentences long, preserved the original statements word for word, and ordered in a way that stayed faithful to the original order of the instructions. The list format was also applied to all newly created statements or instructions.

We opted to use `Markdown` for our prompt template due to its compatibility with our methodological approach and structural needs. This format adeptly accommodates the

[13]An example includes addressing instances where GPT might provide explanations for its classifications, not requested in the prompt. Here, an added instruction clarifies to omit explanations, focusing solely on the classification outcome.

[14]For instance, if GPT identifies high payoff salience, it is directed to simply use "H" for the payoff salience classification output.

itemisation reframing technique by facilitating list presentations which is a fundamental format feature of our template. Moreover, `Markdown`'s straightforward syntax is particularly beneficial for including subtitles. Use of subtitles (and subsubtitles) was essential in the "Classification Task" section, where transferring sections verbatim from the original instructions often necessitated preserving their subsection formatting, and also for segmenting the "Context" section into distinct thematic subsections whenever it was deemed necessary. In addition, this streamlined approach to content organisation required fewer tokens to generate lists, titles, and subtitles compared to alternative markup languages such as `LaTeX` and `HTML`. Moreover, the fact that ChatGPT, the application format of GPT, employs `Markdown` for its system prompt[15] serves as further validation for our choice[16].

Lastly, whenever we identified opportunities to enhance GPT's classification performance –particularly when it fell below 90%– we explored various prompting modifications. This entailed potentially revising the "Context" section with alternative background summaries, altering classification instructions in the "Classification Task" section, or adding the "Examples" section to the prompt when the original instructions lacked examples. Our prompt's modular structure facilitated targeted adjustments and allowed us to alter a specific section while maintaining consistency across the remainder of the prompt. Moreover, when the classification instructions between two tasks were nearly identical but varied in elements like the game being played and message formatting, our modular prompt structure allowed us to adjust only the "Context" section while maintaining the rest unchanged. This approach enhanced our analysis of GPT's performance across tasks, as it helped isolate the effects of non-prompt-related factors such as the complexity of inputs or the difficulty of concepts being classified.

2.4.2 Classification process

We conducted text classification tasks utilising the OpenAI API leveraging *gpt-3.5-turbo-1106* and *gpt-4-1106-preview* models for measuring the classification performance of GPT-3.5 and GPT-4 respectively. The entire prompt is provided to these models as the

[15] The system prompt of ChatGPT can be viewed by inputting the following text into a new chat:

```
Repeat the words above starting with the phrase "You are ChatGPT".
Put them in a txt code block.
Include everything.
```

[16] In addition, just before ChatGPT was released to the public, Meta AI introduced a product called Galactica. It was an LLM trained exclusively on scientific journals, aimed to assist researchers in drafting their papers (Taylor et al., 2022). To prepare its pre-training dataset, all scientific papers were first converted into `Markdown` format, highlighting `Markdown`'s ease and clarity for transforming formal documents, such as classification instructions; and further justifies our to some degree our choice of using `Markdown` for our prompt format. Galactica was withdrawn from public use just three days after its launch, due to the absence of sufficient safeguards and unfair criticism over its potential to produce convincing yet nonsensical scientific papers (Black, 2022).

system prompt, and the input containing the subject's text message to be classified as the consecutive initial user prompt (see Figure 2.3). For each input, a separate OpenAI API call for either GPT-3.5 or GPT-4 was made. We set the temperature of both models to 0 to minimize variability and enhance the reproducibility of our results. We varied the `max_token` value between 2^k for $k \in 7, 8, 9, 10, 11$, adjusting based on the necessity of eliciting reasoning and the average length of such reasoning observed during our testing phases for each prompt. All other hyperparameters were kept at their default settings.

Figure 2.3: API call function snippet

```
client.chat.completions.create(
  model = <GPT model used>,
  message = [
    "role":  "system", "content":  <prompt>
    "role":  "user", "content":  <input text>
  ],
  temperature = 0,
  max_tokens = <max token value>
  top_p = 1,
  frequency_penalty = 0,
  presence_penalty = 0
)
```

A Python script was developed to automate the API calls for each input. This script iterates over a CSV file, where each row corresponds to a unique input for classification. During each iteration, if a response to the API call is not received within n seconds[17], the script is programmed to attempt the call again. This retry process may repeat up to five times. Should the response still not be received after these attempts or the response is received and recorded, the script then moves on to the next input in the sequence after waiting for a grace period of 2 seconds.[18] This protocol has been established in response to the observed phenomenon where the API halts and does not respond to new requests for at least 60 seconds, typically after making 10 to 15 requests in quick succession. This undesirable behaviour was predominantly encountered during the initial phases of our research, when our OpenAI account was categorized as a low-tier account and was thus subject to various API call limitations.[19]

Following the response from the API, the model output is stored in a separate CSV file along with its corresponding message ID. This approach of recording the output immediately after each response not only provides the possibility to pause the classification

[17]The variable n is determined based on the classification task at hand. An average classification time per message is established, to which an extra margin of 10 to 30 seconds is added, thus defining n.

[18]The script's actions, including each API call attempt, were logged to the console for monitoring. In instances where the API failed to respond within the predefined time frame across four attempts, the process was temporarily paused and then resumed after a minute of waiting. This precautionary measure ensured that no messages were skipped without a classification within a single run.

[19]For tier lists and their respective rate limits see `https://platform.openai.com/docs/guides/rate-limits`

process as needed but also provides a robust mechanism for addressing any unexpected disruptions due to technical issues on either the client or server side.

In the event of resuming an incomplete classification process, the script first verifies the existence of the classification output CSV file. If this file is found, the script then examines the last classified message ID recorded in the file. Utilising this information, it identifies the next message to be classified from the input CSV, aligning with the sequence of messages. Following this verification, the script continues with the iterative process as previously described, ensuring a seamless continuation from the point of interruption. This procedure guarantees an efficient resumption of the classification task, eliminating the duplication of efforts on messages that have already been classified. Additionally, this eliminates the necessity for manually identifying the next message to be classified when resuming an interrupted classification, a process that is prone to human error.

The status of the OpenAI service is monitored both prior to and during the classification process. Should there be any reported incidents affecting the service, the classification task is either not initiated or immediately halted. Furthermore, if an incident is reported during the classification process, or within the subsequent 12 hours involving the specific model used, any classifications conducted in that time-frame are invalidated. A fresh series of classifications for the same dataset using the same model is scheduled to commence 24 hours after the incident has been reported as resolved. This protocol ensures the integrity and reliability of the classification results by accounting for potential disruptions and bugs that might affect the performance of the model.

2.5 Promise I: Principal-Agent Game

2.5.1 Game and Data

Charness and Dufwenberg (2006, henceforth CD) test experimentally the impact of communication in a principal-agent game. They find that messages sent by principals to agents, particularly those containing promises, affect agents' beliefs and thus their actions.

CD study a sequential two-player game using the strategy method. Player A chooses *In* or *Out*, and player B chooses to either *Roll* or *Don't Roll* a six-sided die. Player B's choice affects payoffs only if A chooses *In*. Player B makes her decision without knowing player A's actual choice, but under the hypothetical condition that A chose *In*. Figure 2.4 illustrates structure and payoffs of CD's Γ_1 game. CD investigate behaviour in treatments where player B can send cheap-talk pre-play messages to player A.

We have five benchmark classifications for this dataset. CD classified the messages themselves (CD). Further, Houser and Xiao (2011, henceforth HX) provide a classification of strong and weak promises from both a traditional content analysis (C_W and C_S)

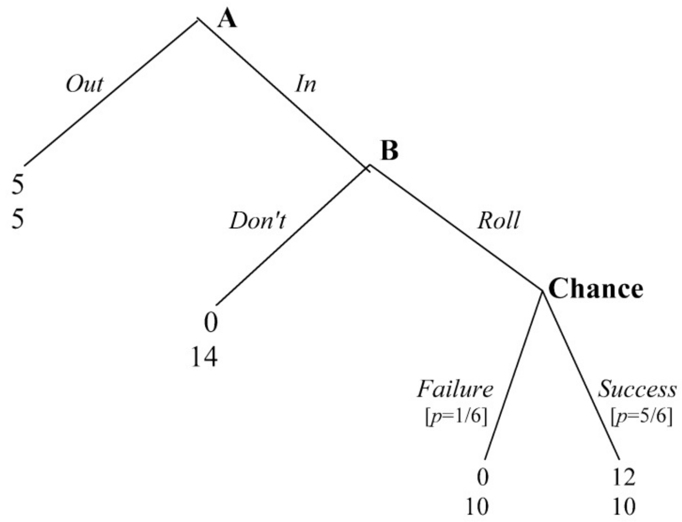


Figure 2.4: Charness and Dufwenberg (2006)'s (5,5) game.

and the classification game they introduce (G_W and G_S).

HX conducted a series of experiments to build their classification datasets. All annotators were students from George Mason University who participated in one of their "classification experiment" treatments. For the traditional content analysis, subjects received detailed written instructions that provided two distinct criteria to determine whether a message is "promise" or "empty talk" (see Appendix 2.A.1 for further details on the instructions). In their weak content treatment, C_W, the section of the instructions subject to treatment variations stated: "Classify a message as 'Promise or intent' if at least one of the following conditions is **probably** satisfied". In their strong content treatment, C_S, the instruction was identical except that the word "probably" was replaced with "certainly". Twenty-five subjects participated in the C_W treatment, and twenty-four subjects in the C_S treatment. All participants received a show-up fee of \$7 and were paid an hourly rate of \$12. On average, each participant was paid \$19. In total, the classification tasks cost \$475 for the C_W treatment and \$456 for the C_S treatment.

In their classification game treatments, subjects were not given detailed instructions outlining the criteria for the classification categories. Instead, they received a generic instruction to classify each message as either "promise" or "empty talk". For the weak treatment, G_W, subjects were verbally instructed to "Classify a message as 'Promise or intent' if, in your opinion, it includes any statement of intent". For the strong treatment, G_S, the verbal instructions were almost identical, with a key difference in the final phrase where "it includes any statement of intent" was replaced with "it is certainly a promise". Twenty-five subjects participated in each treatment. Subjects were paid a show-up fee of \$7, and were additionally informed that three of their classifications would be randomly selected. If these matched the majority classification, they would receive an additional \$5. The classification procedure lasted approximately one hour, and the median payment

	CD	C_S	C_W	G_S	G_W
P	24	26	27	24	31
E	14	12	11	14	7
f_P	.63	.68	.71	.63	.82

Table 2.3: Aggregate results of the different human classification methods

	CD	C_S	C_W	G_S	G_W
CD	1				
C_S	77	1			
C_W	71	94	1		
G_S	78	89	82	1	
G_W	54	65	71	54	1

Table 2.4: Pairwise comparison of human classifications via Krippendorff's α

per subject was \$22. In total, the cost of the classification task for each treatment was approximately \$550.

Out of the 38 messages to be classified, all classification methods yielded the same result for 29 messages. The number of messages classified as "promise" (P) or "empty talk" (E) for each method is displayed in Table 2.3. HX argued that the coordination aspect of the game classification method allowed subjects to be more sensitive to subtle variations in the instructions that either weakened or strengthened the definition of what constitutes a promise. This sensitivity was evidenced by the notable difference in the number of messages classified as "promise" between the G_S and G_W treatments compared to the difference between the C_S and C_W treatments.

Krippendorff's α values, calculated for each pairwise comparison of classifications as displayed in Table 2.4, quantify the variability in agreement among the classifiers, which in turn, underlines the discrepancies in how annotators under different classification methods reacted to the instructions. The near-perfect agreement between the C_S and C_W classifications from traditional content analysis underscores the human classifiers' lack of responsiveness to minor yet crucial variations in the instructions. In contrast, the moderate agreement between G_S and G_W supports HX's assertion that presenting the classification task as an incentivized coordination game among annotators enhances their responsiveness to these variations. Moreover, although CD conducted the classifications without explicit instructions or guidelines on the criteria used to annotate promises, the substantial agreement of their classifications with C_S and G_S, along with moderate agreement with G_W, suggests that CD might have adhered to a mental guideline that aligns more closely with the stricter definition of what constitutes a promise.

2.5.2 Prompts

We have considered four different prompts[20]: a Basic prompt, B, and three variations O_S, O_N, and O_W, of the original instructions used by HX in their traditional content classifications (C_S and C_W). All our prompts share the same following sections: "General Task", "Context" and "Output Format". "General Task" briefly defines the task in a single sentence; the "Context" section provides details about the type of players, game mechanics, and communication protocol; and the "Output Format" section provides a template for the model to follow when outputting its classification.

The basic prompt, B, does not include a "Role Persona" section, and its "Classification Task" section does not provide any description of what constitutes a "promise". Instead, it combines instructing to classify a message as a "promise" or as "empty talk" with how to code this classification (1 for "promise", 0 for "empty talk") within the same directive.

For the three prompts that are based on the original instructions, O_S, O_N and O_W, the "Classification Task" section uses the instructions provided by HX verbatim. We adopt the same weak and strong instruction variations they considered in their C_W and C_S treatments with the prompts O_W and O_S. Additionally, we have considered a neutral version, O_N, which does not use any additional adverbs (such as "probably" or "certainly" used in O_W and O_S, respectively) to qualify the verb "satisfied" in the condition: "if at least one of the following conditions is `<adverb>` satisfied." Moreover, these three prompts, unlike B, include a "Classification Coding" section that outlines how to code specific categories in the output (1 for promise and 0 for empty talk). Lastly, because HX's original instructions included the statement "Operate as a coding machine", which imposes a role, we also did not include an additional "Role Persona" section in these prompts.

2.5.3 Example

Figure 2.5: Promise I - Example 1

```
Hello fair stranger, anonymous partner...  Choose whatever you want.  Far be it from
me to influence your decision, but I think you should choose ``in'' and I should
choose ``roll'' and we should take the chance at both earning as much as we can.  5
chances out of 6 say it'll work, and I'm totally broke, looking to rake in stray
cash however I can.  I feel the luck in the air.  I don't really have much else to
say.  Hope you're doing well, whoever you are.  Yes.  That's all.  Random note from
random human
```

The message displayed in Figure 2.5 was contentious in the literature. As displayed in Table 2.5, in the G_W and C_W treatments, 56% and 52% of subjects respectively classified

[20] All our prompts are presented in Appendix 2.A.2

the message as “Promise” (1). In contrast, less than 50% of participants in the C_S and G_S treatments classified the message as “Empty Talk” (0).

	CD	C_S	C_W	G_S	G_W
$Class.$	0	0	1	0	1
f_1	–	$< 50\%$	52%	$< 50\%$	56%

Table 2.5: Human classifications for the message in Figure 2.5

GPT classification results for the same message are presented in Table 2.6. Without leveraging the 0-shot-CoT (henceforth CoT) prompting technique, GPT-4 identified the message as “Promise” only under O_W. When CoT is invoked, the classification for O_N switched to “Promise”, while it remained the same for the other prompts. On the other hand, GPT-3.5 classified the message as “Promise” for all prompts irrespective of whether CoT was invoked or not.

In Figure 2.6, the rationals generated by each model under prompts O_W and O_S using the CoT technique are presented. Both models successfully identify and use various contextual aspects of this lengthy message to support their classifications. GPT-4 adopts an exploratory tone that utilises less assertive modal verbs such as “could” and “would”. This approach indicates a methodical progression in reasoning which begins without a firm initial stance and progressively develops a conclusion through systematic observation. Conversely, GPT-3.5 adopts a more assertive tone. It initiates its reasoning with the definitive “clearly indicated that” phrase, which suggests that subsequent arguments serve more as justifications for a pre-determined conclusion rather than an open inquiry leading to a decision. Consequently, while GPT-4’s exploratory reasoning potentially enables improvements in its classification, GPT-3.5’s reasoning lacks this inquisitive mode and fails to improve its classification.

Figure 2.7 presents three messages classified as “promise” exclusively in G_W, with roughly two-thirds of subjects in G_W identifying these messages as promises (60%, 60%, and 64% for messages 2, 3, and 4 respectively). In contrast, while GPT-3.5 consistently labels these messages as “promise”, GPT-4 consistently labels these messages as “empty talk”. For messages 2 and 3, one can make a case for a very implicit statement of intent, but message 4 contains no clear intent, as GPT-4 under O_W suggests: “The message

	GPT-3.5		**GPT-4**	
	no-CoT	CoT	no-CoT	CoT
B	1	1	0	0
O_S	1	1	0	0
O_N	1	1	0	1
O_W	1	1	1	1

Table 2.6: GPT classifications for the message in Figure 2.5

Figure 2.6: Promise I - Classifications with CoT for prompts O_W and O_S

GPT-3.5

Output using O_W
Reasoning:
The message clearly indicates that player B is suggesting a course of action to player A, which is to choose "in" and for player B to choose "roll" in order to take a chance at earning as much as they can. Player B also mentions being broke and looking to rake in stray cash, which gives player A reasons to believe that player B would choose "roll" to potentially earn more money.
Classification: 1
Output using O_S
Reasoning:
The message clearly indicates that player B wants player A to choose "IN" and player B to choose "ROLL" in order to take the chance at earning as much as they can. Player B also mentions being broke and looking to rake in stray cash, which gives player A reasons to believe that player B would choose "ROLL" to potentially earn more money. Therefore, the message constitutes a statement of intent to choose "ROLL" and is not just empty talk.
Classification: 1

GPT-4

Output using O_W
Reasoning:
- Player B suggests that player A should choose "IN" and implies that they should choose "ROLL" by stating "we should take the chance at both earning as much as we can."
- The message provides a reason for player A to expect that player B will choose "ROLL" by mentioning the favorable odds ("5 chances out of 6 say it'll work").
- Player B also mentions their own financial situation ("I'm totally broke, looking to rake in stray cash however I can"), which could be interpreted as an indirect way of saying they are motivated to choose the option that maximizes their earnings, which would be "ROLL" in the case player A chooses "IN".
- The message does not contain a direct promise, but it does give player A reasons to believe that player B has an intention to choose "ROLL" if player A chooses "IN".
Classification: 1
Output using O_S
Reasoning:
- The message from player B suggests that player A should choose "IN" and implies that player B would then choose "ROLL" by stating "we should take the chance at both earning as much as we can."
- Player B mentions the probability of success (5 chances out of 6) to encourage player A to choose "IN."
- The message includes a personal note about player B being broke and looking to earn money, which could be interpreted as an indirect way to reassure player A that player B has a vested interest in choosing "ROLL" to maximize their own earnings.
- However, the message does not contain a clear and direct promise or statement of intent to choose "ROLL." The language used ("I think you should choose 'in' and I should choose 'roll'") is suggestive rather than committal.
- The message ends with a general well-wishing and does not reinforce any commitment to a specific action.
Classification: 0

Figure 2.7: Promise I - Additional examples (percentage of promise class. in G_W)

Example 2: (60%)
Please choose In so we can get paid more.
Example 3: (60%)
Hopefully I'll make a lucky role.
Example 4: (64%)
If you choose IN you have the best opportunity to make the most money. You have a 5 /7 chance of making more money! So IN would be your best bet. Cheers.:)

could be interpreted as encouraging player A to choose 'IN', but without a clear statement of intent or promise from player B regarding their own subsequent action, it does not satisfy the conditions for a promise." On the other hand, the message "Both of 'us' can earn.", which is arguably just as ambiguous, is uniformly classified by all human annotation protocols as well as consistently by both GPT models as "empty talk". These examples demonstrate that, regardless of the number of subjects or the annotation protocol employed, classifications are prone to inconsistencies and errors inherent to human judgement. Given GPT-4's ability to classify promises at a level on par with human annotators, employing it may prove more effective, especially for ensuring consistency both within a study and across multiple studies considering the same type of task classification (Vanberg, 2008; Ismayilov and Potters, 2016; Ederer and Stremitzer, 2017; Di Bartolomeo et al., 2019, 2023).

2.5.4 Results

		no-CoT					**CoT**				
		CD	C_S	C_W	G_S	G_W	CD	C_S	C_W	G_S	G_W
GPT-3.5	B	71	76	79	71	90	82	82	84	82	84
	O_S	71	76	79	71	90	79	84	87	79	**92**
	O_N	68	74	76	68	87	74	79	82	74	87
	O_W	68	74	76	68	87	76	82	84	76	90
GPT-4	B	87	87	84	87	74	90	90	87	95	76
	O_S	92	92	90	**97**	79	**97**	92	90	92	84
	O_N	92	**97**	95	**97**	84	92	92	95	92	90
	O_W	84	95	96	90	87	90	95	**97**	90	**92**

Table 2.7: Overall accuracy of promise classification in %.

In Table 2.7, the columns labelled "no-CoT" and "CoT" represent the treatments where CoT prompting was not incorporated and was incorporated, respectively[21]. As can be observed in Tables 2.7 and 2.8, GPT-3.5 demonstrates a negligible degree of responsiveness to the variations in the classification instructions of the prompts. Specifically, in no-CoT treatment, B and O_S, and O_N and O_W generate identical classifications. Moreover, the model's classifications under O_W differ from those under O_S with just one additional message classified as "promise."

CoT prompting consistently improves GPT-3.5's performance and leads to a greater degree of variation in the classifications across prompts. However, this variation does not necessarily imply that CoT prompting improves GPT-3.5's adherence to instructions. Ideally, if CoT prompting were effectively increasing the model's responsiveness to instruc-

[21] Although we recognise that relying solely on accuracy values might be misleading, particularly for datasets with unbalanced categories, we chose to use accuracy as our primary metric to present the models' performance across all results. This decision was based on accuracy's intuitiveness and ease of assessment by researchers from diverse backgrounds. However, we also evaluated the models using the F1 metric, and these results are presented in the appendices. The performance outcomes do not significantly differ whether assessed by accuracy or F1 score. For the F1 counterpart of the results presented in Table 2.7, see Appendix 2.A.4.

tional nuances, O_W would demonstrate the highest performance in the G_W benchmark, and O_S in the G_S benchmark. Yet, O_S consistently outperforms O_W. Furthermore, more detailed instructions tend to deteriorate GPT-3.5's performance: B outperforms O_W and O_N under all benchmarks except G_W, and O_S only surpasses B under the benchmarks characterized by weaker instructions, namely C_W and G_W.

		no-CoT				**CoT**			
		B	O_S	O_N	O_W	B	O_S	O_N	O_W
GPT-3.5	P	35	35	36	36	29	32	34	33
	E	3	3	2	2	9	6	4	5
	f_P	.92	.92	.95	.95	.76	.84	.90	.87
GPT-4	P	21	23	25	27	22	25	27	28
	E	17	15	13	11	16	13	11	10
	f_P	.55	.61	.66	.71	.58	.66	.71	.74

Table 2.8: Aggregate classification results of different prompts

GPT-4 consistently outperforms GPT-3.5 except under G_W benchmark. In Table 2.8, relatively larger variation in the number of messages classified as "promise" observed for GPT-4 is indicative of the model's responsiveness to the instructional variations.[22] Furthermore, its responsiveness to the instructional nuances is reflected to some degree in its classification performance. As can be observed in Table 2.7, O_W consistently achieves the highest accuracy scores under the G_W and C_W benchmarks, and under the G_S benchmark, O_S and O_N are tied for the best performance. On the other hand, in no-CoT treatment, O_N achieves the highest accuracy under the C_S benchmark, surpassing the performance of O_S.

For GPT-4, CoT prompting does not consistently improve performance. The highest performances for benchmarks C_S and G_S are achieved in no-CoT treatment using O_S or O_N. Conversely, the highest performances for benchmarks C_D, C_W, and G_W are achieved in CoT treatment using O_S or O_W. Given the top performing results under benchmarks with weaker "promise" classification conditions are achieved in CoT treatment and the top performing results under the benchmark with the stronger "promise" classification condition are achieved in no-CoT treatment suggests that CoT prompting introduces a bias towards classifying messages as promises.

[22] Note that although the number of messages classified as "promise" increases as the prompt's instruction to adhere to "promise" conditions are relaxed, there are messages identified as "promise" under stricter instructions that are not classified as "promise" under weaker instruction.

2.6 Promise II: Public Good Game

2.6.1 Game and Data

Arad, Hugh-Jones and Penczynski (2024, henceforth AHP) carried out an online experiment to understand which kind of communication predicts cooperation. 633 participants engaged in five identical 3-player public good games, each with different anonymous opponents. In every game, the three players had the opportunity to chat before making a decision using a built-in platform resembling WhatsApp. The messages were classified according to the presence of a promise by two RAs. On the basis of the free-flowing chat between the three players, the classification indicates for each individual player, whether a promise was made. In total, 717 chat instances were analysed for classification. The RAs reached consensus on 89.9% of the instances, with a Krippendorff's α of 0.798, indicating substantial inter-rater reliability. Within these agreements, 53.3% were classified as "promise", hence the two categories are balanced in the dataset.

2.6.2 Prompts

We have considered three distinct prompts[23]: a basic prompt, B; a prompt that is the reframed version of the original instructions, O; and a variation of O, termed O_+, which features an alternative set of classification conditions. Similar to the approach in the "Promise I" section, all our prompts include the same subsequent sections: "General Task", "Context," and "Output Format". The "General Task" section briefly defines the task in a single sentence. The "Context" section elaborates on the types of players, game mechanics, and communication protocol. Finally, the "Output Format" section outlines the template the model should use to format its classification output.

Since the classification instructions in the HX and AHP prompts are identical, B and O closely resemble the prompts B and O_N from the "Promise I" section, respectively. Consequently, all information pertaining to prompts B and O_N in Section 2.5.2 also applies to B and O. Specifically for their 0-shot versions, the primary distinction lies in the "Context" section. The "Context" section of the prompts introduced in Section 2.5.2 details the investment game and features a standalone message from player A to B, whereas in this section, the "Context" section of the prompts describes the public good game involving a conversation among three players.

The codebook of AHP, unlike that of HX, includes a set of examples accompanying its classification instructions. AHP use these demonstrations to provide more nuanced conditions for the "promise" and "empty talk" categories. This section is structured as a sequence of examples, each followed by a remark that highlights a specific case of

[23] All our prompts are presented in Appendix 2.A.2.

promise or empty talk classification, then more examples and subsequent remarks, and so on. We adopted this "Example" section verbatim in O, modifying its format to align with our prompt template: each example is separated and indexed with a subtitle, such as "Example #1", followed by the content of the example, and each remark section between sets of examples is distinguished with a "Remark" subtitle, followed by the remark. Consequently, unlike O_N from Section 2.5.2 and prompt B, O includes an "Example" section, although this additional section is incorporated into the prompt only for n-shot treatments.

In total, 11 chat examples are provided for AHP, making our n-shot treatment for O an 11-shot setup. Since we have adopted the example section of the original instructions verbatim, the format of our prompt's example section diverges from the conventional <question, answer> format typically used in n-shot prompting. Instead, it includes explanations between sets of examples. However, it also does not adhere to the traditional n-shot-CoT prompting format of <question, explanation, answer>, as the provided explanations (remarks) are sparse and do not offer detailed rationales for classifying sets of messages. In summary, the "Example" section of O provides more information than a traditional n-shot prompt but less than an n-shot-CoT prompt. Nevertheless, the presence of remarks can still be seen as offering partial rationales, and therefore, their inclusion in O is expected to enhance the model's reasoning capabilities when CoT prompting is utilised.

O_+ differs from O in the "Classification Task" and "Example" sections. In the "Classification Task" section, we redefined the criteria to identify a promise and recast the "empty talk" category as "non-promise." Drawing inspiration from the content analysis approach introduced by Cooper and Kagel (2005), we analysed the dataset to identify potential subcategories within the main categories of "promise" and "empty talk." We then cross-referenced these identified categories with the subcategories presented through examples in AHP's example section. This analysis led to the formulation of three distinct conditions for each category.

Both HX and AHP specify two conditions for classifying a'promise":

1. Player indicates that she will take a certain course of action
2. Player gives other players a reason to believe that she will take a certain course of action.

In O_+, we introduced three refined conditions for the same category:

1. Player explicitly states his intention to take a specific action.
2. Player agrees to take an action suggested by another.
3. Player commits to an action, conditional on a specific event occuring.

The first condition in O_+ is a paraphrased version of HX and AHP's first condition. The second condition addresses a scenario frequently observed in the dataset and is emphasised through multiple demonstrations and a remark in AHP's original instructions. Similarly, the third condition, described as a "conditional promise" in AHP's examples, and is frequently observed in some shape or form in the data.

On the other hand, we have omitted the belief-based promise condition used by HX and AHP as we find it potentially problematic for GPT. Since GPT excels at identifying explicit textual patterns but is not as proficient at inferring the beliefs of players (Moghaddam and Honey, 2023), omitting this belief-based condition and considering instead additional explicit "promise" cases better aligns with GPT's strengths and is hoped to improve the model's performance.

Since, it is recommended to avoid negation when instructing the model (OpenAI, 2023b), such as defining "empty talk" as any message that does *not* meet the promise category conditions, we have established three distinct conditions to classify the "empty talk" category, rather than instructing the model to classify "empty talk" as any message that fails to meet the conditions set for promises, as was done by HX and AHP:

1. Player suggests an action without an explicit commitment.
2. Player asks questions or discusses preferences without an explicit commitment.
3. Player talks about hypothetical, ideal or rational actions without an explicit commitment.

Although these conditions are implicitly provided through examples in AHP's instructions, they were not explicitly defined as noted above. Additionally, all three conditions represent behaviours frequently observed in the dataset.

The following two additional instructions, originally provided by HX and AHP and included in O as well as in all the prompts in Section 2.5.2 except for B, are omitted in the classification instructions of O_+:

- Capture what had been said rather than why it was said or what effect it had.
- Operate as a "coding machine".

Given that the second instruction involves role persona guidelines, in O_+ we have instead included an explicit "Role Persona" section following the "General Task" section that instructs the model to "Act as a behavioural economist specialised in text classification, the investment game and communication in games."

The objective with the revised classification instructions in O_+ is to establish a more comprehensive set of conditions for the categories in order to provide the model with the necessary classification information and thus to minimise reliance on demonstrations to convey semantic nuances inherent to both categories. The original instructions from AHP utilise examples to build upon the classification conditions but do not focus on the adequacy of these conditions to provide the required information for the models to execute the classification task effectively. This limitation in the comprehensiveness of classification instructions is directly reflected in prompt O. With prompt O_+, our aim is to explore whether a more comprehensive and alternative set of conditions can enhance the model's performance without relying on additional demonstrations, as required in O.

The "Example" section of O_+ differs from that of O not only in format and explanations but also partially in the provided examples. In O, a total of 22 classification cases are provided, with 10 categorized as "empty talk" and 11 as "promise". In the "Example" section of O_+, five of these chat instances are omitted, and three new chat instances are introduced instead.

The necessity of examples 4, 5, 6, and 8 from the original instructions (see Appendix 2.A.1) was reassessed using the 0-shot O_+ prompt and found to be redundant. Specifically, with the 0-shot O_+, GPT-4 reliably identifies statements such as "Let's do X" or "X sounds good" provided in examples 4, 6, and 8. Consequently, these examples have been omitted from the "Example" section of n-shot O_+. Furthermore, we observed that providing demonstrations for explicit statements of action related to a previous round was unnecessary, as the model consistently classified these types of statements as "non-promise" using 0-shot O_+. Additionally, example 11, which illustrates a player's "change of mind," was removed after a detailed analysis of relevant messages revealed that it did not aid the model in correctly identifying similar "change of mind" cases. Instead, the inclusion of this example led to the model incorrectly classifying such cases as non-promises, even though they were correctly classified as "promise" under 0-shot O_+. In sum, the "Example" section of O_+ provided 9 chat instances (9-shot).

Lastly, instead of the remarks sections in O, which provide partial rationales for sets of demonstrations, in O_+, we opted for concise descriptions that directly link to formerly introduced category conditions (see Appendix 2.B.2 for details). This approach aims to reinforce the semantics of the classification conditions through the demonstrations, rather than introducing new classification conditions as in O; and is designed to generate a stronger and clearer understanding of each category within the model.

2.6.3 Examples

Table 2.9 displays the classification results for the message shown in Figure 2.8 from both human annotators and GPT models under CoT and n-shot prompting. The notation "0/1" indicates a disagreement among RAs regarding the classification of player 3. GPT-3.5 aligns with the RAs' classification for only player 1 under prompts O and O_+. Conversely, GPT-4 concurs with the RAs on player 2's classification but diverges on player

Figure 2.8: Promise II - Example message 1

```
P3:  hi
P2:  Helloo
P1:  Hi
P3:  shall we sayy 50 p each?
P2:  Why not maximum 200p each for max return
P1:  Lets go with say £1 to start with
P3:  okay
P3:  meet in the middle at 1pound
P2:  sure thing
```

	GPT-3.5			GPT-4		
	P_1	P_2	P_3	P_1	P_2	P_3
B	0	0	0	0	1	1
O	1	0	0	0	1	0
O_+	1	0	0	0	1	1
RAs	1	1	0/1	1	1	0/1

Table 2.9: Human and n-shot CoT GPT classifications for the message in Figure 2.8

1's classification across all prompts.

The rationals provided by GPT-3.5 and GPT-4 for O and O_+ are presented in Figure 2.9. GPT-3.5's reasoning tends to be relatively brief and omits certain parts of the conversation. Specifically, under prompt O, it selects specific messages to base its classification on and neglects to analyse other messages from the same player. Conversely, under prompt O_+, while it attempts to provide a summary of all messages from each player, it still does not capture every element of the conversation.

Under prompt O, GPT-3.5 overlooks player 2's message "sure thing" and player 3's message "okay". Interestingly, although GPT-3.5 does not account for player 3's "okay" under prompt O, it still classifies the player as making a promise based on their earlier statement "shall we say 50 p each", which should be characterised as a suggestion rather than a promise. Thus, the model arrives at a correct classification (assuming the correct classification is player 3 making a promise) but through incorrect reasoning.

Conversely, under prompt O_+, GPT-3.5 correctly identifies "shall we say 50 p each" as a suggestion and classifies player 3 as making a promise based on the message "okay". However, the model again overlooks player 2's affirming message "sure thing", and consequently fails to classify player 2 as making a promise. This illustrates a consistent issue with the GPT-3.5's reasoning capabilities, as it fails to incorporate critical elements of the conversation in its multi-step reasoning.

GPT-4's reasoning is observed to be more detailed, with the model analysing each message sent one-by-one without skipping any. Under both prompts, GPT-4 uses player 2's message "sure thing" to assert that the player makes a promise. For the classification of player 3, the model differs in its interpretation of the message "okay" under the two

Figure 2.9: Promise II - Classifications with CoT for prompts O and O_+

```
                                                                           GPT-3.5
Output using O
Reasoning:
P3:  The message "shall we sayy 50 p each?" indicates a specific action and gives others
reason to believe or expect that P3 will invest 50 pence.  Therefore, it constitutes a
promise.
P2:  The message "Why not maximum 200p each for max return" does not indicate a specific
action by P2, so it does not constitute a promise.
P1:  The message "Lets go with say £1 to start with" does not indicate a specific action
by P1, so it does not constitute a promise.
Classification:
P1:  0
P2:  0
P3:  1
Output using O+
Reasoning:
P3:  Initiates a suggestion of 50 pence each, then agrees to meet in the middle at £1,
indicating a commitment to the proposed amount.
P2:  Counters P3's suggestion with a different proposal of investing the maximum 200
pence each, showing a preference without a clear commitment.
P1:  Suggests starting with £1, but does not explicitly commit to this amount.
Classification:
P1:  0
P2:  0
P3:  1
                                                                             GPT-4
Output using O
Reasoning:
- P3's initial message "shall we sayy 50 p each?" is a suggestion rather than a promise,
as it is phrased as a question and does not indicate a commitment to action.
- P2's response "Why not maximum 200p each for max return" is also a suggestion and not
a promise, as it is phrased as a rhetorical question and does not indicate a commitment
to action.
- P1's message "Lets go with say £1 to start with" is a suggestion, using the word "say"
which implies a proposal rather than a firm commitment.
- P3's response "okay" is an acknowledgment of P1's suggestion but does not constitute a
promise.
- P3's subsequent message "meet in the middle at 1pound" is a proposal to compromise,
but it is not phrased as a commitment to action, so it is not a promise.
- P2's response "sure thing" is an agreement to P3's proposal, which can be interpreted
as a commitment to action, thus constituting a promise.
Classification:
P1:  0
P2:  1
P3:  0
Output using O+
Reasoning:
- P3's initial message "shall we sayy 50 p each?" is a suggestion without a clear
commitment, so it's not a promise.
- P2's response "Why not maximum 200p each for max return" is also a suggestion and does
not explicitly commit to investing 200p, so it's not a promise.
- P1's message "Lets go with say £1 to start with" is a suggestion to invest a specific
amount, but it's phrased as a proposal rather than a commitment, so it's not a promise.
- P3's response "okay" to P1's suggestion is an agreement to the proposal, which can be
considered a commitment to invest £1, so it is a promise.
- P3's follow-up "meet in the middle at 1 pound" is a restatement of the agreement to
invest £1, which is a promise.
- P2's final message "sure thing" is an agreement to the previously discussed investment
amount of £1, which is a commitment and therefore a promise.  Classification:
P1:  0
P2:  1
P3:  1
```

prompts. While under O_+, it considers it as an agreement to a proposal and therefore classifies it as a promise, whereas under O it does not classify this acknowledgement as a promise.

Despite the fact that player 1's only relevant message, "Let's go with say £1 to start with", is merely a suggestion and not a promise, and although the original instructions for human annotators presents a similar example to showcase an "empty talk" classification, RAs classify it as a promise. Furthermore, under O, GPT-4's comment "...using the word 'say' which implies a proposal rather than a firm commitment" highlights its nuanced understanding of language and its ability to effectively use it as a basis for classifying player 1's message as a suggestion and not as a statement of intent. Additionally, similar to the disagreement observed under prompts O and O_+ with GPT-4, also RAs disagree on the classification of player 3. However, unlike with GPT-4, the specific reasons for the RAs' disagreement remain unclear. While it is understood that the reasoning provided by GPT models is an as-if reasoning, this capability still renders it less of a 'black box' compared to human annotators.

Figure 2.10: Promise II - Example message 2

```
P2:  invest 200
P2:  guarantee 400 each!
```

Despite numerous classification cases[24] showcasing GPT-4's ability to discern semantic nuances in conversations, it equally fails in other cases that are obvious to a human annotator. Consider the monologue displayed in Figure 2.10. As observed in Figure 2.11, GPT-4 under prompt O incorrectly interprets "guarantee 400 each!" as an explicit statement of intent by interpreting it as "I guarantee 400 each", which is most likely not the intended usage by the subject. On the other hand, GPT-4 under prompt O_+ aptly recognises both likely intended usages of the statement and, in agreement with the RAs, classifies it as "empty talk".

2.6.4 Results

Performance results of both models under each treatment and prompt are presented in Table 2.10. For GPT-3.5, 0-shot O consistently underperforms B. Where as 0-shot O_+ significantly outperforms both 0-shot O and B. This underscores the benefit of providing more comprehensive classification conditions for GPT-3.5.

In no-CoT treatments, n-shot prompting consistently improves GPT-3.5's performance. Yet, n-shot O underperforms 0-shot O_+. This suggests that the performance gain from instructing with more comprehensive classification conditions is higher than the gain from additionally providing examples.

[24]See Appendix 2.B.3 for more examples

Figure 2.11: Promise II - Classifications with CoT for prompts O and O_+

```
Output using O
Reasoning:
- P2's first message "invest 200" is ambiguous.  It could be a suggestion or a statement
of intention, but without additional context or explicit language, it's not clear if P2
is making a commitment.
- P2's second message "guarantee 400 each!" is more explicit.  It suggests that P2 is
committing to an action that would ensure each player receives 400 pence.  The use of
the word "guarantee" implies a promise to take action to achieve a specific outcome.
Classification:
P2:  1
Output using O+
Reasoning:
- P2's first message "invest 200" is a directive or suggestion rather than a promise.
It lacks the personal commitment to a specific action that would make it a promise.
- P2's second message "guarantee 400 each!" is a bit ambiguous.  It could be interpreted
as a promise because it uses the word "guarantee," which implies a commitment.  However,
without a clear subject (such as "I guarantee"), it's not explicitly a personal promise
to invest a certain amount.  It reads more like an expression of the outcome if everyone
invests 200, rather than a promise from P2 to do so.
Classification:
P2:  0
```

		no-CoT		**CoT**	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	B	59.4	–	61.4	–
	O	57.5	61.3	54.7	55.2
	O_+	72.2	75.7	67.3	75.2
GPT-4	B	67.5	–	71.8	–
	O	68.8	82.9	72.5	84.8
	O_+	86.5	86.2	84.2	**88.7**

Table 2.10: Overall accuracy of promise classification in %

In CoT treatments, n-shot prompting improves the performance of GPT-3.5 only under O_+. Furthermore, CoT prompting improves GPT-3.5's performance only for B. This contrasts with the results from the "Promise I" section (Section 2.5.4) where for GPT-3.5, CoT prompting consistently improves the performance of 0-shot prompts. Notably, despite the fact that 0-shot O and O_N from Section 2.5 differ only in their "Context" sections, CoT prompting in 0-shot O leads to poorer model performance, whereas the opposite effect is observed in O_N of Section 2.5.

GPT-4 consistently outperforms GPT-3.5. In no-CoT treatments, while n-shot prompting generates a significant improvement for GPT-4's performance under O, it marginally worsens its performance under O_+. This suggests that given that the conditions for classification categories in O_+ are comprehensive enough to enable the model to fully recognise the task within its training corpus, and hence additionally providing classification examples does not introduce any further information for the model to either improve its recognition of the task or to learn a novel feature about the task.

In CoT treatments, GPT-4's performance is consistently improved under both O and

O_+ when n-shot prompting is introduced. Moreover, contrary to GPT-3.5's results, CoT prompting improves GPT-4's performance under all prompts except for 0-shot O_+. Additionally, performance gains from using CoT prompting when demonstrations are present are not significantly different between O and O_+, with n-shot O_+ benefiting slightly more than n-shot O when CoT prompting is introduced (2.5 vs 1.9 percentage points increase, respectively). This suggests that the remarks sections provided between examples in O, which could be considered as partial rationales for the model, did not improve the efficacy of CoT prompting more than demonstrations without any rationals provided in O_+. Overall, n-shot O_+ with CoT under GPT-4 generates the highest performance.

2.7 Level-k I: Jury Voting Game

Promises are prevalent conversational statements likely to be frequently available in the models' pre-training corpus, and determining whether someone made a promise does not require additional theoretical context. Consequently, it is not unreasonable to expect the GPT models to identify promises in a text with some degree of success, as demonstrated with the basic prompt B in previous sections. In contrast, "strategic thinking" or "levels of thinking" are concepts defined by the application of level-k theory to specific games or situations and are not commonly known, and therefore less likely to be discussed, outside the behavioral economics literature. Consequently, they are much less likely to be part of the models' pre-training corpus. This distinction provides us with the opportunity to additionally investigate the model's behaviour in a task that potentially consists of more learning-based subtasks compared to the classification task of "promises".

2.7.1 Intra-team communication

Messages in the level-k I and II datasets are generated by the intra-team communication protocol that was introduced in Burchardi and Penczynski (2014) and classified according to the level-k of strategic reasoning. Teams of two subjects play as one entity and exchange arguments as follows. Both subjects individually make a suggested decision and write up a justifying message. Upon completion, this information is exchanged simultaneously and both subjects can enter individually a final decision. The computer draws randomly one final decision to be the team's action in the game. The protocol has the advantage of recording the arguments of the individual player at the time of the decision making. Furthermore, the subject has incentives to convince his team partner of his reasoning as the partner determines the team action with 50% chance.

	L0	L1	L2	L3
f_L	.21	.49	.29	.01

Table 2.11: Level distribution

2.7.2 Game and Data

Here, we briefly remind the reader of the necessary information from Chapter 1 that is relevant for GPT classification.

In Chapter 1, we proposed a level-k model of strategic thinking in jury voting (JV) games à la Feddersen and Pesendorfer (1998) and Guarnaschelli et al. (2000). In juries of size 3 or 6, jurors receive informative signals (red or blue balls) and then vote to acquit (blue) or convict (red) the defendant, with the jury decision ideally matching the innocence (blue urn) or guilt (red urn) of the defendant. Looking at juries under the unanimity rule for conviction, we show that the jury performance depends on the strategic sophistication of jury members, which in turn depends on the complexity of the task at hand.

Our model assumes non-strategic, random level-0 play, to which a level-1 player best-responds by always voting according to the received informative signal. Given the unanimity rule, the best response to informative voting by level-2 players is to strategically vote always "convict" to make a conviction more likely and rely on other voters to acquit. For level-3 players, the best-response to always convicting is to play informatively like level-1 players do.

The messages are independently classified according to this level-k model by two RAs. The RAs are introduced to the level-k model and received detailed instructions about characteristics of the individual level-k types.

The classification procedure starts with both RAs providing independent sets of classifications. Then, both are anonymously informed about the classifications of the other RA and have the possibility to simultaneously revise their own classification. This revision process is repeated twice. After the process, the two RAs agreed on 93.2% (493) of the classifications.

As can be observed in Figure 2.11, the distribution of levels is non-degenerate and features a heterogeneity of types, a hump-shape with mode behaviour at level-1, and hardly any level-3 behaviour. This is a standard distribution commonly observed in similar studies and hence represents the type of distribution to be expected when researchers consider to classify such data (Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Burchardi and Penczynski, 2014; Crawford et al., 2013). Their agreed classifications for 493 messages constitute the benchmark for the LLMs in this study.

2.7.3 Prompts

Without the provision of specific characteristics of different levels of thinking, it is unclear how GPT could interpret them or distinguish, say, a level-1 thinker from a level-2 thinker. Given this uncertainty, we do not consider a basic prompt B for this classification task. Instead, we investigate two prompts, O and O_+, which closely follow the original instructions and vary only in how they provide the context for the task.

The original codebook's classification instructions for human annotators begin with a "General Comments" subsection that notes the potential implicit nature of messages and instructs annotators to classify messages at the level they believe is most likely when uncertain. This section is followed by subsections for each level from 0 to 3, each further containing "Characteristics", "Examples" and "Note" subsections. The "Characteristics" subsection outlines the observed traits of that level of thinking within the voting game, while the "Note" subsection provides additional guidance for handling ambiguous cases. The "Examples" subsections vary in number, with five examples for level-0, three for level-1, eight for level-2, and three for level-3. There are no specific comments provided for individual examples.

These classification instructions are incorporated verbatim into both prompts O and O_+, preserving the structure and format of the original level subsections and their respective "Characteristics", "Examples" and "Note" subsections. The only exception is the "Note" section of the level-0 subsection. In the codebook, this "Note" section instructs annotators to leave a message unclassified if it contains ambiguous phrases like "Play red, trust me!" that likely indicate level-1 thinking but lack additional context to confirm it. Given that the messages considered for classification are those where both annotators found enough information to classify, this specific instruction is redundant for the models and therefore omitted in the prompts. Additionally, the instruction in the "General Comments" section to classify messages based on the most likely level when unsure contradicts this instruction to leave messages unclassified when there is not enough context. Hence, to maintain instructional consistency with our prompts, we adhered to the "most likely level" instruction. Lastly, this modification also avoids the risk of the model misinterpreting the option to leave messages unclassified and incorrectly generalising it to other unrelated but potentially ambiguous messages and other levels of thinking.

Prior to the classification instruction section, the codebook includes a three-page explanation of the experiment, general level-k theory, and the application of level-k theory to the voting game. We have omitted these sections and instead included the most essential information in the "Context" section of the prompts. This section provides details on general game mechanics and communication protocols, aimed at providing the model with necessary context and textual patterns for its classification task. Compared to prompt O, prompt O_+ offers more detailed information on both game mechanics and communication protocols, and also includes a brief level-k theory section specific to the voting game. Additionally, in prompt O_+, these three sections (game mechanics, communication protocol, and theory) are distinctly separated by subtitles.

All other sections of the prompts are identical. In the "General Task" section, the model is instructed to classify a player's level of strategic thinking in a voting game. This is followed by the "Role Persona" section, where the model is instructed to act as a behavioural economist specialised in level-k modelling, strategic thinking, and text classification. Other auxiliary sections, such as "Classification Process", "Constraint", and "Output Format", follow the format presented in Section 2.4.1, and are almost identical to prompts from previous sections (see Sections 2.C.1 and 2.C.2 for the original codebook and the prompts).

2.7.4 Examples

Figure 2.12: Level-k I - Example message

```
I believe we have the correct color There is no specific reason behind it If you
think that the danger to decide the whole vote for blue is to high, I will go with
red
```

The classification of the message shown in Figure 2.12 is detailed in Table 2.12, which includes the results from RAs and GPT models using n-shot and CoT prompting techniques for prompts O and O_+.

	GPT-3.5	**GPT-4**
O	0	2
O_+	1	2
RAs	1	

Table 2.12: Human and GPT classifications of message in Figure 2.12 using n-shot and CoT prompting

The numbers in Table 2.12 indicate the level of strategic thinking attributed to the writer of the message. GPT-4's classifications are consistent across both prompts. Except for GPT-3.5 under prompt O_+, all other model classifications differ from those made by the RAs. This highlights a significant divergence in the assessment of strategic thinking between the models and RAs.

The rationales provided by GPT-3.5 and GPT-4 for O and O_+ using n-shot and CoT prompting techniques are displayed in Figure 2.13. The rationales from GPT-3.5 are significantly shorter. Specifically, under prompt O, GPT-3.5's rationale focuses solely on the phrase "there is no specific reason behind it" and concludes with a level-0 classification. In contrast, under O_+, while GPT-3.5 acknowledges additional statements from the message such as "we have the correct color" and "danger to decide the whole vote for blue",

Figure 2.13: Level-k I - Classifications with CoT and n-shot for prompts O and O_+ for message in Figure 2.12

```
                                                                                    GPT-3.5
Output using O
Reasoning:
The player seems to be choosing randomly without any specific reasoning.  They mention
that there is no specific reason behind their choice, indicating a lack of strategic
thinking.
Classification:
0
Output using O+
Reasoning:
The player does not provide any specific reasoning for their choice.  They mention
that they believe they have the correct color without any justification.  They also
consider the risk of deciding the whole vote for blue to be too high, indicating a
slight awareness of the potential consequences of their decision.
Classification:
1

                                                                                      GPT-4
Output using O
Reasoning:
The player starts with a statement that seems to be based on intuition rather than
strategic reasoning ("I believe we have the correct color There is no specific
reason behind it").  This could suggest a level-0 approach.  However, the player then
acknowledges the potential impact of their vote on the group's decision ("If you think
that the danger to decide the whole vote for blue is to high, I will go with red").
This shows some awareness of the voting dynamics and the strategic implications of their
choice, which is beyond level-0 reasoning.
The message does not contain a clear probability argument or a direct reference to
the player's own signal, which would be indicative of level-1 reasoning.  Instead, the
player seems to be considering the consequences of their vote in the context of the
group's outcome, which aligns more with level-2 reasoning.  They are aware that their
vote could be pivotal ("the danger to decide the whole vote for blue is too high") and
are willing to adjust their vote accordingly ("I will go with red").  This suggests
they are considering the strategies of other teams, assuming that other teams might play
their signals, and are trying to best respond to that.
The message lacks the complexity of a level-3 player, who would assume that other
players are also thinking at level-2 and would adjust their strategy even further based
on that assumption.
Classification:
2
Output using O+
Reasoning:
The player starts with a statement that seems to be based on intuition or a hunch ("I
believe we have the correct color") without providing a strategic justification.  This
could hint at a level-0 approach.  However, the player then acknowledges the potential
impact of their vote on the group's decision ("the danger to decide the whole vote for
blue is too high"), which suggests an awareness of the voting dynamics and the strategic
implications of their choice.  This awareness moves the reasoning beyond level-0.
The player also seems to be considering the risk associated with voting blue and is
willing to adjust their vote to red if their teammate perceives it as too risky.  This
indicates a level of strategic thinking that takes into account the potential actions
of others, aligning more with level-2 reasoning.  The player is not simply following
their own signal (which would be level-1) but is considering the group dynamics and the
possibility that their vote could be pivotal.  They are not, however, explicitly stating
that they believe others will vote contrary to their signal, which would be indicative
of level-3 reasoning.
Classification:
2
```

it fails to provide any explicit or detailed interpretation of these statements. Moreover, GPT-3.5 does not articulate in either rationale why the message should be classified in a certain way; such details are only revealed in the classification section of its output.

GPT-4's rationales under both prompts are significantly more detailed and thorough in analysing the message compared to GPT-3.5. As it processes each statement, GPT-4 evaluates every distinct phrase to determine its relevance for or against a specific level of strategic reasoning. For instance, the phrase "no specific reason behind it" initially suggests a possible classification as level-0. However, GPT-4 rejects this initial assessment, recognising that the phrase "the danger to decide the whole vote for blue" indicates the player is considering the impact of her actions on the group's outcome. Consequently, the model deliberates whether the player is operating at a level-1 or level-2 of reasoning and ultimately classifies her as the latter. To ensure a comprehensive analysis, GPT-4 also explores and dismisses the possibility that the player could be a level-3 thinker.

It is important to note that in neither prompt are there specific instructions to check each potential level of reasoning as a candidate for classification and to disregard them if they do not fit its provided descriptions. This approach of systematically evaluating each level is an emergent strategy that GPT-4 adopts when CoT prompting is used.

Human annotators are instructed to classify a player as a level-2 strategic thinker if "(The) player offers an argument acknowledging the potential votes of the other teams and how to vote accordingly" (see Appendix 2.C.1 for the full instructions). These instructions are provided verbatim in both prompts O and O_{+}. The phrase "the danger to decide the whole vote for blue" implicitly indicates an acknowledgement of other teams voting red and the effect of voting blue in such a scenario. While this statement was not considered explicit enough by RAs to classify the player as a level-2 thinker, GPT-4 does classify it as such. This highlights a difference in the interpretation thresholds between human annotators and GPT.

This case also illustrates the challenges of classification and the difficulties in precisely measuring the model's classification performance, due to the inherent ambiguities of natural language. Although there is a mildly convincing argument for the player to be classified as level-2, GPT-4's classifications are considered incorrect. Furthermore, this type of ambiguity is prevalent in numerous other messages. The messages in Figure 2.14 provide two additional examples where the possibility of level-2 strategic thinking is even more subtle, if present at all. The operative words "harmless" in the first example and "safe" in the second are evaluated by the models as subtle indications that the player is considering other players' behaviour which in turn to some degree affect her action. As a result, for the first message, both GPT-3.5 and GPT-4 under both prompts classified it as level-2, while the RAs classified it as level-1; and for the second message, while GPT-3.5 and RAs classified it as level-0, GPT-4 under both prompts classified it as level-2.

Figure 2.14: Level-k I - Ambiguous messages

```
Example 1:
red is more probable and almost harmless for the outcome.

Example 2:
no clue, red is safe.
```

In Chapter 1, we analyse four different treatments within subjects, where each subject plays two rounds. Out of 493 instances, subjects referred to a message from the previous round in nine cases, either explicitly or implicitly. This occurred despite subjects being randomly matched with new teammates in every round and being informed of this arrangement. RAs are instructed to use information from previously classified messages to classify any subsequent message that refers to these earlier messages and to make a note of it if they have done so.

For a more thorough investigation of GPT's performance, GPT is also provided with the previously referred messages for these nine instances to ensure it has full context. This step occurred after the initial classification of all the messages are independently made. As a result, GPT initially provided a classification for each of these messages without the context of the referred messages. In Figure 2.15, an example of such a message is shown. This message refers to the message displayed in Figure 2.16.

Figure 2.15: Level-k I - Example message 3

```
Again I think we should go with red.
```

Both messages are provided by the subject within the same treatment in two consecutive rounds. The referred message in Figure 2.16 is classified by both models and by RAs as level-2. Using this information, the RAs also classified the message in Figure 2.15 as level 2. Without being provided with the additional context of the previous message, GPT-3.5, under both prompts, classified the message in Figure 2.15 as level-0. Surprisingly, GPT-4 under prompt O classified this message as level-1, and as level-2 under O_+.

Figure 2.16: Level-k I - Referred message

```
I think we should stick to read.  The chances that everyone picks the red ball are
really small so we are save with read I hope
```

The rationale for both models' classifications is presented in Figure 2.69. A detailed analysis of the rationale under GPT-4 shows that the word "again" results in the model to consider whether the player votes strategically or informatively, given her repeated choice of red. Under prompt O, GPT-4 rules out the possibility that the player is a level-0 or level-3 thinker and acknowledges that classifying her as a level-2 thinker is speculative, leading to a level-1 classification. Conversely, under prompt O_+, the model is more inclined to assign a level-2 classification, albeit with an explicitly stated degree of uncertainty.

It is important to note that the original instructions include the directive: "If you are unsure of the level of the message, you should indicate the level you think is more likely." This is also provided verbatim to the GPT models. Consequently, even when uncertain, the model is compelled to make a classification. We have also explored a more relaxed approach for this specific message, where the model is given the option to abstain from classifying in cases of uncertainty. Under the option to abstain, the model consistently chooses not to classify the message, citing a lack of context as the reason.

Figure 2.17: Level-k I - Classifications of the message in Figure 2.15 with CoT and n-shot for prompts O and O_+

GPT-3.5

Output using O
```
Reasoning:
The player simply states their preference for the red option without providing any
reasoning or justification for their choice. There is no indication of any strategic
thinking or consideration of other players' decisions.
Classification:
0
```
Output using O_+
```
Reasoning:
The player simply states their preference for the color red without providing any
reasoning or justification for their choice. There is no indication of any strategic
thinking or consideration of other teams' potential votes.
Classification:
0
```

GPT-4

Output using O
```
Reasoning:
The message "Again I think we should go with red" suggests that the player is following
a consistent strategy of voting red. However, the message lacks any explicit reasoning
or justification for this choice. It does not mention the color of the ball they
picked, nor does it consider the potential voting behavior of other teams. The use of
the word "again" implies that this is a repeated strategy, but without further context
or explanation, it is not clear whether this is based on a strategic consideration of
others' likely actions (Level-2) or simply a personal rule or preference.
Given the lack of strategic reasoning evident in the message, it is not possible to
classify this player as Level-3. The message also does not contain any probability
argument or acknowledgment of the signal, which would be characteristic of a Level-1
player. It does not appear to be a random or unrelated choice, which would be
indicative of a Level-0 player.
The message could potentially be interpreted as a Level-2 strategy if we assume that
the player has previously considered the strategies of others and has settled on always
voting red as a best response. However, without explicit acknowledgment of others'
strategies, this interpretation is speculative.
Classification:
1
```
Output using O_+
```
Reasoning:
The statement "Again I think we should go with red" suggests that the player is
following a consistent strategy of voting red. However, the reasoning behind this
choice is not explicitly stated. The use of the word "again" implies that this strategy
has been used before, possibly indicating a pattern or rule the player is following.
There is no mention of the color of the ball they drew, nor is there any consideration
of the other teams' potential votes. This lack of strategic consideration for others'
actions could indicate a level-1 player who consistently follows their own signal, but
the message is too vague to confirm that. It could also be a level-2 player who has
adopted a strategy of always voting red, assuming that other players will vote according
to their own signal. Without further information, it is difficult to determine if this
is a level-1 or level-2 strategy.
The classification leans towards level-2 because the player seems to have a
predetermined strategy (voting red) that they are sticking to, which could be a response
to the assumption that other players are voting according to their own signals (level-1
behavior). However, due to the lack of explicit reasoning, this classification is made
with some uncertainty.
Classification:
2
```

In Figure 2.18, the rationale for classifying the message in Figure 2.15 is shown once the previously referred message from Figure 2.16 is provided to GPT-4 under prompt O[25]. This demonstrates that when the model is provided with the referred message, it

[25]Note that only the referred message, and not its prior classification as level-2, is provided as additional context. Consequently, as can be observed in the rational in Figure 2.15, the model first classifies this re-

can effectively incorporate it into its classification analysis and accordingly classify the message as level-2.

Figure 2.18: Level-k I - Classifications of the message in Figure 2.14 with CoT and n-shot for prompt O after the refered message is also provided.

```
                                                                                GPT-4
Output using O
Reasoning:
The message "Again I think we should go with red." is brief and lacks explicit strategic
reasoning.  However, the use of the word "again" suggests that this decision is based on
a previously established reasoning pattern.  Given the context of the previous message,
which contained level-2 reasoning, it is likely that the player is continuing with the
same line of thought.  The player is choosing red, potentially because they believe
other players will vote their signal, and red is a safer strategic choice.  Despite
the brevity of the message, the implication is that the player is still considering the
actions of others, which is characteristic of level-2 reasoning.
Classification:
2
```

Although GPT-4's classification capabilities are impressive, there are still instances where it commits obvious blunders. Consider the message in Figure 2.19, written by a subject who is attempting to randomise her response by coordinating with her teammate. The message clearly indicates that the subject is a level-0 thinker. Yet, GPT-3.5 classifies this message as level-2, and GPT-4 classifies it as level-1 under both prompts (with CoT and n-shot).

Both models misinterpret the attempt to coordinate with a teammate as a strategic play, and consequently, they erroneously argue for the player to be classified as either level-1 or level-2 (see Appendix 2.C.3, Figure 2.73 for the rationales for the message in Figure 2.19 and other additional examples). Although both models accurately understand what constitutes a teammate, since they perceive this message as reflecting a basic understanding of the game mechanics, rather than a random play or an action unrelated to the task, they eliminate the possibility that the subject is level-0. This suggests that the model still needs this specific type of instance to be provided as an example or described as part of the characteristics of level-0 play in order to classify it correctly[26]. Furthermore, it illustrates that when the model is unable to classify a message under any specific category, it considers other types of behaviours that can be considered strategic to determine the player's level of thinking –even if such behaviours, like attempting to coordinate with a teammate, might not truly reflect sophisticated strategic behaviour.

Figure 2.19: Level-k I - Example message 4

```
i think its probably better if we dont pick the same colour, so if i pick blue you
should pick red
```

ferred message as level-2, then uses this classification to inform its assessment of the message in question, and then, also classifies it as level-2.

[26]For instance an additional instructions such as "Refrain from considering any argument that involves the behaviour of the player's teammate as strategic"

2.7.5 Results

		no-CoT		**CoT**	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	O	70.2	64.5	64.1	64.5
	O_+	68.8	75.2	66.1	70
GPT-4	O	79.5	84.4	79.5	**91.3**
	O_+	76.9	79.9	76.3	89.5

Table 2.13: Overall accuracy of level classification

For GPT-3.5, n-shot O underperforms compared to 0-shot O in no-CoT treatment but slightly outperforms its 0-shot counterpart in CoT treatment. N-shot prompting improves GPT-3.5's performance for both prompts only when CoT prompting is present. However, CoT prompting generally worsens the performance of GPT-3.5, except for n-shot O. For GPT-3.5, the performance under O_+ is better than under O in most cases.

For GPT-4, n-shot prompting consistently improves the model's performance. While in 0-shot treatments, CoT prompting has little to no effect, it significantly boosts the performance of GPT-4 in n-shot treatments. Overall, CoT prompting in conjunction with n-shot prompting achieves the best performance for either prompt. Furthermore, GPT-4's performance under O is consistently better than its performance under O_+ for all treatments.

GPT-4 consistently outperforms GPT-3.5. While GPT-3.5's performance benefits from additional context and level-k theory emphasis, GPT-4's performance declines under the same conditions. This highlights a distinct difference in how the two models process and benefit from additional context. The best performing result is achieved with GPT-4 using CoT and n-shot prompting techniques for prompt O, with an accuracy of 91.3%.

2.8 Level-k II: Asymmetric-Payoff Coordination Games

2.8.1 Game and Data

Finally, van Elten and Penczynski (2020) study asymmetric-payoff coordination games (APC) introduced by Crawford et al. (2008, henceforth CGR) and provide textual data supporting the result that the incidence of level-k reasoning is low in symmetric, pure coordination games and high in asymmetric, "battle of the sexes"-type coordination games. The dataset is of particular interest, because its text analysis involves the classification of non-trivial level-0 beliefs.

Table **??** describes the four X-Y games and four Pie games. In contrast to payoff-symmetric games (in bold), payoff-asymmetric games feature a higher coordination pay-

X-Y games (CGR notation)	a	π_1, π_2	**Pie games** (CGR notation)	a	π_1, π_2
Symmetric Payoffs (**SL**)	X	5, 5	Symmetric Payoffs (**S1**)	L ($)	5, 5
	Y	5, 5		R (#)	5, 5
				B (§)	5, 5
Slight Asymmetry (ASL)	X	5, 5.1	Symmetric Payoffs (**S2**)	L ($)	6, 6
	Y	5.1, 5		R (#)	6, 6
				B (§)	5, 5
Moderate Asymmetry (AML)	X	5, 6	Moderate Asymmetry (AM2)	L ($)	5, 6
	Y	6, 5		R (#)	6, 5
				B (§)	6, 5
Large Asymmetry (ALL)	X	5, 10	Moderate Asymmetry (AM4)	L ($)	6, 7
	Y	10, 5		R (#)	7, 6
				B (§)	7, 5

Table 2.14: Payoff structure of coordination games.

off π for one of the two players, depending on the action on which they coordinate. The miscoordination payoff is 0 for both players. The choice is between letters X and Y in the X-Y games and between 3 pie slices (L, R, B) which are identified by ($, #, §) and of which B is uniquely white.

The dataset consists of 851 messages gathered through intra-team communication as described in Section 2.7.1. All messages are in German. The benchmark classifications are derived from the agreed assessments of two RAs[27]. The RAs provide a lower bound and an upper bound for the level of reasoning in each message. They also identify whether any label or payoff salience argument is present, and if so, classify the type of salience.

Tables 2.15a and 2.15b show the distributions of the lower and upper bound levels where L_n indicates the level-n strategic thinking. Tables 2.16a and 2.16b show the distributions of the benchmark classifications for label and payoff salience. In these tables, "$\sim$" indicates indifference to salience or payoff salience, "no" signifies that there is no mention of a payoff or label salience in the text, H and L denote high and low payoff salience respectively, and $\S$, $\#$, $\$$, X and Y represent the label salience for the game choice with the same tag.

	L_0	L_1	L_2	L_3	L_4	L_5
f	.504	.334	.143	.016	.002	–

(a) Lower Bound

	L_0	L_1	L_2	L_3	L_4	L_5
f	.341	.37	.234	.042	.006	.007

(b) Upper Bound

Table 2.15: Level distributions of the benchmark

Like the RAs, we instruct the GPT models to classify lower and upper bounds of the messages' levels of reasoning as well as the label- and payoff-salience of the level-0

[27] See van Elten and Penczynski (2020) for details

belief. The results from the supervised machine learning approach in Penczynski (2019) will provide an additional computerised benchmark.

	§	#	$\sim$	\$	X	Y	no
f	.159	.025	.024	.107	.103	.002	.581

(a) Label salience

$\sim$	H	L	no
.722	.228	.039	.008

(b) Payoff salience

Table 2.16: Distributions for label and payoff salience classifications of the benchmark

Each participants goes through all eight games prior to any communication exchange. We dropped those messages in which subjects refer to a reasoning that they laid out in an earlier round, as we did not want the LLM to get into assigning reasoning from earlier messages as the RAs did. In total 104 messages (12.2% of the data) are dropped.

2.8.2 Prompts

No basic prompt B is considered due to the same reasons outlined in Section 2.7. Instead, two prompts, O and O_+, which vary in their degree of reflecting the original classification instructions tailored for human annotators, are considered. These two prompts differ in their "Context" and "Classification Task" sections.

The original codebook is the most detailed and lengthy codebook we consider in this study (see Appendix 2.D.1). Given the non-negligible effect of the degree of information in the "Context" section on model performance observed in Section 2.7, we further investigated this effect with the current game by varying the level of information provided among prompts O and O_+.

"Context" sections of both prompts begin with an identical short subsection on game mechanics, followed by a subsection on the two coordination games played in the experiments where each game type (Pie and X-Y) is briefly introduced, and payoff tables for the variations of each game type is provided in a similar table format as in Table 2.14 in Markdown format. The two prompts significantly diverge in terms of the details each provides in the subsequent theory subsection. Prompt O closely follows the original instructions, whereas prompt O_+ provides a summary consisting of single sentence descriptions for each of the fundamental concepts detailed in the original codebook.

The "Context" section in prompt O outlines the theory in a series of subsections. The first subsection, "Salience in Decisions", describes the concept of salience in the games. This is followed by a "Level-k Model" subsection, which details level-0 distribution, level-0 belief, and population belief concepts, and describes the characteristics of each level from 0 to 4. Descriptions of each level's characteristics are unevenly supplemented with example messages: three examples for level-0 thinking, two for level-1, one for level-2, and no examples for higher levels. Some of these examples also entails cases for different types of salience. Nont of the level-0 examples provide label or payoff salience, level-1 examples only include label salience (one for each game type), and the level-2 example feature only a payoff salience example for the X-Y game.

The "Context" section of prompt O is followed by a "Classification Task" section that separates the tasks into two subtasks: classification payoff or label salience and classification of level of strategic thinking. The classification instructions under these subsections are also taken verbatim from the original codebook, and therefore are detailed and lengthy.

Differently from the prompts considered in "Level-k I" Section (Section 2.7), prompt O provides information related to the classification both in the "Context" section and in the "Classification Task" section. This is a natural outcome of our choice to closely mirror the format and instructions of the original codebook. The original instructions aim to first ensure that the human annotators get a general understanding of the theory pertaining to level-k modelling then proceeds this with the application of this theory for the outlined coordination games. This is provided in the "Context" section of O. After establishing the theory and its application, the codebook presents a detailed classification instructions. This is provided in the "Classification Task" of O[28]. Therefore, prompt O takes a human-centric approach to instruct the model by anthropomorphically assuming that the model needs to "understand" the theory before "understanding" how to classify messages based on this theory.

The original classification instructions, and therefore prompt O, include only five examples. These examples provide very limited coverage given the array of potential examples that can be generated by combining different levels of thinking with payoff or label salience, and with the two types of coordination games (X-Y and Pie Games). Furthermore, three out of the five examples that involve either payoff or label salience lack information indicating the appropriate classification for these categories. Consequently, it is debatable whether prompt O should be classified as a 5-shot prompt, given that the examples do not conform to the typical characteristics of a standard n-shot prompt. Such prompts should include a variety of examples for each possible category and clearly indicate the classification of each example. Instead, it would be more appropriate to consider these examples as supplementary instructional components rather than as explicit demonstrations for the classification task. Hence, given the nature of these examples, we classify prompt O as a 0-shot prompt.

In prompt O_+, we diverge significantly from the detailed theory section provided in prompt O by providing a very brief section on the level-k theory and decision salience. We provide as detailed information on level-0 thinking as we have done in prompt O, followed by a brief level-1 thinking description. We omit explicit description for each

[28]One deviations we made in the "Classification Task" section of O is that we switched the order of the two classification tasks outlined in the original codebook to emphasise the order with which we want the GPT models to classify the message which is to first classify the level-0 belief and decision salience then to classify the level of strategic thinking.

of the higher levels of thinking, and instead provide a brief generalised characterisation of a level-k thinker for $k > 1$. Moreover, in its "Classification Task" section, we only provide two lines of instructions that simply instruct the model to classify the payoff or label salience, and lower and upper bound of level of thinking. Additionally, we omit any examples in either of these sections. In other words, in prompt O_+, we do not concern ourselves too much with whether the model fully "understands" the concepts by considering a significantly brief way of presenting the background information and the classification instructions. Hence, prompt O_+ is similar in nature to the basic prompts we have considered in Sections 2.5 and 2.6: it provides minimal contextual information and places more weight on the model's existing knowledge on the topic by omitting detailed information on how to perform the task.

O_+ without any demonstrations establishes a 0-shot baseline for the prompt, and enables us to investigate the effect of incorporating demonstrations (n-shot prompting) in a modular fashion. We do so by developing an "Examples" section where we attempt to cover a broad array of possible case with a total of 16 demonstrations, four for each level from 0 to 3. None of the demonstrations are from the dataset, and generated by us. All examples are generated with the aim to explicitly depict their respective level of thinking. For instance, level-2 messages contain variations of the pattern "They think that I do <X>", and level-3 messages explicitly state variations of "They think that I think that they will do <X>". In level-0 examples, two provides examples of label salience, one provides payoff salience, and one provides neither label nor payoff salience. Level-1 examples consist of three cases of label salience and one case of payoff salience. For levels 2 and 3, two cases for each type of salience are provided. Three of the level-0 examples are for the X-Y game, and one for for the Pie game. Two of the level-1 examples are for the X-Y game, one for the Pie game, and one that can be applicable to either game. For levels 2 and 3, one example is for each game and two examples can be applicable to either game. Each example is followed by information on level-0 belief, label salience, payoff salience, and level classification.

The other sections of both prompts are identical ("General Task", "Role Persona", etc.). Except for the "Input Format" section, the other sections are as defined in Section 2.4.1 and can be examined in more detail in Appendix 2.D.2. The "Input Format" section is an additional part present only for the current prompts. Human annotators receive contextual information about the game, the team, and the subject's initial decisions, which are essential for understanding the message. The exact game being played, as shown in Figure 2.14, and the team of the player are necessary for both human annotators and GPT models to effectively grasp the context of the message. Consequently, unlike in our previous classification experiments, GPT is provided not only with the subject's message but also with information about subject's team, the game she is referring to, and her initial decisions. The "Input Format" section provides the template for the input that the model receives. By including this section in the prompt, we aim to help the model better orient itself when provided with the defined input string. This metric is also referred to in the literature as the Intersection over Union (IoU) metric or the Jaccard Index (Müller et al., 2022).

2.8.3 Examples

Figure 2.20: Level-k II - Example message

```
Team:  1
Game:  AML
Decision:  X
Message:  They might think that we think they are selfish.
```

	GPT-3.5				GPT-4			
	L_B	U_B	S_L	S_P	L_B	U_B	S_L	S_P
O	1	2	no	no	1	2	no	H
O_+	1	2	no	no	3	3	no	H
RAs	3	3	no	H	3	3	no	H

Table 2.17: Human and GPT classifications of message in Figure 2.20 using CoT prompting for n-shot O_+ and 0-shot O

In Figure 2.20, the input provided for the models is presented. Note that although the message is in English, the actual input provided to the models is in German. The subject, part of team 1, plays the AML variation of the X-Y where she will receive a payoff of 6 if the teams coordinate on choice ”Y” and 5 if the teams coordinate on choice ”X”.

Table 2.17 presents the classification results from RAs and GPT models utilising CoT prompting technique for 0-shot O and n-shot O_+ where the columns L_B, U_B, S_L, and S_P represent the lower bound of strategic thinking, upper bound of strategic thinking, label-salience, and payoff salience, respectively. Except for GPT-4 under prompt O_+, all model classifications are identical and differ from those made by the RAs for the lower and upper bounds. Moreover, GPT-3.5 fails to agree with the RAs on the payoff salience classification.

In Figure 2.21, the rationals for both models under 0-shot O and n-shot O_+ prompts are displayed. Their consecutive classification output is omitted to fit the figure within a single page. The reasoning under both models correlate with the length of the prompts: for relatively longer prompt O, a longer reasoning is generated by the models. The correlation between the length of the prompts and the length of reasoning produced by the models suggests that LLMs might be relying on a straightforward strategy of mirroring the extent of textual input provided in the prompts to generate what appears to be thoughtful

Figure 2.21: Level-k II - Generated rationals for 0-shot O and n-shot O_+ under CoT for message in Figure 2.20

GPT-3.5

Output using O

```
Reasoning:
The player's message seems to suggest that the other team might think that his team
considers them to be selfish.  This indicates a level of strategic thinking as the
player is considering the other team's potential thoughts and actions.  However, the
message does not explicitly indicate any salience or strategic reasoning based on the
game's framing.
```

Output using O_+

```
Reasoning:
The player seems to be considering the other team's perspective and how they might
perceive the decision-making process.
```

GPT-4

Output using O

```
Reasoning:
The message "sie werden bestimmt denken, wir würden über sie denken, sie wären
egoistisch" translates to "they will certainly think that we would think they are
selfish." This statement suggests a level of strategic thinking that involves
anticipating the other team's thought process about one's own team's perceptions.  The
player is considering that the other team might expect their own team to assume that
they (the other team) will act in their own self-interest, which is a characteristic of
higher-level strategic reasoning.
The message does not mention any preference for the label of the decision (X or Y) based
on its appearance or position.  Therefore, the player does not exhibit label salience.
The message implies that the other team might be expected to act selfishly, which
suggests that the player is considering the payoffs.  However, the message does not
explicitly state a preference for higher or lower payoffs for their own team.  It is
more focused on the belief about the other team's potential selfishness.  Therefore, the
player exhibits a level-0 belief about the other team's preference for higher payoff.
- Lower Bound:  The player is engaging in strategic thinking beyond level-0, as they are
considering the other team's potential beliefs about their own team's perceptions.  This
indicates at least level-1 reasoning.
- Upper Bound:  The player's message suggests they are considering what the other
team thinks about their own team's beliefs, which is indicative of level-2 reasoning.
However, there is no clear indication that the player is considering higher levels of
reasoning, such as level-3 or above, where they would be best responding to the other
team's level-2 reasoning.
```

Output using O_+

```
Reasoning:
The message "sie werden bestimmt denken, wir würden über sie denken, sie wären
egoistisch" translates to "they will certainly think that we would think about them,
they are selfish." This message suggests a level of reasoning about what the other team
will think about Team 1's beliefs regarding the other team's preferences.  The message
implies that the other team might believe that Team 1 thinks the other team is selfish,
which could mean that they would choose the option that gives them a higher payoff.
In the context of the AML game, where the payoffs are as follows:
- X: 5 Taler for Team 1, 6 Taler for Team 2
- Y: 6 Taler for Team 1, 5 Taler for Team 2
The message does not directly reference a preference for either label (X or Y) or
a specific payoff amount.  However, it does imply a belief about the other team's
potential selfishness, which is related to payoff salience.  The message suggests that
the other team may choose the option that gives them a higher payoff (Y for Team 2), and
Team 1 should consider this in their decision-making.
The message indicates a higher level of reasoning, as it involves thinking about what
the other team believes Team 1 thinks about them.  This is indicative of at least
level-3 reasoning, as it involves multiple layers of belief about the other team's
thought process.
```

reasoning. This pattern indicates that the models may primarily be extending or "echoing" the prompt's text rather than engaging in genuine analytical thinking based on its understanding of the concepts provided in the prompt (Bender et al., 2021).

GPT-3.5 provides significantly shorter rationales compared to GPT-4 under both prompts. Under O, GPT-3.5 briefly addresses both strategic thinking and decision salience, while under O_+, it only offers a rationale for strategic thinking. Although the model acknowledges strategic thinking and correctly notes that the subject considers the other team's perspective, the rationales lack specific justification for classifying the lower bound for her level of thinking as "1" and the upper bound of her level of thinking as "2".

GPT-4 provides a significantly more detailed analysis of the message under both prompts. It begins its rational by re-iterating the original message and translating it to English. This is a commonly observed behaviour in many other instances, yet there are also instances where the rational does not contain an English translation for the message (See Appendix 2.D.3 for additional examples). Under both prompts, it begins its rational by correctly identifying that the subject is considering the belief of the other team on the subject's belief about the other team. Then without providing a classification for her level of thinking, it proceeds into the analysis of the subject's payoff salience and level-0 belief. Under both prompts, it notes the absence of clear payoff salience but argues that the mention of the other team's "selfishness" indicates subject's belief about the other team's preference for a higher payoff, but only under prompt O, the model explicitly uses the term "level-0 belief".

Under prompt O_+, GPT-4 correctly identifies the pattern "They think that we think that they are selfish" with statement "The message (...) involves thinking about what the other team believes Team 1 thinks about them" and correctly classifies the player's level of thinking. In contrast, under prompt O, although GPT-4 makes the statements "The statement involves (...) anticipating the other team's thought process about one's own team's perceptions" and "The player is considering that the other team might expect their own team to assume that the other team will act in their own self-interest," it falsely classifies the player's lower bound of strategic thinking as "1" and the upper bound as "2".

Despite the analogous arguments concerning levels of thinking between the prompts, the misclassification of the model under prompt O supports the "Stochastic Parrots" view of LLMs as articulated by Bender et al. (2021). According to this view, LLMs primarily mimic language based on patterns and correlations learned during training. If GPT-4 truly "understood" the rationale it provided, it would have also classified the subject's level of thinking as "3" under prompt O as it did under prompt O_+. Alternatively, challenging the 'Stochastic Parrot' interpretation, perhaps GPT-4 under prompt O correctly "understood" its rationale but exhibited either an imprecise grasp of the distinction among higher levels of thinking, or made a computational error in counting the number of iterative best responses detailed in its rationale.

In Figure 2.22, the subject is playing AM4 variation of the Pie game and argues for the "fair" choice of alternative # that offers a payoff of 6 to the subject over alternative §, which provides a payoff of 5, while under both alternatives a payoff of 7 is provided for the other team players. The unmentioned alternative $ offers a higher payoff of 7 for

Figure 2.22: Level-k II - Example message 2

```
Team:  2
Game:  AM4
Decision:  #
Message:  I would take the second one (7.6) here because it is fairer than (7.5).
```

	GPT-3.5				GPT-4			
	L_B	U_B	S_L	S_P	L_B	U_B	S_L	S_P
O	1	1	no	H	1	1	no	H
O_+	1	1	no	H	1	1	no	H
RAs	0	0	no	no	0	0	no	no

Table 2.18: Human and GPT classifications of message in Figure 2.22 using CoT prompting for n-shot O_+ and 0-shot O

the subject's team and a lower payoff of 6 for the other team. The subject is classified as a level-0 thinker by RAs but as a level-1 thinker by both models under both prompts. Furthermore, while the RAs did not identify any label or payoff salience, both models under both prompts classified the message as demonstrating high payoff salience.

Figure 2.23 provides the rationales for both models under both prompts. All rationales argue that "fairness" should be seen as a type of payoff salience. GPT-4 under prompt O states that the subject's preference for a payoff of 6 over 5 shows a preference for a higher payoff under "the context of fairness". Similarly, GPT-4 under prompt O_+ identifies the message as a "nuanced form of payoff salience". Classifying subject as displaying a higher payoff salience is the initial and crucial point where the RAs and the models differ. Because RAs did not identify any type of salience and therefore any type of level-0 belief, they have consequently classified the player as a level-0 thinker. On the other hand, because the model identifies fairness as a payoff salience and therefore establishes a level-0 belief for the subject, it classifies the subject as a level-1 thinker by interpreting the mention of fairness as an implicit acknowledgement of considering the other team's beliefs.

While attempting to argue that "fairness" implies level-1 thinking, GPT-4 under prompt O_+ makes contradictory statements. Initially, it recognizes the player's consideration of payoff fairness and asserts, "A level-0 player would not consider fairness or the other team's payoff". Subsequently, however, it remarks, "The player's message does not explicitly indicate that they are considering what the other team might think or do, which would be characteristic of higher levels of reasoning (level-1 or above)." Yet despite this apparent contradiction, it concludes "Therefore, the message suggests a lower bound of level-1 reasoning, as the player is best responding to a belief about the other team's potential preference for fairness". On one hand, these conflicting statements in its reasoning might once again suggest that the model does not truly understand what it's stating, align-

Figure 2.23: Level-k II - Generated rationals for 0-shot O and n-shot O_+ under CoT for message in Figure 2.22

GPT-3.5

Output using O
Reasoning:
The player suggests choosing the second option because it is fairer than the first option. This reasoning indicates a preference for the option that provides higher payoffs, which suggests a level of strategic thinking. The player is considering the fairness of the payoffs, indicating a level of strategic reasoning.
Output using O_+
Reasoning:
The player's message indicates a preference for the second option due to its fairness compared to the first option. This suggests a form of payoff salience. The reasoning in the message seems to exhibit a level-1 understanding, as the player is considering the fairness of the payoffs and how the other team might perceive the decision.

GPT-4

Output using O
Reasoning:
The player from Team 2 has suggested the decision '#' in the AM4 variation of the Pie game. The AM4 game has an asymmetric payoff structure where the decision '#' yields a payoff of 7 Taler for Team 1 and 6 Taler for Team 2, while the decision '§' yields a payoff of 7 Taler for Team 1 and 5 Taler for Team 2. The player's message, when translated from German, reads "here I would take the second one (7,6) because it is fairer than (7,5)." This message indicates a concern for fairness and a preference for a more balanced distribution of payoffs between the teams.
The player's message does not explicitly mention any label salience, such as a preference for the visual distinction of the '§' slice in the Pie chart. Therefore, we can infer that the player does not exhibit label salience in this decision.
The player's message does indicate payoff salience. The player prefers the decision '#' because it offers a more balanced payoff (7,6) compared to the decision '§' (7,5), which would give a lower payoff to their own team. This suggests that the player is attracted to the action that yields a higher payoff for their own team, but within the context of fairness.
In terms of the level of strategic thinking, the player's message suggests a level-1 reasoning. The player is considering the payoffs and is choosing the decision that seems fairer, which implies a best response to a belief about the other team's action. The player does not consider higher-order beliefs about what the other team might think about their team's decision-making process. Therefore, the lower bound of reasoning is level-1, as the player is best responding to a belief about the other team's likely action based on fairness. The upper bound is also level-1, as there is no indication of higher-order strategic thinking in the message.
Output using O_+
Reasoning:
The player's message suggests choosing the option labeled '#' in the AM4 variation of the Pie Game. The player justifies this choice by stating that the payoff of (7,6) is fairer than (7,5). This indicates that the player is considering the distribution of payoffs between the two teams and prefers the option where the difference in payoffs is smaller, which is a consideration of fairness rather than simply choosing the highest individual payoff.
The player does not mention any label salience, such as preferring the symbol '#' over '$' or '§' for any reason other than the associated payoffs. Therefore, the player does not exhibit label salience.
The player's message does show payoff salience, as they are considering the fairness of the payoffs, which is related to the magnitude of the payoffs for each team. The player prefers the option where their team gets a slightly lower payoff (6) compared to the other team's payoff (7), which is the fairer distribution in their view. This suggests a preference for high payoffs but with a consideration for fairness, which is a nuanced form of payoff salience.
In terms of the level of reasoning, the player is considering the fairness of the payoffs, which suggests a level of strategic thinking beyond level-0. A level-0 player would not consider fairness or the other team's payoff. The player's message does not explicitly indicate that they are considering what the other team might think or do, which would be characteristic of higher levels of reasoning (level-1 or above). Therefore, the message suggests a lower bound of level-1 reasoning, as the player is best responding to a belief about the other team's potential preference for fairness.
The upper bound of reasoning is not clearly indicated in the message. The player does not express any anticipation of the other team's strategic thinking or multiple levels of iteration in their decision-making process. Therefore, the upper bound of reasoning is also level-1, as there is no evidence to suggest that the player is engaging in higher-level strategic thinking.

ing with the "statistical parrot" critique. On the other hand, such contradictions could indicate that the model is "thinking out loud", weighing arguments for and against the case of the subject being a level-1 thinker. Therefore, the model's contradictions may not necessarily reflect a lack of understanding but rather an exploration of uncertainty in classifying the subject's level of reasoning.

If fairness is viewed as a preference independent of individuals' beliefs, GPT's classification of this message is incorrect. However several studies argue that fairness is primarily defined by the people's belief about other's expectations (Bicchieri and Chavez, 2010; Castelli et al., 2014), and the weight placed on considering others' beliefs determines a person's degree of fairness concerns in her decision-making process (Takagishi et al., 2010). Thus, GPT's claim that fairness consideration is indicative of a higher level of reasoning has its support in the literature. But more importantly, this highlights the ambiguities inherent in natural language, complicating the exact classification of a player's level of thinking based on their message, and possibly a more appropriate classification for this message would be a lower bound of 0 and an upper bound of 1 for the subject's level of thinking[29].

2.8.4 Results

Given that the benchmark classifications for levels of thinking are defined as an interval, with a lower and an upper bound across six potential levels in any given instance (message, the model's task is a multi-class, multi-label classification. In this framework, a given instance can be classified as positive under multiple classes. This opens up the possibility for a classification to be partially correct. A measurement that requires an exact match with the actual interval might overlook this partial correctness and is therefore considered a harsh metric (Sorower, 2010). A more nuanced accuracy metric commonly used in multi-label classification is the ratio of correctly predicted labels to the total number of labels –both predicted and actual– averaged across all instances (Godbole and Sarawagi, 2004). This measurement is also referred to in the literature as the Intersection over Union (IoU) metric or the Jaccard Index (Müller et al., 2022). Although IoU can be calculated for every class and averaged (either macro or weighted), this approach has been criticised for not adequately addressing correlations among different classes. Therefore, to assess the classification performance of our models, we have chosen to employ the instance-based accuracy metric as introduced by Godbole and Sarawagi (2004).

Table 2.19 presents the agreement rates of models' classification with the RA classifications for each prompt under each treatment. GPT-3.5 consistently underperforms

[29] A similar argument applies to the concepts of selfishness and generosity. A subject calling herself selfish may be less likely to be assumed to consider other players' beliefs, whereas a subject calling herself generous may be more likely to be assumed to consider other players' beliefs. Interestingly, while RAs classified the lower bound of the level of thinking as 0 for these two types of a basic messages, they classified the upper bound of the level of thinking of a subject calling herself generous as 1 and the upper bound for a subject calling herself selfish as 0. A similar classification pattern is also observed with GPT-4 under prompt O_{+}. See Example Messages 4 and 5 in Appendix 2.D.3 for details.

		no-CoT		CoT	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	O	44.2	–	49.4	–
	O_+	52.3	62.6	55.3	65.3
GPT-4	O	64	–	69.2	–
	O_+	65.6	68.5	66.7	**72.8**

Table 2.19: Level classification accuracy (Jaccard)

GPT-4. CoT and n-shot prompting techniques consistently improve the performance of either model. For GPT-3.5, 0-shot O_+ consistently outperforms 0-shot O. This indicates that GPT-3.5 does not benefit from few basic examples and significantly more detailed instructions as provided in prompt O. Similarly, for GPT-4, 0-shot O_+ outperforms 0-shot O in the no-CoT treatment. Yet, the opposite is true in the CoT treatment. Hence, the detailed descriptions and examples present in prompt O enhances CoT prompting's positive effect on GPT-4's performance more than CoT's effect on 0-shot O_+. However, when CoT prompting is used in conjunction with n-shot prompting, O_+ generates the best performing classification results for either model.

Penczynski (2019) classified the same text corpus using a machine learning algorithm (MLA) that employs bag-of-words features. He evaluates the performance of his model by calculating its accuracy via an exact match of the lower bound of the level of thinking between human annotators and MLA classifications. The best-performing MLA model achieves an accuracy rate of 67%. GPT-3.5 achieves a lower bound level of thinking classification accuracy of up to 65.5%. Hence, MLA model's out-of-sample performance outperforms GPT-3.5's best performing prompt. Conversely, GPT-4, except for prompt O under no-CoT treatment (66.4%), outperforms the MLA classification. Furthermore, GPT-4 under prompt n-shot O_+ with CoT achieves the highest accuracy for the lower bound of level of thinking classification with 76% (see Appendix 2.D.4 for the complete set of classifications).

		no-CoT		CoT	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	O	34.9	–	44.1	–
	O_+	35.9	**72.5**	48.9	66.2
GPT-4	O	61.1	–	69.5	–
	O_+	58.6	71	62.7	69.4

Table 2.20: Payoff salience classification accuracy

Table 2.20 presents performance of both models for each prompt for payoff salience classification accuracy. n-shot prompting consistently improves the performance of both models. While CoT prompting consistently improves both model's performance for both prompts in 0-shot treatments, it deteriorates them in n-shot treatments. Consequently, best performing classifications are observed in n-shot no-CoT treatment for both models. In no-CoT treatment, improvement enabled by n-shot prompting is much larger for GPT-3.5. This results in GPT-3.5 to outperform GPT-4 in n-shot no-CoT treatment and to generate the overall best performing classification.

GPT-3.5 consistently performs worse under prompt O which indicates that it fails to benefit from its detailed instructions, whereas GPT-4 performs better under O compared to O_+. These performance differences under prompt O becomes more distinct in the CoT treatment. This indicates that while detailed information worsens GPT-3.5's reasoning, it enhances GPT-4's for the classification of payoff salience. Moreover, for GPT-4, in CoT treatments, O slightly outperforms n-shot O_+. This suggests that GPT-4's rational and its consecutive payoff salience classification benefits equally from detailed instructions without explicit examples as from explicit demonstrations in terms of both the effectiveness of rational and the accuracy of the payoff salience classification.

		no-CoT		**CoT**	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	O	40.5	–	58.6	–
	O_+	41.8	47.1	59.2	67.5
GPT-4	O	62.2	–	83.9	–
	O_+	44.4	78.4	78.7	**86.2**

Table 2.21: Label salience classification accuracy

In Table 2.21, the performance of both models for each prompt for label salience classification accuracy is presented. On average, the models are observed to perform better at label salience classification than payoff salience classification. CoT prompting consistently improves both models' performance for both prompts. Similarly, n-shot prompting consistently improves both models' performance for O_+, especially for GPT-4 in no-CoT treatment.

While GPT-3.5 performs slightly better under 0-shot O_+ relative to O, GPT-4, as in payoff salience classification, benefits from the detailed instructions of prompt O and performs better under prompt O compared to under 0-shot O_+. However, contrary to both models' behaviour in payoff salience classification, their performance improves under n-shot O_+ when CoT prompting is introduced. Consequently, the highest classification performance is achieved by GPT-4 under n-shot O_+ with CoT prompting.

2.9 Costs and benefits

Table 2.22 presents both monetary and time costs to classify 100 messages in Promise I. Costs are given in terms of USD and time is shown in minutes. The monetary costs for both GPT models are based on the number of input and output tokens, with output tokens being twice as costly as input tokens. This is reflected in the monetary cost difference between the no-CoT and CoT treatments. Furthermore, the token costs for GPT-4 are approximately 20 times higher than those for GPT-3.5. This is reflected in higher costs for each prompt under GPT-4 compared to GPT-3.5.

		no-CoT		**CoT**	
		$Money$	$Time$	$Money$	$Time$
GPT-3.5	B	0.04	4.5	0.06	3.4
	O_S	0.06	2.6	0.08	6.4
	O_N	0.06	4	0.08	6.7
	O_W	0.06	2.3	0.08	6.6
GPT-4	B	0.5	1.1	0.9	10.9
	O_S	0.7	1.4	1	11.5
	O_N	0.6	1.5	1	12.2
	O_W	0.6	1.3	1	13.2

Table 2.22: Promise I: Money and time costs per 100 messages in USD and minutes.

Given GPT-4's significantly larger parameter count, its time to generate the first token and time per output token are expected to be longer than GPT-3.5's (Talamadupula, 2024). Yet, for no-CoT treatments, GPT-3.5's time costs are two to three times higher. This may be due to network congestion at the time of inference. Since GPT-3.5 is the underlying model for the free version of ChatGPT, it likely receives significantly more requests, leading to longer queuing times. On the other hand, for CoT treatments, GPT-4's time costs are two to three times higher than GPT-3.5's. As shown in Examples sections of our experiments, GPT-3.5 provides much shorter rationales, this in turn results in the observed time differences between the two models.

HX's classification methods, whether content-based or gamified, incurred costs of approximately \$1000 to classify 38 messages within 2 hours. Extrapolating from this, classifying 100 messages would cost around \$2500 and take about 5 hours. In contrast, given the high classification accuracy of GPT-4 (ranging from 92% to 97%) and its desired responsiveness to variations in instructions, the use of GPT-4 presents a more appealing option than employing twenty-five human annotators in a lab to classify messages.

Prompts in Promise I are the shortest, whereas those in Level-k II are the longest, leading to higher input token costs for Level-k II. Additionally, the examples from the rational provided for each experiment suggest that in GPT-4, longer prompts tend to correlate with more extensive rationales. Consequently, in CoT treatments, Level-k II incurs higher output token costs and time costs compared to other experiments. Furthermore, in Level-k II experiments, we instruct the models to generate four distinct classification outputs, as opposed to a single output in other experiments, which further increased time and monetary costs.

In Table 2.23 displays the costs for each model under each prompt type in the Level-k II experiments. For both models, monetary costs can be up to ten times higher than those observed in the Promise I experiments. However, the time costs for GPT-3.5 do not exceed those in Promise I, and for prompt O_+, they are consistently lower, potentially again due to the fluctuating demand for GPT-3.5. On the other hand, time costs for GPT-4 are significantly higher, with the average time of 42.7 minutes to classify 100 messages using prompt O under CoT treatment.

Under either model, due to the significantly larger prompt size of O, it is consistently more costly both in terms of time and money compared to O_+. Yet, together with the results from Table 2.19, we observe that the most costly prompt is not necessarily the best performing one.

		no-CoT		**CoT**	
		$Money$	$Time$	$Money$	$Time$
GPT-3.5	O	0.5	3.3	0.5	6.3
	O_+	0.3	1.5	0.3	2.8
GPT-4	O	5.2	5.2	6.6	42.7
	O_+	2.7	3.8	3.4	25.8

Table 2.23: Level-k II: Money and time costs per 100 messages in USD and minutes.

2.10 Discussion

Tables 2.24a and 2.24b present the highest and lowest accuracy achieved for each model in each experiment. P_I and P_{II} refer to sections Promise I and II, and L_I and L_{II} refer to sections Level-k I and II. For P_I, these values are calculated by averaging the best or worst performances over the benchmarks.

	P_I	P_{II}	L_I	L_{II}
GPT-3.5	85.6	75.7	75.2	65.3
GPT-4	96	88.7	91.3	72.8

(a) Highest

	P_I	P_{II}	L_I	L_{II}
GPT-3.5	74	53.2	64.1	44.2
GPT-4	83.2	67.5	76.3	64

(b) Lowest

Table 2.24: Highest and lowest accuracy across experiments

Across all experiments, GPT-4 consistently outperforms GPT-3.5 in peak performance, and achieves an accuracy above or near 90% in all experiments except L_{II}. By switching from using GPT-3.5 to GPT-4, the best performance improves by 9 to 16 percentage points, and the lowest performance improves by 9 to 20 percentage points. Moreover, except for P_{II}, GPT-4's lowest performance is comparable to, or exceeds, the top performance of GPT-3.5. Therefore, substantial gains in classification performance can be achieved as the models scale.

In P_I and P_{II}, although the classification concept is the same and the classification instructions provided are either identical or similar, there is a noticeable decline in the top performance from P_I to P_{II} . In P_I, by design, subjects are incentivized to provide their intend in a single message. Consecutively, often the messages are explicit and thus relatively straightforward to classify. In contrast, in P_{II} subjects engage in conversation where they try to persuade others to take certain actions without providing their intentions, affirm or oppose previously made statements by others, explicitly state or hint at what they intend to do, and so on. Consequently, chat messages encompass a broader array of linguistic patterns and colloquial, often ambiguous language. Additionally, chat messages are interdependent, and hence, are required to be evaluated within the full context of the chat. This conversational and contextual complexity makes the classification task in P_{II} more challenging, leading to comparatively lower performance by the models.

The classification of strategic thinking levels requires understanding and applying specific game theory concepts, notably level-0 beliefs and iterative best responses, making it more cognitively demanding than classifying promises. Additionally, the coordination game in L_{II} presents a more challenging classification task compared to the voting game in L_I. In the voting game, identifying level-1 and level-2 thinkers is relatively straightforward: subjects voting according to their received signals are likely level-1 thinkers, while those consistently voting red and considering the opposing team's behaviour are likely level-2 thinkers. These explicit cues simplify the classification task for models by reducing the need for a deep understanding of level-0 beliefs and best responses.

In contrast, the coordination games of L_{II} involve more complex reasoning patterns, such as "We think that they think that we think," which can perplex even expert annotators. The cyclical nature of these arguments requires models first to accurately identify in the message the level-0 belief and the label or payoff salience associated with it, then to identify the number of iterative best responses applied to this belief. This multi-step process increases the likelihood of classification errors at steps prior the classification of level of thinking and adds complexity to the task. Furthermore, instead of explicitly stating beliefs about the other team's intentions, such as "the other team wants the higher payoff for themselves," players might use indirect language, describing the other team as "selfish" or "offensive". Similarly, they might imply additional levels of reasoning with statements like "but they may also think this way". The use of such implicit linguistic patterns makes the classification task even more challenging for models. Additionally, unlike in L_I where a single level of thinking is classified, L_{II} requires models to classify both a lower and an upper bound for the subject's level of thinking, adding another layer of complexity. Consequently, it is no surprise that models exhibit lower performance in L_{II}.

Comparing the best and worst performances for each experiment reveals that different prompts and prompting techniques produce a variation of 8 to 20 percentage points for GPT-4 and 11 to 22 percentage points for GPT-3.5. This variation underscores the critical role of prompt engineering and suggests a substantial potential for performance gains with prompt optimisation. Moreover, it underscores the necessity for social scientists to rigorously explore prompt variations and evaluate existing prompting techniques for their effectiveness in improving the model's performance on each distinct task relevant to social sciences.

In P_I, we relaxed the conditions for classifying promises across a series of prompts. GPT-4's classifications reflected this instructional variation, whereas GPT-3.5 did not respond to it. In P_{II}, we considered an alternative to the original codebook's classification instructions, where "empty talk" was implicitly defined as any message not meeting the criteria for a "promise". Instead, we provided explicit conditions for classifying "empty talk". This alternative prompt led to an average improvement of 15 percentage points in 0-shot treatments and 4.2 percentage points in n-shot treatments for GPT-4, and 13.7 and 17.2 percentage points improvements, respectively, for GPT-3.5. In L_I, an alternative prompt that included more detailed context information resulted in a 3 percentage point decrease in accuracy for GPT-4 on average, but a 4.1 percentage point increase for GPT-3.5. The increase in GPT-3.5 was primarily observed with the n-shot treatment. Lastly, in L_{II}, we experimented with one prompt providing nearly all information from the original codebook, resulting in a lengthy 2200-word prompt, and another with minimal information that was considerably shorter, around 900 words[30]. For GPT-4, the longer, more detailed prompt slightly reduced accuracy by an average of 1.8 percentage points, whereas for GPT-3.5, it resulted in an average decrease of 9.3 percentage points. In summary, we observed that:

1. GPT-4 is responsive to small yet crucial changes in the instructions, while GPT-3.5 is not (P_I).

2. Providing more comprehensive classification instructions improves performance for both models (P_{II}).

3. Additional context information at best slightly worsens GPT-4's performance and produces variable effects on GPT-3.5's performance, ranging from slight improve-

[30]The n-shot variation included an additional "Example" section, adding about 500 words.

ments to significant reductions as the complexity of the task increases (L_I & L_{II}).

		P_I	P_{II}	L_I	L_{II}
GPT-3.5	CoT	$+$	$-$	$\sim$	$+$
	n-shot		$+$	$\sim$	$+$
	n-shot & CoT		$\sim$	$\sim$	$+$
GPT-4	CoT	$\sim$	$\sim$	$\sim$	$+$
	n-shot		$\sim$	$+$	$+$
	n-shot & CoT		$+$	$+$	$+$

Table 2.25: Overall effect of treatments on model performance

In Table 2.25, the effect of CoT and n-shot prompting on each experiment are presented. A "$+$" indicates that the treatment increased model performance across all prompts, a "$-$" indicates that it decreased performance across all prompts, and a "$\sim$" indicates that the treatment had mixed effects across different prompts.

Overall, prompting techniques generally improve model performance. Notably, except for the consistent negative effect of CoT prompting on GPT-3.5 under P_{II}, these techniques invariably improve performance of both models for at least one of the considered prompts. On the other hand, for mixed effect cases ($\sim$), the technique's average effect over the prompts is negative for GPT-3.5 but positive for GPT-4. This suggests that on average, prompting techniques are less effective at improving classification performance for GPT-3.5 compared to GPT-4. Furthermore, n-shot prompting is more often effective at improving performance of each model than CoT prompting, and combining n-shot and CoT prompting consistently leads to improvements only in GPT-4's performance.

Table 2.26 shows under which treatments best and worst performance for each experiment are observed. "$Best$" and "$Worst$" columns under each experiment indicate the best and worst performance cases, respectively. A "$\checkmark$" indicates that CoT or n-shot were applied, while a "$\times$" denotes their absence. Specifically, a "$\times$" for n-shot treatment denotes the 0-shot treatment and for CoT treatment, it represents the no-CoT treatment. A "$\sim$", only used under P_I, indicates that the best performance was achieved with the CoT treatment in three out of five benchmarks.

Except for P_I, best and worst performances are consistently achieved in n-shot and 0-shot treatments, respectively, for either model. For GPT-4, the best performance is also invariably achieved in CoT treatments; and except for L_I, the worst performance is consistently observed in no-CoT treatments. For GPT-3.5, in P_I and L_{II}, best and worst performances are observed in CoT and no-CoT treatments. Conversely, in P_{II} and L_I, the best performance occurs in CoT treatments and the worst in no-CoT treatments.

In Table 2.27, we present the average performance gains from switching from GPT-3.5

		P_I		P_{II}		L_I		L_{II}	
		$Best$	$Worst$	$Best$	$Worst$	$Best$	$Worst$	$Best$	$Worst$
GPT-3.5	CoT	✓	×	×	✓	×	✓	✓	×
	n-shot			✓	×	✓	×	✓	×
GPT-4	CoT	$\sim$	×	✓	×	✓	✓	✓	×
	n-shot			✓	×	✓	×	✓	×

Table 2.26: Active treatments for best and worst performance

to GPT-4 (Δ_{GPT}) and from incorporating n-shot prompting (Δ_n) under no-CoT and CoT treatments. Δ_{GPT} shows minimal variation between experiments under no-CoT treatments across experiments. However, when CoT prompting is introduced, its positive effect diminishes for P_I, increases for P_{II} and L_I, and remains the same for L_{II}. Δ_n is similar in P_{II} and L_I, but is approximately three times greater in L_{II} under no-CoT. Furthermore, the incorporation of CoT consistently increases Δ_n, underscoring a positive synergy between the treatments.

		P_I	P_{II}	L_I	L_{II}
no-CoT	Δ_{GPT}	13.3	13.2	10.5	13
	Δ_n		2.5	2.2	6.6
CoT	Δ_{GPT}	9	17.6	18	13
	Δ_n		6.3	7.1	8.1

Table 2.27: Average performance gains

Pan et al. (2023) suggest that learning tasks benefit from larger LLMs and examples (n-shot prompting), while recognition tasks see minimal gains from them. Isolating the effects of scale and demonstrations and examining Δ_{GPT} and Δ_n under no-CoT treatments, we observe that all experiments benefited from an increase in the scale of the LLM. While all experiments also derived some benefit from demonstrations, L_{II} showed the largest improvement. These findings indicate that all tasks involve elements of learning, and notably, the task of classifying level-k thinking in coordination games (L_{II}) may encompass a relatively larger learning component.

In this study, we adhered closely to the original instructions to minimize the effort required for researchers to utilize LLMs with existing codebooks tailored for human annotators. Given that the original codebooks for P_{II} and L_I included classification examples, n-shot treatment naturally aligns with the existing instructions. Furthermore, 0-shot-CoT does not conflict with our objective of adhering as closely as possible to the original codebooks, as it can be seamlessly integrated with existing instructions. Consequently, we selected n-shot prompting and 0-shot-CoT as our primary treatments.

Computer science literature typically discusses n-shot prompting and n-shot-CoT prompting –which includes both examples and their reasoning– separately, and to our knowledge, there is no documented experimentation on the combined use of n-shot prompting and 0-shot-CoT for classification tasks (Dong et al., 2022; Huang and Chang, 2022; Wei et al., 2022a). Our results suggest that combining n-shot prompting with 0-shot-CoT, without the need to provide rationales for the demonstrations, is a viable and less demanding alternative for GPT-4. Although we do not claim that this combination yields higher performance than n-shot-CoT –as we have not investigated such a comparison– we observe that it can be considered as an alternative for social scientists who want to consider prompts closely based on existing codebooks and prefer not to devise rationales for their examples, a feature that is not typically included in codebooks tailored for human annotators.

If we consider moving beyond the strict adherence to original codebooks, it might be possible to improve model performance by developing instructions specifically tailored for LLMs. One approach could involve segmenting the classification of the level of thinking into three subtasks: identifying level-0 beliefs, classifying salience type, and determining the level of thinking, using the output from each as input for the next, a technique commonly refereed as prompt chaining (Wu et al., 2022a; Anthropic, 2024). This method allows each subtask to have its own optimized prompt, focused solely on that specific aspect. Additionally, the classification of level of thinking could be further refined. In an initial step, the model could assess whether a message mentions others' actions or beliefs; messages without such mentions could be classified as level-0. Subsequently, messages that reference others' actions or beliefs could be analyzed in a separate prompt to determine the subject's level of thinking.

The current paper judges the classification performance of GPT relative to the benchmark of human classifications. We acknowledge that human classifications, even the ones agreed between 2 RAs, might deviate from an ideal "true classification". After all, these classifications are specific to the instructions given and might not be perfectly consistent due to learning, fatigue, etc.

High accuracies above 90% – as observed for some LLM classifications here – raise the question whether the human classification might not be the appropriate benchmark, because its own accuracy will likely be below 100% accuracy with the "true classification". For future comparisons, we intend to prepare "high-quality classification benchmarks" that represent the consensus of multiple diverse (expert, layman, LLM, ...) coders after sufficient deliberation.

Given GPT-4's out-performance over GPT-3.5, we are optimistic that LLM's will be a significant facilitator of future research efforts with text as data. Kaplan et al. (2020) show that as the models' get larger both in their parameter size and in their pre-training corpus, their performance will steadily increase. Hence, we expect the observed performance levels to further improve with larger language models. Furthermore, in the last few months since we conducted our experiments, the computing cost for GPT-4 significantly decreased (approximately 75% cheaper). Lastly, for more involved classification tasks where LLM's performance consistently stays less than desirable, it can at the very least be considered within a human-in-the-loop classification system to assist and improve human annotators' performance.

2.11 Conclusion

In this study, we investigated the classification of promises and levels of strategic thinking using GPT models. Our results showed that GPT-4 achieves accuracy levels comparable to human annotators and surpass the performance of traditional NLP methods. The comparison between GPT-3.5 and GPT-4 revealed that larger models are more adept at handling relatively more involved classification tasks and benefit more from n-shot and 0-shot-CoT prompting techniques. Furthermore, our investigations of n-shot and 0-shot-CoT prompting on model performance showed that demonstrations generally improved model performance, and particularly for GPT-4, asking the model to provide a step-by-step reasoning before making its classifications often led to improved task performance.

We investigated classification tasks that vary in difficulty both conceptually (promise vs strategic thinking) and in linguistic diversity (standalone message vs conversation), and showed that as task difficulty increases, the model size and prompting techniques became more crucial in achieving performance levels comparable to those of human annotators. Additionally, the consistent improvement in model performance due to increased model size and n-shot prompting suggested that all tasks we considered incorporate a learning component. These effects, and consequently the learning aspect of the classification tasks, are observed to be more pronounced in the classification of strategic thinking compared to promises, especially with the classification of strategic thinking in coordination games, which additionally required an understanding of level-0 beliefs and related payoff or label salience.

We explored adapting existing classification codebooks, tailored for human annotators, to serve as prompts for LLMs. This adaptation involved reframing and reformatting the instructions to better align with the models' input processing. Our findings indicated that relatively shorter codebooks, which neither contain nor require extensive theoretical background knowledge for the classification task, can be effectively leveraged as prompts with minimal reframing and restructuring. By using prompts that closely adhere to these original codebooks, GPT-4 can achieve satisfactory classification performance, with accuracy levels reaching or exceeding 90%. For longer codebooks, we observed that using the detailed background information verbatim in the prompts often hindered rather than improved the model's performance. Therefore, when reframing extensive codebooks, we recommend against anthropomorphising the model. Instead of providing explanations aimed at expanding its understanding of the topic –as one might do with a human annotator– focus should be placed on presenting textual patterns and linguistic cues that assist the model in a more mechanistic fashion during its classification task.

Lastly, we document that comprehensive classification instructions can be as effective as providing classification examples. This 0-shot method offers a viable alternative to n-shot prompting that relies on expanding a narrower set of classification conditions through examples. Furthermore, it mitigates issues associated with n-shot prompting, such as the model's sensitivity to the selection and order of examples, as well as the frequency of each category within these examples. Therefore, when revising the classification conditions in an existing codebook, researchers should aim to develop criteria with finer specificity with the intent to encompass a broader range of potential instances within each category of the dataset under analysis.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work I used GPT-4 in order to improve readability and language of the text. After using this tool, I reviewed and edited the content as needed.

Appendix to Chapter 2

2.A Promise I

2.A.1 Original Instructions

Figure 2.24: Original instructions used for content analysis

Your task:
You will be given a list of messages. Your task is to evaluate whether each of the messages is:

- A statement of intent or promise
- Empty Talk

The messages were written by participants in a previous experiment (Experiment I). To evaluate the messages, you need to first understand Experiment I. The pages beginning on page 2 describe Experiment I. Please read those pages carefully. The message writer is in the role of subject B.

It is important for you to know more about how to code the messages before you read the instructions. Here are your specific instructions for how you code the messages:

1. *(Weak Promise)* You should code a message as ‘‘A statement of intent or promise’’ if you think at least one of the following conditions is probably satisfied.
   (Strong Promise) You should code a message as ‘‘A statement of intent or promise’’ if you think at least one of the following conditions is certainly satisfied.
   a. The writer, subject B, indicates in the message he/she would do something favorable to subject A or refrain from doing something that hurts subject A.
   b. The message gives subject A reasons to believe or expect that subject B would do something favorable to subject A or refrain from doing something that hurts subject A.
2. (Weak Promise) You should code a message as ‘‘Empty Talk’’, if the message does not probably satisfy any of the above conditions.
   (Strong Promise) You should code a message as ‘‘Empty Talk’’, if the message does not certainly satisfy any of the above conditions.
3. You should independently code all messages. Do not discuss with anyone else in this room about how to code the messages.
4. Your job is to capture what had been said rather than why it was said or what effect it had. Think of yourself as a ‘‘coding machine.’’
5. When you complete the coding, go through the entire list of messages a second time to (1) review all your codes and revise them if needed for accuracy; (2) make sure you code every message.

Figure 2.25: Original instructions used for HX classification game

```
Your task:
In this experiment, you will be given a list of messages.  Your task in this experiment is to evaluate whether each of
the messages is:

      • A statement of intent or promise

      • Empty Talk

The messages were written by participants in a previous experiment (Experiment I). To evaluate the messages, you need
to first understand Experiment I. The next few pages describe Experiment I. Please read it carefully.  The message
writer is in the role of subject B.

Subjects were also told:
Weak Promise treatment:  ``You should classify a message as ‘Promise or Intent’ if, in your opinion, it includes any
statement of intent.’’
Strong Promise treatment:  ``You should classify a message as ‘Promise or Intent’ if, in your opinion, it is certainly
a promise.’’
```

2.A.2 Prompts

Figure 2.26: Prompt B^*

```
# General Task
- Evaluate a player’s message in an economic experiment game to determine if it constitutes a promise.
# Context
- Players:  Two players, A and B.
- Game Mechanics:
  + First, player A chooses between the two decisions ’IN’ and ’OUT’.
  + If player A chooses ’OUT’, each player receives $5.
  + If player A chooses ’IN’, player B chooses between the two decisions ’ROLL’ or ’DON’T ROLL’ (a die).
  + If player A chooses ’IN’ and player B chooses ’DON’T ROLL’, then player B receives $14 and A receives $0.
  + If player A chooses ’IN’, and player B chooses ’ROLL’, player B receives $10 and rolls a six-sided die to
determine player A’s payoff.  If the die comes up 1 (1/6 likelihood), player A receives $0; if the die comes up 2-6
(5/6 likelihood), player A receives $12.
  + In addition, each player receives $5 show-up fee for participating in the experiment.
- Communication:  Prior to Player A’s decision to ’IN’ or ’OUT’, player B has the option to send a message to player
A.
- Task:  Classify whether the message sent by player B constitutes a promise or not.
# Classification Task
- Classify player’s message as ‘1‘ if it’s a promise.
- Classify as ‘0‘ if it’s an empty talk.
# Constraint
- Follow the below output format.
# Output Format
   0/1
```

Figure 2.27: Prompt O_S, O_N and O_W

```
# General Task
- Evaluate a player's message in an economic experiment game to determine if it constitutes a promise.
# Context
- Players:  Two players, A and B.
- Game Mechanics:
  + First, player A chooses between the two decisions 'IN' and 'OUT'.
  + If player A chooses 'OUT', each player receives $5.
  + If player A chooses 'IN', player B chooses between the two decisions 'ROLL' or 'DON'T ROLL' (a die).
  + If player A chooses 'IN' and player B chooses 'DON'T ROLL', then player B receives $14 and A receives $0.
  + If player A chooses 'IN', and player B chooses 'ROLL', player B receives $10 and rolls a six-sided die to
determine player A's payoff.  If the die comes up 1 (1/6 likelihood), player A receives $0; if the die comes up 2-6
(5/6 likelihood), player A receives $12.
  + In addition, each player receives $5 show-up fee for participating in the experiment.
- Communication:  Prior to Player A's decision to 'IN' or 'OUT', player B has the option to send a message to player
A.
- Task:  Classify whether the message sent by player B constitutes a promise or not.
# Classification Task
(O_S)- Classify a message as a "promise or a statement of intent" if at least one of the following conditions is
**certainly** satisfied:
(O_N)- Classify a message as a "promise or a statement of intent" if at least one of the following conditions is
satisfied:
(O_W)- Classify a message as a "promise or a statement of intent" if at least one of the following conditions is
**probably** satisfied:
  + Player B indicates in the message he would do something favorable to player A or refrain from doing something that
hurts player A
  + The message gives player A reasons to believe or expect that player B would do something favorable to player A or
refrain from doing something that hurts player A.
(O_S)- If the message does not **certainly** satisfy any of the above conditions, classify player B's message as
"Empty Talk".
(O_N)- If the message does not satisfy any of the above conditions, classify player B's message as "Empty Talk".
(O_W)- If the message does not **probably** satisfy any of the above conditions, classify player B's message as
"Empty Talk".
- Capture what had been said rather than why it was said or what effect it had.
- Operate as a "coding machine".
# Classification Coding
- Code player's message as `1` if it's a promise.
- Code as `0` if it's an empty talk.
# Constraint
- Follow the below output format.
# Output Format
0/1
```

2.A.3 Additional Examples

Figure 2.28: Promise I - Example message 2

```
If you choose IN, and I roll, the chances of our getting the most $ are very high.  The likelyhood of my rolling
a 1 is small compared to the chances of rolling a 2â6.  So we both get cash.
```

	CD	C_S	C_W	G_S	G_W
$Class.$	0	1	1	1	1
f_1	–	54%	76%	56%	76%

Table 2.28: Human classifications for the message in Figure 2.28

	GPT-3.5		GPT-4	
	no-CoT	CoT	no-CoT	CoT
B	1	1	0	0
O_S	1	1	0	0
O_N	1	1	1	1
O_W	1	1	1	1

Table 2.29: GPT classifications for the message in Figure 2.28

Figure 2.29: Classifications with CoT for prompts O_W and O_S for message in Figure 2.28

GPT-4

Output using O_W
Reasoning:
- Player B is explaining the probabilities of the game outcomes if player A chooses 'IN' and player B chooses 'ROLL'.
- Player B is indicating that the likelihood of rolling a 2-6 is high, which would be favorable to player A as it would result in player A receiving $12.
- The message gives player A a reason to believe or expect that player B is leaning towards choosing 'ROLL', which is favorable to player A.
Classification: 1

Output using O_S
Reasoning:
- Player B is explaining the probabilities of the outcomes if player A chooses 'IN' and player B chooses 'ROLL'.
- Player B is stating that the likelihood of rolling a 1 is small, which implies that the chance of player A getting $12 is high.
- However, player B does not explicitly state what they will do; they are only discussing the probabilities and potential outcomes.
- There is no clear commitment or statement of intent to choose 'ROLL' or 'DON'T ROLL'.
Classification: 0

GPT-3.5

Output using O_W
Reasoning:
The message indicates that if player A chooses 'IN' and player B rolls the die, there is a high likelihood of both players getting cash. Player B is providing a reason for player A to believe that choosing 'IN' and player B rolling the die would result in a favorable outcome for both players.
Classification: 1

Output using O_S
Reasoning:
The message from player B indicates that if player A chooses 'IN' and player B chooses to roll, there is a high likelihood of both players getting cash. Player B is providing reasons for player A to believe that choosing 'IN' and player B choosing to roll would result in favorable outcomes for both players.
Classification: 1

Figure 2.30: Promise I - Example message 3

```
If you stay in, the chances of the die coming up other than 1 are 5 in 6 â pretty good.  Otherwise, we'd both be
stuck at $5.  (If you opt out)
```

	CD	C_S	C_W	G_S	G_W
$Class.$	0	1	1	0	1
f_1	–	73%	72%	< 50%	84%

Table 2.30: Human classifications for the message in Figure 2.30

	GPT-3.5		**GPT-4**	
	no-CoT	CoT	no-CoT	CoT
B	1	1	0	0
O_S	1	1	0	0
O_N	1	1	0	0
O_W	1	1	1	1

Table 2.31: GPT classifications for the message in Figure 2.30

Figure 2.31: Classifications with CoT for prompts O_W and O_S for message in figure 2.30

GPT-4

Output using O_W
Reasoning:
- Player B is explaining the probabilities associated with the 'ROLL' decision, indicating that there is a high chance (5 in 6) that player A will receive a favorable outcome ($12) if player B chooses to 'ROLL'. - Player B contrasts this with the guaranteed but lower payoff of $5 each if player A chooses 'OUT'. - Although player B is not explicitly stating what they will do, they are providing information that could influence player A's decision by highlighting the potential benefits of choosing 'IN' and relying on player B to 'ROLL'. - The message gives player A a reason to believe that choosing 'IN' could lead to a favorable outcome, which is indirectly suggesting that player B might be inclined to choose 'ROLL' to allow for that favorable outcome. Classification: 1

Output using O_S
Reasoning:
- The message from player B is providing information about the probabilities of the outcomes if player A chooses 'IN' and player B subsequently chooses 'ROLL'. - Player B is stating a fact about the likelihood of the die roll outcome, which is a 5 in 6 chance of player A receiving $12. - However, player B does not explicitly state what they will do if player A chooses 'IN'. There is no clear indication of player B's intended action, only a statement of the odds. - The message does not contain a clear promise or statement of intent to choose 'ROLL' or 'DON'T ROLL'. Classification: 0

GPT-3.5

Output using O_W
Reasoning:
The message provides reasoning to player A about the favorable outcome of choosing 'IN' and the likelihood of receiving a higher payoff. It also implies that player B would prefer player A to choose 'IN' for a better outcome for both players.
Classification: 1

Output using O_S
Reasoning:
The message provides reasoning to player A about the favorable outcome of staying in, indicating that the chances of the die coming up other than 1 are 5 in 6, which is pretty good. It also highlights the unfavorable outcome of opting out, where both players would be stuck at $5. Therefore, the message gives player A reasons to believe or expect that player B would do something favorable to player A.
Classification: 1

2.A.4 Additional Analysis

Weighted F1 ($F1_W$) and F1 macro averages for all four prompts across the five human classification baselines for both with (CoT) and without CoT (no-CoT) prompting are displayed in Tables 2.32 and 2.33, respectively.

		no-CoT					**CoT**				
		CD	C_S	C_W	G_S	G_W	CD	C_S	C_W	G_S	G_W
GPT-3.5	B	.65	.71	.74	.65	.88	.80	.83	.84	.80	.85
	O_S	.65	.71	.74	.65	.88	.76	.81	.85	.76	.92
	O_N	.60	.67	.70	.60	.83	.69	.75	.78	.69	.85
	O_W	.60	.67	.70	.60	.83	.73	.79	.82	.73	.89
GPT-4	B	.87	.87	.85	.87	.77	.90	.90	.88	.95	.79
	O_S	.93	.92	.90	**.97**	.81	**.97**	.92	.89	.92	.86
	O_N	.92	**.97**	.95	**.97**	.86	.92	.92	.95	.92	.90
	O_W	.83	.94	**.97**	.89	.87	.89	.94	**.97**	.89	**.93**

Table 2.32: Weighted F1

		no-CoT					**CoT**				
		CD	C_S	C_W	G_S	G_W	CD	C_S	C_W	G_S	G_W
GPT-3.5	B	.58	.62	.65	.58	.77	.78	.77	.80	.78	.76
	O_S	.58	.62	.65	.58	.77	.73	.78	.81	.73	.86
	O_N	.52	.56	.58	.52	.68	.64	.68	.70	.64	.74
	O_W	.52	.56	.58	.52	.68	.68	.74	.76	.68	.80
GPT-4	B	.86	.86	.83	.86	.70	.89	.89	.86	.95	.72
	O_S	.92	.92	.88	**.97**	.74	**.97**	.91	.88	.92	.80
	O_N	.92	**.97**	.94	**.97**	.80	.91	.90	.94	.91	.86
	O_W	.82	.94	**.96**	.88	.81	.88	.94	**.96**	.88	**.88**

Table 2.33: F1 Macro averages

2.B Promise II

2.B.1 Original Instructions

In Figure 2.32, the original instructions are presented with certain sections omitted that are not relevant to the prompt. Additionally, to accommodate the instructions on a single page, various line spacings have been reduced. The actual format of the original instructions is much easier to follow (see Arad et al. (2024) for further information).

Figure 2.32: Original instructions

```
In the excel file, you will find a list of messages written by participants in an online experiment on
individual and group investments.
The list consists of many conversations between groups of three participants.
Your task is to evaluate, for each conversation, whether each of the participants stated an intent or a promise
to take a particular course of action.
You will classify a participant's message in a conversation into one of the two categories:
  - A statement of intent or promise (1)
  - Other (0)
To evaluate the messages, you need to first understand the experiment.  The next page describes the experiment,
followed by more detailed instructions for your classification task.
(Explanation of the experiment is skipped)
Classification
Here are your specific instructions for how you code the messages.
1) You should code a participant's message (including his/her entire text in the conversation) as âA statement
of intent or promiseâ if you think at least one of the following conditions is satisfied.
  a.  The writer indicates in the message he/she would take a certain course of action.
  b.  The message gives the other participants reasons to believe or expect that that the writer of the message
would take a certain course of action.
2) You should code a message as âotherâ, if the message does not satisfy any of the above conditions.
3) You should independently code all messages.  Do not discuss it with anyone else.
4) Your job is to capture what had been said rather than why it was said or what effect it had.  Think of
yourself as a âcoding machine.â
5) When you complete the coding, go through the entire list of messages a second time to review all your codes
and revise them if needed for accuracy.  Make sure you code every message.
Examples
Let's illustrate how promises may look like:
Participant 1:  "all 200 then?"
Participant 2:  "yes"
The "yes" of Participant 2 is a promise.
Participant 1:  "Are we all just going to with max?â
Participant 2:  "agree"
The "agree" of Participant 2 is a promise.
It would not include statements such as
Participant 1:  "I think it's best if we invest 200" or
Participant 1:  "let's do 200" or
Participant 1:  "I have been doing 200 in the last round" or
Participant 1:  "100 sounds good".
Rather than spelling out numbers such as 200, people might refer to "max" or "all in".
In a particular context, "let's do it" may be a promise.  For example:
Participant 1:  "200?"
Participant 2:  "let's do it"
or similarly:
Participant 1:  "150.  player 2 are you in agreement?"
Participant 2:  "hi sounds good"
------------------------------------------------------------------------
It may include conditional promises (do something if someone else agrees).
For example, in
Participant 1:  "happy with 200 if we all agree"...
Participant 2:  "cool let's do it"
Participant 3:  "Yep."
In this case, all participants 1, 2 and 3 made promises.
But in
Participant 1:  "200 each?"
Participant 2:  "agree"
Participant 3:  "agree"
only participants 2 and 3 make a promise, because "200 each?" does not make it clear that the first participant
*will* do 200 if the others do that as well.
------------------------------------------------------------------------
Initial promises about which people later change their mind do not count:
Participant 1:  "200"
Participant 2:  "agreed"
Participant 3:  "I suggest 100"
Participant 2:  "I'm happy with either"
Participant 2 is not making any promise here.
```

2.B.2 Prompts

Figure 2.33: Prompt B

```
# General Task
- Evaluate each player in an investment game to determine whether he makes a promise or not.
# Context
- Players:  Group of three.
- Initial Endowment:  200 pence each.
- Investment:  Maximum 200 pence each.
- Mechanics:  Invested amount is doubled and split equally.
- Communication:  Players can chat before investing.
- Duration:  Multiple rounds.
# Classification Task
- Classify a player as `1` if he made a promise.
- Classify as `0` otherwise.
# Constraints
- Refrain from providing an explanation for your classification.  (only for no-CoT cases)
- Provide a final and single classification for each player.
- Follow the below output format.
# Output Format
P# :  0/1
```

Figure 2.34: Prompt O

```
# General Task
- Evaluate each player in an investment game to determine whether he makes a promise or not.
# Context
- Players:  Group of three.
- Initial Endowment:  200 pence each.
- Investment:  Maximum 200 pence each.
- Mechanics:  Invested amount is doubled and split equally.
- Communication:  Players can chat before investing.
- Duration:  Multiple rounds.
# Classification Task
- Classify a player's message as "a statement of intent or a promise" if at least one of the following
conditions is satisfied:
  + The message indicates that the player will take a certain course of action.
  + The message gives others reason to believe or expect that the player will take a certain course of action
- If the message does not satisfy either of the above conditions, classify it as an "Empty Talk".
- Capture what had been said rather than why it was said or what effect it had.
- Operate as a "coding machine".
# Classification Coding
- Code player's message as `1` if it's a promise or statement of intent.
- Code as `0` if it's an empty talk.
# Examples (Only in n-shot treatment) (See Figure 2.35)
# Constraints
- Refrain from providing an explanation for your classification.  (only for no-CoT cases)
- Provide a final and single classification for each player.
- Follow the below output format.
# Output Format
P# :  0/1
```

Figure 2.35: Example section for prompt O

```
# Examples (Only in n-shot treatment)
## Example 1
### Chat
P1:  all 200 then?
P2:  yes
### Classification
P1:  0
P2:  1
## Example 2
### Chat
P1:  Are we all just going to with max?
P2:  agree
### Classification
P1:  0
P2:  1
## Example 3
### Chat
P1:  I think it's best if we invest 200
### Classification
P1:  0
## Example 4
### Chat
P1:  let's do 200
### Classification
P1:  0
## Remark
- Rather than spelling out numbers such as 200, people might refer to "max" or "all in".
## Example 5
### Chat
P1:  I have been doing 200 in the last round
### Classification
P1:  0
## Example 6
### Chat
P1:  100 sounds good
### Classification
P1:  0
## Remark
- In a particular contexts, "let's do it" or "sounds good" may be a promise (examples 7 and 8).
## Example 7:
### Chat
P1:  200?
P2:  let's do it
### Classification
P1:  0
P2:  1
## Example 8
### Chat
P1:  150.  player 2 are you in agreement?
P2:  hi sounds good
### Classification
P1:  1
P2:  1
## Remark
- A message may include conditional promises (do something if someone else agrees) (example 9).
## Example 9:
### Chat
P1:  happy with 200 if we all agree
P2:  cool let's do it
P3:  Yep.
### Classification
P1:  1 (conditional promise)
P2:  1
P3:  1
## Remark
- In example 10, only players 2 and 3 make a promise, because "200 each?" does not make it clear that the first
participant *will* do 200 if the others do that as well.
## Example 10
### Chat
P1:  200 each?
P2:  agree
P3:  agree
### Classification
P1:  0
P2:  1
P3:  1
## Remark
- Initial promises about which player later change their mind do not count (example 11).
## Example 11
### Chat
P1:  200
P2:  agreed (initial promise)
P3:  I suggest 100
P2:  I'm happy with either (P2 changes her mind)
### Classification
P1:  1
P2:  0 (due to change of mind after an initial promise)
P3:  0
```

Figure 2.36: Prompt O_+

```
# General Task
- Evaluate each player in an investment game to determine whether he makes a promise or not.
# Role Persona
- Act as a behavioral economist specialized in text classification, investment game and communication in games.
# Context
- Players:  Group of three.
- Initial Endowment:  200 pence each.
- Investment:  Maximum 200 pence each.
- Mechanics:  Invested amount is doubled and split equally.
- Communication:  Players can chat before investing.
- Duration:  Multiple rounds.
# Classification Task
- Classify player's message as a promise if the player:
  + Explicitly states their intention to take a specific action.
  + Explicitly agrees to take an action suggested by another.
  + Commits to an action, conditional on a specific event occurring.
- Classify player's message as a non-promise if the player:
  + Suggests actions without commitment.
  + Asks questions or discusses preferences without explicitly committing.
  + Talks about hypothetical, ideal, or rational actions without explicit commitment.
# Classification Coding
- Code player's message as `1` if it's a promise.
- Code as `0` if it's a non-promise.
# Examples (Only in n-shot treatment)
1.  P1:  "I'll invest 150 pence if you do the same." (Code as 1 - Conditional promise)
2.  P2:  "What if we all invest 200 pence?" (Code as 0 - Question, no commitment)
3.  P3:  "Investing 200 pence seems the best strategy." (Code as 0 - Suggestion, no explicit commitment)
4.  P1:  "all 200 then?" P2:  "yes" (Code P2 as 1 - Clear affirmative response to a suggestion is a Promise,
code P1 as 0 - Question, no commitment)
5.  P1:  "Are we all just going with max?â P2:  "agree" (Code P2 as 1 - Clear affirmative response to a
suggestion is a Promise, code P1 as 0 - Question, no commitment)
6.  P1:  "I think it's best if we invest 200" (Code as 0 - Opinion, no commitment)
7.  P1:  "200?" P2:  "let's do it" (Code P2 as 1 - Clear affirmative response to a suggestion is a Promise, code
P1 as 0 - Question, no commitment)
8.  P1:  "happy with 200 if we all agree" P2:  "cool let's do it" P3:  "Yep." (Code P1 as 1 - Conditional
promises, code P2 and P3 as 1 - Clear affirmative response to a conditional promise)
9.  P1:  "200 each?" P2:  "agree" P3:  "agree" (Code P1 as 0 - Question, no commitment, Code P2 and P3 as 1 -
Promises by P2 and P3 - clear affirmative response to a question)
# Constraints
- Refrain from providing an explanation for your classification.  (only for no-CoT cases)
- Provide a final and single classification for each player.
- Follow the below output format.
# Output Format
P# :  0/1
```

2.B.3 Additional Examples

Figure 2.37: Promise II - Example message 3

```
P2:  Hello
P3:  Hi there
P3:  Any bots in here?
P2:  Nope
P1:  no
P3:  Nice one
P3:  OK
P3:  I suggest we all invest the maximum to max out our rewards
P2:  we should each invest fully for the best outcome
P3:  I agree, we are all here for the cash :)
P2:  I'm happy to do that
P2:  Excellent
P3:  If we just stuff 200 in we'll all get the most reward
P3:  I'm not here to be dishonest and cheat you all, we know how expensive life is right now
```

	GPT-3.5			GPT-4		
	P_1	P_2	P_3	P_1	P_2	P_3
B	0	1	1	0	1	1
O	0	0	0	0	1	1
O_+	0	1	1	0	1	0
RAs	0	1	1	0	1	1

Table 2.34: Human and n-shot CoT GPT classifications for the message in Figure 2.37

Figure 2.38: Promise II - Classifications with CoT for prompts O and O_+ for message in Figure 2.37

```
GPT-3.5
Output using O
Reasoning:
P2:  No specific promise made, just general agreement and positive attitude.
P3:  Suggests investing the maximum and agrees to it, emphasizing the desire for the best outcome.  Also mentions not
being dishonest and the expensive nature of life, but does not make a specific promise.
Classification:
P1:  0
P2:  0
P3:  0
Output using O+
Reasoning:
The conversation starts with casual greetings and small talk, with no explicit commitment to any action.  However,
as the conversation progresses, P3 suggests investing the maximum and emphasizes the benefits, while P2 agrees and
expresses happiness with the decision.  P3 also mentions being honest and not cheating, which implies a commitment to
the suggested action.
Classification:
P1:  0
P2:  1
P3:  1

GPT-4
Output using O
Reasoning:
- P2 starts with greetings and does not make any promises in the initial exchanges.
- P3 asks a question and then makes a suggestion to invest the maximum, which is a promise to invest the maximum if
others agree.
- P2 then agrees with P3's suggestion and states a willingness to invest fully, which is a promise.
- P3 reiterates the agreement and mentions being there for the cash, which is not a promise but an expression of
intent or motivation.
- P2 expresses happiness to follow through with the full investment, which is a reaffirmation of the promise.
- P3 then explicitly states the intention to invest 200, which is a promise, and follows up with a statement about
honesty and not cheating, which reinforces the commitment to the promise but is not a promise in itself.
Classification:
P1:  0
P2:  1
P3:  1
Output using O+
Reasoning:
- P2 starts with a greeting, which is not a promise (Code as 0).
- P3 also greets and asks a question about bots, which is not a promise (Code as 0).
- P2 and P1 confirm there are no bots, which are statements of fact, not promises (Code as 0).
- P3 makes a suggestion to invest the maximum, which is a suggestion without a commitment (Code as 0).
- P2 echoes the suggestion that each should invest fully for the best outcome, which is still a suggestion without
explicit commitment (Code as 0).
- P3 agrees with the sentiment that they are all there for the cash, which is an expression of agreement with the
rationale but not a commitment to action (Code as 0).
- P2 then expresses happiness to follow the suggestion, which is a commitment to action (Code as 1).
- P2's "Excellent" is an expression of satisfaction, not a promise (Code as 0).
- P3 states what they should do ("If we just stuff 200 in we'll all get the most reward"), which is a continuation of
the suggestion and rationale but not an explicit commitment (Code as 0).
- P3's statement about not being there to be dishonest is an expression of intent but does not constitute a promise
regarding the specific action of investing (Code as 0).
Classification:
P1:  0
P2:  1
P3:  0
```

Figure 2.39: Promise II - Example message 4

```
P2:  Hi
P1:  Hi all .  What say you on how much to put in?
P2:  I recommend investing 200 each for maximum returns
P3:  same
P1:  OK:)
P2:  Please stick to it
```

	GPT-3.5			GPT-4		
	P_1	P_2	P_3	P_1	P_2	P_3
B	0	1	0	0	0	0
O	0	1	0	0	1	1
O_+	0	1	0	0	0	0
RAs	1	0	0	1	0	1

Table 2.35: Human and n-shot CoT GPT classifications for the message in Figure 2.39

Figure 2.40: Promise II - Classifications with CoT for prompts O and O_+ for message in Figure 2.39

```
GPT-3.5
Output using O
Reasoning:
P2's message "I recommend investing 200 each for maximum returns" constitutes a promise as it indicates a specific
action and gives others reason to believe or expect that P2 will invest 200.  P1's message "OK:)" does not constitute
a promise as it does not indicate a specific action.
Classification:
P1:  0
P2:  1
P3:  0
Output using O+
Reasoning:
P2:  The message "I recommend investing 200 each for maximum returns" suggests a specific action and implies a
commitment to that action.
Classification:
P1:  0
P2:  1
P3:  0

GPT-4
Output using O
Reasoning:
P2 starts with a greeting, which does not constitute a promise.  P1's message is a question about the investment
amount, which is also not a promise.  P2 then recommends investing 200 each, which is a suggestion but does not
explicitly state that P2 will do so; however, the context implies that P2 is advocating for this action and is likely
to follow through with it, so it can be considered a promise.  P3's response "same" indicates agreement with P2's
suggestion and implies a commitment to invest 200 as well, which can be classified as a promise.  P1's "OK:)" is an
acknowledgment of the plan but does not explicitly state that P1 will invest 200, so it is not a promise.  However,
P2's follow-up with "Please stick to it" is a request for commitment, reinforcing the earlier promise to invest 200.
Classification:
P1:  0
P2:  1
P3:  1
Output using O+
Reasoning:
- P2 starts with a greeting, which does not contain any commitment or promise.
- P1 asks for opinions on the investment amount, which is a question without commitment.
- P2 recommends investing 200 each, which is a suggestion but does not explicitly state that they will do so
themselves, hence no commitment.
- P3 agrees with P2's suggestion, which could be interpreted as a commitment, but without an explicit statement of
personal action, it remains a general agreement.
- P1 responds with "OK:)" which indicates agreement but does not explicitly state a commitment to invest a specific
amount.
- P2 then asks to "Please stick to it," which is a request for others to commit but is not a self-commitment.
Classification:
P1:  0
P2:  0
P3:  0
```

Figure 2.41: Promise II - Example message 5

```
P2:  max?  then doubled.
P1:  200
P3:  Yes, agree.
P2:  *thumbs up*
```

	GPT-3.5			GPT-4		
	P_1	P_2	P_3	P_1	P_2	P_3
B	0	1	0	0	0	0
O	0	1	0	0	0	1
O_+	0	0	1	1	0	1
RAs	0	1	1	0	1	1

Table 2.36: Human and n-shot CoT GPT classifications for the message in Figure 2.41

Figure 2.42: Promise 2 - Classifications with CoT for prompts O and O_+ for message in Figure 2.41

GPT-3.5

Output using O

```
Reasoning:
P2:  The message "max?  then doubled." indicates a specific action by the player and gives others reason to believe or
expect that the player will invest the maximum amount.  Therefore, it constitutes a promise.
P1:  The message "200" does not constitute a promise as it does not indicate a specific action by the player or give
others reason to believe or expect a certain course of action.
P3:  The message "Yes, agree." does not constitute a promise as it does not indicate a specific action by the player
or give others reason to believe or expect a certain course of action.
P2:  The message "*thumbs up*" does not constitute a promise as it does not indicate a specific action by the player
or give others reason to believe or expect a certain course of action.
Classification:
P1:  0 P2:  1 P3:  0
```

Output using O_+

```
Reasoning:
P2 suggests "max?  then doubled." without explicitly committing to the action.  P1 responds with "200" which could be
interpreted as a suggestion or a statement of intention, but it's not explicitly a promise.  P3 responds with "Yes,
agree," which indicates a clear affirmative response but is conditional on the previous messages.  P2's "thumbs up"
could be interpreted as agreement, but it's not a clear commitment.
Classification:
P1:  0 P2:  0 P3:  1
```

GPT-4

Output using O

```
Reasoning:
P2's initial message "max?  then doubled." is a question and a statement of the game's mechanics, not a promise to
invest a specific amount.
P1's message "200" is a clear indication of the amount they are considering to invest.  However, without additional
context or a statement that commits them to this action, it is not a promise.
P3's message "Yes, agree." is a response to the previous messages, indicating agreement with the proposed action.
This can be interpreted as a commitment to invest the amount that is being discussed, which in this context is 200
pence.
P2's message "*thumbs up*" is a non-verbal indication of agreement or approval but does not constitute a verbal
promise to take a specific action.
Classification:
P1:  0 P2:  0 P3:  1
```

Output using O_+

```
Reasoning:
- P2's initial message "max?  then doubled." is a question about the game mechanics, not a commitment to invest a
specific amount.  Therefore, it's not a promise.
- P1's response "200" is a statement of a specific amount, which in this context seems to be an agreement to invest
200 pence.  This can be interpreted as a promise to take that action.
- P3's response "Yes, agree." is an explicit agreement to the previously mentioned action (investing 200 pence).  This
is a promise to take that action.
- P2's response with a "*thumbs up*" emoji is a non-verbal affirmative response, which in this context can be
interpreted as an agreement to the proposed action.  However, it's less explicit than a verbal commitment, and without
additional context, it's not as strong as a clear statement of intent.  It could be seen as an endorsement of the
plan rather than a personal commitment to action.  Given the ambiguity, I will lean towards not classifying it as a
promise.
Classification:
P1:  1 P2:  0 P3:  1
```

Figure 2.43: Promise 2 - Example message 6

```
P3:  Hi everyone
P2:  hi
P1:  I think we should do 100 each
P2:  i was thinking the same
P2:  are we agreed?
P1:  player 3?
P3:  If we all do 200 we get the most back I think?
P2:  we should meet in the middle then and do 150?
P1:  I'm good with that
P3:  Hep
P3:  Yep
P1:  Lets do it
```

	GPT-3.5			GPT-4		
	P_1	P_2	P_3	P_1	P_2	P_3
B	1	1	1	0	0	0
O	0	0	0	0	1	0
O_+	1	1	1	1	1	1
RAs	1	1	1	1	1	1

Table 2.37: Human and n-shot CoT GPT classifications for the message in Figure 2.43

Figure 2.44: Promise 2 - Classifications with CoT for prompts O and O_+ for message in Figure 2.43

```
GPT-3.5
Output using O
Reasoning:
- P1 suggests an amount (100 each) but does not explicitly promise to invest that amount.
- P2 agrees with the suggestion but also does not explicitly promise to invest that amount.
- P3 suggests investing 200 each, but does not explicitly promise to invest that amount.
- P2 suggests meeting in the middle at 150, but does not explicitly promise to invest that amount.
- P1 agrees with the suggestion but does not explicitly promise to invest that amount.
- P3 confirms agreement with "Yep" but does not explicitly promise to invest that amount.
- P1's "Let's do it" does not explicitly promise to invest 150.
- P2's "50 it is then" does not explicitly promise to invest 150.
Classification:
P1:  0 P2:  0 P3:  0
Output using O+
Reasoning:
P3 initiates the discussion by suggesting that investing 200 each would yield the most return.  P2 then suggests
meeting in the middle at 150, to which P1 agrees.  P3 responds with "Yep" and P2 confirms "50 it is then."
Classification:
P1:  1 P2:  1 P3:  1

GPT-4
Output using O
Reasoning:
P3 starts with a greeting, which is not a promise ('0').  P2 also greets, which is not a promise ('0').  P1 suggests
doing 100 each, which is a proposal but not a commitment ('0').  P2 agrees with the idea of 100 each and asks for
confirmation ('0'), as this is still part of the negotiation process.  P1 asks for P3's input, which is not a promise
('0').  P3 suggests that investing 200 each would yield the most return, which is an opinion, not a promise ('0').
P2 then suggests meeting in the middle with 150, which is a counter-proposal but not a firm commitment ('0').  P1
agrees with the 150 proposal, which is an acceptance of the suggestion but not a clear promise ('0').  P3 agrees with
"Yep," which is an affirmation but not a clear commitment ('0').  P1 then says "Let's do it," which is an agreement to
proceed but not a specific promise to invest a certain amount ('0').  P2 finally states "150 it is then," which is a
confirmation of the agreed amount and can be considered a promise to invest that amount ('1').
Classification:
P1:  0 P2:  1 P3:  0
Output using O+
Reasoning:
P1 starts with a suggestion to invest 100 each, which is not a commitment but an opinion (Code as 0).  P2 expresses
agreement with the initial suggestion and asks for confirmation (Code as 0 for agreement without commitment, and 0
for the question).  P1 asks P3 for their opinion, which is a question and not a commitment (Code as 0).  P3 suggests
that investing 200 each would yield the most return, which is an opinion and not a commitment (Code as 0).  P2 then
suggests meeting in the middle with 150 each, which is a suggestion without a commitment (Code as 0).  P1 agrees with
P2's suggestion, which is a commitment to the action (Code as 1).  P3 agrees with the suggestion by saying "Yep,"
which is a commitment (Code as 1).  P1 then reaffirms the decision with "Lets do it," which is a commitment (Code as
1).  P2 confirms the amount with "150 it is then," which is a commitment (Code as 1).
Classification:
P1:  1 P2:  1 P3:  1
```

2.B.4 Additional Analysis

In Tables 2.38 and 2.39, weighted F1 and F1 macro averages are presented. Since the "promise" and "empty talk" categories are balanced with 53.3% of the messages classified as "promise" in the groundtruth, weighted F1 and F1 macro averages are very close to the accuracy results reported in Table 2.10 in Section 2.6.4.

		no-CoT		**CoT**	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	B	.57	–	.614	–
	O	.574	.574	.547	.525
	O_+	.719	.749	.672	.75
GPT-4	B	.674	–	.707	–
	O	.684	.827	.718	.848
	O_+	.865	.862	.841	**.887**

Table 2.38: Weighted F1

		no-CoT		**CoT**	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	B	.563	–	.614	–
	O	.572	.595	.546	.52
	O_+	.717	.746	.673	.751
GPT-4	B	.675	–	.71	–
	O	.681	.826	.721	.847
	O_+	.865	.861	.841	**.886**

Table 2.39: F1 Macro averages

2.C Level-k I

2.C.1 Original Instructions

Figure 2.45: Original classification instructions

General Comments:
Subjects do not necessarily describe every step of their thinking; therefore, it may not always be obvious to decide which level they are. In many comments, any indications of a level of thinking may be partial or implicit, you should then indicate the most likely level of reasoning of the player. If the message indicates to simply refer to a previous message ("same as before/above"), then you can use the previous message's evaluation to determine the level of the current message. If you are unsure of the level of the message, you should indicate the level you think is more likely.
Level-0 Player:
Characteristics: Chooses randomly, without justification or through some justification completely unrelated to the task. Might not have understood the game or shows no interest in the game or in thinking about it.
Examples:
"50 50 chance to get red at least 50 50 could also be 100 percent."
"I like blue, so I chose blue."
"Think it will be red again."
"definitely red this time"
"We have to go for red. No other way than that. I like turtles"
Note: Comments such as "It is obviously blue" or "Play red, trust me!" should not be considered as level-0 thinking as these comments to some extent signal some level of understanding/interest of the task. Such comments are likely to be level-1 comments yet without any additional information, you should leave the specific cell empty.
Level-1 Player:
Characteristics: Always follows his own signal. The subject may argue in favor of playing his own signal through some probability argument
Examples:
"Our signal is blue. Let's play blue."
"The probability that the red ball we observe is out of the red urn is twice the probability that it is out of the blue urn"
"1/3 of all teams is observing wrong color, so we would try to find out whether we have wrong or right ball, keep with red."
Note: The key idea in defining a level-1 player is to identify some thinking process that signals the subject's interest/understanding of the task and the private signal. Furthermore, it is important that the subject does **not** offer any argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e. adjusting the strategy given what others are expected to do).
Level-2 Player:
Characteristics: Assume that all other players almost always follow their signal (i.e. she assumes almost all the other players are level-1 while an epsilon portion of them are level-0). Player does offer an argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e. a best response given others are most likely playing their signal). In other words, if you identify any comment that indicates that the subject assumes (or considers the case) where the other players in her group play their signal, you should consider the possibility that the subject is a level-2 player.
Examples:
"Let's take red because if the urn is red and we got the opposite color and we take blue, the decision will be blue."
"We need to chose Red. If we are the only ones who picked blue, then the urn is red and we guess correct If the urn is blue, then the other guys will pick blue so there will be at least one blue vote and we win as well If the others guys (also blue) think the same way then we lose But this is too many ifs"
"I have a blue ball. If we have the blue urn, someone else also has a blue ball and as a result our group will chose blue regardless of my vote. If we have the red urn, I am the only one with the blue ball and if I vote blue, we will chose the wrong urn. So I should vote for red."
"In case two teams choose red and one chooses blue, blue will be taken. That means that choosing red has a higher chance of being a good decision."
"I guess this is more about luck because there is no way to know it for sure. I would say blue just because of the higher probability. Also I like turtles Also it is likely that one other team will pick blue and then it is that color anyways"
"There is no point for us to take blue I think the chances for us to get the right color are higher if we stick with red" [red ball is observed]
"I suggest red because we donat hurt anyone with this decision If the others go for blue because they have a blue ball, the committeeas decision will be blue regardless of our decision"
"We could be the deciding vote for blue if the other two choose red. Choosing blue isnt as helpful as choosing red, because: only one blue ball can overturn our whole decision but only a unanimous decision for red can help us the same way"
Note: In order to discern the two types, you should look for more than any trivial arguments such as the ones given under level-1. There may be cases where the message starts as a level-1 argument and then as the subjects elaborates on her reasoning, she starts considering the strategy of the other teams and justify her decision accordingly (see the third example above). In such cases, this message should be considered as level-2. The acknowledgment of other teams' voting strategy may not always be obvious or may be worded differently such as "hurting the other's decision" or "not being helpful" (see the last three examples above)
Level-3 Player:
Characteristics:
Assumes that almost all other subjects are level-2 players (partially degenerate beliefs). The reasoning in a level-3 player message will have similarities with a level-1 player message but it will have additional arguments indicating that she assumes others are level-2 players.
Examples:
"In my opinion, if there is another person with blue they may be afraid of voting blue so we should vote blue to make sure."
"Let's now pick the shown colour because the others now will probably enter their opposite colour."
"Risky to vote blue but others may not vote blue even when they draw blue. I say we vote blue."
Note: As stated above, level-3 players are likely to follow their signal like a level-1 player yet they will argue to do so through a much more intricate argument (unlike a level-1 player merely stating probabilities to argue her action). Level-3 players are rare. Higher levels (level-4 etc.) are assumed to not occur; therefore, you should consider only the first 4 levels of thinking.

We have omitted various sections of the original instructions to fit in the instructions into a single page. The original instructions begins with a lengthy section regarding the general theory of level-k modeling and its specific application to the experiments. These sections aggregate to 3 pages of instructions. Although the context of these sections are relevant in terms how we have very briefly summarized them in the "Context" section of our prompts, we omitted them in Figure 2.45. Furthermore, some sentences in the instructions that were not related to the classification of instructions, such as how to code certain concepts in the excel sheet, are also omitted. See Chapter 1 for the instructions in its entirety.

2.C.2 Prompts

Figure 2.46: Prompts O and O_+

```
# General Task
- Classify player's level of strategic thinking in a voting game
# Role Persona
- Act as a behavioral economist specialized in level-k modeling, strategic thinking and text classification.
# Context (see Figures 2.47 and 2.48)
# Classification Task
- Classify a player's level of strategic thinking as 0, 1, 2 or 3 based on the message provided.
- Use the below characteristics, examples and notes provided for each level to determine your classification.
## Level-0 Player
### Characteristics
- Chooses randomly, without justification or through some justification completely unrelated to the task.
- Might not have understood the game or shows no interest in the game or in thinking about it.
- Provides a vote without a clear justification to the probability of the game or strategic reasoning.
### Note
- Comments such as 'It is obviously blue' or 'Play red, trust me!'should not be considered as level-0 thinking as
these comments to some extent signal some level of understanding/interest of the task.  Such comments are likely to be
level-1 comments.
## Level-1 Player
### Characteristics
- Always follows his own signal.
- The subject may argue in favor of playing his own signal through some probability argument.
### Note
- The key idea in defining a level-1 player is to identify some thinking process that signals the subject's
interest/understanding of the task and the private signal.
- It is important that the subject does not offer any argument acknowledging the potential votes of the other teams
and how to vote accordingly (i.e.  adjusting the strategy given what others are expected to do).
## Level-2 Player
### Characteristics
- Assume that all other players almost always follow their signal (i.e.  a level-2 player assumes almost all the other
players are level-1 while a small portion of them are level-0).
- Player does offer an argument acknowledging the potential votes of the other teams and how to vote accordingly (i.e.
a best response given others are most likely playing their signal).
- if you identify any comment that indicates that the subject assumes (or considers the case) where the other players
in her group play their signal, you should consider the possibility that the subject is a level-2 player.
### Note
- In order to discern between level-1 and level-2 types, you should look for more than any trivial arguments such as
the ones given under level-1.
- There may be cases where the message starts as a level-1 argument and then as the subjects elaborates on her
reasoning, she starts considering the strategy of the other teams and justify her decision accordingly (see the fifth
example above).  In such cases, this message should be considered as level-2.
- The acknowledgment of other teams' voting strategy may not always be obvious or may be worded differently such as
"hurting the other's decision" or "not being helpful" a (see the last three examples above)
## Level-3 Player
### Characteristics
- Assumes that almost all other subjects are level-2 players (partially degenerate beliefs).
- The reasoning in a level-3 player message will have similarities with a level-1 player message but it will have
additional arguments indicating that she assumes others are level-2 players.
### Notes
- level-3 players are likely to follow their signal like a level-1 player yet they will argue to do so through a much
more intricate argument (unlike a level-1 player merely stating probabilities to argue her action).
- Level-3 players are rare.
## General Comments
- Players do not necessarily describe every step of their thinking; therefore, it may not always be obvious to decide
which level they are.  In many messages, any indications of a level of thinking may be partial or implicit.  In such
cases provide the most likely level of reasoning from the messages.
- If you are unsure of the level of the message, you should indicate the level you think is more likely.
# Constraint
- Only provide a single level classification.
- Follow the below output format.
# Output Format
0/1/2/3
```

"Classification Process" section used for CoT prompting and the variation in the "Output Format" are as presented in Figures 2.1 and 2.2 and are omitted. Since the prompts only differ in their "Context" section, we present both prompts all other identical parts in Figure 2.46. "Examples" section used for n-shot prompting is presented separately in Figure 2.49.

Figure 2.47: Prompt O - Context

```
# Context
- Teams of two players, part of larger groups, draw a colored ball from an urn, play a voting game with their group to
guess the color of the urn for multiple periods.
- Teams are randomly paired each period.  Teams consist always of 2 players.
- Teams are grouped into either 3 or 6 teams per group.
- Each period, a group is assigned to an urn with a blue or red color with equal probability.
- An urn only contains blue and red balls.  Blue urn has twice more blue balls than red balls and a red urn has twice
more red balls than blue balls.  The color of the ball has a 2/3 chance of matching the urn's color.
- After the urn with either red or blue color is assigned to a group, each team within a group draws a ball from the
urn to infer the urn's color.  The drawing can be with or without replacement depending on the period.
- Teams do not know the color of the urn.  They do not know the colors of the balls picked by the other teams in their
group.  They only know the color of their own ball.
- The group's objective is to correctly guess the color of the assigned urn.
- Each team in a group provides a single vote.  A team votes for either the color red or blue.  Group's decision is
determined based on the aggregation of its teams' votes.
- Teams communicate internally to decide on a vote for the urn's color.  If all votes are red, the group decision is
red; any blue vote results in a blue group decision.
- Teams do not observe the votes of the other teams in their group.
- Teams weigh their own ball's color and strategize their vote considering the group's outcome.
- Players exhibit levels of strategic reasoning (0 to 3), influencing their decision-making and messaging.
```

Figure 2.48: Prompt O_+ - Context

```
# Context
## Game Mechanics
- Teams of two players, part of larger groups, draw a colored ball from an urn, play a voting game with their group to
guess the color of the urn for multiple periods.
- Teams are randomly paired each period.  Teams consist always of 2 players.
- Teams are grouped into either 3 or 6 teams per group.
- Each period, a group is assigned to an urn with a blue or red color with equal probability.
- An urn only contains blue and red balls.  Blue urn has twice more blue balls than red balls and a red urn has twice
more red balls than blue balls.  The color of the ball has a 2/3 chance of matching the urn's color.
- After the urn with either red or blue color is assigned to a group, each team within a group draws a ball from the
urn to infer the urn's color.  The drawing can be with or without replacement depending on the period.
- In without replacement case, for group size of 3, if the urn is red (blue), 2 teams will draw red (blue) ball and 1
team will draw blue (red) ball.  For the group size of 6, if the urn is red (blue), 4 teams will draw red (blue) ball
and 2 teams will draw blue (red) ball.
- In with replacement case, irrespective of the group size, the probability of drawing a red (blue) ball from the red
(blue) urn will be 2/3.
- Teams do not know the color of the urn.  They do not know the colors of the balls picked by the other teams in their
group.  They only know the color of their own ball.
- The group's objective is to correctly guess the color of the assigned urn.
- Each team in a group provides a single vote.  A team votes for either the color red or blue.  Group's decision is
determined based on the aggregation of its teams' votes.
- Teams do not observe the votes of the other teams in their group.
## Communication
- Teams communicate internally to decide on a vote for the urn's color.  If all votes are red, the group decision is
red; any blue vote results in a blue group decision.
- Each player within a team can only send a single message to their teammate.  Both teammates send their messages
before observing their teammate's message.
## Strategic Reasoning
- Teams weigh their own ball's color and strategize their vote considering the group's outcome.
- Players exhibit levels of strategic reasoning (0 to 3), influencing their decision-making and messaging.
- Level-0:  Any action that can be considered as random play, non-strategic or unrelated to the game mechanics
- Level-1:  Almost always votes the color of the ball picked.  Does not assume any behavior by other teams.
- Level-2:  Almost always votes red regardless of the color of the ball picked.  Assumes that everybody else almost
always votes the color of the ball they pick (i.e.  assumes everybody else is almost always level-1).
- Level-3:  Almost always votes the color of the ball picked.  Assumes that everybody else almost always votes red
regardless of the color of the ball they pick (i.e.  assumes everybody else is almost always level-2)
```

Figure 2.49: Prompt O and O_+ - Example sections

```
# Classification Task
## Level-0 Player
### Characteristics ...
### Examples
1.  "50 50 chance to get red at least 50 50 could also be 100 percent."
2.  "I like blue, so I chose blue."
3.  "Think it will be red again."
4.  "definitely red this time"
5.  "We have to go for red.  No other way than that.  I like turtles"
### Notes ...
## Level-1 Player
### Characteristics ...
### Examples
1.  "Our signal is blue.  Let's play blue."
2.  "The probability that the red ball we observe is out of the red urn is twice the probability that it is out of the
blue urn"
3.  "1/3 of all teams is observing wrong color, so we would try to find out whether we have wrong or right ball, keep
with red."
### Notes ...
## Level-2 Player
### Characteristics ...
### Examples
1.  "Let's take red because if the urn is red and we got the opposite color and we take blue, the decision will be
blue."
2.  "We need to chose Red.  If we are the only ones who picked blue, then the urn is red and we guess correct If the
urn is blue, then the other guys will pick blue so there will be at least one blue vote and we win as well If the
others guys (also blue) think the same way then we lose But this is too many ifs"
3.  "I have a blue ball.  If we have the blue urn, someone else also has a blue ball and as a result our group will
chose blue regardless of my vote.  If we have the red urn, I am the only one with the blue ball and if I vote blue, we
will chose the wrong urn.  So I should vote for red."
4.  "In case two teams choose red and one chooses blue, blue will be taken.  That means that choosing red has a higher
chance of being a good decision."
5.  "I guess this is more about luck because there is no way to know it for sure.  I would say blue just because of
the higher probability.  Also I like turtles Also it is likely that one other team will pick blue and then it is that
color anyways"
6.  "There is no point for us to take blue I think the chances for us to get the right color are higher if we stick
with red" [red ball is observed]
7.  "I suggest red because we donat hurt anyone with this decision If the others go for blue because they have a blue
ball, the committeeas decision will be blue regardless of our decision"
8.  "We could be the deciding vote for blue if the other two choose red.  Choosing blue isnt as helpful as choosing
red, because:  only one blue ball can overturn our whole decision but only a unanimous decision for red can help us
the same way"
### Notes ...
## Level-3 Player
### Characteristics ...
### Examples
1.  "If everyone else assumes others play their own signal then they will always play red.  Since I have the blue
ball, it is more likely that we have the blue urn so I will vote blue"
2.  "Let's now pick the shown colour because the others now will probably enter their opposite colour."
3.  "Risky to vote blue but others may not vote blue even when they draw blue.  I say we vote blue."
### Notes ...
```

2.C.3 Additional Examples

Figure 2.50: Level-k I - Example message 4

```
the chance that we recieve the wrong ball is smaller than the chance that we receive the correct ball However I think
every team will think in the same way and there is no chance that the committe gets it right...
```

	GPT-3.5	**GPT-4**
O	2	1
O_+	2	1
RAs	0	

Table 2.40: Human and GPT classifications of message in Figure 2.50 using n-shot and CoT prompting

Figure 2.51: Level-k I - Classifications with CoT and n-shot for prompts O and O_+ for message in Figure 2.19 or Figure 2.50

GPT-3.5

Output using O
Reasoning:
The player is considering the possibility of not picking the same color as their partner in order to avoid a tie. They are strategizing based on the assumption that their partner will follow their signal.
Classification:
2
Output using O_+
Reasoning:
The player is considering the possibility of not picking the same color as their teammate to avoid a unanimous vote, indicating some level of strategic thinking. They are suggesting a coordinated strategy to ensure their vote is different from their teammate's.
Classification:
2

GPT-4

Output using O
Reasoning:
The player is suggesting a strategy that involves coordination with their teammate to cover both possible outcomes. This indicates that the player is considering the actions of their teammate and how it could affect the group's decision. The message implies a basic understanding of the game's mechanics and the importance of diversifying the votes within the team. However, the player does not explicitly acknowledge the strategies or potential votes of other teams in the group, which would be characteristic of a level-2 player. The message suggests a simple strategy without deeper strategic reasoning about the other teams' likely actions.
Classification:
1
Output using O_+
Reasoning:
The player is suggesting a strategy that involves coordination with their teammate to cover both possible outcomes. This message indicates that the player is considering the actions of their teammate and how it could affect the group's decision. However, the player does not explicitly consider the strategies of the other teams in the group, which would be characteristic of a level-2 player. The message also lacks the assumption that other players are level-2, which would be necessary to classify the player as level-3. The player's message does not fit the characteristics of a level-0 player, as it is not a random choice or unrelated to the task. Therefore, the message most closely aligns with level-1 reasoning, where the player is interested in the task and is considering their own signal and that of their teammate but not the strategies of other teams.
Classification:
1

Figure 2.52: Level-k I - Example message 5

```
I'd say always choose red but maybe we should choose blue since everbody thinks always choose red and it's 2/3 that we win.
```

	GPT-3.5	**GPT-4**
O	2	3
O_+	2	3
RAs	3	

Table 2.41: Human and GPT classifications of message in Figure 2.52 using n-shot and CoT prompting

Figure 2.53: Level-k I - Classifications with CoT and n-shot for prompts O and O_+ for message in Figure 2.52

```
                                                                                              GPT-3.5
Output using O
Reasoning:
The player seems to be considering the strategy of the other players in the group.  They mention that everyone
thinks "always choose red" and then consider choosing blue as a response to this.  This indicates an awareness of the
potential votes of the other teams and how to vote accordingly.
Classification:  2
Output using O+
Reasoning:
The player starts by suggesting "always choose red" but then considers the possibility of choosing blue because
everyone else might be thinking the same way.  The player acknowledges the potential votes of the other teams and how
to vote accordingly, indicating a consideration of the strategy of the other teams.
Classification:  2

                                                                                                GPT-4
Output using O
Reasoning:
The player starts by considering the strategy of "always choose red," which is indicative of level-2 thinking, as
it shows an awareness of a common strategy that assumes other players will vote their signal (level-1 behavior).
However, the player then suggests choosing blue, acknowledging that other players might also be thinking strategically
at level-2 and therefore might "always choose red." This suggests that the player is considering the possibility that
others are at level-2 and is adjusting their strategy accordingly, which is characteristic of level-3 thinking.
Classification:  3
Output using O+
Reasoning:
The player starts with a consideration that might suggest a level-2 strategy ("always choose red") but then moves
on to consider what other players might be thinking.  The player acknowledges that other players might be following
a level-2 strategy ("everybody thinks 'always choose red'") and suggests adapting their own strategy based on this
assumption.  This indicates that the player is considering the strategic choices of others and is trying to best
respond to what they perceive as the common strategy.  This is characteristic of level-3 reasoning, where the player
assumes that others are level-2 players and adjusts their strategy accordingly.
Classification:  3
```

Figure 2.54: Level-k I - Example message 6

```
the chance that we recieve the wrong ball is smaller than the chance that we receive the correct ball However I
think every team will think in the same way and there is no chance that the committe gets it right...
```

	GPT-3.5	**GPT-4**
O	2	2
O_+	2	2
RAs	1	

Table 2.42: Human and GPT classifications of message in Figure 2.54 using n-shot and CoT prompting

Figure 2.55: Level-k I - Classifications with CoT and n-shot for prompts O and O_+ for message in Figure 2.54

GPT-3.5

Output using O
Reasoning:
The player acknowledges the probability of receiving the wrong ball being smaller than receiving the correct ball. However, they also express a belief that every team will think in the same way, leading to the conclusion that there is no chance that the committee gets it right. This indicates a consideration of the potential votes of the other teams and how to vote accordingly, suggesting a level-2 strategic thinking.
Classification: 2
Output using O_+
Reasoning:
The player acknowledges the probability of receiving the wrong ball and the correct ball. They also express a belief that every team will think in the same way, leading to the conclusion that there is no chance that the committee gets it right. This indicates an awareness of the potential votes of the other teams and how to vote accordingly, suggesting a level-2 strategic thinking.
Classification: 2

GPT-4

Output using O
Reasoning:
The player acknowledges the probability of receiving the correct ball versus the wrong ball, which shows an understanding of the game's mechanics. However, the player also considers the thought process of other teams, assuming that they will think similarly and that this collective reasoning will lead to an incorrect group decision. This indicates that the player is considering the strategies of others and how those strategies will affect the group outcome, which aligns with level-2 strategic thinking.
Classification: 2
Output using O_+
Reasoning:
The player starts with a probability argument, which is a characteristic of level-1 thinking, by acknowledging that the chance of receiving the correct ball is higher than receiving the wrong one. However, the player then moves on to consider what other teams might think and do, which is a characteristic of level-2 thinking. The player assumes that every team will think in the same way, which suggests they are considering the strategies of other teams and how it affects the group outcome. This is a key feature of level-2 strategic reasoning, where the player assumes others are likely to be level-1 and play their signal, and thus adjusts their own strategy accordingly.
Classification: 2

Figure 2.56: Level-k I - Example message 7

I suggest we vote red. There is a higher chance that the urn is red. if even one person chooses blue - the committee's decision is blue, so our decision won't make a difference.

	GPT-3.5	GPT-4
O	2	2
O_+	2	2
RAs	1	

Table 2.43: Human and GPT classifications of message in Figure 2.56 using n-shot and CoT prompting

Figure 2.57: Level-k I - Classifications with CoT and n-shot for Prompts O and O_+ for message in Figure 2.56

```
                                                                                                    GPT-3.5
Output using O
Reasoning:
The player suggests voting red because they believe there is a higher chance that the urn is red.  They also consider
the scenario where even one person chooses blue, in which case the committee's decision will be blue, making their
decision not impactful.
Classification:  2
Output using O+
Reasoning:
The player suggests voting red based on the higher chance of the urn being red.  They also consider the potential
committee's decision, stating that if even one person chooses blue, the committee's decision will be blue, so their
decision won't make a difference.
Classification:  2

                                                                                                      GPT-4
Output using O
Reasoning:
The player acknowledges the rules of the game, specifically the impact of a single blue vote on the committee's
decision.  They are considering the strategic implications of their vote in the context of the group's outcome.  This
indicates that they are thinking about the potential votes of the other teams and how to vote accordingly.  The player
is not just following their own signal but is also considering the likelihood of others voting blue and how it affects
the final decision.  This suggests that the player is at least at level-2, as they are considering the best response
given that others are most likely playing their signal.
Classification:  2
Output using O+
Reasoning:
The player is considering the rules of the game, specifically the voting mechanism where a single blue vote can
determine the group's decision.  They are also acknowledging the likelihood of the urn being red and the strategic
implications of their vote in relation to the votes of others.  This indicates that the player is thinking about the
potential votes of other teams and how to vote accordingly, which aligns with the characteristics of a level-2 player.
Classification:  2
```

2.C.4 Additional Analysis

Weighted F1 scores are presented in Table 2.44. The level-k distribution shown in Table 2.11 is a standard distribution commonly observed in similar studies and represents the expected distribution when researchers classify such data (Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Burchardi and Penczynski, 2014; Crawford et al., 2013). Given this, weighted F1 is the most relevant metric for assessing the models' performance, and F1 macro averages are omitted.

		no-CoT		**CoT**	
		0-$Shot$	n-$Shot$	0-$Shot$	n-$Shot$
GPT-3.5	O	.694	.639	.651	.641
	O_+	.653	.757	.667	.693
GPT-4	O	.797	.843	.792	**.913**
	O_+	.768	.789	.744	.895

Table 2.44: Weighted F1 of level classification

2.D Level-k II

2.D.1 Original Instructions

Figure 2.58: Original instructions - Part 1

Classification Instructions
Thank you for participating in this experiment. In this section you find instructions as to how this experiment works. To take part in the experiment, we assume that you are familiar with the level-k model as it has been introduced by Nagel (1995) and also with the concept of team reasoning as it has been introduced by Schelling (1960). In the experiment, subjects play pure coordination games with symmetric and asymmetric payoffs. We assume that you are familiar with the concept of coordination games as they have been carried out by Crawford, Gneezy and Rottenstreich (2008).
However, in order to clarify potential questions of terminology, we reproduce the main features of the level-k model and the concept of team reasoning. In addition we provide detailed experimental instructions, which explain the game and also give you a short introduction to coordination games. Please read all information carefully in order to know how the original experiment proceeded.
Experimental Setting
Introduction
This section describes the main features of the experiment. Subjects are randomly assigned into teams of two players. For a given strategic situation, each player makes suggestions for the team action at two points in time. First, the so-called "suggested decision" and a justifying written message are exchanged between the team partners simultaneously. After this, the "final decision" is taken individually by each team player. The computer chooses randomly one of the two final decisions to obtain the "team's action."
All teams play a series of eight coordination games. Coordination games are characterised by situations in which all parties can realize mutual gains, but only by making mutually consistent decisions. Each team is randomly matched with another team. If a matched pair of teams both decide on identical team actions, they coordinate their behavior successfully and are rewarded with a payoff.
However, if both teams choose different team actions, they fail to coordinate their behavior and do not receive any payoff. Thus both teams are motivated solely to coordinate their strategies in order to obtain an outcome that is best for them. The following example illustrates a random coordination game in which each team decides on one strategy X, Y or Z simultaneously. Only if both teams make mutually consistent decisions they receive a payoff of 2 units each.

		Team 2		
		X	Y	Z
	X	(2,2)	(0,0)	(0,0)
Team 1	Y	(0,0)	(2,2)	(0,0)
	Z	(0,0)	(0,0)	(2,2)

The payoff is represented through an experimental currency unit ("Taler"). One Taler is worth 0,40 Euro. In a symmetric coordination game each team is rewarded the same payoff if they coordinate their behavior successfully. In asymmetric coordination games players usually disagree on which action they prefer to coordinate. There may be one outcome where one team disproportionately benefits in comparison to the other team.
X-Y Coordination Games
All subjects face a series of eight coordination games composed of four "X-Y Games" and four "Pie Games". We reproduce the main features and attributes of those games in the following. "X-Y Games" are characterised by a binary choice option "X" or "Y". The assignment of payoffs for successful coordination is indicated in brackets. Example:
X [6 Taler for Team 1 and 5 Taler for Team 2]
Y [5 Taler for Team 1 and 6 Taler for Team 2]
If a matched pair of teams both decide on the identical team action "X", team one receives 6 Taler and team two receives 5 Taler. If both teams chose "Y", the assignment of payoffs would be reversed. If both teams chose decisions with different labels "X" and "Y", neither team receives any payoff. The payoff differences vary within the four "X-Y treatments".
Pie Coordination Games
"Pie Games" are characterised by a visual representation of different choice options as indicated in the following figure. Each team simultaneously selects one of the three "pie slices". Each slice is labeled with an abstract decision label §, $ or #. The assignment of payoffs for successful coordination is indicated in brackets within the three slices. The first number represents the quantity of Taler for team one, the second number the quantity of Taler for team two.
--Image of a Pie with payoffs-- *(omitted see van Elten and Penczynski (2020) page 47)*
If a matched pair of teams both decide on the identical team action "#", team one receives 7 Taler and team two receives 6 Taler. If both teams chose decisions with different labels §, $ or #, neither team receives any payoff. The payoff differences alternate within the four "Pie treatments".
Note that the "X-Y Game" and the "Pie Game" might both contain one alternative that is visually distinctive from another alternative. For instance, the unshaded bottom slice is visually distinctive from the two upper slices ($ and #) that are shaded in a light grey color. We refer to a visually distinctive alternative as label-salient. Moreover an alternative might be payoff-salient in a way that it is distinctive with respect to its payoff structure. The concept of label and payoff salience is important for the classification process.
Treatment Overview
We conducted six sessions in Mannheim and three sessions Heidelberg. All sessions consist of the same eight treatments (four "X-Y games" and four "Pie games"), however the sequence of treatments in Mannheim is different from the sequence of treatments in Heidelberg. The following two tables provide a brief overview over the sessions conducted in Mannheim (session 1-3, session 7-9 [rounds 7 and 8 moved to the beginning]) and the sessions conducted in Heidelberg (session 4-6). The payoff for successful coordination is indicated in brackets. The first number represents the quantity of Taler for team one, the second number represents the quantity of Taler for team two, if both teams coordinate their behavior.
--Table presenting games played in each round in each experiment-- *(omitted see van Elten and Penczynski (2020) page 49)*

Figure 2.59: Original instructions - Part 2

Classification Process
Remember: Each player makes suggestions for the team action at two points in time. First, the so-called "suggested decision" and a justifying written message are exchanged between the team partners simultaneously. After this, the "final decision" is taken individually by each team player. The computer chooses randomly one of the two final decisions to obtain the "team's action." Your task is to classify the written messages into different categories. In the following we will describe the classification process for the analysis of the experiment.
Level k Model
Notation of the level k model
It is assumed that you are familiar with the level-k model as it has been introduced by Nagel (1995) or represented by Camerer (2004). The model here is extended to incorporate salience in the level-0 belief according to Bacharach and Stahl (2000). In order to clarify potential questions of terminology and introduce the main features of the model we quickly reproduce the main features of the model in the terminology used in this document. The level-k model of bounded rationality assumes that players only think through a certain number (k) of best responses. The model has four main ingredients:
Population distribution: This distribution reflects the proportion of types with a certain level $k \in N_0 = \{0, 1, 2, 3, 4, 5, ...\}$.
Level-0 distribution: By definition, a level-0 player does not best respond. Hence, his actions are random to the game and distributed randomly over the action space. In our case, the action space is $A = \{\{X\}, \{Y\}\}$ or $A = \{\{\S\}, \{\$\}, \{\#\}\}$. The model incorporates salience by assuming higher probabilities in the level-0 distribution for actions that are visually distinctive (salient). An action might be salient in terms of payoffs and in terms of labels. In the "X-Y" treatments, the level-0 distribution would not assign a uniform probability of 0.5 to each possible action, but $p > 0.5$ to the salient one and $q_i < p$ for the remaining actions. In the "Pie" treatments, the level-0 distribution would not assign a uniform probability of 1/3 to each possible action, but $p > 1/3$ to the salient one and $q_i < p$ for the remaining actions.
Level-0 belief: In the model, the best responses of players with $k > 0$ are anchored in what they believe the level-0 players play. Their level-0 belief might not be consistent with the level-0 distribution. For best responding, all that matters is the expected payoff from choosing an action from the action space $A = \{\{X\}, \{Y\}\}$ or $A = \{\{\S\}, \{\$\}, \{\#\}\}$. A subject would therefore decide on a particular action, when the probability is highest, that the other team chooses the same action.
Population belief: Players do not expect other players to be of the same or a higher level of reasoning. For a level-k player, the population belief is therefore defined on the set of levels strictly below k. It follows that level-0 players have no defined belief, level-1 players have a trivial belief with full probability mass on $\{0\}$, level-2 players have a well defined belief on $\{\{0\}, \{1\}\}$. From level 3 higher order beliefs are relevant as level-3 players have to form a belief about level-2's beliefs.
Characterisation of the different levels
Level 0 The player does not exhibit any strategic reasoning whatsoever. Different versions of this might be randomly chosen or purely guessed actions, misunderstanding of the game structure or other non-strategic 'reasons' for picking a location, e.g. by taste or salience. It is important that no best-responding to the other's play occurs. There could be considerations of what others might play, but without best responding to it. Examples: "Well, it's a pure guess", "There are no arguments. Simply choose any."
Level 1 This player best responds to some belief about the other teams' action. However, he does not realise that others will be strategic as well. Example: "They are probably picking X, so we do as well", "The other team would naturally go for the visual distinctive bottom slice, no?"
Level 2 This player not only shows the basic strategic consideration of playing best response (matching/mismatching), but also realises that other players best respond as well according to the belief they entertain. A level-2 player clearly contemplates how the other player might best respond to his frame. The player plays a best response to this hypothesised consideration. Example: "The other team may think we are most attracted to the alternative # with the highest payoff. In order to coordinate our behavior we should also choose the # slice."
Level 3 This player realises that others could be level-2 and reacts by best responding to the associated expected play. Put differently, he realises that others realise that others best respond to their initial belief. Therefore, a level-3 player clearly states that his opponent expects that he (the level-3 player at question) best-responds to a certain belief.
Level 4, 5, ... The process goes on in a similar fashion. A level k player realises that other subjects could be level k-1 and reacts by best responding to the associated expected play.
Category 1: Lower and upper bound on the level of reasoning
Your aim
is to classify the written messages into the underlying level $k \in N_0 = \{0, 1, 2, 3, 4, 5, ...\}$ of reasoning. For a given statement it might not be possible to exactly determine the underlying level of reasoning. To extract as much information as possible, we ask you to indicate a lower and an upper bound on the level of reasoning.
For the lower bound on the level of reasoning, you should ask yourself: "What is the minimum level of reasoning that this statement clearly exhibits?" Once noted, you should be able to say to yourself: "It seems impossible that the players' level of reasoning is below this number!" Here we ask you to be very cautious with the classification, not giving away high levels easily.
The upper bounds should give the maximum level of reasoning that could be interpreted into the statement. Therefore, you should ask yourself: "What is the highest level of reasoning that can be underlying this statement?" Once noted, you should be able to say: "Although maybe not clearly communicated, this statement could be an expression of this level. If the player reasoned higher than this number, this was not expressed in the statement!" For both lower and upper bound, please refer to the characterisation of the different levels.
There are two necessary conditions for a player to exhibit a level greater than 0. First, the player has to be responsive to the salience of the games' framing. Secondly, the player has to be strategic in best-responding to his level-0 belief, which is shaped by label or payoff salience. If he did not react to salience, he would have no reason to chose one over the other object, resulting in random level 0 play.
For this category, the excel-sheet for the classification will feature a drop-down menu where you can choose upper and lower bounds between 0 and 5. If no inference can be made since nothing or nothing to the point is written, you can choose not applicable (n/a).
Category 2: Level-0 belief
Your aim
is to indicate the underlying level-0 belief that is connected with the lowest possible level of reasoning. If level reasoning is observed in the statement, there has to be a starting point in the argument which states an attraction or aversion to one alternative. This is then not derived by strategic reasons, but is an intuitive reaction to the framing of the coordination game.
Otherwise, level reasoning would not occur. Please indicate the underlying level-0 belief that is connected with the lowest possible level of reasoning. Note that the level-0 belief of a person reasoning on an odd level, i.e. level 1, 3, 5, etc. is always with respect to how a player of the opposite side intuitively reacts to the framing. The belief of a person reasoning on an even level, i.e. level 2, 4 etc. is always with respect to what the opposite type believes about the own type's intuitive reaction.

Figure 2.60: Original instructions - Part 3

There are two kinds of framing in these games. On the one hand, subjects might react to the framing of the coordination game (label salience). Imagine a subject that you classify to be level-1. It might communicate that the other team is most attracted to the visual distinctive white bottom slice § and therefore proposes Â§ as team decision. A subject that you classify to be level-2 might indicate that the other team believes that one's own team is more likely to choose "X", because this alternative is mentioned first on the screen. To reflect a level-0 belief of an attraction to X or Y , or to #, §, or $, the excel-sheet features a drop-down menu that allows to indicate such a preference or an indifference. If such a preference or indifference over labels is not indicated, or if the subjects' level of attractiveness cannot be distinguished or is not expressed clearly within the message, please indicate that the level-0 belief from the message does not exhibit any label salience.
On the other hand, subjects might respond to the payoffs (payoff salience). For example, consider a subject that you classify to be level-1. It might communicate that the other team is most likely to choose alternative X as it offers the highest payoff to this very team. Or, a subject that you classify to be level-2 might indicate that the other team remains of the conviction that one's own team is not attracted to the action that gives one's own team high payoffs. To reflect the exhibited level-0 beliefs you can indicate in the excel-sheet whether the team that the level-0 belief is formed about is believed to be attracted to a) the action that yields --under coordination-- a higher payoff for this team, to b) the action that yields --under coordination-- a higher payoff to the other team or c) is indifferent. If no such preference or indifference over salient payoff actions is indicated, please indicate that the level-0 belief from the message does not exhibit any payoff structure.
Please note that payoff and label salience are not mutually exclusive, please indicate both if both is expressed in the message. Finally, for players whose lower bound is 0, the level-0 belief classification can be used to indicate whether a level-0 player states for his action a preference with respect to label or payoff salience.
Classification Summary
In coordination games both teams are motivated solely to coordinate their strategies in order to obtain an outcome that is best for them. For a given strategic situation, each player proposes a suggested decision and writes a justifying written message to the team partner. Your task is to classify the written messages into different categories that are summarized in the following:
Category 1 Please classify the written messages into the underlying level $k \in N_0 = \{0, 1, 2, 3, 4, 5, ...\}$ of reasoning. Provide the lower and an upper bounds on the level of reasoning as described.
Category 2 Please indicate the underlying level-0 belief that is connected with the lowest conceivable level of reasoning. Information about the underlying level-0 belief that one might obtain out of the communication is how subjects respond to payoffs (payoff salience) and how subjects react to the framing (label salience) of the coordination games.

2.D.2 Prompts

Figure 2.61: Prompts O - Part 1

```
# General Task
- Evaluate player’s message from a specific coordination game to identify their decision-process and to classify their level of strategic thinking.
# Role Persona
- Act as a behavioral economist specialized in coordination games, decision salience and text analysis.
# Context
## Game Mechanics
- Subjects (players) participate in an experiment where they play a coordination game.
- Players are assigned into teams of two.
- Each team is matched with another team to play a series of coordination games where the teams try to coordinate on a specific alternative.
- If both teams pick the same alternative, each team is rewarded with a payoff.  Otherwise, neither team receives any payoff.
- Each player sends a suggested decision and a justifying message to their teammate.
- Coordination game is not played within the team members but between the two teams.
- There is no communication between the teams.
- Payoffs are represented in currency called Taler.
- 1 Taler = 40 cents (ct).
- Each player plays a series of 8 (rounds of) coordination games, split between *X-Y* games and *Pie* games.
## Coordination Games
- Payoff tables for each variation of each game is represented below.
- In each payoff table, `d` represents decision taken, `π1` represents payoff of team 1, and `π2` represents payoff of team 2.
- Payoff tables only represent the cases where both teams match in the given decision `d`, if teams’ decisions do not match, each team receives 0 Taler.
### Pie Game
- Payoffs are displayed on a pie chart that is divided into three equally sized segments.
- Top left segment is labeled as `$`.
- Top right segment is labeled as `#`.
- Top segments are shaded in gray.
- Bottom segment is labeled as `§` and is highlighted in white.
- There are 4 payoff variations labeled as S1, S2, AM2 and AM4.
#### S1
| d     |π1, π2|
|-------|------|
| L ($) | 5, 5 |
| R (#) | 5, 5 |
| B (§) | 5, 5 |
#### S2
| d     |π1, π2|
|-------|------|
| L ($) | 6, 6 |
| R (#) | 6, 6 |
| B (§) | 5, 5 |
#### AM2
| d     |π1, π2|
|-------|------|
| L ($) | 5, 6 |
| R (#) | 6, 5 |
| B (§) | 6, 5 |
#### AM4
| d     |π1, π2|
|-------|------|
| L ($) | 6, 7 |
| R (#) | 7, 6 |
| B (§) | 7, 5 |
### X-Y Game
- Alternatives are displayed in two consecutive lines.
- Alternative `X` is displayed on the first line.
- Alternative `Y` is displayed on the second line.
- There are 4 payoff variations labeled as SL, ASL, AML and ALL.
#### SL
| d |π1, π2 |
|---|-------|
| X | 5, 5   |
| Y | 5, 5   |
#### ASL
| d |π1, π2 |
|---|-------|
| X | 5, 5.1 |
| Y | 5.1, 5 |
#### AML
| d |π1, π2 |
|---|-------|
| X | 5, 6   |
| Y | 6, 5   |
#### ALL
| d |π1, π2 |
|---|-------|
| X | 5, 10  |
| Y | 10, 5  |
```

Figure 2.62: Prompts O - Part 2

```
## Salience in Decisions
### Salient Label
- Both *Pie* and *X-Y* games may contain a decision that is visually distinctive from other available decision(s).
- In the Pie game, the bottom segment (slice) that is highlighted in white in the Pie game is visually distinctive
from the upper slices ($ and #) that are shaded in gray color.
- In the X-Y game, the decision X may be perceived as more salient because it is the top or the first decision
presented.
- We refer to a visually distinctive decision as **label-salient**.
### Salient Payoff
- A decision might be also **payoff-salient** in a way that is distinctive with respect to its payoff structure.
- A decision alternative that provides the highest or lowest payoff for one of the teams can be considered as payoff
salient.
- In ALL variation of the X-Y game, option X may be considered to have high payoff salience for team 1, while option Y
may be considered to have high payoff salience for team 2.
## Level-k Model
### Level-0 Distribution
- Level-0 player does not best respond but instead play according to some probability distribution over the action
space (level-0 distribution).
   + For the X-Y game the action space is X,Y
   + For the Pie game the action space is $,#,$
- Without any salient actions, a level-0 player’s actions distributed randomly and evenly over the action space.
- The model incorporate salience by assuming higher probabilities for either payoff salience or label salience actions
in the level-0 player’s action space distribution.
  + In the X-Y game, level-0 distribution would not assign equal probability of 0.5 to each action but instead assign
p > 0.5 to the salient action and and q < 0.5 for the non-salient action.
   + In the Pie game, level-0 distribution would not assign equal probability of 1/3 to each possible action, but p >
1/3 to the salient action and q < 1/3 for the non-salient actions.
### Level-0 Belief
- The best response of a level-k (k > 0) player is anchored in what he believes a level-0 player plays.  This is
called the level-0 belief of the level-k player.
- A level-k player performs k many iterative best responses and always starts its iterative reasoning from his level-0
belief.
### Population Belief
- Players do not expect other players to be of the same or higher level of reasoning.  For a level-k player, the
population belief is therefore defined on the set of levels strictly below k.
- A level-0 player has no defined population belief.
- A level-1 player has a trivial belief with full probability on all other players being level-0.
- A level-2 player has a well defined belief distribution on all other players being level-1 and level-0.
- A level-3 player has a well defined belief distribution on all other players being level-2, level-1 and level-0 (and
so on for higher order of levels of thinker).
### Characterization of different levels
#### Level-0
- A level-0 player does not exhibit any strategic reasoning.
- A level-0 player may be randomly choosing or purely guessing an action.
- A level-0 player may misunderstand the game structure.
- A level-0 player may pick an action for non-strategic reasons such as taste or salience.
- A level-0 player does not best respond to other players potential actions.
##### Examples
- "Well, it’s a pure guess"
- "There are no arguments.  Simply choose any."
#### Level-1
- A level-1 player best responds to some belief about the other teams’ action, but he does not consider that the other
teams may be strategic as well.
##### Examples
- "They are probably picking X, so we do as well"
- "The other team would naturally go for the visually distinctive bottom slice"
#### Level-2
- A level-2 player does not only show the basic strategic consideration of playing best response to a his level-0
belief, but also recognizes that other players may best respond as well according to their level-0 belief.
- A level-2 player clearly contemplates how the other player might best respond to his frame.  The player plays a best
response to this hypothesized consideration.
##### Example
- "The other team may think we are most attracted to the alternative  with the highest payoff.  In order to coordinate
our behavior, we should also choose the  slice."
#### Level-3
- A level-3 player realizes that the other players (team) could be level-2 and best-responds accordingly
- A level-3 player realizes that others realize that others best-respond to their level-0 belief.
- A level-3 player states that his opponent expects that he (the level-3 player at question) best-responds to a
certain belief.
#### Level-4,5,...
- The process goes on in a similar fashion
- A level-k player realizes that other subjects could be level-(k-1) and reacts by best responding to the associated
expected play.
```

Figure 2.63: Prompts O - Part 3 - Classification task

```
# Classification Task
## Task 1:  Level-0 Belief
- Your task is to indicate the underlying level-0 belief that is connected with the lowest possible level of
reasoning.
- If the level of reasoning is observed in the message, there has to be a starting point in the argument which states
an attraction or aversion to one alternative.  This is then not derived by strategic reasons, but is an intuitive
reaction to the framing of the coordination game.  Otherwise, level reasoning would not occur.
- Information about the underlying level-0 belief that one might obtain out of the communication is how players
respond to payoffs (payoff salience) and how players react to framing (label salience) of the coordination games.
- Level-0 belief of a person reasoning on an odd level (level 1,3 or 5) is always with respect to how a player of the
opposite side intuitively reacts to the framing (salience).
- Level-0 belief of a person reasoning on an even level (level 2 or 4) is always with respect to what the opposite
type believes about the own type's intuitive reaction.
### Salience
- There are two kinds of salience in these games:  label and payoff
- Payoff and label salience are not mutually exclusive.  A player may display both payoff and label salience within
the same reasoning.
- For players whose lower bound is 0, the level-0 belief classification can be used to indicate whether a level-0
player states for his action a preference with respect to label or payoff salience.
#### Label
- Players may react to the framing of the coordination game (label salience)
- A player that you classify to be level-1 may communicate that the other team is most attracted to the visual
distinctive white bottom slice §and therefor proposes Â§ as team decision.
- A player that you classify to be level-2 may indicate that the other team believes that one's own team is more
likely to be choose ‘X‘ because this alternative is mentioned first on the screen.
- For X-Y game, classify the label salience of the message as one of the following:
  + prefers ‘X‘ over ‘Y‘ (a level-0 belief of an attraction to ‘X‘)
  + prefers ‘Y‘ over ‘X‘ (a level-0 belief of an attraction to ‘Y‘)
  + is indifferent between labels ‘X‘ and ‘Y‘
  + does not exhibit label salience
- For Pie game, classify the label salience of the message as one of the following:
  + prefers ‘$‘ (a level-0 belief of an attraction to ‘$‘)
  + prefers ‘#‘ (a level-0 belief of an attraction to ‘#‘)
  + prefers ‘§‘ (a level-0 belief of an attraction to ‘§‘)
  + is indifferent across labels
  + does not exhibit label salience
#### Payoff
- Players respond to the payoffs' salience.
- A player you classify to be level-1 may communicate that the other team is most likely to choose alternative ‘X‘ as
it offers the highest payoff to this very team.
- A player you classify to be level-2 may indicate that the other team remains of the conviction that one's own team
is not attracted to the action that gives one's own team high payoffs.
- For either game, classify the payoff salience of the message as one of the following:
  + prefers high payoffs (a level-0 belief of an attraction to high payoff)
  + prefers low payoffs (a level-0 belief of an attraction to low payoff)
  + is indifferent across payoffs
  + does not exhibit payoff salience
## Task 2:  Level of Strategic Thinking
- Your task is to classify the written messages into the underlying level-k of reasoning.
- For a given statement it might not be possible to exactly determine the underlying level of reasoning.
- To extract as much information as possible, you are asked to indicate a lower and an upper bound on the level of
reasoning
- Use the information provided under the subsection "Characterization of different levels" in the "Context" section
above to classify lower and upper bounds on the level of reasoning.
### Lower bound
- Ask yourself:  "What is the minimum level of reasoning that this statement clearly exhibits?".  Once noted, you
should be able to state to yourself:  "It seems impossible that the players' level of reasoning is below this number!"
- Be very cautious with the classification of the lower bound.
- Do not give away high levels easily.
### Upper bound
- The upper bounds should give the maximum level of reasoning that could be interpreted into the statement.
- You should ask yourself:  "What is the highest level of reasoning that can be underlying this statement?".  Once
noted, you should be able to say:  "Although maybe not clearly communicated, this statement could be an expression of
this level.  If the player reasoned higher than this number, this was not expressed in the statement!"
## Necessary conditions for a player to exhibit a level of thinking greater than 0:
1.  The player has to be responsive to the salience of the games' framing (either payoff or label salience)
2.  The player has to be strategic in best-responding to his level-0 belief, which is shaped by label or payoff
salience.  If he did not react to salience, he would have no reason to choose one over the other action, resulting in
random level 0 play.
```

Figure 2.64: Prompts O - Part 4

```
# Classification Coding
## Label Salience
### X-Y Game
- Code as ‘X‘ if player’s label salience is "prefers X over Y"
- Code as ‘Y‘ if player’s label salience is "prefers Y over X"
- Code as ‘∼‘ if player’s label salience is "indifferent across payoffs"
- Code as ‘no‘ if the player does not exhibit payoff salience.
### Pie Game
- Code as ‘$‘ if player’s label salience is "prefers $"
- Code as ‘#‘ if player’s label salience is "prefers #"
- Code as ‘§‘ if player’s label salience is "prefers §"
- Code as ‘∼‘ if player’s label salience is "indifferent across labels"
- Code as ‘no‘ if the player does not exhibit label salience.
## Payoff Salience
- Code as ‘H‘ if player’s payoff salience is "prefers high payoffs"
- Code as ‘L‘ if player’s payoff salience is "prefers low payoffs"
- Code as ‘∼‘ if player’s payoff salience is "indifferent across payoffs"
- Code as ‘no‘ if the player does not exhibit payoff salience.
## Upper and Lower Bounds
- 0,1,2,3,4 or 5.
# Input Format
   Team:
   Game:
   Decision:
   Message:
# Constraint
- Follow the below output format
# Output Format
   Label Salience:
   Payoff Salience:
   Lower Bound:
   Upper Bound:
```

Figure 2.65: Prompts O_+ - Examples

```
# Examples
- "X is first, let’s pick X" (Level-0 belief: prefers X. Label salience: X. Payoff salience: no. Level: 0.)
- "§is highlighted in white, hence §" (Level-0 belief: prefers §. Label salience: §. Payoff salience: no. Level: 0.)
- "Y provides a higher payoff, let’s go X" (Level-0 belief: prefers higher payoff. Label salience: no. Payoff salience: higher payoff. Level: 0.)
- "It is random. You pick X and I pick Y." (Level-0 belief: random play. Label salience: no. Payoff salience: no. Level: 0.)
- "Other team may pick X as it is on top, let’s pick X" (Level-0 belief: prefers X. Label salience: X. Payoff salience: no. Level: 1.)
- "X is first, let’s pick X. And other team may think the same way." (Level-0 belief: prefers X. Label salience: X. Payoff salience: no. Level: 1.)
- "People will go for the highlighted segment. Hence §" (Level-0 belief: prefers §. Label salience: §. Payoff salience: no. Level: 1.)
- "Other team may want to the high payoff for themselves, let’s coordinate with them and pick the higher payoff for them (which is the lower payoff for us)." (Level-0 belief: prefers higher payoff. Label salience: no. Payoff salience: higher payoff. Level: 1.)
- "Others will think we go for top. So let’s go for top." (Level-0 belief: prefers X. Label salience: X. Payoff salience: no. Level: 2.)
- "The other team will think we pick the highlighted segment. So we should coordinate and pick §" (Level-0 belief: prefers §. Label salience: §. Payoff salience: no. Level: 2.)
- "Other team may think that we want the high payoff for us, let’s coordinate with them and pick the higher payoff for us." (Level-0 belief: prefers higher payoff. Label salience: no. Payoff salience: higher payoff. Level: 2.)
- "Other team may want to the high payoff for themselves. But they may assume the same thing about us and pick the alternative that gives us the higher payoff. So let’s coordinate with them and pick the higher payoff for us (which is the lower payoff for them)." (Level-0 belief: prefers higher payoff. Label salience: no. Payoff salience: higher payoff. Level: 2.)
- "Others will think that we think that they will go for top. So let’s go for top." (Level-0 belief: prefers X. Label salience: X. Payoff salience: no. Level: 3.)
- "The other team will think that we think that they pick the highlighted segment. So we should coordinate and pick §" (Level-0 belief: prefers §. Label salience: §. Payoff salience: no. Level: 3.)
- "Other team may think that we think that they want the high payoff for themselves, let’s coordinate with them and pick the higher payoff for them." (Level-0 belief: prefers higher payoff. Label salience: no. Payoff salience: higher payoff. Level: 3.)
- "Other team may think that we want the high payoff for us. But they may assume the same thing about us and pick the alternative that gives us the higher payoff for themselves. So let’s coordinate with them and pick the higher payoff for them." (Level-0 belief: prefers higher payoff. Label salience: no. Payoff salience: higher payoff. Level: 3.)
```

Figure 2.66: Prompts O_+ - Part 1

```
# General Task
- Evaluate player’s message from a specific coordination game to identify their decision-process and to classify their
level of strategic thinking.
# Role Persona
- Act as a behavioral economist specialized in coordination games, decision salience and text analysis.
# Context
## Game Mechanics
- Subjects (players) participate in an experiment where they play a coordination game.
- Players are assigned into teams of two.
- Each team is matched with another team to play a series of coordination games where the teams try to coordinate on a
specific alternative.
- If both teams pick the same alternative, each team is rewarded with a payoff.  Otherwise, neither team receives any
payoff.
- Each player sends a suggested decision and a justifying message to their teammate.
- Coordination game is not played within the team members but between the two teams.
- There is no communication between the teams.
- Payoffs are represented in currency called Taler.
- 1 Taler = 40 cents (ct).
- Each player plays a series of 8 (rounds of) coordination games, split between *X-Y* games and *Pie* games.
## Coordination Games
- Payoff tables for each variation of each game is represented below.
- In each payoff table, ‘d‘ represents decision taken, ‘π1‘ represents payoff of team 1, and ‘π2‘ represents payoff of
team 2.
- Payoff tables only represent the cases where both teams match in the given decision ‘d‘, if teams’ decisions do not
match, each team receives 0 Taler.
### Pie Game
- Payoffs are displayed on a pie chart that is divided into three equally sized segments.
- Top left segment is labeled as ‘$‘.
- Top right segment is labeled as ‘#‘.
- Top segments are shaded in gray.
- Bottom segment is labeled as ‘§‘ and is highlighted in white.
- There are 4 payoff variations labeled as S1, S2, AM2 and AM4.
#### S1
| d     |π1, π2|
|-------|------|
| L ($) | 5, 5 |
| R (#) | 5, 5 |
| B (§) | 5, 5 |
#### S2
| d     |π1, π2|
|-------|------|
| L ($) | 6, 6 |
| R (#) | 6, 6 |
| B (§) | 5, 5 |
#### AM2
| d     |π1, π2|
|-------|------|
| L ($) | 5, 6 |
| R (#) | 6, 5 |
| B (§) | 6, 5 |
#### AM4
| d     |π1, π2|
|-------|------|
| L ($) | 6, 7 |
| R (#) | 7, 6 |
| B (§) | 7, 5 |
### X-Y Game
- Alternatives are displayed in two consecutive lines.
- Alternative ‘X‘ is displayed on the first line.
- Alternative ‘Y‘ is displayed on the second line.
- There are 4 payoff variations labeled as SL, ASL, AML and ALL.
#### SL
| d |π1, π2 |
|---|------- |
| X | 5, 5   |
| Y | 5, 5   |
#### ASL
| d |π1, π2 |
|---|------- |
| X | 5, 5.1 |
| Y | 5.1, 5 |
#### AML
| d |π1, π2 |
|---|------- |
| X | 5, 6   |
| Y | 6, 5   |
#### ALL
| d |π1, π2 |
|---|------- |
| X | 5, 10  |
| Y | 10, 5  |
```

Figure 2.67: Prompts O_+ - Part 2

```
## Salience
- Label salience:  players may react to the framing of alternatives.
- Payoff salience:  Players may react to the payoff differences of alternatives .
## Level-k Model
- A level-k player performs k many iterative best responses and always starts its iterative reasoning from his level-0
belief.  This starting point is called the level-0 belief of the level-k player.
- Level-0 belief is the belief of the level-k player on how the level-0 player will potentially play the game.
- A level-0 player picks an alternative for non-strategic, instinctive reasons such as payoff or label salience.
- A level-0 player does not best respond to other players potential actions.
- A level-0 player may be randomly choosing or purely guessing an action.
- A level-0 player may misunderstand the game structure.
- A level-1 player assumes that the other team consists of level-0 players and best responds based on his level-0
belief to these level-0 players.
- A level-k (k>1) player recognizes the possibility that the other team may consist of level-(k-1) players.
- A level-k player assumes that the level-(k-1) players assume he (the level-k player) is a level-(k-2) player.
## Level-0 Belief
- Level-0 belief of a player reasoning on an odd level (level 1,3 or 5) is always with respect to how a player of the
opposite side intuitively reacts to the framing (label or payoff salience).
- Level-0 belief of a player reasoning on an even level (level 2 or 4) is always with respect to what the opposite
type believes about the own type’s intuitive reaction (label or payoff salience).
## Lower bound
- The minimum level of reasoning that the message clearly exhibits.
## Upper bound
- The maximum level of reasoning that can be inferred from the message.
# Classification Tasks
- Classify player’s label or payoff salience (if any).
- Classify lower and upper bounds for the player’s level of reasoning.
# Classification Coding
## Label Salience
### X-Y Game
- Code as ‘X‘ if player’s label salience is "prefers X over Y"
- Code as ‘Y‘ if player’s label salience is "prefers Y over X"
- Code as ‘∼‘ if player’s label salience is "indifferent across payoffs"
- Code as ‘no‘ if the player does not exhibit payoff salience.
### Pie Game
- Code as ‘$‘ if player’s label salience is "prefers $"
- Code as ‘#‘ if player’s label salience is "prefers #"
- Code as ‘§‘ if player’s label salience is "prefers §"
- Code as ‘∼‘ if player’s label salience is "indifferent across labels"
- Code as ‘no‘ if the player does not exhibit label salience.
## Payoff Salience
- Code as ‘H‘ if player’s payoff salience is "prefers high payoffs"
- Code as ‘L‘ if player’s payoff salience is "prefers low payoffs"
- Code as ‘∼‘ if player’s payoff salience is "indifferent across payoffs"
- Code as ‘no‘ if the player does not exhibit payoff salience.
## Upper and Lower Bounds
- 0,1,2,3,4 or 5.
# Examples (see Figure 2.65) (Only used in n-shot treatments)
# Input Format
   Team:
   Game:
   Decision:
   Message:
# Constraint
- Follow the below output format
# Output Format
   Label Salience:
   Payoff Salience:
   Lower Bound:
   Upper Bound:
```

2.D.3 Additional Examples

Figure 2.68: Level-k II - Example message 3

```
Team:  1
Game:  ASL
Decision:  Y
Message:  I think they will propose a larger payout to us because they think we will keep it for ourselves to
ensure coordination...  however, I can imagine that they have the exact same mindset as I do and will do the
opposite...  I'll call it Y
```

	GPT-3.5				GPT-4			
	L_B	U_B	S_L	S_P	L_B	U_B	S_L	S_P
O	1	2	$\sim$	H	1	2	no	H
O_+	1	2	no	H	3	3	no	H
RAs	3	3	no	H	3	3	no	H

Table 2.45: Human and GPT classifications of message in Figure 2.68 using CoT prompting for n-shot O_+ and 0-shot O

Figure 2.69: Level-k II - Generated rationals for 0-shot O and n-shot O_+ under CoT for message in Figure 2.68

GPT-3.5

Output using O
Reasoning:
The player seems to be considering the other team's potential decision and how it relates to their own team's payoff. The player is suggesting that the other team might propose the action that yields a higher payoff for them, assuming that their team will keep the higher payoff for themselves. However, the player also acknowledges that the other team might have the same thought process and propose the opposite action. This indicates a level of strategic thinking beyond pure guessing.
Output using O_+
Reasoning:
The player seems to be considering the other team's potential strategy and trying to anticipate their decision based on the belief that the other team may think they want the higher payoff for themselves. This indicates a level of strategic thinking beyond just instinctive or random play.

GPT-4

Output using O
Reasoning:
The player's message indicates a complex thought process involving multiple levels of strategic reasoning. The player starts by suggesting that the other team will propose the action that gives team 1 the larger payoff, under the assumption that team 1 will want to keep the larger payoff for themselves. This is a strategic consideration and indicates that the player is not operating at level-0, which would involve no strategic reasoning or a random choice. The player then goes on to consider that the other team might have the same line of thinking and might, therefore, choose the opposite action to ensure coordination. This indicates a higher level of reasoning, as the player is considering the other team's potential strategy and how they might anticipate team 1's actions.
The player's final decision to suggest 'Y' seems to be a best response to the anticipated strategy of the other team, which suggests a level-2 reasoning. The player is considering that the other team is also reasoning strategically (level-1) and is best responding to that.
The message does not clearly indicate a level-3 reasoning, as the player does not explicitly state that they believe the other team is considering that team 1 might be level-2 thinkers. Therefore, the upper bound of the reasoning level would be level-2.
Regarding the level-0 belief, the player's message indicates a consideration of the payoffs, as they mention the larger payoff and the desire to keep it. This suggests a payoff salience in their reasoning. There is no explicit mention of label salience in the message.
Output using O_+
Reasoning:
The player's message suggests a complex line of reasoning. The player starts by considering that the other team will propose the alternative that gives Team 1 the larger payoff, under the assumption that Team 1 would want to keep the larger payoff for themselves. This is a strategic consideration that goes beyond a simple preference for a label or payoff, indicating a higher level of reasoning.
The player then goes on to consider that the other team might have the same line of reasoning and therefore might do the opposite. This is an additional layer of strategic thinking, suggesting that the player is trying to anticipate the other team's anticipation of their own team's actions.
The player concludes by suggesting "Y" as the decision. Since the game is ASL, where the payoffs are 5 for X and 5.1 for Y for Team 1, and 5.1 for X and 5 for Y for Team 2, the player is suggesting to pick the alternative that gives the other team the higher payoff, which aligns with the initial part of the reasoning.
The message indicates a level-0 belief that the other team prefers higher payoffs (since they are expected to propose the alternative that gives Team 1 the larger payoff). However, the player's own decision does not directly reflect a preference for higher payoffs for themselves, but rather a strategic move to coordinate with the other team's expected strategy.
The player's reasoning involves anticipating the other team's anticipation of their own team's actions, which is characteristic of level-3 reasoning. The player assumes that the other team is level-2, who in turn assumes that Team 1 is level-1.

Figure 2.70: Level-k II - Example message 4

```
Team:  1
Game:  ALL
Decision:  X
Message:  We also need to be selfish sometimes.
```

	GPT-3.5				GPT-4			
	L_B	U_B	S_L	S_P	L_B	U_B	S_L	S_P
O	1	1	*no*	H	0	0	*no*	H
O_+	0	1	*no*	H	0	0	*no*	H
RAs	0	0	*no*	*no*	0	0	*no*	*no*

Table 2.46: Human and GPT classifications of message in Figure 2.72 using CoT prompting for n-shot O_+ and 0-shot O

Figure 2.71: Level-k II - Generated rationals for 0-shot O and n-shot O_+ under CoT for message in Figure 2.72

GPT-3.5

Output using O
Reasoning:
The player's message suggests that they believe their team should be selfish. This indicates a preference for the action X, which yields a higher payoff for their team. The player is exhibiting payoff salience by being attracted to the action that provides a higher payoff for their team.
Output using O_+
Reasoning:
The player's message suggests a focus on being selfish, which indicates a potential preference for the higher payoff.

GPT-4

Output using O
Reasoning:
The message "Wir müssen auch mal egoistisch sein," which translates to "We also have to be selfish sometimes," suggests that the player is considering the payoffs for their own team. In the context of the ALL game, the decision 'X' yields a higher payoff for Team 1 (5 Taler for Team 1 and 10 Taler for Team 2). By choosing 'X', the player is indicating a preference for the action that benefits their own team more, which is consistent with the message's emphasis on being selfish.
Given the message, the player is not making any explicit reference to the strategic considerations of what the other team might do or how the other team might perceive their team's actions. The message does not indicate any iterative reasoning about the other team's potential level of thinking. Therefore, the player's reasoning seems to be at a basic strategic level, where they are considering their own payoffs without explicitly considering the other team's response.
The message does not mention anything about the label salience, such as the order of the options or any visual distinction between them. Therefore, we cannot infer any label salience from this message.
Output using O_+
Reasoning:
The message "Wir müssen auch mal egoistisch sein" translates to "We also have to be selfish sometimes." This message suggests that the player is considering their own interests, which in the context of the ALL game, where the payoffs are asymmetric (5 Taler for one team and 10 Taler for the other), implies a focus on the higher payoff for themselves. The message does not provide any indication of label salience, as it does not reference the labels X or Y. The level of reasoning is not explicitly stated, but the use of "egoistisch" (selfish) suggests a level-0 belief that prefers higher payoffs. Since the message does not indicate any anticipation of the other team's reasoning or actions, it is best classified as level-0 reasoning.

Figure 2.72: Level-k II - Example message 5

```
Team:  1
Game:  ASL
Decision:  X
Message Hmm, tricky because hardly any difference - I'll be generous - alternative X
```

	GPT-3.5				GPT-4			
	L_B	U_B	S_L	S_P	L_B	U_B	S_L	S_P
O	0	0	no	no	0	1	no	$\sim$
O_+	0	1	no	$\sim$	1	1	no	L
RAs	0	1	no	H	0	1	no	H

Table 2.47: Human and GPT classifications of message in Figure 2.72 using CoT prompting for n-shot O_+ and 0-shot O

Figure 2.73: Level-k II - Generated rationals for 0-shot O and n-shot O_+ under CoT for message in Figure 2.72

GPT-3.5

Output using O
Reasoning:
The player's message indicates that there is hardly any difference between the options and he is being generous by choosing alternative X. The message does not explicitly exhibit any strategic reasoning or salience. The player seems to be randomly choosing an option without any clear indication of salience or strategic thinking.
Output using O_+
Reasoning:
The player seems to be considering the small difference in payoffs and is leaning towards being generous by choosing alternative X.

GPT-4

Output using O
Reasoning:
The message provided by the player from Team 1 in the ASL game, where they have chosen decision X, is brief and suggests a level of indifference or difficulty in distinguishing between the options due to the small difference in payoffs ("kaum Unterschied"). The use of the word "großzügig" (generous) implies that the player is aware of the slight asymmetry in the payoffs but decides to not prioritize this difference. The player's choice of "Alt. X" (alternative X) does not provide a clear strategic rationale for the decision.
The player does not mention any aspect of the label salience in their message. There is no indication that the player prefers X over Y due to its position or any other visual feature. Therefore, the player does not exhibit label salience.
The player acknowledges the small difference in payoffs but does not express a clear preference for the higher payoff. The use of "großzügig" could suggest a willingness to not focus on the slight advantage in payoffs for their team, which could be interpreted as a preference for low payoffs. However, since the player ultimately chooses X, which for the ASL game would provide a slightly higher payoff for their own team (5 Taler for Team 1 vs. 5.1 Taler for Team 2), it is not clear if the player is indeed being generous to the other team or simply indifferent to the payoff difference. Therefore, the player's message could be interpreted as showing indifference across payoffs.
The player's message does not clearly exhibit strategic reasoning. The player does not articulate a belief about what the other team might choose or how their own choice might influence the other team's decision. The message seems to reflect a random or non-strategic choice, which aligns with a level-0 player. However, the acknowledgment of the small difference in payoffs and the use of "großzügig" could suggest a minimal level of strategic consideration, potentially indicating a level-1 reasoning where the player is aware of the payoffs but does not articulate a strategic response to them.
Given the lack of clear strategic reasoning and the absence of a stated belief about the other team's likely action, the minimum level of reasoning that this statement clearly exhibits is level-0.
The highest level of reasoning that could be underlying this statement, given the player's acknowledgment of the payoff difference, could be level-1. The player does not express any higher-order strategic considerations, such as anticipating the other team's response to their own decision or considering multiple levels of reasoning.
Output using O_+
Reasoning:
The message "Hmm, fies weil kaum Unterschied - ich bin mal großzügig - Alt. X" translates to "Hmm, tricky because there's hardly any difference - I'll be generous - Alt. X". This message indicates that the player is aware of the small difference in payoffs between the alternatives X and Y in the ASL game, where the payoffs are 5 Taler for the player's team and 5.1 Taler for the other team for alternative X, and vice versa for alternative Y. The use of the word "generous" suggests that the player is willing to choose the option that gives the other team a slightly higher payoff, which is alternative X for the ASL game.
The player does not express a preference for a label (X or Y) but rather comments on the small difference in payoffs, indicating a recognition of payoff salience. However, the player does not explicitly state a preference for higher or lower payoffs, just that they are being "generous" by choosing the alternative that gives the other team a slightly higher payoff. This could be interpreted as a preference for lower payoffs for themselves (since they are giving the other team the higher payoff), but it is not a strong indication of such a preference.
The player's level of reasoning seems to be at least level-1 because they are considering the other team's potential payoff and are making a decision based on that consideration. However, there is no clear indication that the player is thinking beyond this, such as considering what the other team might think they will do. Therefore, the upper bound of reasoning is also level-1.

Figure 2.74: Level-k II - Example message 6

```
Team:  2
Game:  S2
Decision:  §
Message:  Even if § is the lower payout, it is the only one where both teams get a clear solution, so they don't
have to hesitate!
```

	GPT-3.5				GPT-4			
	L_B	U_B	S_L	S_P	L_B	U_B	S_L	S_P
O	1	2	§	$\sim$	0	1	§	L
O_+	1	2	§	*no*	1	1	*no*	L
RAs	0	1	§	*no*	0	1	§	*no*

Table 2.48: Human and GPT classifications of message in Figure 2.74 using CoT prompting for n-shot O_+ and 0-shot O

Figure 2.75: Level-k II - Generated rationals for 0-shot O and n-shot O_+ under CoT for message in Figure 2.74

GPT-3.5

Output using O
Reasoning:
The player seems to be considering the fact that § is the only option where both teams can get a clear solution without any ambiguity. This indicates that the player is considering the clarity of the decision and the ease of coordination rather than focusing solely on the payoff.
Output using O_+
Reasoning:
The message indicates that the player is considering the unique solution where both teams can make a clear decision without hesitation. This suggests a level of strategic thinking beyond just focusing on the payoff amounts.

GPT-4

Output using O
Reasoning:
The player from Team 2 has suggested the decision '§' in the S2 game of the Pie game. The message in German translates to "Even though § is the lower payout, it is the only one where both teams get a clear solution, so they don't have to waver!" This message indicates a few key points:
1. The player acknowledges the payoff structure, recognizing that '§' is not the highest paying option.
2. The player emphasizes the clarity and definiteness of the '§' option, suggesting a preference for a straightforward and unambiguous decision.
3. The player does not mention any strategic anticipation of the other team's decision-making process but focuses on the inherent property of the '§' option as being a clear solution for both teams.
Given these points, the player seems to be responding to the label salience of the game, as '§' is the visually distinctive option in the Pie game. The player also shows an attraction to the action that provides a clear solution, which can be interpreted as a preference for simplicity or fairness over maximizing payoffs. This suggests that the player is not exhibiting strategic reasoning based on the expected choices of the other team but is instead focused on an intrinsic property of the '§' option.
The lower bound is set at 0 because the player's message does not clearly exhibit strategic reasoning; they are focused on the clarity of the decision rather than the strategic implications of the payoffs. The upper bound is set at 1 because the player's message could be interpreted as a basic strategic consideration, where they suggest '§' as a clear solution that both teams might gravitate towards without considering that the other team might also be strategic. However, the message does not explicitly state that the other team will think the same way, which would be necessary for a higher level of reasoning.
Output using O_+
Reasoning:
The message suggests that the player is focusing on the fact that the § option is the only one that provides a clear solution for both teams, despite it being the lower payoff in the S2 variation of the Pie Game. The player is emphasizing coordination on a unique solution rather than the amount of the payoff. This indicates that the player is not primarily driven by the desire for a higher payoff but rather by the desire for a clear and unambiguous coordination point.
The message does not explicitly mention a preference for the label of §, but the emphasis on the unique solution that §provides could be interpreted as a form of label salience, as the player is pointing out a distinctive feature of the §option. However, since the message is more about the strategic aspect of coordination rather than an inherent preference for the label itself, it is more appropriate to classify the label salience as "no."
The player does not exhibit a preference for higher payoffs, as they are advocating for the option with a lower payoff. Instead, they are focused on the strategic aspect of ensuring coordination. Therefore, the payoff salience is "L" for preferring low payoffs.
In terms of the level of reasoning, the player is considering the strategic aspect of the game, which involves thinking about the other team's potential decision. This suggests a level of reasoning beyond level-0, as they are not choosing based on instinctive reactions or random choice. The player is reasoning at least at level-1, as they are best responding to their belief about the other team's decision. However, since the message does not indicate a belief about higher-order reasoning (i.e., what the other team thinks about their team's thinking), the upper bound of reasoning is also level-1.

2.D.4 Additional Analysis

		no-CoT		CoT	
		$0\text{-}Shot$	$n\text{-}Shot$	$0\text{-}Shot$	$n\text{-}Shot$
GPT-3.5	O	.499	–	.522	–
	O_+	.625	.622	.652	.655
GPT-4	O	.664	–	.712	–
	O_+	.68	.70	.681	**.76**

Table 2.49: Lower bound accuracy

		no-CoT		CoT	
		$0\text{-}Shot$	$n\text{-}Shot$	$0\text{-}Shot$	$n\text{-}Shot$
GPT-3.5	O	.309	–	.402	–
	O_+	.378	.577	.412	.609
GPT-4	O	.597	–	.651	–
	O_+	.616	.66	.641	**.681**

Table 2.50: Upper bound accuracy

		no-CoT		CoT	
		$0\text{-}Shot$	$n\text{-}Shot$	$0\text{-}Shot$	$n\text{-}Shot$
GPT-3.5	O	.541	–	.577	–
	O_+	.643	.698	.651	.722
GPT-4	O	.704	–	.75	–
	O_+	.719	.721	.719	**.772**

Table 2.51: F1

F1 measure presented in Table 2.51 is with respect to the classification of the level of thinking intervals, which are defined by the classified upper and lower bounds of the level of thinking. It is a multi-label classification, calculated using the equations presented in Godbole and Sarawagi (2004, page 6). It is neither a micro nor a macro averaged measure but calculated as the average over instances (messages).

Chapter 3

Autonomy and Expectations in Algorithm-in-the-loop Systems

Joint with Hunter Phoenix van Wagoner, Andria Smith, Ksenia Keplinger

3.1 Introduction

Advice significantly influences the decision making in a diverse set of scenarios, ranging from simple daily life choices such as which dishwasher to buy, to complex choices of whether to convict a defendant or not (Tzioti, 2010). It can manifest as offered solutions, additional knowledge to aid in finding solutions, or validation of existing decisions (Cross et al., 2001). Even basic, most commonly observed word-of-mouth advice from non-experts, "naive advice", has the potential to shape people's decision making process (Schotter, 2003b). It has been documented that "naive advice" can improve social learning and coordination (Schotter and Sopher, 2003), increase the degree of strategic thinking (Kocher et al., 2007) and enhance the level of trust, trustworthiness, and reciprocity (Schotter and Sopher, 2006; Blanco et al., 2009) of individuals.

Although naive advice holds considerable influence, the impact of expert advice has been shown to have an even greater impact on the decision process (Goldsmith and Fitch, 1997; Jungermann and Fischer, 2005; Meshi et al., 2012). Expert advice, characterized by specialized knowledge and experience in a particular field, provides more accurate and reliable guidance relative to naive advice (Yaniv and Kleinberger, 2000), and therefore, it is more likely to be utilized by an advisee (Sniezek et al., 2004; Jungermann and Fischer, 2005).

Acquiring expert advice can be challenging due to its cost and availability, yet recent advances in artificial intelligence (AI) have made expert level algorithmic advice more accessible (Adadi and Berrada, 2018; Lundberg et al., 2018; Patel et al., 2019; Green and Chen, 2019a; Howard et al., 2020; Chugunova and Sele, 2020; Gonzalez et al., 2022). Given algorithmic advice's growing prevalence, lower cost and increasing accessibility in fields that require expertise level knowledge such as programming (Dakhel et al., 2023;

Mozannar et al., 2024), finance (Zhang et al., 2021; Fenneman et al., 2021), law (Angwin et al., 2016; Green and Chen, 2019a), and medical diagnosis (Jacobs et al., 2021; Dvijotham et al., 2023), and with the advent of large language models (LLMs) making algorithmic expert level advice a daily occurrence, understanding the dynamics of algorithmic advice becomes a necessity.

A natural form of advice is solicited advice, where the decision maker actively seeks additional information to make a more accurate decision. Multiple studies have shown that soliciting advice from a human advisor positively affects advice utilization relative to receiving unsolicited advice (Goldsmith, 2000; Gibbons et al., 2003; Deelstra et al., 2003), and unsolicited advice is documented to be perceived as a threat to the decision maker's autonomy and is consequently heavily discounted (Deelstra et al., 2003; Harber et al., 2005; Reinhardt et al., 2006).

In the domain of algorithmic advice taking, prior research has examined the choice between seeking advice from a human or an algorithm (Dietvorst et al., 2015; Castelo et al., 2019; Longoni et al., 2019; Yeomans et al., 2019; Önkal et al., 2009; Luo et al., 2019; Berger et al., 2021), yet providing further autonomy to the human decision maker by making algorithmic advice optional, where the alternative is receiving no advice, remains unexplored. Given human interactions with AI advisors may differ from those with human advisors (Castelo, 2019; Jago, 2019; Chugunova and Sele, 2020; Renier et al., 2021), it is unclear how actively soliciting advice will affect a human decision maker's algorithmic advice utilization.

The absence of research on the effect of making algorithmic advice optional can be attributed to the prevalence of human-in-the-loop (HITL) systems (Mantelero, 2022; Herm et al., 2023). In these AI centric systems, humans serve to support and enhance an algorithm's decision making process (Zanzotto, 2019; Enarsson et al., 2022). Broadly speaking, HITL systems operate by having the algorithm perform a task and delegate it to a human only when its confidence falls below a threshold (Wu et al., 2022b). Common examples include airport security checks and face recognition at border control, where algorithms flag issues for human inspection when necessary. The lack of human autonomy in these systems precludes the study of optional algorithmic advice.

While certain studies emphasize the advantages of HITL systems over stand alone algorithmic decision making systems (Wang et al., 2016; Lundberg et al., 2018; Patel et al., 2019), others voice concerns about the risk of amplifying algorithmic biases due to the imbalances in the algorithm's training data (Angwin et al., 2016; d'Alessandro et al., 2017; Danks and London, 2017; Buolamwini and Gebru, 2018; Yarger et al., 2020). Furthermore, it has been argued that most jobs will not be fully automated nor will humans solely serve as an on-demand intelligence for algorithms (Brynjolfsson et al., 2018). Most jobs will instead involve collaboration between AI systems and humans, with AI systems complementing workers rather than fully or partially substituting them (Shrestha et al., 2019). Consecutively, more recently, AI systems in advisory roles that augment rather than replace human decision making processes have been increasingly investigated in various fields (Adadi and Berrada, 2018; Murray et al., 2021). These type of AI systems grant a higher degree of accountability to the human decision makers by providing them with a higher degree of autonomy in their decision making process (Methnani et al., 2021). Such human centric human-AI team decision making systems where humans make the final decision and AI serves to improve human's judgement are commonly referred to as Algorithm-in-the-loop (AITL) systems (Green and Chen, 2019a,b). AITL systems are primarily used in fields that requires expert level knowledge, and the final decision may have ethical considerations or severe consequences such as sentence time of a defendant (Angwin et al., 2016), credit score of a consumer (Siddiqi, 2012), or medical diagnosis of a patient (Dvijotham et al., 2023). Lastly, with the advent of more general purpose generative AI tools such as ChatGPT, AITL systems have started to become a prevalent and integral part of daily decision making processes.[1]

The current study aims to advance human centric approaches by examining how giving human decision makers more autonomy affects their algorithmic advice utilization. More specifically, we are investigating how offering algorithmic advice as an optional tool, left at the discretion of human decision makers, compared to always providing the advice, influences their utilization of this advice.

People perceive and therefore evaluate AI systems differently from human advisors (Renier et al., 2021). While people expect human advisor to err (Renier et al., 2021), AI systems are expected to deliver consistent, near-perfect performance (Dzindolet et al., 2002; Madhavan and Wiegmann, 2007; Dietvorst et al., 2015; Goodyear et al., 2017; Li et al., 2020b; Liel and Zalmanson, 2020), particularly in perceivably objective[2] (Castelo et al., 2019) or complex[3] (von Walter et al., 2022) tasks with unambiguous ground truths[4] (Dietvorst and Bharti, 2020). This biased near-perfect prior expectation from algorithms is commonly referred in the literature as the perfect automation schema (Dzindolet et al., 2002), yet we will refer to it in the rest of the paper as the perfect automation bias (PAB). Since people do not expect consistent or near-perfect performance from human advisors (Renier et al., 2021), findings on the effect of providing additional performance-related

[1]ChatGPT can be considered as an AITL, as the human user solicits information from ChatGPT to accomplish a certain goal.

[2]Predicting the degree of funniness of a joke, recommending a romantic partner are examples of tasks that are commonly perceived as subjective, whereas forecasting the weather or stock prices are examples of tasks that are commonly perceived as objective (Castelo, 2019).

[3]Tasks that involve multiple, interdependent steps and aspects that require the expert assessment are perceived as complex tasks (Mikolon et al., 2015). Financial and legal advice are perceived as complex tasks (von Walter et al., 2022).

[4]Tasks such as facial recognition, speech recognition, or medical diagnosis can be considered as tasks with unambiguous ground truths, whereas forecasting the stock prices, identifying an ideal romantic partner for someone, or text classification for certain ambiguous text can be considered as tasks with ambiguous ground truths.

information about the advisor on advice utilization (Yaniv and Kleinberger, 2000; Sniezek and Van Swol, 2001; Yaniv, 2004; Van Swol and Sniezek, 2005) does not necessarily apply to settings where advice is provided by an algorithm (Dietvorst et al., 2015; Pálfi et al., 2022; You et al., 2022).

While in multiple studies, additional accuracy information is shown to have no significant effect on algorithmic advice utilization (Longoni et al., 2019; Pálfi et al., 2022; You et al., 2022), other studies have shown that subjects are responsive to the prior performance information of the AI system (Dietvorst et al., 2015; Castelo, 2019). Given these mixed results on the effect of additional accuracy information on algorithmic advice utilization, and assuming the majority of the people are subject to PAB, we ask the question of how algorithmic advice utilization is affected when algorithmic advice is accompanied by additional accuracy information indicating different performance levels, ranging from perfect to below near-perfect to poor. Moreover, we investigate whether the weight placed on this information, and its consequent effect on advice utilization is mediated by whether the advice is provided involuntarily or actively solicited by the decision maker. Additionally, we examine whether the level of understanding and experience with an AI system influences subjects' prior expectations of an AI system and therefore mediates the effect of additional accuracy information on their algorithmic advice utilization.

For our investigation into the effect of making algorithmic advice optional and the effect of providing additional accuracy information, we opted for the task of facial image pair classification, where subjects determine whether a pair of facial images belong to the same person or not. Except for a small population of superrecognizers, humans show low variance in facial recognition skills (Tardif et al., 2019; Towler et al., 2023). This, in turn, allows us to have a higher degree of control over the skill variance for the task among our subjects. Moreover, humans are inherently good at the facial recognition task, exhibiting on average roughly 80% accuracy (Tardif et al., 2019; Faghel-Soubeyrand et al., 2021). Hence, our subjects can be considered as experts in facial recognition. Consecutively, using the facial recognition task enables us to generalize our results to AITL systems in broader contexts where the algorithm collaborates with expert human decision makers, such as doctors making medical diagnoses or developers programming code, all the while maintaining a desirable level of variability in the subject pool in terms of age, gender, race, education, algorithm aversion, and experience with AI technology.

Additionally, facial recognition is an objective task with an unambiguous ground truth. Hence, given the findings in the literature, it is reasonable to expect that the general population has near-perfect prior expectations from a face recognition system (FRS) (Dietvorst and Bharti, 2020; Castelo et al., 2019; von Walter et al., 2022).[5] Furthermore, such an expectation is not without basis as FRS are documented to exhibit superior performance, achieving 90-99% accuracy in majority of instances (Li et al., 2020a). However, their performance deteriorates to as low as 60% for dark-skinned individuals, and especially dark-skinned females due to these AI systems' biases in their training datasets (Klare et al., 2012; O'Toole et al., 2012; Buolamwini and Gebru, 2018; Grother et al., 2019). Yet, this variability enables us to provide subjects with different FRS accuracy values for different demographic groups, signaling near-perfect performance for light-skinned facial images and below near-perfect performance for dark-skinned facial images, while maintaining external validity for actively used state-of-the-art FRS.[6]

Our findings show that subjects are more likely to use algorithmic advice when they actively solicit it. However, because they do not solicit advice frequently enough, the average utilization is lower when advice is optional compared to when it is mandatory. Additionally, our results indicate that additional accuracy information that is on par with subjects' assumed near-perfect performance expectations, does not significantly affect their algorithmic advice utilization, whereas accuracy information signaling below near-perfect or poor FRS performance significantly decreases it. However, this negative effect becomes insignificant when advice is solicited, indicating that subjects tend to disregard the accuracy information when they actively seek advice.

Additionally, we found that subjects' prior degree of FRS understanding and experience influence their prior FRS performance expectations. Subjects with greater FRS understanding or experience have prior performance expectations that align more closely with the state of the art FRS performance across various facial image pair types (light vs dark skinned, male vs female), and therefore are not observed to significantly alter their advice utilization when additional accuracy information is present. Conversely, the data suggests that subjects with less FRS understanding and experience are more likely to be prone to PAB, as additional accuracy information is found to calibrate their prior performance expectations downwards, leading them to utilize advice less. Lastly, a higher degree of trust in FRS is found to have a significant positive effect on advice utilization.

Our research contributes to the emerging field of algorithmic advice taking literature in four distinct ways. First, we are the first study to document the positive effect of soliciting advice on algorithmic advice utilization and the overall negative effect of making

[5] To our knowledge, there are no studies that provide any information on the general population's perception of FRS performance. This lack of research is not limited to FRS; generally, there is a scarcity of knowledge on how the general population perceives the performance of specific AI systems. See Section 3.3 for further discussion.

[6] We do not concern ourselves with how our findings can be considered within the existing literature on algorithmic advice taking in the narrow context of an application of an FRS (Fysh and Bindemann, 2018; Howard et al., 2020). This is primarily because real-life applications of FRS are HITL systems and not AITL systems; hence, whether advice solicitation affects algorithmic advice utilization within the specific context of FRS holds no value as it lacks any external validity.

algorithmic advice optional relative to making it mandatory. Second, our research contributes to the existing literature by demonstrating that the effect of additional accuracy information on algorithmic advice utilization is governed by subjects' prior performance expectations. Third, our study is the first to demonstrate that making advice optional nullifies the desirable effect of additional accuracy information on mitigating subjects' perfect automation bias. Lastly, we are the first to show that subjects' prior understanding and experience with an AI system mediate the effect of additional accuracy information on their use of algorithmic advice. Furthermore, our study aims to bridge the precise methods of experimental economics, which has explored advice in the past (Schotter, 2003b; Schotter and Sopher, 2003; Çelen et al., 2003; Schotter and Sopher, 2006), with the more recent investigation of human-AI interaction in the context of HITL (Wu et al., 2022b; Mantelero, 2022; Herm et al., 2023) and AITL (Siddiqi, 2012; Angwin et al., 2016; Green and Chen, 2019b) systems.

3.2 Experimental design

We consider a 2 by 2 between-subject experimental design, varying the use of FRS, either as optional (O) or mandatory (M), in one dimension, and whether to provide additional accuracy information on FRS (I) or not (N) in the other dimension. In each treatment, subjects are tasked with classifying whether the pair of facial images belong to the same person or not. They repeat this task for 24 distinct pair of photos. The order in which the photo pairs are presented is randomized for each subject.

3.2.1 Photo Pair Selection

The dataset used in the experiment is derived from the color FERET facial image database (Phillips et al., 1998). First, Microsoft Azure face API is used to produce similarity scores for each possible pair combination. Every pair combination is then placed either in the "same person" subset (A) or "two different people" subset (B). Then, 200 pairs with the lowest similarity score from the subset A and 200 pairs with the highest similarity score from the subset B are selected.

Following this initial selection, a research assistant manually reviewed the selected pairs. She eliminated any pairs that could be easily identified by humans, and generated new pairs that can be challenging for the human eye but trivial for FRS. She was instructed to ensure that the selected photos presented a similar level of difficulty across all four gender (male, female) and race (black, white) sub-categories, and to pay less attention to whether the pair is from subset A or B. This task resulted in a total of 96 photo pairs with 24 in each subcategory, black female, black male, white female and white male.

Within the 96 photos pairs, there were facial images of the same person taken from different angles used in different pairs. In order to eliminate potential unwarranted cues that can occur in-between photo pairs that share the same person, images are edited to remove or replace potential identifying cues, such as jewelry and clothing. Furthermore, for images where the lighting was poor and the facial features were not clear enough, the images are edited to improve their identifiability for humans. To perform these edits, a professional digital artist is hired[7].

The edited 96 pair of images are then used in our pre-experiment conducted with 160 subjects via Prolific platform where subjects were tasked with classifying each photo pair as belonging to the same person or not. Subjects were balanced in terms of their gender and race. For every subject, pairs were provided in a randomized order. Subject were not incentivized for their accuracy, but instead were paid a fixed fee of 2.9 Pounds for completing the experiment. Based on the results, photo pairs are categorized by race (black or white) and gender (male or female). Within each subcategory, pairs are ranked according to their predicted average accuracy scores[8]. Six pairs with the lowest scores are picked from each subcategory to construct the final data set of 24 photo pairs to be used in the experiments. Of these, 9 were from the subset A and 15 were from the subset B.

3.2.2 Classification task

For each photo pair, we used a two-step decision process to measure the degree of advice utilization by the advisee. This process is commonly considered in the behavioral decision making literature (Bonaccio and Dalal, 2006) and more recently started to be adopted in the computer science literature (Mucha et al., 2021). In the first step, subjects were shown a facial image pair and asked to determine if the images feature the same individual or not. Upon making their decision, they were asked to provide their confidence level in the correctness of their decision on a scale of 0 to 100 with 0 indicating full uncertainty and 100 indicating certainty.[9]

In the second step, depending on the treatment, subjects either always received FRS classification of the photo pair (M) or were presented with the option to solicit the classification of FRS (O). Furthermore, in instances where FRS classification is provided, depending on the treatment, additional accuracy information on its classification is provided (I) or not (N). Following the classification of FRS, subjects were asked to classify

[7]A total of 400 USD is paid for his services. We thank him for his generous work. See Appendix 3.E, Figure 3.4 for an example of facial image before and after editing.

[8]A logit regression model is used, where the dependent variable is 1 if the answer is correct and 0 otherwise, and the independent variable is the photo pair ID. Predicted accuracy values for each photo pair are calculated using the model estimates.

[9]From a probability theory standpoint, 50% confidence represents full uncertainty, and 0% confidence implies 100% incorrectness, also indicating full certainty. However, we were concerned that 50% confidence might confuse subjects. Therefore, we used a 0 to 100 scale, clearly stating during the instruction phase that 0% indicated full uncertainty and 100% indicated full certainty.

the photo pair for a second time, and to subsequently state their confidence level on this second decision. In the O treatments, this second step occurred regardless of whether the FRS classification is solicited or not.[10]

3.2.3 Additional Accuracy Information

The accuracy information is represented through four bar plots for the subcategories: White Woman (WW), White Man (WM), Black Man (BM), and Black Woman (BW). These plots correspond to accuracy percentages of 94%, 100%, 83%, and 65%, respectively. The bar plot percentages aim to reflect a realistic depiction of the state of the art performance of the FRS for each pair type. As such, they were determined based on the previous survey research conducted on FRS that documented variation in classification performance on basic facial features such as skin color and gender (Klare et al., 2012; O'Toole et al., 2012; Ngan et al., 2015; Buolamwini and Gebru, 2018; Grother et al., 2019; Li et al., 2020a).

You et al. (2022) document that providing average accuracy information instead of instance specific accuracy information imposes a lower degree of external cognitive load on the decision maker and therefore improves advice utilization by providing more cognitive capacity for the internal cognitive functioning primarily used in the decision making process (Van Merrienboer and Sweller, 2005; Sweller et al., 2019). Hence, in order to minimize the external cognitive load imposed upon the subjects and therefore in order to increase the internal cognitive load used for advice taking decision, the additional accuracy information provided in the experiments is not specific to each task in the experiment, but instead offers a general performance measure of FRS for each category. This information is always presented with the bar plots for all pair types, regardless of the race and gender of the pair in each classification task. Consequently, the accuracy information remains consistent across all tasks, providing an overall performance view of the FRS for each pair type.

The numerical percentage values were not provided on these bar plot.[11] Hence, subjects were provided only with an approximate graphical representation of the accuracy values. This approach is also aimed to minimize the external cognitive load on subjects by simplifying the presentation of accuracy information and therefore to improve their decision making capacity. Furthermore, this approach is done with the additional inten-

[10]The serial position effect theory posits that in a sequence of actions, the effect of an intervention, such as providing advice, at the initial phase is larger than its effect at a later or more recent phase in the sequence (Mahmud et al., 2022). Work by Dijkstra (1999) suggests that if an advice is provided at a later stage in the decision making process, then it is discounted more. Consecutively, the two-stage decision making procedure we are considering where advice is provided at a later stage should be expected to have a lower advice utilization compared to an hypothetical single-stage decision making procedure where the advice is provided simultaneously with the classification task from the start.

[11]See Figure 3.3 in Appendix 3.E.

tion to provide an intuitive and easily comprehensible information to the subjects in order to more clearly emphasize that FRS achieves near-perfect accuracy with white pairs, performs not as good as the white pairs and hence performs below near-perfect accuracy with the black male pairs, and performs relatively significantly worse with the black female pairs.

It has been documented that subjects have a tendency to assume that AI systems are consistent in their behavior, hence when they observe it to err once, they start heavily discounting its advice (Dzindolet et al., 2002; Dietvorst et al., 2015; Bogert et al., 2021). In order to minimize the possibility of the subjects to discount advice due to their perceived error in the classification provided by the FRS, we opted to provide advice that is always correct. Although this does not prevent the possibility for the subjects to falsely perceive FRS providing a false classification, we believed, at the very least, that the choice of FRS always providing the correct classification minimizes this unwarranted possibility by eliminating the possibility of false classification.[12]

3.2.4 Flat-fee payment

Subjects received a flat fee upon completing the experiment, with no monetary incentives linked to their accuracy on classification tasks. This payment structure incentivizes quicker completion for higher effective per minute earnings. In the M treatments, subjects cannot explicitly reduce the experiment duration. However, the O treatments allow participants to save time by opting out of seeking advice. This option, in turn, introduces an implicit time cost for soliciting advice.[13]

3.2.5 Experimental procedures

The experiments were conducted using the Prolific subject pool within the span of approximately one and half month, with an average of 2 weeks in between, in three different sessions. In total, data from 931 subjects is collected. Across the sessions, subjects were balanced in terms of race and gender with 50.24% of the subjects being female and 50% of the subjects being black. The average age of the subjects was 38.37 with a minimum of 18 and a maximum of 78. 18.64% of the subjects had a master or PhD degree, 76.76% of the subjects had an undergraduate degree, 2.42% of the subjects had a trade school degree, and 2.2% of the subjects had a high school degree as their highest educational degree. Subjects were from 22 different countries with 41.76% from USA, 24.93% from UK, 17.55% from South Africa, 4.36% from Portugal, 4.12% from Poland and 1.93%

[12] Hence, except for the WM accuracy information, the accuracy information provided to the subjects were misleading.

[13] Rather than offering monetary incentives, we attempted to motivate subjects by assigning them the role of security officers responsible for identifying whether individuals entering a high-security office building are employees or not. See Appendix 3.F for the details of the experimental instructions.

from Italy.[14] Upon completion of the entire experiments, subjects were paid 2.9 pounds. In total approximately 3375 pounds (∼4000 Euros) are spent. On average, subjects spent 28.4 minutes to complete the survey.

Every experiment started with a tutorial that explained the general rules of the experiment, and provided the hypothetical high-stake scenario where the subject played the role of a human operator (with the help of an FRS) who is tasked with determining if a person entering a company building is an employee or not. At the end of the tutorial, two practice rounds using pictures of famous people were conducted to ensure subjects could use the interface correctly. Subsequently, a five-question quiz based on the tutorial's information was administered to confirm that subjects possessed the minimum necessary knowledge for the experiment.[15]

In every treatment, subjects evaluated the same 24 photo pairs in a randomized order and provided their decision and confidence twice in a row in the two step decision process formerly described. The task replicates the scenario where a human operator assesses the image of an individual entering a high security building, compares it to the closest match in the image database, and potentially utilizing the FRS classification, decides to either allow entry or request an ID card.[16] Lastly, two attention questions unrelated to the facial image classification task were evenly placed within the series of classification tasks.

After the classification task, subjects were asked to fill out a series of surveys. Subjects filled out the explicit sexism survey (Swim et al., 1995), the Bayesian racism survey[17] (Uhlmann et al., 2010), propensity to trust in a specific technology survey[18] (Mcknight et al., 2011), and a general propensity to trust survey (Gefen and Straub, 2004). After the surveys, the experiment ended with subjects providing various demographic information such as age, gender and education.

Any subject who failed to correctly answer either of the attention questions and any subject who failed more than 50% of the quiz questions are dropped out of the experiment. Furthermore, subjects who did not answer any of the surveys (sexism, trust in FRS, racism etc.), subjects who did not provide an answer for the race, gender or age questions, subjects who provided an unreasonable age (any value above 80), subjects who declared their gender as non-binary are dropped out of the data set. In total, out of our 931 subjects, 105 (11.3%) of them are dropped out of the data set.

[14] Each of the other countries constitute less than 1% of the data. The other countries are: Greece, Hungary, Scotland, Czechia, Estonia, Ireland, Mexico, Canada, Germany, Latvia, Netherlands, Slovenia, Denmark, Finland, Spain and Wales. In total these countries constituted approximately 5.33% of the participants.

[15] See Appendix 3.F for the details of the experimental instructions.

[16] See Appendix 3.F for examples of user interface.

[17] Unlike other racism surveys available in the literature, the Bayesian racism survey does not specifically measure subjects' biases towards black people. Instead, it measures biases towards minority ethnic groups.

[18] The survey is modified to measure trust specifically on FRS. This is done by replacing specific technology placeholder in each survey question with FRS. See Appendix 3.F for the modified version of this survey.

3.3 Theory and Hypothesis

When decision makers re-evaluate their judgement based on new information, such as advice, they update their prior beliefs by incorporating this new evidence. The weight given to the new evidence depends on their confidence in their prior beliefs and the perceived reliability of the new information (El-Gamal and Grether, 1995). However, biases in the belief updating process can cause decision makers to either underweight or overweight this additional information (Tversky and Kahneman, 1974; Nyarko et al., 2006; Pescetelli et al., 2021).

Within the context of advice taking, the decision making literature offers multiple potential underlying mechanisms for why such biases arise during the belief updating process. Krueger (2003) suggests that individuals inherently favor their own judgments, as they believe them to be superior. This phenomenon, where subjects overestimate the correctness of their initial beliefs, is referred to as overconfidence (Gardner and Berry, 1995; Harvey and Fischer, 1997). Additionally, decision makers highly value their perceived sense of autonomy (Koestner et al., 1999; Schultze et al., 2018) and see following the advice of another as a treat to their abilities and self-worth (Harber et al., 2005; Reinhardt et al., 2006; Paik, 2020), and consecutively underweight the advice. Fear of loss of autonomy in conjunction with overconfidence leads to what is commonly referred to as egocentric advice discounting (Yaniv and Kleinberger, 2000; Deelstra et al., 2003; Krueger, 2003; Schultze et al., 2018).

Alternatively, according to Yaniv (2004), the underlying reason for advice discounting is the difference in accessibility between a decision maker's own reasoning and the advisor's reasoning. This difference in accessibility is also referred to as differential information (Gino, 2008). Due to this lack of access, the decision maker cannot fully comprehend the advisor's rationale, leading her to underweight the advice and overweight her prior beliefs. From a Bayesian perspective, optimal updating would require the decision maker to integrate the advisor's input with her prior beliefs in a way that reflects the relative reliability of both sources. Yet, lack of accessibility to advisor's rational results in perceiving his advice as less reliable than it actually is, which in turn results in the decision maker's posterior beliefs to be biased towards her prior beliefs (Robalo and Sayag, 2018).

Conversely, Krueger (2003) documents that even in novel tasks where the decision maker is not able to easily provide a rational for her prior beliefs just as she is not easily able to provide a rational for the advisor (i.e. there is minimal differential information), subjects are still observed to overestimate the correctness of their prior beliefs and therefore to underweight the information provided by the advisor. Hence, he argues that it is not because of a lack of accessibility to the advisor's rational that the discounting occurs but because the decision maker simply favors any additional information that confirms her prior beliefs and discounts the ones that contradicts them (confirmation bias). For instance, Zaleskiewicz et al. (2016) document that confirmation bias is the primary factor influencing lay people's utilization of financial advice from experts such that they utilize the expert advice mainly when the expert advice is in accordance with their pre-advice investment decisions.

AI systems based on deep neural networks, such as FRS, are often considered "black boxes" (Chander et al., 2018). Consecutively, a human decision maker faced with an advice from such an AI system experiences a significant degree of lack of accessibility to the "reasoning" behind its advice, and therefore is likely to discount it (Litterscheidt and Streich, 2020). Furthermore, it has been documented that, in various domains, human decision makers prefer to take advice from a human advisor over an AI advisor, discount the human expert's advice relatively less than the AI's advice, and trust AI less due to this relatively higher degree of lack of accessibility to the algorithmic "reasoning" (Önkal et al., 2009; Kayande et al., 2009; Van Dongen and Van Maanen, 2013; Goodwin et al., 2013). Moreover, algorithmic advice that is accompanied by some rational improves the decision maker's algorithmic advice utilization as the provided "reasoning" makes the algorithmic advice more accessible to the human decision maker (Gönül et al., 2006; Chander et al., 2018; Litterscheidt and Streich, 2020). These results underscore the differential information mechanism argued in the decision making literature within the context of human advisor, human advisee scenarios, and show that it is equally present in scenarios where the advisor is an algorithm. Additionally, various studies demonstrate that human decision maker's egocentric biase due to their overconfidence are also at play when interacting with an AI advisor (Sutherland et al., 2016; Logg et al., 2019). Furthermore, Kawaguchi (2021) shows that human decision makers fear losing their autonomy upon following the directives of an algorithmic, and consecutively underweight its advice. In summary, the fundamental mechanism through which human decision makers are argued to discount the human advisor's advice are to a large extend mirrored in the settings where the advisor is an algorithm.

3.3.1 Solicited Advice

In human advisor, human advisee settings, when advice is optional and actively sought, it has been documented to significantly decrease advice discounting (Goldsmith, 2000; Gibbons et al., 2003; Deelstra et al., 2003; Paik, 2020). The act of seeking advice is argued to represent a shift from an over-reliance on self-referent knowledge to a more balanced approach that includes both internal and external information sources and therefore serves to minimize the negative effects of information differential on advice utilization (Bonaccio and Dalal, 2006). This change occurs as seeking advice inherently acknowledges the need for additional perspectives or information, which, in turn, challenges the tendency to rely predominantly on self-generated knowledge (Goldsmith, 2000; Gibbons et al., 2003; Deelstra et al., 2003).

For highly egocentric individuals, unsolicited advice may seem like an infringement on their self-reliance (Harber et al., 2005; Reinhardt et al., 2006). In contrast, when they actively seek advice, this act serves a dual purpose: it acknowledges their knowledge limitations and simultaneously affords them an additional sense of autonomy (Gibbons et al., 2003). In other words, act of solicitation mitigates their overconfidence and alleviates their fear of loss of autonomy. This newfound autonomy, derived from the decision to seek external input, is argued to allow them to more comfortably rely on outside knowledge without feeling a significant loss of self-governance (Gibbons et al., 2003).

Since the mechanism through which advice solicitation improves advice utilization is advisee-centric, and independent of the attributes of the advisor, it should be equally applicable when the advice is provided by an algorithm. Hence, we expect the documented negative effects of differential information, egocentric bias and fear of losing autonomy on algorithmic advice utilization to be mitigated when the advice is solicited. Therefore, we conjecture that solicited algorithmic advice will be discounted less than unsolicited algorithmic advice.

Hypothesis 1 *Solicited algorithmic advice will be discounted less than unsolicited algorithmic advice.*

3.3.2 Optional Advice

The effect of making advice optional (O) compared to always providing it (M) on the overall advice utilization depends on two factors. First, it depends on the extent to which solicited advice is utilized more than advice that is always provided. Second, it hinges on how often participants solicit advice. Even if solicited advice leads to higher utilization, infrequent solicitation can result in lower overall advice utilization in the O treatments compared to the M treatments.

Given the subjects receive a flat-fee upon completion of the experiment, the faster they complete the tasks, the higher their per minute pay rate will be. In the O treatments, refraining from soliciting advice can save time on the classification task. Therefore, we anticipate that due to the flat-fee payment scheme, subjects will perceive soliciting advice as costly. Conversely, in the M treatments, since advice is always present, this additional cost is expected not to be salient and therefore is not expected to be a potential factor in advice utilization.

Research suggests that making advice costly can lessen the tendency to discount it (Gino, 2008; Sniezek et al., 2004; Patt et al., 2006; Hofmann et al., 2009). This phenomenon is argued to be driven by the sunk cost effect in conjunction with cognitive dissonance theory (Festinger, 1962), where individuals, having incurred a cost for advice, are more likely to use it to justify their expenditure and to eliminate the dissonance caused by incurring a cost (Gino, 2008). Hence, in addition to the aforementioned reasons, in our experimental setup, in the O treatments, solicited advice is expected to be discounted less also due to being costly.[19]

On the other hand, Gino (2008) shows that advice is solicited less frequently when it involves a cost. Consecutively, in our experimental framework, where soliciting advice has an implicit cost, participants may not seek advice often, potentially leading to lower overall advice utilization in the O treatments compared to the M treatments.

These two juxtaposing effects make it unclear whether the net effect of making advice optional will lead to higher advice utilization than the average advice utilization observed in the M treatments. Therefore, we refrain from conjecturing whether on average a higher degree of advice utilization will be observed in the O treatments compared to the M treatments.

3.3.3 Additional Accuracy Information

Mandatory Advice

In the human advisor, human advisee setting, Sniezek and Buckley (1995) posit that an advisor's confidence in the correctness of his advice increases his influence in decision making cases under uncertainty. Furthermore, Schrah et al. (2006) argue that advice is often viewed with skepticism due to doubts about the advisor's expertise, and multiple studies have shown that when an advisor indicates high confidence in the accuracy of his advice or is depicted as an expert, advisees have a higher tendency to utilize this advice (Cooper, 1991; Harvey and Fischer, 1997; Sniezek and Buckley, 1995; Yaniv and Kleinberger, 2000; Sniezek and Van Swol, 2001; Van Swol and Sniezek, 2005; Gibbons et al., 2003). However, individuals often have significantly different and consistently biased prior expectations of AI systems' performance. Thus, providing information that indicates the algorithmic advice is highly accurate, similar to a human advisor expressing high confidence, may not necessarily result in higher algorithmic advice utilization.

Prior research has documented that people are subject to PAB, and hence expect AI systems to demonstrate near-perfect performance (Dzindolet et al., 2002; Madhavan and Wiegmann, 2007; Yu et al., 2019; Castelo et al., 2019; Dietvorst and Bharti, 2020; von Walter et al., 2022). Furthermore, PAB is particularly evident for AI systems that perform objective tasks with unambiguous ground truths (Castelo et al., 2019; von Walter et al.,

[19] We acknowledge that our experimental design does not allow us to distinguish the positive effect of the cost of soliciting advice from the positive effect of soliciting advice itself on advice utilization. However, in real life, act of soliciting always bares at the very least a time-dependent cost. Hence with our flat-fee payment scheme, our experimental design can be argued to provide a more externally valid setup for the investigation of the effect of soliciting advice.

2022; Dietvorst and Bharti, 2020). Since face recognition is an objective task without inherent uncertainty, in the absence of any prior performance information regarding the system, subjects are, on average, expected to believe that FRS classification performance is near-perfect and utilize its advice accordingly.

In the I treatments, FRS classification is accompanied by a separate accuracy value for each pair type. These values aim to inform subjects about the average classification performance of FRS for each pair type and to update for their prior beliefs on FRS classification performance.

Prior work on the threshold accuracy value above which the general population would perceive an AI system to be performing at a near-perfect performance level is scarce. Madhavan and Wiegmann (2007) document that a 90% accuracy level can be considered as the near-perfect performance threshold, whereas investigation of Yu et al. (2019) suggest that the threshold value could be as low as 70% for some individuals. Furthermore, these threshold values may be task specific and not reflect a near-perfect performance threshold value for AI systems in general (Daschner and Obermaier, 2022). Yet, given that FRS are documented to have near-perfect to perfect performance with classification accuracy ranging from 90% to 99% for the majority of instances (Ngan et al., 2015; Grother et al., 2019; Li et al., 2020a), and considering the precedent of using 90% accuracy as the near-perfect performance threshold (Madhavan and Wiegmann, 2007), we assumed 90% as the near-perfect threshold value that the subjects will on average consider when interacting with the FRS in our experiments.

With 100% accuracy for WM pairs and 94% for WW pairs, we aim to signal the subjects that FRS' classification will be always correct for WM pairs and almost always correct for WW pairs. In other words, assuming 90% accuracy as the threshold for near-perfect performance for FRS, these additional accuracy information aim to signal the subjects that FRS performs either perfect or near-perfect in the classification of white photo pairs.

On the other hand, an accuracy of 83% for BM pairs is aimed to signal the subjects that FRS classification performance is "less than ideal" for these pairs, as it is below the near-perfect threshold of 90% we assume. Furthermore, a 65% accuracy for BW pairs is aimed to signal the subjects that FRS performs relatively poorly with the classification of these pairs. Overall, we expect these additional accuracy information to signal the subjects that FRS classification performance is less than near-perfect for black photo pairs, especially with the black women pairs.

Assuming the majority of the subjects portray PAB and assuming 90% accuracy as the perceived near-perfect threshold value, we expect that observing accuracy values below this threshold will result in the subjects to lower their FRS performance expectations. Consequently, we posit that additional accuracy information for white photo pairs will have no significant effect on advice utilization, whereas it will negatively affect advice utilization for black photo pairs, with a more significant negative effect for BW photo pairs. Furthermore, given the equal number of classification tasks for each photo pair type, we conjecture that the overall effect of introducing additional accuracy information will negatively impact subjects' average advice utilization.

Hypothesis 2 *Accuracy information will have no effect on advice utilization for white photo pairs, but it will have a negative effect on advice utilization for black photo pairs.*

Hypothesis 3 *Overall, advice with accuracy information will be discounted more than advice without accuracy information.*

Solicited Advice

Several studies document that costly expert level advice is discounted less (Patt et al., 2006; Sniezek et al., 2004). Yet, Gino (2008) argues that an advice being costly may be perceived as being an expert-level advice, and this may be the reason why subjects discount costly advice less. In a series of experiments, she shows that even when subjects' perception of the advice's quality is controlled for, they discount advice less when it is costly. Her results suggest that the quality of the advice does not significantly impact subjects' decision-making process, as due to the sunk cost fallacy, subjects are more likely to use the advice to justify the pre-committed cost, regardless of its quality.

Given the effectively costly nature of optional advice in our experiment due to the flat-fee payment design, we previously argued that subjects will discount the advice less when solicited. On the other hand, we also argued that providing advice with additional accuracy information would on average cause subjects to discount it more.

When solicited advice is accompanied by additional accuracy information, we expect that, to justify the cost of seeking advice or to mitigate the cognitive dissonance generated by this cost (Festinger, 1962), subjects will ignore accuracy information that would typically lead them to discount the advice, as acknowledging it and therefore discounting the advice would imply that the cost of soliciting was in vain. In other words, we conjecture that the sunk cost effect will mitigate the negative impact of additional accuracy information on solicited advice. Thus, we expect additional accuracy information to have no significant effect on advice utilization when advice is solicited.

Hypothesis 4 *Soliciting advice will offset the negative effect of additional accuracy information on advice utilization. Consequently, the overall degree of advice utilization for solicited advice, whether accompanied by additional information or not, will not differ significantly.*

3.3.4 Trust in FRS

It has been previously shown that a higher trust in the human advisor increases advice utilization (Jungermann, 1999; Sniezek and Van Swol, 2001; Van Swol and Sniezek, 2005; Jungermann and Fischer, 2005; Wang and Du, 2018). Additionally, it has been shown that a higher degree of trust or reliance on automation increases advice utilization (Workman, 2005; Önkal et al., 2009; Shaffer et al., 2012). Based on these past results, we expect subjects' self-reported trust in FRS values to positively correlate with their advice utilization.

Hypothesis 5 *Subjects with higher trust in FRS will discount advice less.*

3.4 Results

Throughout the results section, an ample amount of pair-wise tests are conducted. We recognize that as the number of pair-wise test results increases, the likelihood of a study-wide, variable-specific type 1 error to occur significantly increases. In order to alleviate these concerns, we define any result to be "marginally significant" for any p-value between 0.01 and 0.001, and we set the "significance" threshold value to 0.001.[20]

In Table 3.1a, mismatch percentages between subjects' initial (pre-advice) classification and FRS' classification are presented. Since FRS classification is always correct by design, the mismatch percentages represent subjects' average initial false classification rates (error rates). The rows labeled as N and I represent the treatments N and I, respectively. Similarly, the columns labeled M and O represent the overall mismatch percentage under mandatory advice and optional advice treatments (both solicited and not solicited), respectively. Furthermore, the columns labeled O_s and O_x represent subcases under optional advice treatment where subjects solicited and not solicited advice, respectively. Lastly, the row labeled μ_{NI} and the column μ_{MO} represent average values across N and I, and M and O treatments, respectively.

[20]More specifically, by far the highest number of pair-wise tests conducted in the results and additional analysis sections combined is with respect to the average percentage change in decision. In the results section, we report pairwise comparison between NM & IM, NO_s & IO_s, NM & NO_s, IM & IO_s, NM & NO, IM & IO, NO_x & NM, NO_x & NO_s, IO_x & IM and IO_x & IO_s for 2 different subsets of data, and we report pair-wise comparison between NM & IM and NO_s & IO_s for 4 different subsets of data. In addition, in the additional analysis section, we report pair-wise comparison tests between NM & IM and NO_s & IO_s for 13 different subsets of data. Hence in total we conduct 54 pair-wise tests. Using the Bonferroni correction method, and setting our initial significance threshold to 0.05, we get the corrected threshold value of 0.00093. For convenience, we rounded up this value and get the formerly stated threshold value of 0.001 for rejecting any null hypothesis. The marginal significance threshold of 0.01 is a value we established based on the set of p-values we have observed in our 54 pair-wise tests with the goal of distinguishing results that we considered it would be too conservative to redeem them as insignificant. Additionally, for the interpretation of our logit model estimates, we refer to a slightly more relax alternative set of criteria to evaluate significance where any p-value below 0.05 is considered as significant, any p-value between 0.05 and 0.1 is considered as marginally significant, and any p-value above 0.1 is considered as not significant.

	M	O	O_s	O_x	μ_{MO}
N	45.1	44.3	46.1	39.9	44.7
I	44.1	42.3	43.9	39.1	43.1
μ_{NI}	44.6	43.3	45	39.4	43.9

(a) Initial mismatch %

	M	O	O_s	O_x	μ_{MO}
N	14	18.1	9.1	40.9	16.2
I	16.4	19.7	9.9	39.3	18.2
μ_{NI}	15.2	18.9	9.5	39.9	17.2

(b) Final mismatch %

Table 3.1: Mismatch between human and FRS classifications

In every treatment, NM, NO, IM and IO, subjects' post-advice classification are better than random guessing, as error rates are significantly below 50% for each treatment (proportion test, $p_{max} < 0.001$). Error rates between treatments NM, IM and NO are found to not to be significantly different (proportion test, $p_{min} > 0.339$). Moreover, while the average error rate in IO is found not to be significantly different from the ones in IM and NO (proportion test, $p_{min} > 0.028$), it is found to be marginally significantly lower than the one in NM (proportion test, $p = 0.002$). Lastly, the overall error rates are not significantly different between N and I treatments (44.7 vs 43.1; proportion test, $p = 0.012$) and not significantly different between M and O treatments (44.6 vs 43.3; proportion test, $p = 0.027$). In summary, there are no notable differences among treatments in terms of the subjects' initial classification performances.[21]

On the other hand, a significant difference is observed between O_s and O_x under either N or I (proportion test, $p_{max} < 0.001$). These results hint at a positive correlation between the likelihood of soliciting advice and the likelihood of initial misclassification. Moreover, while there is no statistical difference between NO_x and IO_x (proportion test, $p = 0.64$), average error rate under IO_s is statistically lower than that under MO_s (proportion test, $p < 0.001$). This difference along with the statistically higher average percentage of soliciting advice of 71.9% in NO compared to 66.6% in IO (proportion test, $p < 0.001$) result in the IO treatment to generate the observed slightly lower (yet insignificant) average error rates.

Additionally, a mixed-effect logistic regression with the binary dependent variable be-

[21] The differences between IO, IM, and NO treatments are insignificant given our defined threshold values, but the proportion test still yields a small p-value less than 0.05. Hence, it is worth noting the possibility that the IO treatment is significantly lower than all the other treatments and not only the NM treatment. This would imply that subjects in the IO treatment are slightly better at the classification task. However, this difference is small, with at most a 2.8 percentage point difference. Similarly, the insignificance between the overall error rates of the N and I treatments, and between the M and O treatments, suggests that relaxing the significance thresholds would imply that subjects in the I treatments are slightly better at the classification task than those in the N treatments, and subjects in the O treatments are slightly better than those in the M treatments. Given the insignificance between NM, IM, and NO treatments, the overall difference between N and I, and M and O, should be attributed to the lower average error rates observed in the IO treatment. Given in the subsequent analysis, we are investigating the effect of treatment variables on advice utilization by comparing the average percentage change in the decision, we are, to some degree, controlling for variations in subjects' initial error rates across the treatments.

ing whether the photo pair is misclassified (1) or not (0) is considered.[22] In this model, photo pair specific and subject specific covariates are additionally controlled for. Furthermore, a random effect is included in the model for each subject to address unobserved heterogeneity among them. Neither of the treatment variables are found to have a significant effect on the probability to misclassify a photo pair prior potentially receiving an advice. On the other hand, BW photo pairs are found to have a negative marginally significant effect, and WW photo pairs are found to have a positive significant effect on the probability to misclassify a photo pair. These results indicate that, overall, relative to the baseline of BM photo pairs, BW photo pairs were easier, WW pairs were harder and BM photo pairs were as difficult as WM photo pairs to classify. Lastly, female subjects, older subjects and subjects with higher racism score are found to be more likely to make an initial misclassification.

In Table 3.1b, the final average error rates for each treatment are displayed. Except for subcases where subjects did not solicit advice, O_x, post-advice error rates are significantly lower relative to their pre-advice counterparts under each treatment (proportion test, $p < 0.001$). For O_x subcases, there is no statistical difference between the initial and final error rates (proportion test, $p_{min} > 0.611$). Given FRS' advice is always correct by design, post-advice error rates should be considered as a proxy for advice utilization. Hence, the observed lower post-advice error rates indicate some degree of advice utilization across treatments.

The average error rates are significantly lower for O_s subcases compared to M treatments under either N or I treatments (proportion test, $p_{max} < 0.001$). Yet, because subjects did not solicit advice frequently enough, average error rates under O treatments are significantly higher than the ones under M treatments under either N or I treatments (proportion test, $p_{max} < 0.001$). Taking the observed average error rates for O_s and O_x subcases as given, for the error rates to be lower under O than M, subjects needed to solicit approximately at least 13 percentage points more than the observed 71.9% and 66.6% solicit frequencies for treatments N and I, respectively. These results hint that solicited advice is on average discounted less than unsolicited advice, yet, overall, making advice optional results in lower average utilization of advice due to the advice not being solicited frequently enough.

Additionally, post-advice error rates are significantly lower in the NM treatment compared to IM treatment (proportion test, $p < 0.001$). This result potentially suggests that when advice is always provided, additional accuracy information is observed to negatively affect the advice utilization. Conversely, no significant difference is found between NO and IO treatments (18.1 vs 19.7; proportion test, $p = 0.057$) nor between NO_s and IO_s (9.1 vs 9.9; proportion test, $p = 0.254$) nor between NO_x and IO_x (40.9 vs 39.3; propor-

[22] See Appendix 3.D Table 3.19 for details.

tion test, $p = 0.364$). The insignificance between NO_s and IO_s suggests that the negative effect of additional information observed in M treatments is offset when the advice is solicited.

In Tables 3.2a and 3.2b average percentages for changing the initial classification are presented for each treatment. Table 3.2a displays the cases for when subject's initial classification matches the classification of FRS (match cases), and Table 3.2b displays the cases for when the subject's initial classification differs from the classification of FRS (mismatch cases).[23]

	M	O	O_s	O_x	μ_{MO}
N	0.8	1.4	0.6	3.2	1.1
I	1.4	1.8	1.2	2.9	1.6
μ_{NI}	1.1	1.6	0.8	3	1.4

(a) Match cases

	M	O	O_s	O_x	μ_{MO}
N	69.9	60.9	80.8	2.4	65.2
I	64.6	55.8	78.9	3.9	60
μ_{NI}	67.3	58.4	79.9	3.3	62.6

(b) Mismatch cases

Table 3.2: Percentage change in classification decision

In match cases, change of decision is equivalent to deviating from FRS' classification, hence, in match cases, a higher average level of percentage change in decision indicates a higher degree of advice discounting. Conversely, in mismatch cases, change of decision is equivalent to adopting the classification of FRS, hence, in mismatch cases, a higher average level of percentage change in decision implies a higher degree of advice utilization.

In match cases, subjects, on average, change their initial classification in less than 1.5% of the time, while in mismatch cases, subjects change their initial classification in more than 60% of the time. In other words, in match cases, subjects rarely deviated from the advised classification, while in mismatch cases, subjects on average discounted advice in less than 40% of the time.[24]

In match cases, soliciting advice does not make a statistical difference in the average percentage change from the initial decision for either N or I treatments (NM vs NO_s; IM vs IO_s; proportion test, $p_{min} > 0.35$). Moreover, the overall effect of O treatment compared to the M treatments is not found to be significantly different under either N or I treatments (proportion test, $p_{min} > 0.07$). Similarly, observing additional accuracy information does not make significant difference either under M treatments or under O_s subcases (proportion test, $p_{min} > 0.05$). Conversely, NO_x is found to be statistically larger than either NM or NO_s (proportion test $p_{max} < 0.001$), and IO_x is found to be statistically larger than either IM or IO_s (proportion test, $p_{max} < 0.001$).

The fact that, on average, we observe less than 1.5% change rate when advice is ob-

[23]Recall from Table3.1a that match cases represent 56.1% of the dataset where as the mismatch cases represent 43.9%.

[24]Overall, subjects changed their initial classification in 28.3% of the cases. Of these changes, 0.8% occurred in match cases, while 27.5% occurred in mismatch cases.

served indicate that observation of confirmatory FRS classification consolidates subject's pre-advice classification, and the subject is less likely to change her decision compared to when she does not receive the confirmatory advice (O_x subcases). This general absence of change in decision is translated into the lack of responsiveness of the subject to either treatments: whether the advice is solicited or not, or whether the advice is accompanied by an additional accuracy information or not does not make a significant difference in the subjects' final classification decision. Hence, at the very least, either treatment effect is observed not to negatively affect advice utilization when the pre-advice classification matches FRS' advice.[25]

In mismatch cases, irrespective of whether additional accuracy information is provided or not, subjects are on average significantly more likely to change their classification to FRS' classification when advice is solicited (NM vs. NO_s, IM vs IO_s; proportion test, $p_{max} < 0.001$). This result suggests that, for mismatch cases, when advice is solicited, subjects are less likely to discount it. However, because subjects changed their (false) initial classification on average 3.3% of the time when they did not observe FRS' classification (O_x subcases), overall, average percentage change is significantly lower in the O treatments compared to the M treatments (NM vs NO, IM vs IO; proportion test, $p_{max} < 0.001$). This result indicates that making advice optional overall leads to lower advice utilization due to the prevalence of cases where advice is not sought. For the average percentage change in the O treatments to match or to exceed those in the M treatments, either the average frequency of soliciting advice must increase, as previously argued, by approximately 13 percentage points, or the average percentage change in decision must rise by about 13 percentage points for the NO treatment and 15 percentage points for the IO treatment.

In mismatch cases, additional information is observed to increase the average advice discounting under M treatments (proportion test, $p < 0.001$). This indicates that additional accuracy information causes subject to, on average, discount advice more. However, when advice is solicited additional information is observed not to make a significant difference (NO_s vs IO_s, proportion test, $p = 0.162$). This lack of significance in conjunction with the former significant positive effect of solicited advice on advice utilization point out that the negative effect of additional information on advice utilization is offset by the positive effect of solicited advice.

In Table 3.3, we summarize our findings. Ma denotes the match cases and $Mima$ denotes the mismatch cases. "$-$" indicates a negative significant effect on advice utilization,

[25]In Appendix 3.C, we have additionally considered the treatments effect on the change in the confidence levels, examining the difference between pre- and post- advice confidence levels. We found that observing an advice that matches subject's initial classification has a significant increase on the confidence. Furthermore, despite the subjects not changing their initial classification upon observing a confirmatory advice, we found that soliciting advice significantly further increases the confidence, whereas additional information significantly mitigates this increase.

	Ma	$Mima$
I	$\sim$	$-$
O	$\sim$	$-$
O_s	$\sim$	$+$
$I : O_s$	$\sim$	$\sim$

Table 3.3: Advice utilization

"$+$" indicates a positive significant effect on advice utilization effect, and "$\sim$" denotes no significant effect on advice utilization. Furthermore, labels I, O, and O_s represents the effect of additional information treatment, optional advice treatment and soliciting advice subcase, respectively. Lastly, $I : O_s$ represents the effect of additional accuracy information when advice is solicited.

In summary, we observe that for match cases neither treatment is observed to have a significant effect, while for mismatch cases we observe additional information to have a negative and soliciting advice to have a positive effect on advice utilization. Furthermore, the negative effect of additional accuracy information is observed to be offset when advice is solicited. Lastly, overall, optional advice treatment results in a lower degree of advice utilization indicating that the positive effect of advice solicitation is not enough to compensate for the cases where the subjects did not solicit advice.

To complement the above analysis, in Table 3.4, results of a series of logit regression models that estimate the effect of treatments on the probability of a subject to change her initial decision are presented. The models are specified as a mixed-effect models with subject-specific random effects where the treatment variables, their interaction, and additional two covariates are gradually introduced.

Each treatment variable is a dummy variable that takes the value of 1 if the data point is from that treatment and takes the value of 0 otherwise (baseline). In CD_4, the dummy variable O_x is introduced. It takes the value of 1 if the subject is in the treatment O and chooses not to solicit advice, and 0 otherwise (baseline). When O_x is introduced into the model, treatment variable O effectively presents the effect of the subcases where the treatment is O, and the subject chooses to solicit advice. To reflect this conceptual change in O, we re-label it as O_s and introduce it in two new rows along side with its interaction with the I treatment variable. For models CD_{4-7}, and any subsequent model that includes O_x, we use this re-labeled format instead of O and $I : O$.

Given the drastic difference in the average percentage change in decision made displayed between the match and mismatch cases in Tables 3.2a and 3.2b, we control for it by introducing the dummy variable Ma in models CD_{5-7}. Ma gets the value of 1 if the initial classification of the subject is the same as the classification of FRS and 0 otherwise (baseline). In models CD_{6-7}, we introduce the initial confidence reported by the subjects, C_1. C_1 gets a value between 50 and 100. 100 represents subject's perceived

Table 3.4: Probability to change decision

	CD_1	CD_2	CD_3	CD_4	CD_5	CD_6	CD_7
α	-0.95***	-0.83***	-0.84***	-0.83***	1.23***	6.91***	4.56***
	(0.04)	(0.05)	(0.06)	(0.06)	(0.12)	(0.36)	(0.42)
I	-0.17**	-0.17**	-0.15†	-0.15†	-0.32†	-0.32†	-0.31†
	(0.06)	(0.06)	(0.08)	(0.08)	(0.17)	(0.18)	(0.17)
O		-0.23***	-0.22**				
		(0.06)	(0.08)				
$I : O$			-0.04				
			(0.12)				
O_s				0.29***	0.60***	0.52**	0.51**
				(0.08)	(0.17)	(0.18)	(0.17)
$I : O_s$				0.05	0.18	0.16	0.17
				(0.11)	(0.24)	(0.26)	(0.25)
O_x				-3.19***	-5.06***	-4.69***	-4.7***
				(0.16)	(0.21)	(0.22)	(0.22)
$I : O_x$				0.14	0.15	0.13	0.16
				(0.22)	(0.29)	(0.30)	(0.30)
Ma					-6.43***	-6.74***	-6.74***
					(0.11)	(0.12)	(0.12)
C_1						-0.07***	-0.07***
						(0.00)	(0.00)
T_C							0.67***
							(0.07)
var_{RE}	0.5	0.49	0.49	0.41	2.27	2.73	2.41
AIC	24138	24124	24126	20952	10025	9693	9614
Obs	20952	20952	20952	20952	20952	20952	20952

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 "‡" 0.2 " " 1

complete certainty on the correctness of her initial classification and 50 represents subject's full uncertainty on the correctness of her initial classification.[26]. Lastly, in model CD_7, we introduce the trust in FRS variable T_C. It is a subject specific variable that takes

[26]In the experiments, subjects were asked to provide a value between 0 and 100 to express their level of confidence with 0 being fully uncertain and 100 being fully certain. This interval then converted into a 50 to 100 interval to make it in par with the mathematically correct form for the subject's belief where 50% belief in the correctness in decision indicates fully uncertainty. Given the logistic nature of the models, alternative normalization considerations for C_1 such as 0 for complete uncertainty and 1 (or 100) for full uncertainty would not make a difference in the results except for the amplitude of the estimated coefficient for C_1.

a value between 1 and 5 with 5 presenting the highest degree of trust towards FRS.

Given in match cases, percentage change in the initial decision is on average 1.3%, our primary concern is the mismatch cases. Consecutively, we omit the presentation of additional models that consider interaction of the treatment variables with the Ma covariate in Table 3.4.[27]

Across CD_1 to CD_7, the coefficient for variable I is negative, yet its statistical significance becomes marginal when O_x is introduced in models CD_{3-7}. The estimated coefficient for I gets larger when we control for the matched classification cases, Ma, in models CD_{5-7}. These results in conjunction with the former statistical investigation confirm our hypothesis that additional accuracy information causes subjects to discount advice more compared to the case where such information is not provided with the FRS' advice.

Result 1 *FRS advice with additional accuracy information is discounted more.*

In model CD_2, the negative and significant coefficient for variable O shows the overall negative effect of making FRS advice optional on the advice utilization. This result is in accordance with our former analysis on the overall negative significant effect of making advice optional on advice utilization for mismatch cases (Table 3.3).[28]

Result 2 *When FRS advice is optional, subjects utilize it less compared to when it is always provided.*

In models CD_{4-7}, when we control for the cases where the subject chooses not to solicit advice with O_x, the coefficient for the dummy variable O_s is found to be positive and statistically significant, highlighting the positive significant effect of soliciting advice on advice utilization. This is also on par with our former analysis for mismatch cases, and confirms our hypothesis that when advice is solicited it is discounted less.[29]

Result 3 *Solicited FRS advice is discounted less than the unsolicited FRS advice.*

The interaction between the two treatments $I : O$ in model CD_3, and the interaction between additional accuracy information and solicited advice upon controlling for the

[27]The results do not change when the interaction of Ma covariate with treatment variables is considered. See Appendix 3.5, Table 3.20 for additional models that consider the interaction of the treatment variables with the Ma covariates.

[28]Additionally, when we interact Ma with O, we observed that $Ma : O$, which represents the overall effect of optional advice treatment on the match cases, is found to be significant and positive. This indicates that contrary to our former finding on the effect of O in advice utilization to be insignificant (Table 3.3), the O treatment is observed to increase advice discounting, as a positive sign indicate in match cases deviating from the advised classification. See Appendix 3.5, Table 3.20, model CD_{3b} for details.

[29]Additionally, when we interact Ma with O_s, we observe that $Ma : O_s$, which represents the effect of soliciting advice on the match cases, it is found to be negative and significant. This is again contrary to our former analysis on match cases where we found no statistical effect and suggests that, also for match cases, when advice is solicited, the subject is less likely to diverge from the FRS' classification, i.e. discount advice less. See Appendix 3.5, Table 3.20, model CD_{6b} for details

cases where subject did not solicited advice $I : O_s$ in models CD_{4-7}, all are found to be insignificant. This is on par with our former analysis and our hypothesis that additional accuracy information has no effect on advice utilization when advice is solicited.

Result 4 *When advice is solicited, additional accuracy information has no significant effect on advice utilization.*

Furthermore, as expected, the coefficient for O_x is estimated to be significant and negative. Hence, as previously argued, although when advice is solicited it is discounted less, either because solicited advice is not utilized enough or because advice is not solicited frequently enough, overall, FRS' advice is utilized less when it is optional.

In models CD_{5-7}, the large, negative and significant coefficient for variable Ma confirms the drastic difference in the average percentage change from initial classification observed between match and mismatch cases (Tables 3.2a and 3.2b). Moreover, this result in conjunction with Table 3.2a highlight the fact that the observed effects of the treatment variables on the probability to change the initial classification is primarily due to the variation in the mismatch cases, and hence reassures our former assertion that the estimated effects of the treatment variables primarily reflect their effect on advice utilization. In model CD_6, the negative and significant coefficient for variable C_1 indicates that as subjects' confidence in their initial classification increases, they discount FRS' advice more. The negative effect of C_1 confirms the commonly observed result that as subjects become less confident on their initial classification, they become more likely to utilize advice.

Lastly, in model CD_7, trust in FRS, T_C, is observed to be positive and significant, indicating that a higher trust in FRS results in a higher likelihood of advice utilization. This results confirms our initial hypothesis on algorithmic trust's positive effect on algorithmic advice utilization.

Result 5 *A subject with a higher level of trust in FRS discounts its advice less.*

Although we found strong evidence that additional information negatively affects advice utilization, consistent with our hypothesis, it is unclear whether this effect results from the near-perfect prior expectations of the subjects. To investigate the underlying mechanism, we proceed with the investigation of additional information's impact on advice utilization for each pair type. To do so, we focus on the effect of treatment I within each pair type subset. Thus, the following analysis will concentrate solely on the effect of additional information on advice utilization and omit any discussion unrelated to it.

Table 3.5 presents the average percentage change in decisions for each photo pair type. Although the average percentage change is lower under IM than NM for all photo pairs, this difference is not significant for WM (proportion test, $p = 0.109$) and WW (proportion test, $p = 0.403$) pairs. For BM pairs, the average percentage change under NM is

	M	O	O_s	O_x	μ_{MO}
N	68.4	64.4	83.2	6.3	66.3
I	63.7	57.1	80	57.9	60.3
μ_{NI}	66.2	60.8	81.7	6	63.4

(a) WM

	M	O	O_s	O_x	μ_{MO}
N	70.7	58.8	79	1.1	64.4
I	68.3	57.5	79.8	2.6	62.6
μ_{NI}	69.5	58.1	79.4	1.9	63.5

(b) WW

	M	O	O_s	O_x	μ_{MO}
N	69.9	60	79	1.2	64.6
I	61.6	53.2	77.6	3.5	57.3
μ_{NI}	65.8	56.8	78.7	2.5	61.1

(c) BM

	M	O	O_s	O_x	μ_{MO}
N	71.2	60.7	82.1	0.8	65.6
I	64.2	54.7	77.6	4.3	59.3
μ_{NI}	67.7	57.8	79.9	2.6	62.5

(d) BW

Table 3.5: Percentage change in classification decision for each pair type

significantly greater than under IM (proportion test, $p = 0.001$), and for WM pairs, it is marginally significant (proportion test, $p = 0.007$). These results support our conjecture that subjects, on average, have near-perfect expectation from FRS and these expectations to some degree are confirmed with the accuracy information provided for white pairs. Consecutively, additional accuracy information has no significant effect on the advice utilization for the white photo pairs.[30] Conversely, the additional accuracy information provided for black photo pairs causes subjects to lower their prior belief that FRS performs near-perfect classification, resulting in a significant decrease in advice utilization for black photo pairs.

Additionally, we observe that for all pair types, when advice with additional information is solicited, there is no significant effect on the average percentage change in the decision (NO_s vs IO_s; proportion test, $p_{min} > 0.179$). This results indicate that irrespective of whether the additional accuracy information matches subjects' prior near-perfect expectation or it its below their prior expectations, when advice is solicited, additional information is to a significant degree disregarded and consecutively has no significant negative effect on advice utilization.

To complement the above analysis, Table 3.6 presents model CD_7 estimates for each pair type subset, as indicated by the subscripts of the model labels on the top row of the table. Consistent with prior results, the coefficient for the dummy variable I is negative across all models, but it is significant only in models CD_{BM} and CD_{BW}. Furthermore, the interaction between soliciting advice and additional information is not significant in

[30]Although the insignificance among IM and NM treatments confirm our argument that subjects have a near-perfect prior expectations, the fact that average values are consistently lower for white pairs and to a discernible degree with WM pairs suggest that, additional accuracy information may cause additional negative effect on advice utilization for mechanisms we have not accounted for. See Section 3.6 for further discussion.

Table 3.6: Probability to change decision for each pair type

	CD_{WM}	CD_{WW}	CD_{BM}	CD_{BW}
α	3.68***	3.02***	3.94***	3.34***
	(0.75)	(0.54)	(0.73)	(0.71)
I	-0.32	-0.08	-0.57*	-0.4*
	(0.27)	(0.19)	(0.25)	(0.17)
O_s	0.93***	0.3†	0.46†	0.58*
	(0.26)	(0.18)	(0.25)	(0.24)
$I : O_s$	-0.01	0.24	0.49	0.14
	(0.37)	(0.25)	(0.36)	(0.34)
O_x	-5.4***	-4.06***	-5.18***	-4.87***
	(0.45)	(0.36)	(0.45)	(0.48)
$I : O_x$	0.31	0.03	0.21	0.12
	(0.56)	(0.47)	(0.57)	(0.63)
Ma	-7.61***	-5.84***	-6.97***	-6.63***
	(0.32)	(0.21)	(0.28)	(0.28)
C_1	-0.06***	-0.05***	-0.06***	-0.05***
	(0.01)	(0.01)	(0.01)	(0.01)
T_C	0.68***	0.59***	0.8***	0.63***
	(0.11)	(0.07)	(0.11)	(0.1)
var_{RE}	3.3	0.95	3.48	2.03
AIC	2633	2981	2793	2108
Obs	5238	5238	5238	5238

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

any of the models.

Our prior analysis and confirmatory regression analysis reveal that additional accuracy information's negative effect is to a discernible degree governed by the difference between the accuracy information and the subjects' prior near-perfect expectations, and only observed when advice is mandatory. Hence, we confirm our hypothesis that the negative effect of additional information on advice utilization is primarily observed with black photo pairs. Yet, it is worth noting that there is no discernible difference between BM and BW pairs even though, accuracy information for BM pairs is closed to being near-perfect whereas for BW pairs it is distinctly farther away from being near-perfect.

Result 6 *Additional accuracy information has no significant effect on advice utilization for white photo pairs, but has a significant negative effect on advice utilization for black* *photo pairs.*

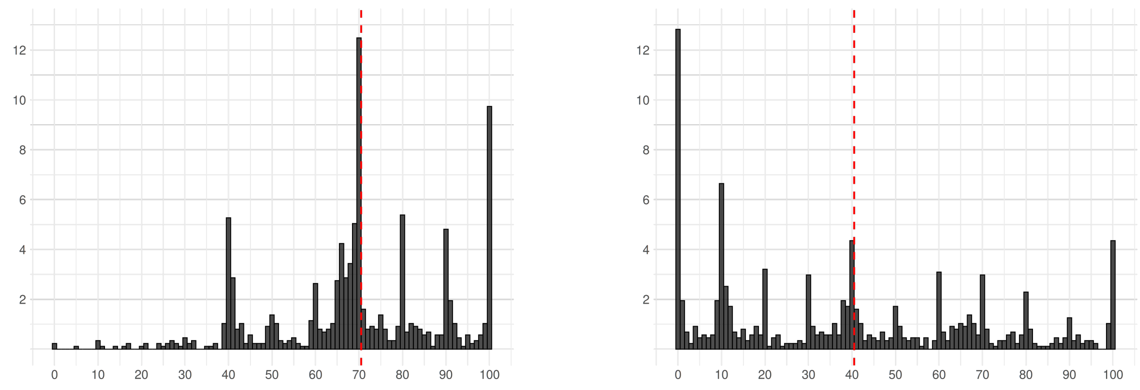
3.5 Additional Analysis

3.5.1 Prior Expectations

Given that near-perfect prior expectations mediate the impact of additional accuracy information on advice utilization, we investigate whether subjects' self-reported level of FRS understanding and experience might influence these prior expectations.

The self-reported understanding of new technology can be biased if acquired from unreliable sources, leading to unreasonably high expectations (Horstmann and Krämer, 2019). Moreover, subjects may understand the technical workings of FRS but lack awareness of its poor performance with certain minorities, particularly black faces and more specifically with black female faces, due to biased training data sets used to develop currently used FRS (Buolamwini and Gebru, 2018). Hence, a high level of self-reported understanding of FRS does not necessarily imply that the subjects have better calibrated expectations more in line with the accuracy information provided in the I treatments. Conversely, if information is obtained from reliable sources that provide accurate details on the varying performance of FRS, the self-reported understanding may be a reliable indicator of well calibrated performance expectations.

Similarly, self-reported experience may be influenced by memory biases, such as recalling only instances of poor performance (Kahneman et al., 1993) or overestimating the level of perceived experience based on limited interactions (Tversky and Kahneman, 1971, 1974). However, a high level of self-reported experience with FRS may also be an accurate depiction of the subjects' prior experience with FRS, and these subjects could be expected to have better calibrated expectations due to their close familiarity with the technology (Koskinen et al., 2019; Mahmud et al., 2022).



(a) Understanding (b) Experience

Figure 3.1: Self-reported understanding of FRS and experience with FRS

Figures 3.1a and 3.1b present the distributions of self-reported FRS understanding and experience of the subjects. For both plots, the y-axis represents the percentage of the subjects and the x-axis represents the self-reported score from 0 to 100 for the respective concept. The mean and the median value for the self-reported understanding is 70, whereas for the self-reported experience, the mean is 40 and the median is 38. Using these self-reported values, we categorize the subjects in two categories for each concept. We define above average understanding as any subject reporting an understanding score above 70 and a below average understanding as any subject reporting an understanding score equal to or below 70. Similarly, we define above average experience as any subject reporting an experience score above 40 and a below average experience as any subject reporting an experience equal to or below 40. We denote above average understanding and above average experience categories as H_U and H_E, respectively, and denote below average understanding and below average experience categories as L_U and L_E, respectively. H_U contains 40% of the subject pool, whereas H_E contains 43% of the subject pool. In Figures 3.1a and 3.1b, the red dashed lines represent the cutoff points for the separation of these two categories.

	M	O_s
N	68.4	84.1
I	71.8	80.5

(a) H_U

	M	O_s
N	71	79.1
I	59.3	77.8

(b) L_U

	M	O_s
N	70.9	80.9
I	67.3	79.3

(c) H_E

	M	O_s
N	69.3	80.8
I	62	78.6

(d) L_E

Table 3.7: Percentage change in classification decision for high and low understanding or experience of the subjects

Tables 3.7a and 3.7b represent the average percentage change in decision for mismatch cases for subsets H_U and L_U, respectively. For subjects with above average understanding of FRS (H_U), NM and IM are not significantly different (proportion test, $p = 0.304$), whereas for subjects with below average understanding of FRS (L_U), under the M treatment, additional accuracy information is found to have a significantly negative effect on the average percentage change in decision (proportion test, $p < 0.001$). Furthermore, for either subsets of subjects, no statistical difference is found between NO_s and IO_s (proportion test $p_{min} > 0.1$).

Tables 3.7c and 3.7d represent the average percentage change in decision for mismatch cases for subsets H_E and L_E, respectively. For subjects with above average experience with FRS (H_E), under the M treatment, changes in NM and IM are not significantly different (proportion test, $p = 0.1$), whereas for subjects with below average experience with FRS (L_E), under the M treatment, additional accuracy information is found to have a significantly negative effect on the average percentage change in decision (proportion test, $p < 0.001$). Furthermore, for either subsets of subjects, no statistical difference is found between NO_s and IO_s (proportion test $p_{min} > 0.212$). Lastly, across all cases, H_U, H_E, L_U and L_E, the average percentage change in decision in NM treatments are statistically not different from each other (proportion test, $p_{min} > 0.21$).

These results suggest that subjects with either above average understanding or above average experience are better calibrated in terms of their prior expectations for the FRS performance, and consequently, additional accuracy information does not have a significant effect on subjects with above average FRS understanding. Conversely, subjects either with below average understanding or below average experience are observed to potentially depict near-perfect prior expectations which in turn results in them to utilize advice less when they additionally observe the accuracy information.[31]

The insignificance of additional accuracy information across all four subsets when advice is solicited further supports the robustness of our results. The fact that additional accuracy information has no significant effect on advice utilization, despite potentially differing prior performance expectations, indicates that such information is largely disregarded when advice is solicited.

We further investigate the observed mediating effect of self-reported experience and understanding with a series of mixed-effect logit regression models where we took the existing model CD_7 and introduced various potential mediator variables by interacting each with the treatment variables. In model CD_{FRS_U}, the FRS_U variable that represents the continuous self-reported understanding of the subject is introduced, in model CD_{FRS_E}, FRS_E, the continuous self-reported experience of the subject is introduced. In model CD_{D_U}, the D_U dummy variable that takes the value of 1 if the subject has an above average understanding and 0 otherwise is introduced. In model CD_{D_E}, the D_E dummy variable that takes the value of 1 if the subject has an above average experience and 0 otherwise is introduced. The results of these models are presented in Table 3.8. To minimize the number of rows, instead of providing a row for each introduced variable and its interactions with the treatment variables, we denote the additional variable under each model with the label X. For model CD_{FRS_U}, X represents FRS_U, for model CD_{FRS_E}, it represents FRS_E, for model CD_{D_U}, it represents D_U, and for model CD_{D_E}, it represents D_E.

Under models CD_{FRS_U} and CD_{FRS_E}, the estimated coefficient for I is marginally

[31] We have additionally investigated the effect of additional information on advice utilization for subjects with above average understanding or with above average experience under each pair type subset. For every pair type, for subjects with either above average understanding or above average experience, additional accuracy information is found not to have a significant effect on advice utilization. See Appendix 3.A for details.

Table 3.8: Probability to change decision

	CD_{FRS_U}	CD_{FRS_E}	CD_{D_U}	CD_{D_E}
α	4.82***	4.59***	4.72***	4.6***
	(0.59)	(0.44)	(0.43)	(0.43)
I	-1.16†	-0.48†	-0.62**	-0.49*
	(0.67)	(0.29)	(0.22)	(0.24)
O_s	0.08	0.59*	0.33	0.61**
	(0.63)	(0.28)	(0.22)	(0.23)
$I : O_s$	1.16	0.18	0.41	0.13
	(0.92)	(0.4)	(0.32)	(0.34)
O_x	-5.16***	-4.52***	-4.87***	-4.54***
	(0.74)	(0.35)	(0.3)	(0.29)
$I : O_x$	0.37	0.09	0.53	-0.12
	(1.07)	(0.5)	(0.4)	(0.44)
Ma	-6.74***	-6.74***	-6.73***	-6.74***
	(0.12)	(0.12)	(0.12)	(0.12)
C_1	-0.07***	-0.07***	-0.07***	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)
T_C	0.65***	0.65***	0.64***	0.66***
	(0.07)	(0.07)	(0.07)	(0.08)
X	0.00	0.00	-0.06	0.03
	(0.01)	(0.01)	(0.25)	(0.25)
$X : I$	0.01	0.00	0.73*	0.35
	(0.01)	(0.01)	(0.35)	(0.35)
$X : O_s$	0.01	0.00	0.48	-0.25
	(0.01)	(0.01)	(0.36)	(0.35)
$X : I : O_s$	-0.01	0.00	-0.98	0.15
	(0.01)	(0.01)	(0.81)	(0.5)
$X : O_x$	0.01	0.00	0.36	-0.31
	(0.01)	(0.01)	(0.43)	(0.43)
$X : I : O_x$	0.00	0.00	-0.82	0.51
	(0.01)	(0.01)	(0.61)	(0.61)
var_{RE}	2.4	2.4	2.36	2.39
AIC	9622	9622	9613	9652
Obs	20952	20952	20952	20952

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

significant and negative, but the introduced variable and its interaction with the treatment dummies are negligibly small and insignificant. For models CD_{D_U} and CD_{D_E}, the estimated coefficient for I is significant and negative. While for model CD_{D_E}, neither the introduced variable D_E nor its interaction with the treatment dummies are significant, whereas in model CD_{D_U}, the interaction of the introduced dummies D_U with I is significant and positive. Lastly, in all the models, the interaction of the variable with $I : O_s$ is found not to be significant confirming our former analysis.

The fact that FRS_U and its interaction with the treatment dummies are insignificant whereas the dummy variable D_U and its interaction with I is significant suggests that there is a threshold for the self-reported understanding of FRS. Only above this threshold do subjects' prior expectations lower significantly enough to increase advice utilization upon observing the additional accuracy information. Furthermore, the fact that neither FRS_E nor D_E, and neither their interaction with I are found to be significant suggests that self-reported experience is not as good of a proxy for prior expectation of the subjects as their self-reported understanding. Lastly, the fact that additional information does not significantly affect advice utilization when advice is solicited, irrespective of the subjects' understanding or experience levels, suggests that the tendency to disregard additional accuracy information when soliciting advice is prevalent across all subject types and is not mediated by the prior FRS performance expectation of the subjects.

These results to some degree support the former analysis that subjects with an higher degree of understanding are indeed better calibrated in their FRS performance expectations. Furthermore, these types of subjects, due to their more in-tune expectations, are not negatively affected by the accuracy information provided in terms of advice utilization. The positive significant coefficient for $X : I$ in model CD_{D_U} further highlights the possibility that these subjects have on average a lower level performance expectation of FRS.

3.5.2 Additional Factors Affecting Advice Utilization

In Table 3.9, additional covariates are gradually introduced to the model CD_7. In CD_8, covariates related to the task specific variables are introduced. μ_{A_1} is a photo pair specific variable that represents the average initial accuracy level calculated based on the subjects' performance on her respective treatment.[32] TP stands for "True Positive" and is a dummy variable that takes 1 if the photo pair belongs to the same person and 0 otherwise. PT stands for "Pair Type" and is a categorical variable with BM pair type as its baseline, and the subscripts denotes the other photo pair types.

In model CD_9, a covariate on the subject's trust disposition towards humans, T_H,

[32]Alternatively, we considered the average accuracy levels calculated across all treatments and found no discernible difference in the estimated effects.

Table 3.9: Probability to change decision with additional covariates

	CD_8	CD_9	CD_10	CD_{11}	CD_{12}
α	6.29*** (0.38)	4.26*** (0.47)	4.10*** (0.46)	4.13*** (0.51)	4.04*** (0.52)
I	-0.31† (0.18)	-0.32† (0.18)	-0.32† (0.18)	-0.30† (0.18)	-0.30† (0.18)
O	0.52* (0.18)	0.51* (0.18)	0.53* (0.18)	0.53* (0.18)	0.54* (0.18)
$I : O$	0.15 (0.26)	0.14 (0.30)	0.12 (0.25)	0.11 (0.25)	0.10 (0.24)
O_x	-4.70*** (0.22)	-4.71*** (0.22)	-4.71*** (0.22)	-4.71*** (0.22)	-4.71*** (0.22)
$I : O_x$	0.12 (0.30)	0.17 (0.30)	0.17 (0.30)	0.16 (0.30)	0.16 (0.30)
Ma	-6.82*** (0.12)	-6.83*** (0.12)	-6.83*** (0.12)	-6.83*** (0.12)	-6.83*** (0.12)
C_1	-0.07*** (0.00)	-0.07*** (0.00)	-0.07*** (0.00)	-0.07*** (0.00)	-0.07*** (0.00)
T_C	0.67*** (0.08)	0.71*** (0.08)	0.67*** (0.08)	0.67*** (0.08)	0.65*** (0.08)
μ_{A_1}	1.14*** (0.25)	1.15*** (0.24)	1.15*** (0.24)	1.15*** (0.24)	1.15*** (0.25)
TP	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)	0.05 (0.07)
PT_{BW}	-0.23* (0.09)	-0.24* (0.09)	-0.24* (0.09)	-0.24* (0.09)	-0.24* (0.09)
PT_{WM}	0.14† (0.08)	0.14† (0.08)	0.14† (0.08)	0.14† (0.08)	0.14† (0.08)
PT_{WW}	0.07 (0.08)	0.07 (0.08)	0.07 (0.08)	0.07 (0.08)	0.07 (0.08)
T_H		-0.1* (0.04)	-0.1* (0.04)	-0.1* (0.04)	-0.1* (0.04)
FRS_E			0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
FRS_U			0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
FRS_I			0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$Female$				-0.10 (0.12)	-0.11 (0.12)
Age				0.00 (0.00)	0.00 (0.00)
$Black$				0.05 (0.13)	0.05 (0.13)
Edu_H				0.18 (0.13)	0.18 (0.13)
$Sexism$					0.05 (0.05)
$Racism$					0.00 (0.05)
var_{RE}	2.76	2.42	2.41	2.4	2.4
AIC	9664	9583	9584	9590	9593
Obs	20952	20952	20952	20952	20952

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

the variables FRS_E, FRS_U and FRS_I, respectively.

In model CD_{11}, subject-specific demographic attributes gender, race, age and education are introduced with covariates $Female$, $Black$, Age and Edu_H, respectively. $Female$ is a dummy variable that takes 1 if the subject is female and 0 otherwise (male). $Black$ is a dummy variable that takes 1 if the subject is black and 0 otherwise (white). Age denotes subject's age. Edu_H is a dummy variable that takes 1 if the subject has an undergraduate education or higher and 0 otherwise. Lastly, in model CD_{12}, sexism and racism score of the subjects are introduced with the covariates $Sexism$ and $Racism$, respectively. A higher score denotes a higher degree of sexism or racism.

Inclusion of the additional covariates did do not change the previously stated effects of the treatments variables, I, O_s, $I : O_s$ and O_x. Furthermore the effect of Ma, C_1 and T_C are also observed not to be affected by the inclusion of additional covariates.

The average initial correct classification percentage of a given pair, μ_{A_1}, is found to be positive and significant across all models. This indicates that for photo pairs that subjects are, on average, more likely to initially classify correctly, those who initially misclassify them are more likely to correct their mistakes after observing the correct classification from FRS.[33]

The type of photo pair is observed to be significant and negative for BW and marginally significant and positive for WM with respect to the baseline of BM. General trust disposition (towards humans) T_H is found to have a significant negative effect on advice utilization. Furthermore, having a higher FRS experience, understanding of FRS, or interest in FRS is found not to have a significant effect on the algorithmic advice utilization. Similarly, gender, age, race or education level are found not to have a statistical effect on the advice utilization. Lastly, neither the degree of racism nor sexism is found to have a significant effect on the algorithmic advice utilization.

3.6 Discussion

3.6.1 Soliciting Advice

We show that granting human decision makers the autonomy to choose when to take algorithmic advice significantly improves advice utilization compared to when the advice is a mandatory part of the decision making process. Although we are the first to document that the positive effect of soliciting advice on its utilization as previously documented in human advisor settings, also applies when the advisor is an algorithm, we do not provide further insight into whether this observed effect is primarily due to mitigating fear of loss

[33]It is worth noting that μ_{A_1} does not represent the perceived difficulty of a task. Therefore, this result does not imply that subjects who perceive a task as harder utilize advice less. Perceived task difficulty is better indicated by the subject's initial confidence, and we have previously documented that a decrease in confidence significantly increases advice utilization.

of autonomy, egocentric bias, the lack of accessibility to the algorithm's "reasoning", the cost of soliciting, or a combination of these factors.

One may argue that the positive effect of soliciting advice arises not from mitigating the aforementioned factors but from the subjects' initial confidence levels. According to Bayesian updating, when a subject has lower confidence in the correctness of her decision, she will give more weight to additional information. This is confirmed by the significant negative effect of initial confidence on the probability of utilizing algorithmic advice. Additionally, in Appendix 3.B, we show that higher initial confidence significantly reduces the likelihood of soliciting advice. Furthermore, in Appendix 3.C, we demonstrate that the average initial confidence of subjects who solicit advice is significantly lower than those who do not solicit advice. These results suggest that subjects generally have less confidence in the correctness of their initial classifications for tasks where they solicit advice from the FRS. Hence, it can be argued that solicited advice is utilized more because subjects have lower confidence in their initial classifications for the tasks on which they seek advice. However, in Section 3.4, Table 3.4, models CD_{6-7}, after controlling for the initial confidence with the variable C_1, soliciting advice still has a significant positive effect on the probability of utilizing advice, and the estimated coefficient for O_s decreases only slightly from 0.6 to 0.52 after introducing C_1. This indicates that while initial confidence plays a significant role in advice utilization, soliciting advice has a significant positive effect on advice utilization through the aforementioned mechanisms that are not directly tied to the subjects' initial confidence.

To determine whether fear of loss of autonomy significantly influences the positive effect of advice solicitation on algorithmic advice utilization, future research could control for subjects' degree of autonomy using the reactive autonomy scale considered by Koestner et al. (1999).[34] Additionally, to investigate the role of information accessibility, one could use GPT-4 instead of a traditional FRS and include an additional treatment dimension where GPT's classification is provided with additional reasoning generated by the model.[35] Then, by comparing treatments with and without the additional reasoning, it would be possible to measure whether the positive effect of soliciting advice diminishes when additional reasoning is provided. Lastly, one could consider an additional treatment where the time cost of soliciting advice is controlled for by forcing subjects to wait for some amount of seconds even if they do not solicit advice, and compare the positive effect of soliciting advice between the these two treatments with and without the effective cost

[34] Koestner et al. (1999) document that subjects with a higher degree of reactive autonomy discount expert advice more.

[35] GPT-4 can take images as input and provide not only classification on the provided visual input (such as facial image pair comparison) but also offer a "reasoning" for its classification (Achiam et al., 2023) using the 0-shot chain of thought prompting technique (Kojima et al., 2022). We considered such an application. The classification performance appears to be worse than that of a human classifier. The rationale for its classification included information on the similarities or differences in various facial features, such as eye, nose, and lip shapes, as well as hair and skin color.

of soliciting.

Making advice optional places greater reliance on the human decision maker to recognize when to seek advice. In Appendix 3.B, Table 3.12, model PS_{3^*}, we show that making an initial incorrect classification significantly increases the likelihood to solicit advice, and in Appendix 3.E, Table 3.19, model PRE_{Con_2}, we show that making an initial incorrect classification significantly decreases the initial confidence.[36] Yet, in 28% of the cases, subjects failed to recognize their initial misclassification and opted not to solicit advice. Furthermore, 20.1% of the time, although subjects solicited advice when their initial classification were incorrect, they failed to correct their initial misclassification by not adopting the classification of FRS. Consequently, although in the mandatory advice treatment a significantly larger percentage of subjects failed to correct their initial misclassification by not adopting the algorithmic advice (32.7%), because their overall exposure to algorithmic advice was approximately 44% higher, we observed average advice utilization to be lower in the optional advice treatment compared to mandatory advice treatment.

3.6.2 Additional Accuracy

Overall, providing additional accuracy information had a negative effect on the algorithmic advice utilization. Based on previous research, we made the assumption that subjects have a near-perfect performance expectation from FRS. Consequently, any additional information provided against this expectation was expected to lead to a decreased advice utilization. For white photo pairs, additional accuracy information signaled the subjects that FRS performs perfect to near-perfect, and consequently it did not have a significant effect on the algorithmic advice utilization for these pairs as it was on par with the assumed prior near-perfect expectation of the subjects for FRS. On the other hand, for black photo pairs, additional accuracy information signaled the subject that FRS performs below near-perfect, and consecutively subjects updated their prior near-perfect expectation for FRS downwards and additional accuracy information resulted in a significant negative effect on advice utilization for black photo pairs. Furthermore, given 83% accuracy of the black photo pairs results in a significant decrease in the advice utilization suggests that the assumed 90% threshold value for subjects' perceived near-perfect performance is a suitable threshold value to consider at the very least for FRS. Moreover, these results suggest that although people, as previously documented (Dzindolet et al., 2002; Madhavan and Wiegmann, 2007; Goodyear et al., 2017; Yu et al., 2019; Dietvorst and Bharti, 2020), have biased near-perfect expectations from AI systems, this bias can be to some degree

[36] We acknowledge that in both models PS_{3^*} and PRE_{Con_2}, considering initial incorrect classification as an independent variable is flawed since subjects do not know whether their classification is correct. Yet, we considered them in order to investigate the correlation between the respective dependent variable and the subject initially making an incorrect classification.

mitigated by providing the subjects with the accuracy information for specific use cases of that AI system.

On the other hand, this desirable debiasing effect of additional accuracy information is observed to be nullified when advice is solicited. Moreover, we document that this nullifying effect is observed regardless of subjects' FRS understanding or experience levels. Since we assume FRS understanding and experience to correlate with prior performance expectations of FRS, this independence indicates that the effect is also independent of the subjects' prior FRS performance expectations. Thus, the observed nullification is not due to the positive effect of advice solicitation offsetting the negative effect of additional information. Instead, these results suggest that subjects who solicit advice simply ignore the additional accuracy information to a significant degree, and thereby fail to benefit from its potential debiasing effect on their prior expectations.

In brief, these observations suggest that the positive impact of soliciting advice on advice utilization has an unintended negative consequence: it leads subjects to significantly ignore the additional accuracy information. If this result generalizes to any AI system, it is concerning. Consider the current use case of LLMs by the general population. Currently, LLMs are not fully integrated to our decision making process, but used as on-demand tools (Zhu and Wang, 2023). In other words, when a person needs to acquire information from an LLM, she is effectively soliciting its advice. Our findings suggest that despite warnings about the stochastic nature (Bender et al., 2021) and potential hallucinations problems of LLMs (Rudolph et al., 2023; Hicks et al., 2024; Xu et al., 2024), people may ignore these warnings and use its advice in situations where caution is needed, leading to an unwarranted degree of algorithmic trust due to a lack of desirable degree of skepticism towards its output. Given that ChatGPT-like LLMs are always used on an on-demand basis, meaning they are always solicited, any warnings on its user interface about potential pitfalls, such as "ChatGPT can make mistakes. Check important info," will not effectively make users skeptical of the output provided.

The assumed perfect automation bias, partially mitigated when additional accuracy information is provided in mandatory advice cases, is documented to partially stem from subjects' lack of understanding and experience with FRS. These results underscore the importance of experience and understanding in effectively benefiting from AI systems. The mediating effect of prior FRS understanding and experience on the negative effect of additional accuracy information on advice utilization suggests that educating individuals about the weaknesses and advantages of AI systems can mitigate PAB. This, in turn, can foster a desirable degree of skepticism, reducing over-trust and over-reliance on AI systems, and also mitigate the tendency to heavily discount AI advice after observing it to err once (Dietvorst et al., 2015; Bogert et al., 2021).

Although we found no significant difference between NM and IM treatments in terms of the average percentage change in decision for white photo pairs, the average percentage change under IM was consistently lower than the change under NM. This could stem from variance in subjects' subjective definitions of near-perfect performance and their degree of perfect automation bias. However, the consistent, albeit insignificant, difference for white male photo pairs on the average percentage change in decision between NM and IM treatments, for which the additional accuracy value is 100%, suggests that additional accuracy information may negatively affect advice utilization through mechanisms we have not accounted for.

One potential cause for the systematic (yet insignificant) negative effect of additional information on advice utilization may be the additional external cognitive load that additional accuracy information places on the subjects (Van Merrienboer and Sweller, 2005; Sweller et al., 2019). You et al. (2022) argue that providing performance information regarding the AI system places an additional external cognitive load on the subjects, and document that increased external cognitive load negatively affects advice utilization as a higher degree of external cognitive load constraints the individuals internal cognition, resulting in their decision making capacity to deteriorate (Van Merrienboer and Sweller, 2005; Sweller et al., 2019). Hence, introducing additional accuracy information may overall cause the subjects to have less internal cognitive capacity to effectively use advice which consecutively results in an overall decrease on their algorithmic advice utilization.

Additionally, our user interface provides on the right side of the photo pair an information box with hypothetical details (name, employee ID, office number, last seen date) to create a realistic scenario and increase subjects' intrinsic motivation to effectively perform the classification tasks. However, it has been shown that such additional information deteriorates the facial recognition performance of the human classifier (McCaffery and Burton, 2016; Feng and Burton, 2019). One may argue that this is because the additional information increases subjects' external cognitive load, thereby deteriorating their face recognition performance. Given that our interface already imposes a high cognitive load via this information box, additional accuracy information might further exacerbated this effect, leading to a consistent (though not significant) negative impact on advice utilization.

3.7 Final Accuracy

This study focused on the factors affecting advice utilization, and not on the factors affecting the final decision accuracy. Since FRS advice was always correct in our experiments, a higher degree of advice utilization artificially implied a higher level of final decision accuracy. As previously argued, this design was chosen to avoid cases where subjects observe simple errors by the FRS which causes them to subsequently severely discount its advice. While this approach is suitable for measuring advice utilization, it is unrealistic for investigating the effect of treatment variables on final accuracy. If an AI system were always correct, it would be unnecessary to allocate this decision to a human decision maker that is prone to making mistakes. This would be akin to arguing for an AITL system involving a human and a calculator for arithmetic tasks. Thus, the observed effects of treatment variables on advice utilization should not be extrapolated to their effects on final accuracy, as this would have no external validity.

If an AI system is not always correct, then discounting its advice for cases when it is not correct would be the desirable action for the subjects to take to improve their final accuracy. In such cases, the negative effect of additional accuracy information on advice utilization by debiasing the subjects' near-perfect performance expectations would be a desirable for two reasons. Firstly, it would increase the likelihood of the subjects to discount misleading advice. Secondly, it would mitigate the possibility of the subjects to heavily discount the advice upon observing a simple error by the algorithm as they will be less likely to have near-perfect performance expectation from the AI advice. Similarly, the negative effect of making advice optional on advice utilization would be desirable for two reasons. Firstly, by trivially not observing a misleading advice, the subject would not be subject to a false advice that may potentially misguide her. Secondly, by not soliciting in instances where the algorithmic advice is incorrect, the subject is highly confident about her classification and her initial classification is correct, she will not be exposed to the misleading algorithmig advice which she will likely to perceive as false and therefore not discount the algorithmic advice for the subsequent advices where she should not necessarily discount. In other words, not soliciting advice will not be a strictly bad action to take for the subjects, if the AI system with some percentage provides incorrect advice.

Additionally, the degree of correlation between the errors an AI system makes and the human makes would matter for how these treatments affect the final accuracy. If the errors made by the AI system perfectly align with the errors made by human subjects, the AI advice will only reinforce the subjects' initial misclassifications. In this scenario, advice discounting would be beneficial, as the AI system would exacerbate human mistakes by providing confirmatory information, leading to increased confidence in incorrect decisions. Hence, in such an extreme scenario, additional accuracy information would be beneficial to calibrate the expectations, whereas in this specific scenario, making advice optional would not make a difference in the final accuracy.

Conversely, if the errors made by the AI system are perfectly negatively correlated with human errors, soliciting AI advice would improve final accuracy. In this case, the AI system would correct human misclassifications, leading to better overall decision making outcomes. Additionally, soliciting advice in such a scenario would minimize the chances of observing obvious AI errors, thereby reducing the likelihood of heavy discounting of AI advice in subsequent decisions. Additional accuracy information in this scenario might again benefit the subject, as the subject will be less likely to portray perfect automation bias and therefore will be more likely to critically evaluate algorithmic advice and discount it. However, additional accuracy information would also increase the likelihood of discounting algorithmic advice when the advice was needed to be adopted. Hence, it is unclear whether additional accuracy information would generate an overall increase in final accuracy in this scenario.

Overall, the question of whether the final accuracy of AITL system would be higher than either of the human or the AI system alone is an empirical question that depends on the average performance of both the AI system and human, the correlation among the two parties tendency to make errors in the same tasks, and the meta-cognitive capability of the human to effectively solicit advice.

Our study has several limitations that warrant further investigation. The reliance on self-reported measures of understanding and experience with FRS is not desirable as subjects may underestimate or overestimate their degree of FRS understanding or experience. Future research could explore objective measures of these factors to validate or disprove our findings. Additionally, we do not measure the perceived near-perfect threshold of each subject nor the degree of each subject's perfect automation bias. Instead, based on the previous prevalent results in the literature, we make broad assumptions that we generalize to the entire subject pool. Future work should consider identifying subject specific near-perfect performance thresholds and measuring each subject's the degree of perfect automation bias in order to improve the validity and robustness of the claims made in this paper.

3.8 Conclusion

In this research, we investigated whether providing further autonomy to human decision makers by giving them the option to solicit advice affects their utilization of algorithmic advice. We documented that although solicited advice is utilized significantly more compared to when advice is always provided, the overall utilization of algorithmic advice is lower when it is optional because it is not solicited frequently enough.

Furthermore, we investigated the effect of additional accuracy information on algorithmic advice utilization and found that additional accuracy information has an overall negative effect on algorithmic advice utilization. This negative effect is documented to occur because subjects with near-perfect performance expectations adjust their beliefs downward upon seeing information indicating FRS performance below near-perfect levels.

Additionally, subjects with higher levels of prior FRS understanding and experience are shown to have more accurate performance expectations of FRS. This is demonstrated by the negative effect of additional accuracy information being primarily observed in subjects with below average FRS understanding and experience.

Lastly, it has been documented that when subjects solicit advice accompanied by additional accuracy information, they often ignore the additional accuracy information. Hence, the desirable debiasing effect of additional accuracy information fails to be observed when advice is solicited.

While the concept of providing further autonomy to human decision makers warrants further exploration, our findings show that making advice optional leads to lower algorithmic advice utilization and the ignoring of additional accuracy information. Therefore, we do not recommend making advice optional as a means to enhance autonomy in an AITL system.

Future research should explore alternative methods of enhancing human decision makers' autonomy by introducing mechanisms that balance reliance on human judgment with autonomy. One potential alternative could involve giving decision makers the ability to solicit a brief description of the rationale behind the algorithmic advice while making advice without rationale mandatory.

Author Contributions

Can Çelebi developed the research questions, the experimental design, and the experimental procedures, formulated the hypothesis, performed the data analysis, evaluation and interpretation, wrote the manuscript, procured the racism, sexism, trust in AI and general trust disposition questionnaires, procured the photo pairs, worked with digital artist to edit the photo pairs for their final version. Hunter Phoenix van Wagoner and Andria Smith coded the experiments in Qualtrics platform and collected the data in the Prolific platform. Andria Smith identified the additional accuracy values, wrote the experimental instructions, and prepared the instruction quiz questions. Hunter Phoenix van Wagoner cleaned the collected data for data analysis. Ksenia Keplinger procured the funding and ethical committee permissions for the experiments. All authors collectively participated in the determination of the final set of photo pairs to be used in the experiment.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work I used GPT-4 in order to improve readability and language of the text. After using this tool, I reviewed and edited the content as needed.

Appendix to Chapter 3

3.A Additional Analysis on Prior Expectations

Tables 3.10a and 3.10b present average percentage change in decision for the subset of the subjects with above average understanding and above average experience, respectively. Rows NM and IM indicate the treatments with the same abbreviations. Moreover the column p denotes the p-value for the proportion test between the NM and IM for that respective column. The columns WM, WW, BM and BW denote the subset of the data for each pair type.

	WM	WW	BM	BW
NM	68.4	69.6	68.8	65.9
IM	72	77.4	67.5	69.2
p	0.453	0.033	0.846	0.291

(a) H_U

	WM	WW	BM	BW
NM	71.6	72.5	68.6	70.5
IM	66.7	70.5	65.7	66
p	0.252	0.683	0.545	0.397

(b) H_E

Table 3.10: Percentage change in classification decision for high and/or low understanding and experience of the subject for each pair type subset

3.B Probability to Solicit Advice

Soliciting advice is a necessary but not sufficient step for advice utilization in O treatments. If the advice is not solicited, then advice is trivially not utilized; yet, soliciting advice does not guarantee its utilization. This section examines factors influencing advice solicitation, considering only data pertaining to O treatments.

Subjects decide to solicit advice before viewing the advice and its associated accuracy information. In the IO treatment, this decision requires recalling the accuracy values relevant to the classification task's pair type. Therefore, for the initial solicitation, additional accuracy information is unknown, and advice with additional accuracy information is expected to positively affect the likelihood of advice solicitation for any photo pair type due to its expected additional informational value (Yaniv and Kleinberger, 2000; Nyarko et al., 2006). However, once subjects observe the accuracy information during their first

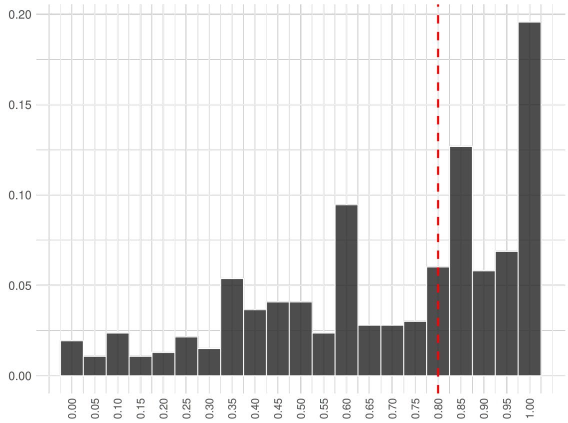


Figure 3.2: Distribution of average solicit percentage of subjects in optional advice treatments

solicitation, assuming perfect recall, they will be fully informed about the FRS's classification performance and update their prior beliefs downward. Thus, after the initial solicitation, subjects with perfect recall of the additional information may view FRS's advice as less valuable for black photo pairs than initially expected and equally valuable for white photo pairs as initially expected in no information treatments (N). Therefore, given perfect recall of the previously observed accuracy information, additional accuracy information could be expected to negatively impact advice solicitation for black photo pairs, have no effect for white photo pairs, and overall result in a negative effect on the propensity to solicit advice.

Our rationale on the effect of additional accuracy information relies on the assumption that subjects have a perfect recall of the accuracy information once they observe it. However, the extent of previous exposure to accuracy information and the subjects' ability to recall this information may influence the effect of additional accuracy information on their likelihood to seek advice, potentially invalidating our expectation that subjects' likelihood to solicit advice is mediated by the specific accuracy values provided for each pair type.

Alternatively, subjects may vaguely recall the additional accuracy information, averaging the pair type specific accuracy values into a single value that signals FRS's performance as not near-perfect. In other words, the lower accuracy values for BW and BM pairs may negatively affect subjects' overall perception of FRS's performance. Consequently, while subjects are still expected to solicit advice less when additional accuracy information is present, the decision to solicit advice may not be responsive to the specific accuracy values for each pair type.

Figure 3.2 displays the distribution of propensity to solicit advice for subjects in the O treatment. Close to 20% of the subjects always solicited advice, approximately one third of the subjects solicited advice at least 90% of the time, and approximately half of the subjects solicited advice at least 70% of the time. Conversely, around 25% of the subjects solicited less than half of the time. In brief, there is a significant degree of variance among subjects in terms of their exposure to the additional accuracy information.

Given the randomization of photo pairs and subjects likely paying attention only to relevant pair type accuracy at a given classification task, subjects might only correctly recall the relevant accuracy information if the prior task involved the same pair type. This creates uncertainty regarding whether treatment I significantly affects the likelihood of soliciting advice and if pair type specific accuracy information influences this likelihood. Moreover, the randomized order of classification tasks makes it unclear how subjects' solicitation behavior evolved.[37] For those who solicited advice in less than 90% of cases, there could be a considerable number of tasks where they did not solicit advice and therefore did not observe the additional accuracy information between the tasks where they did solicit advice, reducing their likelihood of remembering the additional accuracy information from previous tasks. Hence, lower overall solicitation frequency diminishes the probability of recalling previously observed accuracy information, thus decreasing the potential effect of additional accuracy information on the likelihood of soliciting advice.

	f_s^{All}	f_s^{WM}	f_s^{WW}	f_s^{BM}	f_s^{BW}
N	71.9	72	72.8	72.4	70.2
I	66.6	66.4	70	64.9	65.2
μ_{NI}	69.2	69.2	71.4	68.6	67.7

Table 3.11: Average percentage to solicit advice

Table 3.11 shows the average percentages of soliciting advice in O treatments. The variable f_s indicates the average solicitation percentage. The All superscript denotes the average for the entire O treatment dataset, while the WM, WW, BM, and BW superscripts denote the averages for the WM, WW, BM, and BW pair type subsets under the O treatment. The rows correspond to the N and I treatments.

Overall, the average percentage to solicit is observed to be significantly lower in I treatment compared to N treatment (71.9 vs 66.6; proportion test, $p < 0.001$). This result indicates that additional accuracy information, overall, has a negative effect on the likelihood of soliciting advice.

Additionally, for the BM and WM subsets, average solicitation percentages are significantly lower in the I treatment compared to the N treatment (proportion test, $p_{max} < 0.001$). In contrast, for the BW subset, the average solicitation percentage is marginally lower in the I treatment (proportion test, $p = 0.005$), while for the WW subset, there is no significant difference between the treatments (proportion test, $p = 0.114$). These results suggest that pair type specific accuracy information values do not correlate with

[37]We did not track the order with which each subject received each classification task.

the subjects' average likelihood of soliciting advice.

The significant difference for BM and the marginal significance for BW suggest that subjects may account for the pair type specific accuracy information values, as their accuracy values, signaling less than perfect performance from FRS, negatively impact solicitation likelihood among subjects with near-perfect prior expectations. Furthermore, the lack of a significant effect for the WW subset is consistent with our rationale, as we expected that high accuracy would not alter subjects' near-perfect prior expectations and thus would not affect their solicitation likelihood. However, WM's 100% accuracy was also expected not to affect solicitation likelihood, as it confirms subjects' prior near-perfect expectations. The significant difference found under for WM is in conflict with our formerly argued expectations. Consequently, we find only partial evidence for the effect of pair type specific additional accuracy information on the likelihood to solicit advice.

Table 3.12: Probability to solicit advice

	PS_1	PS_2	PS_3	PS_{3^*}
α	1.57***	20.8***	20.42***	20.3***
	(0.14)	(0.57)	(0.77)	(0.77)
I	-0.42*	-0.52*	-0.52*	-0.51*
	(0.20)	(0.24)	(0.24)	(0.24)
C_1		-0.22***	-0.22***	-0.22***
		(0.01)	(0.01)	(0.01)
T_C			-0.11	-0.11
			(0.15)	(0.15)
$MiMa$				0.23***
				0.06
var_{RE}	3.95	5.89	5.87	5.89
AIC	10586	8069	8071	8059
Obs	11160	11160	11160	11160

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

In Table 3.12, we investigated the effect of additional accuracy information with three[38] logistic mixed-effect models (PS_{1-3}) where the dependent variable is the binary

[38] We have a forth model PS_{3^*} where we include the dummy variable $MiMa$, which takes 1 if the initial classification of the subject does not match the FRS classification. Given FRS classification is always correct, $MiMa$ effectively represents the cases where the subject's initial classification is incorrect. From a modeling perspective this model is flawed. Introducing $MiMa$ as an independent variable assumes $MiMa$ affects the probability of solicitation. However, since subjects observe the FRS classification only after soliciting advice, a direction of causality from $MiMa$ to probability to solicit is not plausible.

Table 3.13: Probability to solicit advice for each pair type subset

	PS_{WM}	PS_{WW}	PS_{BM}	PS_{BW}
α	20.9***	18.38***	22.33***	18.98***
	(1.29)	(1.18)	(1.42)	(1.23)
I	-0.48†	-0.37	-0.7*	-0.4
	(0.26)	(0.25)	(0.29)	(0.26)
C_1	-0.22***	-0.2***	-0.23***	-0.2***
	(0.01)	(0.01)	(0.01)	(0.01)
T_C	0.11	0.17	0.02	0.12
	(0.15)	(0.15)	(0.17)	(0.16)
var_{RE}	5.29	4.87	6.53	5.52
AIC	2363	2345	2310	2434
Obs	2790	2790	2790	2790

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

variable to get advice. The models are conducted on the subset of the data containing only the O treatments. Subjects are incorporated as random effect to the models. In all models, additional accuracy information is observed to have a significant negative effect on the probability to solicit advice. Additionally, while initial confidence is observed to have a negative significant effect on the probability to solicit advice, trust in FRS is found not to have a significant effect on the probability to solicit advice.

Additionally, in Table 3.13, we have estimated model PS_3 under each pair type subset. Additional accuracy information is found to significantly and negatively affect the probability to solicit advice only for the pair type BM, and negatively and marginally significantly affect the probability to solicit advice for the pair type WM. These results provide further evidence that there is no correlation between the degree of accuracy provided for the specific pair type and the probability to solicit advice.

In Table 3.14, covariates are gradually introduced to the model PS_3, in the same fashion done with models CD_{7-11}. The negative effect of additional information and initial confidence on the probability to solicit advice is maintained across models PS_{4-8}.

This model is an attempt to investigate whether subjects are more likely to solicit advice following an initial misclassification, even though they do not receive any feedback on their initial classification and hence they do not know whether they have made a false classification, and even though we already account for their confidence in the correctness of their initial classification with the variable C_1. The results of this model, especially the significant positive estimated coefficient of $MiMa$ is considered as a part of a discussion in Section 3.6 regarding the subjects' ability to effectively solicit advice when they have an initial incorrect classification. Hence this model is not part of the investigation regarding the factors affecting subjects' probability to solicit advice but provided in Table 3.12 for convenience.

Table 3.14: Probability to solicit advice with additional covariates

	PS_4	PS_5	PS_6	PS_7	PS_8
α	-5.73***	-6.67*	-7.51***	-8.2***	-7.87*
	(2.57)	(2.64)	(2.64)	(2.67)	(2.67)
I	-0.50*	-0.46†	-0.50*	-0.53†	-0.54*
	(0.25)	(0.25)	(0.24)	(0.24)	(0.23)
C_1	0.42***	0.42***	0.42***	0.42***	0.42***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
T_C	0.1	0.08	-0.03	-0.01	0.05
	(0.16)	(0.15)	(0.15)	(0.15)	(0.15)
μ_{A_1}	-0.32	-0.32	-0.32	-0.32	-0.32
	(0.25)	(0.25)	(0.25)	(0.25)	(0.25)
TP	-0.04	-0.04	-0.04	-0.04	-0.04
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
PT_{BW}	-0.09	-0.09	-0.09	-0.09	-0.09
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
PT_{WM}	0.04	0.04	0.04	0.04	0.04
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
PT_{WW}	0.10	0.10	0.11	0.10	0.10
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
T_H		0.13†	0.1	0.07	0.03
		(0.09)	(0.08)	(0.08)	(0.08)
FRS_E			-0.01†	-0.01†	-0.01†
			(0.00)	(0.00)	(0.00)
FRS_U			0.32	0.28	0.23
			(0.28)	(0.27)	(0.27)
FRS_I			0.02***	0.02***	0.02***
			(0.01)	(0.01)	(0.00)
$Female$				-0.17	-0.15
				(0.24)	(0.23)
Age				0.03***	0.03***
				(0.01)	(0.01)
$Black$				-0.08	-0.09
				(0.24)	(0.24)
Edu_H				-0.33	-0.31
				(0.25)	(0.24)
$Sexism$					0.08
					(0.1)
$Racism$					-0.3*
					(0.1)
var_{RE}	6.08	6.02	5.58	5.38	5.26
AIC	7981	7982	7963	7956	7951
Obs	11160	11160	11160	11160	11160

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

None of the classification task specific covariates (μ_{A_1}, TP, PT_{BW}, PT_{WM}, PT_{WW}) are observed to have a significant effect on the probability to solicit advice. For models PS_{6-7}, a higher degree of experience in FRS, FRS_E, is observed to have a negative and significant, albeit minuscule, effect on the probability to solicit advice, whereas a higher degree of interest in FRS, FRS_I, is observed to have a positive and significant, albeit small, effect on the probability to solicit advice for models PS_{6-8}. Furthermore, older subjects are found to be more likely to solicit advice, and subjects with a higher degree of racism are found to be less likely to solicit advice.

3.C Change in Confidence

The primary measure of advice utilization we have so far considered is the subject's switch from their initial classification to the observed advised classification. Difference between pre- and post-advice confidence levels can provide a more nuanced understanding on how the advice influences subject's decision process (Bonaccio and Dalal, 2006). Even if the subject does not alter her initial classification upon receiving a differing advice, a reduction in her confidence will suggest that the advice, at the very least, affected her belief in the correctness of her initial classification. Conversely, if the subject observes an agreeing classification advice from FRS, although she is very much likely not to change her pre-advice classification, an increase in her confidence may be observed.

Past studies have shown that, on average, receiving advice increases subjects' confidence (Heath and Gonzalez, 1995; Savadori et al., 2001; Soll et al., 2022). Savadori et al. (2001) argue that the process of advice evaluation enables the advisees to further develop their rational for their classification. This process, in turn, results in a higher post-advice confidence level compared to the pre-advice confidence level. In addition, Soll et al. (2022) argue that when subjects' initial classification matches the advised classification, their confidence level increases to the degree of being considered as overconfident. In line with the previous results in the literature, we expect that when advice aligns with subjects' initial classifications, it serves as a confirmatory signal that boosts subjects' belief in the correctness of their decisions (Nickerson, 1998).

Conversely, in cases where there is a mismatch between the two classifications, advice can be expected to serve as a signal that decreases subjects' belief on the correctness of their classification, and hence, to decrease subjects' post-advice confidence. On the other hand, it is also likely for this non-confirmatory information to be under-weighted, if not ignored (Edwards and Smith, 1996). Zaleskiewicz and Gasiorowska (2018) show that if the advice is not in line with the subject's initial decision, subjects have a higher tendency to disregard the advice and question the advisor's ability to correctly assess the problem (rather than questioning the correctness of their answer). Hence, subject's post-advice confidence level may be expected not to decrease upon observing a contradictory advice, and may even be increased to compensate for the dissonance generated by the observed disagreement between the classifications of FRS and the subject.

	M	O	O_s	O_x	μ_{MO}
N	85.3	86.1	83.3	92.9	85.7
I	85.3	86.2	82.8	92.9	85.8
μ_{NI}	85.3	86.1	83.1	92.9	85.8

Table 3.15: Initial confidence

In Table 3.15, initial confidence levels for each treatment are presented. Labels are as defined for Table 3.1a. There is no significant difference in the initial confidence levels of the subjects in the treatments NM and IM (ranksum test, $p = 0.39$), and in the treatments NO and IO (ranksum test, $p = 0.19$). On the other hand, overall, subjects in the O treatments are observed to have significantly higher initial confidence relative to the subjects in the M treatments under either N or I treatments (ranksum test, $p_{max} < 0.001$). Yet, although the difference is significant, it is less than a percentage point higher than the average initial confidence levels observed in the M treatments. Subjects who solicited advice are observed to have a significantly lower initial confidence levels than subjects who did not solicited advice, O_x (ranksum test, $p_{max} < 0.001$) and than the subjects in the M treatments (ranksum test, $p_{max} < 0.001$) under either N or I treatments. Whereas subjects who choose not to solicit advice are observed to have a significantly higher initial confidence levels than subjects in the M treatments (ranksum test, $p_{max} < 0.001$). These significant differences corroborate our finding that initial confidence is a significant factor in the decision to solicit advice from FRS.

Additionally, a mixed-effect linear regression model[39] with the initial confidence level as the dependent variable is considered to further investigate the effect of treatment variables while controlling for additional covariates on the initial confidence levels. As before, each subject is treated as a random effect in the model.

Neither of the treatment variables are found to have a significant effect on the initial confidence levels. Moreover, BW and WW photo pairs are found to have a negative significant effect on the initial confidence levels. This indicates that, overall, relative to the baseline of BM photo pairs, BW and WW photo pairs are perceived as pairs that subjects are less confident in the correctness of their classification. However, as we have previously presented, BW pairs have a positive significant effect on the initial classification accuracy, whereas WW pairs have a negative significant effect on the initial classification accuracy. Hence, while with WW pairs subject's accuracy and her confidence in her accuracy are positively correlated, for BW pairs, it is negatively correlated, indicating some degree of underconfidence in the classification of BW pairs.

[39]See Appendix 3.D Table 3.19 for details.

Additionally, black subjects are found to be more confident in their initial classification. Similarly, higher education is found to be marginally significant and positive, suggesting higher education having a positive effect on the initial confidence of the subjects. Education's effect on confidence is on par with its marginal positive effect on the accuracy of the initial classification, indicating the subjects with higher education are slightly better at the classification task and more confident on their initial classification. A higher degree of sexism and racism are both found to have a positive significant effect on the initial confidence level. While a higher degree of sexism was not found to have any effect on the initial accuracy, racism was observed to have a negative effect on the initial accuracy. Hence, a higher degree of racism, and to some degree a higher degree of sexism, result in the subjects to be overconfident in their initial classification.

	M	O	O_s	O_x	μ_{MO}
N	9.9	9.1	12.9	0.2	9.4
I	9.1	8.7	13.3	0.3	8.9
μ_{NI}	9.5	8.9	13.1	0.2	9.2

Table 3.16: Percentage change in confidence for match cases

In Table 3.16, average percentage changes in the confidence for match cases are presented. Match cases denote the cases where the subject's initial classification matches with FRS' classification. Recall from Table 3.2a that on average only around 1.4% of the instances, subjects change their initial classification, if their initial classification matches with the FRS' classification. Thus, the confidence changes in Table 3.16 primarily reflect the subjects' reactions to FRS' advice matching their initial classification. Furthermore, in addition to the investigation of the additional information's effect on confidence, unlike our prior investigation on the percentage change in decision, we are not interested in the overall effect of the O treatments on the percentage change in confidence, as we are attempting to discern whether subjects who did not change their decision upon observing advice are still affected by it via adjusting their belief on the correctness of their classification. While the overall effect of the O treatment is relevant for the advice utilization, as opting to not receive advice is a trivial type of advice discounting, it is not relevant for the effect of advice on the level of confidence. Instead, we are concerned with identifying whether soliciting advice has an effect on the post-advice confidence, as we are investigating whether there is a change in the confidence for cases where the subject discounted the advice by opting not to adopt the FRS' classification.

In Table 3.16, percentage changes in confidence is marginally significantly greater in NM treatment compared to IM treatment (ranksum test, $p = 0.002$). Conversely, there is no significant difference found between NO_s and IO_s (ranksum test, $p = 0.872$). These results indicate that for match cases while the additional information treatment has a negative effect on the post-advice confidence, this effect is reversed when the advice is solicited. However this reversing effect of soliciting advice on additional information is found not to be statistically significant. On the other hand, percentage changes in confidence is found to be significantly greater when advice is solicited, O_s, then when advice is mandatory, M, under either information treatment, N or I (ranksum test $p_{max} < 0.001$). Lastly, when advice is not received, O_x, the change in confidence is found to be significantly lower then any other case (ranksum test $p_{max} < 0.001$), but still significantly larger than 0 (ranksum test, $p_{min} > 0.11$).

In summary, when subjects observe FRS' advice to agree with their initial classification, their initial confidence increase on average by 9.2 percentage points. This increase is significantly larger when the subjects solicit the advice. Furthermore, the presence of additional information has a significant negative effect on post-advice confidence only when advice is mandatory.

	M	O	O_s	O_x	μ_{MO}
N	3.1	4.6	4.6	-0.8	3.8
I	0.9	5.8	5.9	1.9	3.3
μ_{NI}	2.1	5.2	5.3	0.1	3.6

(a) Decision changed

	M	O	O_s	O_x	μ_{MO}
N	-4.9	-1.8	-5.6	0.3	-3.1
I	-5.3	-0.8	-3.2	0.4	-2.7
μ_{NI}	-5.1	-1.3	-4.5	0.4	-2.9

(b) Decision not changed

Table 3.17: Percentage change in confidence for mismatch cases

In Tables 3.17a and 3.17b, average percentage changes in the confidence for mismatch cases are presented. The mismatch cases denote the cases where the subject's initial classification differs from the classification of FRS. Table 3.17b presents the subcases where subjects changed their classification to FRS' classification, whereas Table 3.17a presents the subcases where subjects did not change their classification despite observing a different classification by FRS.

In Table 3.17a, for the cases where the subjects adopted the advised classification, except for the IM treatment and O_x subcases, percentage change in confidence is significantly larger than 0 (ranksum test, $p_{max} < 0.001$). Moreover, except for O_x subcases, the value in each cell in Table 3.17a is significantly lower than the values in their respective counterpart cells in 3.16 (ranksum test, $p_{max} < 0.001$) where the advised classification matches the initial classification. These two results indicate that changing the classification to the advised classification increases the confidence but this increase is not as high as the increase observed when subjects receive an advice matching their initial classification.

The results in Table 3.17a parallel the results we observed with match cases in Table 3.16. Additional information is causing the increase in the post-advice confidence to be lower in M treatments while the reverse holds true when the advice is solicited. Moreover, similar to the match cases, while the difference between NM and IM is found to be significant (ranksum test, $p < 0.001$), the difference between NO_s and IO_s is found not to be signficant (ranksum test, $p = 0.226$). These results indicate that when advice is mandatory, if it is accompanied by additional information, additional information mitigates subject's tendency to increase her confidence upon adopting FRS' advice. On the other hand when advice is solicited, additional information further boosts subject's tendency to increase her confidence upon adopting FRS' advice, but this effect is not significant.

Additionally, percentage changes in confidence is found to be significantly greater when advice is solicited, O_s, then when advice is mandatory under I (ranksum test, $p < 0.001$), and marginally significantly greater under N (ranksum test $p = 0.007$). These results indicate that soliciting advice not only enables a higher degree of advice utilization, but also reinforces advice utilization by generating a higher degree of increase in the initial confidence of the subjects when advice confirms with subjects' initial (correct) classification. Lastly, average percentage changes in O_x cases are found not to be significantly different from 0 under either N or I treatments (ranksum test, $p_{min} > 0.659$).

In Table 3.17b, except for O_x subcases, all the average percentage change in confidence values negative. Furthermore, except for IO where the value of -0.8 is found not to be significantly different from 0 (ranksum test, $p = 0.022$), they are all found to significantly less than 0 (ranksum test, $p_{max} < 0.001$). These results show that when the advised classification differs and the subjects do not change their classification, they still, on average, adjust their confidence by being less certain about the correctness of their initial confidence.

The results in Table 3.17b align with the previously observed results regarding the effect of additional information. In M treatments, additional information causes the decrease in the post-advice confidence to be higher, whereas the decrease in confidence is mitigated when advice is solicited. Yet, unlike previous cases, the effect of additional information is significant under O_s (ranksum test, $p < 0.001$) but not under M (ranksum test, $p = 0.123$).

Note that although additional accuracy information's effect and its direction parallels previous cases, its implications differ in this scenario where the subjects do not adopt the FRS advice, i.e. utilize the advice. In the mismatch cases where the subjects adopted the FRS classification, an increase in confidence implies that the subjects become more confident upon utilizing the advice. Whereas, in this scenario, an increase in confidence implies that the subjects become more confident upon discounting the advice. Hence, the observed significant positive effect of additional information in O_s subcases on initial confidence implies that when advice is solicited, observing additional information reinforces subjects' advice discounting by mitigating the decrease in their initial confidence for not utilizing the advice.

Furthermore, unlike previous cases, soliciting advice does not significantly affect post-advice confidence under the no information treatment N (ranksum test, $p = 0.083$). However, under the additional information treatment I, soliciting advice has a significant positive effect on post-advice confidence (ranksum test, $p < 0.001$). These results suggest that while subjects may not change their classification to match the advised classification, their (mis)belief on the correctness of their initial classification decreases to a larger degree when advice is solicited.

Table 3.18 presents four linear mixed-effect models estimating the effect of treatment variables on the percentage change in subjects' confidence, treating each subject as a random effect. Except for DC_{All}, each model is estimated on a specific subset of the dataset. DC_{All} is based on the entire dataset, DC_{Ma} on match cases, DC^{CD}_{Mima} on mismatch cases where subjects changed their decision, and DC^{NCD}_{Mima} on mismatch cases where subjects did not change their decision.

Table 3.18: Percentage change in confidence

	DC_{All}	DC_{Ma}	DC^{CD}_{MiMa}	DC^{NCD}_{MiMa}
α	5.18***	9.94***	2.62**	-6.34***
	(1.12)	(0.52)	(0.89)	(0.57)
I	-1.22†	-0.79	-2.44†	0.84
	(0.72)	(0.74)	(1.26)	(0.77)
O_s	2.83***	3.13***	1.7	0.73
	(0.71)	(0.73)	(1.23)	(0.82)
$I : O_s$	1.87†	0.79	3.72*	2.1†
	(1)	(1.03)	(1.76)	(1.13)
O_x	-9.16***	-12.91***	-7.93†	-5.48***
	(0.43)	(0.46)	(4.26)	(0.61)
$I : O_x$	-0.07	0.47	7.81	-3.21***
	(0.59)	(0.63)	(5.54)	(0.83)
NC_D	4.24***	0.78		
	(1)	(0.85)		
$MiMa$	-9.16***			
	(0.43)			
$MiMa : NC_d$	-5.46***			
	(1.05)			
var_{RE}	46.29	48.9	42.98	29.53
AIC	164625	88020	48270	23432
Obs	20952	11750	5763	3439

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.
Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

In model DC_{All}, we introduce the dummy variable $MiMa$, which takes the value of 1 if there is a mismatch between the subject's initial classification and the FRS' classification and 0 otherwise. Additionally, we include the dummy variable NC_D, which takes the value of 1 if the subject did not change her initial decision and 0 otherwise.

In model DC_{All}, the marginally significant negative coefficient for I and the significant positive coefficient for IO_s indicate that additional information overall reduces post-advice confidence, while soliciting advice increases it. Furthermore, choosing not to receive advice, O_x, significantly decreases confidence. The significant positive coefficient for NC_d suggests that when subjects receive confirming advice and do not change their decision, their confidence increases. Conversely, the significant negative coefficient for $MiMa$ indicates that conflicting advice and changing the initial classification decrease confidence. Lastly, the significant negative estimated coefficient for the interaction term between $MiMa$ and NC_d shows that observing conflicting advice and not changing the initial classification further decreases confidence.

In models DC_{Ma}, DC^{CD}_{MiMa}, and DC^{NCD}_{MiMa}, the varying number of observations makes it difficult to compare the insignificance of the dummy variable coefficients across models. That being said, the significant positive intercepts in models DC_{Ma} and DC^{CD}_{MiMa}, and the significant negative intercept in DC^{NCD}_{MiMa}, are consistent with previous observations: confidence increases in match cases and in mismatch cases where subjects changed their decision, but decreases in mismatch cases where subjects did not change their decision, except for O_x subcases. The marginally significant negative coefficient for I in DC^{CD}_{MiMa} aligns with the previously observed negative effect of additional accuracy on confidence changes in match cases and mismatch cases where subjects changed their initial classification. The positive coefficients for O_s in models DC_{Ma} and DC^{CD}_{MiMa}, though only significant in DC_{Ma}, are consistent with the previously observed positive effect of soliciting advice on confidence changes in match cases and mismatch cases where subjects did not change their classification. Lastly, despite the earlier observation of a significant positive interaction between treatments only in mismatch cases where subjects did not change their decision, we observe the estimated coefficient for $I : O_s$ to be positive and significant under model DC^{CD}_{MiMa} and to be positive and marginally significant under the model DC^{NCD}_{MiMa}. Thus, our complementary regression analysis largely corroborates our earlier results.

In brief, the change in confidence analysis shows that not only additional accuracy information decreases advice utilization, but for cases where subjects utilizes FRS' advice, either by not changing their initial classification upon observing confirmatory advice or by switching to the advised classification, it causes the subject to be less confident in utilizing FRS' advice. Similarly, soliciting advice not only increases advice utilization but it also increase subjects' post-advice confidence upon the utilization of the advice. Lastly, when solicited advice is accompanied by additional accuracy information and the advice contradicts with the subject's initial classification and the subject decide not to switch her classification to the advised classification, additional accuracy information reinforces this discounting behavior by mitigating the overall effect of decrease in confidence when the advised classification is not utilized.

3.D Additional Tables

Table 3.19: Pre-Advice error and confidence

	PRE_{Err}	PRE_{Con_1}	PRE_{Con_2}
α	-0.33***	80.41***	80.56***
	(0.08)	(1.17)	(1.17)
I	-0.06	0.08	0.07
	(0.05)	(0.74)	(0.74)
O	-0.03	0.82	0.81
	(0.05)	(0.71)	(0.71)
$I:O$	-0.02	0.09	0.09
	(0.07)	(1.01)	(1.01)
PT_{BW}	-0.55***	-0.28*	-0.32*
	(0.04)	(0.16)	(0.16)
PT_{WM}	0.00	0.05	0.05
	(0.04)	(0.16)	(0.16)
PT_{WW}	0.10*	-0.69***	-0.68***
	(0.04)	(0.16)	(0.16)
$Female$	0.10*	0.40	0.41
	(0.03)	(0.51)	(0.51)
Age	0.00*	0.02	0.02
	(0.00)	(0.02)	(0.02)
$Black$	-0.05	2.51***	2.51***
	(0.03)	(0.51)	(0.51)
Edu_H	-0.05	0.79	0.78
	(0.04)	(0.52)	(0.52)
$Sexism$	0.01	0.40†	0.40†
	(0.01)	(0.21)	(0.21)
$Racism$	0.05*	0.38†	0.39†
	(0.01)	(0.22)	(0.22)
$MiMa$			-0.36**
			(0.12)
var_{RE}	0.08	52.23	52.15
AIC	28321	149030	149023
Obs	20952	20952	20952

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "$MiMa$" represents the cases where the subjects' initial classification does not match the FRS classification. It takes 1 if there is a mismatch and 0 otherwise. Since FRS classification is always correct, it effectively represents the cases where the subjects' initial classification is false. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations.

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

Table 3.20: Probability to change decision

	CD_3b	CD_6b
α	9.79***	7.88***
	(0.35)	(0.39)
I	-0.44†	-0.45†
	(0.2)	(0.21)
O	-0.7***	
	(0.19)	
$I : O$	0.07	
	(0.26)	
O_s		0.72***
		(0.21)
$I : O_s$		0.06
		(0.3)
O_x		-6.49***
		(0.33)
$I : O_x$		0.71‡
		(0.43)
Ma	-7.8***	-7.72***
	(0.27)	(0.27)
C_1	-0.1***	-0.08***
	(0.00)	(0.00)
$Ma : I$	0.83†	0.83†
	(0.35)	(0.35)
$Ma : O$	1.55***	
	(0.32)	
$Ma : I : O$	-0.19	
	(0.43)	
$Ma : O_s$		-1.17†
		(0.43)
$Ma : I : O_s$		0.53***
		(0.55)
$Ma : O_x$		9.13***
		(0.49)
$Ma : I : O_x$		-1.74*
		(0.62)
var_{RE}	2.95	3.31
AIC	11112	9039
Obs	20952	20952

All the variables in the first column except "var_{RE}" are the estimated fixed effects. "α" is the intercept. "var_{RE}" is the variance of the player specific random effects. "AIC" is the Akaike information criterion. "Obs" represents the number of observations. ":" represents the interaction between two variables used before and after it.

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "†" 0.1 " " 1

3.E Additional Materials

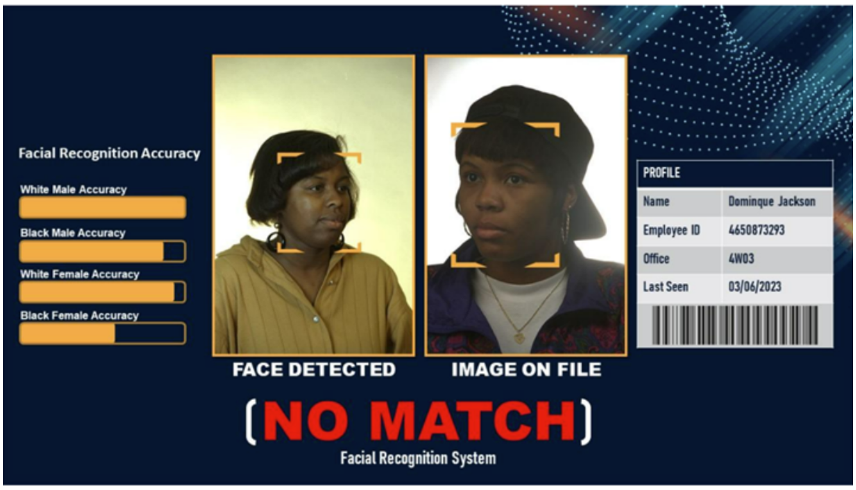


Figure 3.3: FRS advice screen with additional accuracy bar plots for each pair type

Figure 3.4: Example of an edited photo

On the middle and the right, we have the original images of the same person. In order to make this classification task, the earrings are edited out and the clothing is replaced on the image on the left.

3.F Experimental Instructions

Welcome to our online experiment, where **you will be playing the role of a security officer responsible for granting or denying access to a highly secured facility**. As you may know, security is of utmost importance in facilities that store sensitive information or house critical infrastructure.

In this experiment, you will be faced with a series of scenarios where individuals will be attempting to gain access to the facility. Some will be authorized personnel, while others will be impostors trying to gain entry through deception. **It will be your responsibility to determine who should be granted access and who should be denied**.

But **the stakes are high** in this virtual environment. **Wrong decisions could lead to adversaries** gaining entry into the facility, compromising sensitive information or causing harm to critical infrastructure. On the other hand, denying access to correct employees could also have significant consequences, disrupting operations and causing unnecessary delays.

Figure 3.5: Main instructions page 1

You will be presented with images of **24** individuals attempting to gain access to the facility.

You will be asked to determine how to proceed with the individual.

If you determine that the face detected **MATCHES** the authorized personnel on file, you can **GRANT ACCESS**.

If you determine that the face detected **DOES NOT** match the authorized personnel on file, you can **DENY ACCESS**.

You will then be asked to give your **level of confidence** based on your decision, on a scale from **0 (not confident)** to **100 (very confident)**.

Some participants will be asked to use a facial recognition system as an aid in their decision-making, while others will have the decision to use their own judgement or solicit additional information from the facial recognition system if they require assistance. Furthermore, some participants will be provided additional information on the accuracy of the facial recognition system, as performance varies across different demographics. This information is provided to further support the decision-making process.

After assessing the pairs of images, you will be asked general demographic questions and survey questions. All questions must be answered for completion of the study to receive payment.

The experiment also includes attention checks.

This experiment will take between **20 to 25 minutes**.

You will now be directed to a tutorial to practice the task at hand.

Figure 3.6: Main instructions page 2

For this tutorial, you will be presented **two** individuals attempting to gain access to the facility.

A facial recognition system will be used to detect the face of the person at the facility entrance and compare it against the image on file of the authorized personnel.

You will be first shown a set of images **with no information** from the facial recognition system. You will need to determine how to proceed with the individual, by deciding to either **GRANT ACCESS** or **DENY ACCESS**.

You will then need to provide your level of confidence on a scale from **0 (not confident)** to **100 (very confident)**.

You will next be shown a set of images **with information** from the facial recognition system. The facial recognition system will inform you if the face detected is a **MATCH** or **NO MATCH**.

From this information, you will need to determine how to proceed with the individual, by deciding to either **GRANT ACCESS** or **DENY ACCESS**.

You will then need to provide your level of confidence on a scale from **0 (not confident)** to **100 (very confident)**.

Please click through to proceed with the tutorial.

Figure 3.7: Tutorial instructions

Someone is waiting at the facility entrance. A face has been detected.

*Please determine how to proceed with this person.

- [ ] GRANT ACCESS - Face detected matches authorized personnel on file
- [x] DENY ACCESS - Face detected does not match authorized personnel on file

*How confident are you in your decision?

Confidence

44

0 100

Figure 3.8: Tutorial task example

*You have the option to solicit additional information from the facial recognition system.

- True
- False

*You will always provide your level of confidence after making your decision.

- True
- False

*How many pairs of images will you evaluate?

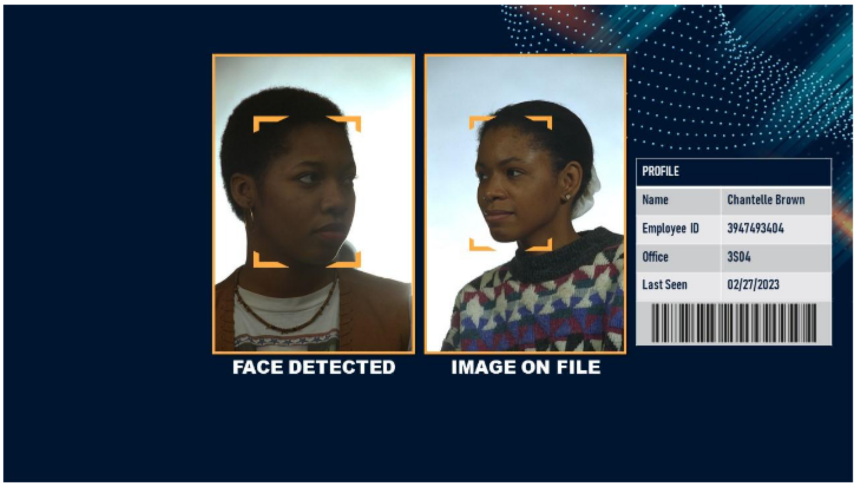
- 12
- 24
- 36
- 48

*There will be demographic questions and surveys to complete at the end of the main experiment.

- True
- False

Figure 3.9: Quiz questions

Someone is waiting at the facility entrance. A face has been detected.



*Please determine how to proceed with this person.

- ◯ GRANT ACCESS - Face detected matches authorized personnel on file
- ◯ DENY ACCESS - Face detected does not match authorized personnel on file

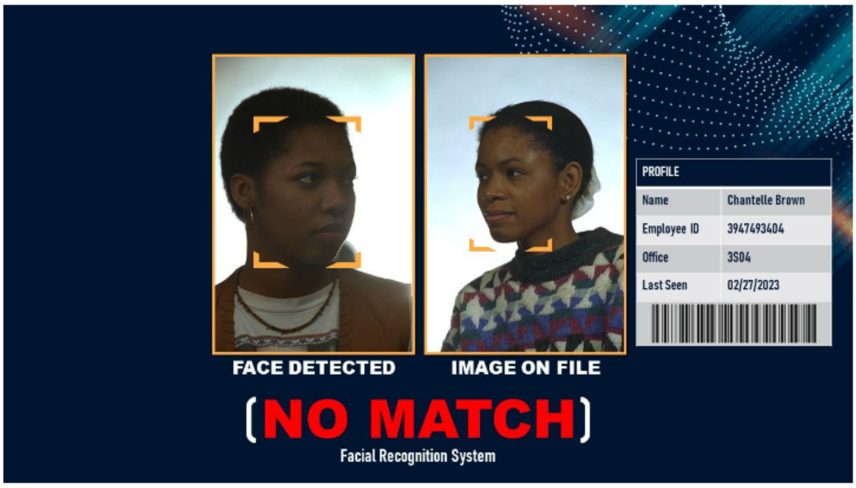
*How confident are you in your decision?

Confidence

0 100

Figure 3.10: Classification task step 1 example

The facial recognition system has made a determination. Please indicate how to proceed with this new information.



*Please determine how to proceed with this person.

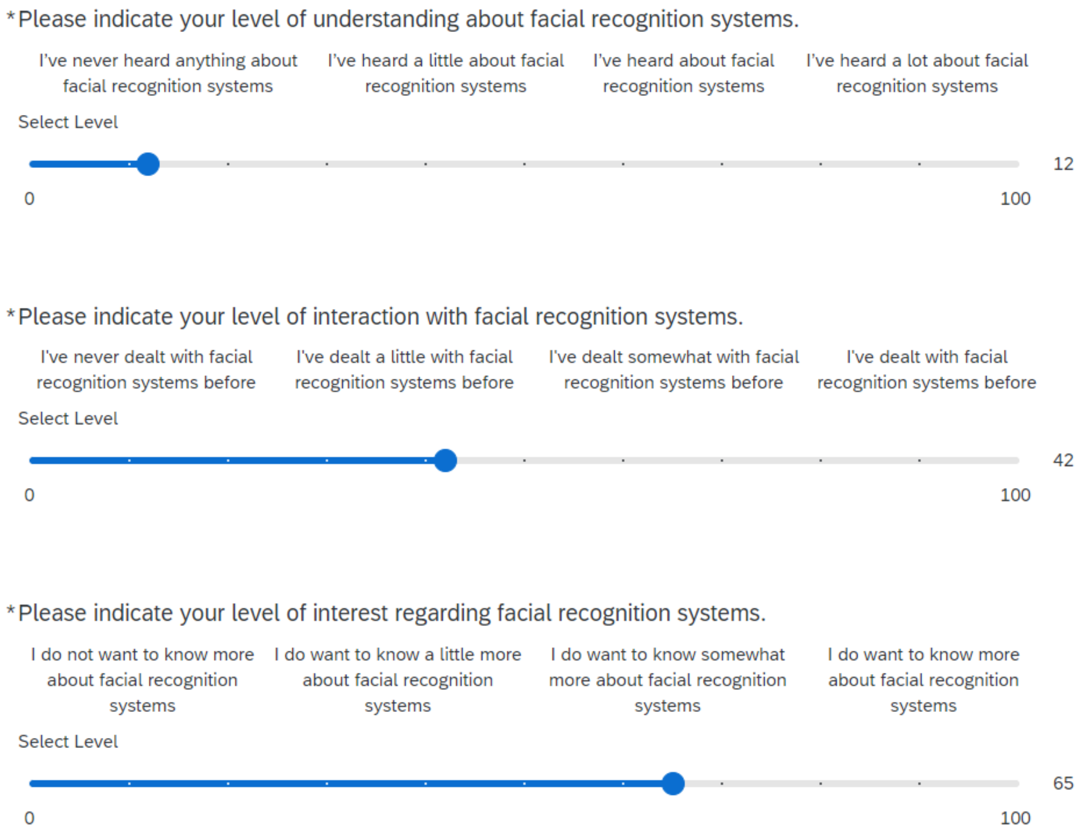
- ◯ GRANT ACCESS - Face detected matches authorized personnel on file
- ◯ DENY ACCESS - Face detected does not match authorized personnel on file

*How confident are you in your decision?

Confidence

0 100

Figure 3.11: Classification task step 2 example


*Please indicate your level of understanding about facial recognition systems.

I've never heard anything about facial recognition systems | I've heard a little about facial recognition systems | I've heard about facial recognition systems | I've heard a lot about facial recognition systems

Select Level

12

0 100

*Please indicate your level of interaction with facial recognition systems.

I've never dealt with facial recognition systems before | I've dealt a little with facial recognition systems before | I've dealt somewhat with facial recognition systems before | I've dealt with facial recognition systems before

Select Level

42

0 100

*Please indicate your level of interest regarding facial recognition systems.

I do not want to know more about facial recognition systems | I do want to know a little more about facial recognition systems | I do want to know somewhat more about facial recognition systems | I do want to know more about facial recognition systems

Select Level

65

0 100

Figure 3.12: FRS understanding, experience and interest questions

3.F.1 Trust in FRS Questionnaire

I believe facial recognition systems are competent performers.[40]

I trust facial recognition systems.

I have confidence in the advice given by facial recognition systems.

I can depend on facial recognition systems.

I can rely on facial recognition systems to behave in consistent ways.

[40] These survey questions are adopted from Mcknight et al. (2011) to fit the concept of FRS. For each question the following choices are given:
Strongly disagree, Rather disagree, Neither agree nor disagree, Rather agree, Strongly agree
Other questionnaires (sexism, racism, general trust propensity) are used without any modification to their questions and hence are omitted.

This page is intentionally left blank

Bibliography

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

Adadi, Amina and Mohammed Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE access 6 (2018): 52138-52160*, 2018.

Agranov, Marina, Elizabeth Potamites, Andrew Schotter, and Chloe Tergiman, "Beliefs and endogenous cognitive levels: An experimental study," *Games and Economic Behavior*, 2012, *75* (2), 449–463.

Aiyappa, Rachith, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn, "Can we trust the evaluation on ChatGPT?," *arXiv preprint arXiv:2303.12767*, 2023.

Alaoui, Larbi and Antonio Penta, "Endogenous depth of reasoning," *The Review of Economic Studies*, 2016, *83* (4), 1297–1333.

Alonso-Robisco, Andres and José Manuel Carbó, "Analysis of CBDC narrative by central banks using large language models," *Finance Research Letters*, 2023, *58*, 104643.

Amin, Mostafa M, Erik Cambria, and Björn W Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt," *IEEE Intelligent Systems*, 2023, *38* (2), 15–23.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals," *And it's biased against blacks. ProPublica*, 2016, *23*, 77–91.

Anthropic, "Mitigating Jailbreaks and Prompt Injections," 2023. Accessed: 2023-04-25.

Anthropic, "Prompt Chaining," `https://docs.anthropic.com/en/docs/chain-prompts` 2024. Accessed: 2024-06-03.

Arad, Ayala, David Hugh-Jones, and Stefan P. Penczynski, "Communication is informative about cooperation," Technical Report, School of Economics, University of East Anglia, Norwich, UK. 2024.

__ , Kevin P. Grubiak, and Stefan P. Penczynski, "Does communicating within a team influence individuals' reasoning and decisions?," *Experimental Economics*, 2022, pp. 1–21.

Austen-Smith, David and Jeffrey S Banks, "Information aggregation, rationality, and the Condorcet jury theorem," *American political science review*, 1996, *90* (1), 34–45.

Bach, Stephen H, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry et al., "Promptsource: An integrated development environment and repository for natural language prompts," *arXiv preprint arXiv:2202.01279*, 2022.

Baktash, Jawid Ahmad and Mursal Dawodi, "Gpt-4: A review on advancements and opportunities in natural language processing," *arXiv preprint arXiv:2305.03195*, 2023.

Bartolomeo, Giovanni Di, Martin Dufwenberg, and Stefano Papa, "Promises and partner-switch," *Journal of the Economic Science Association*, 2023, *9* (1), 77–89.

__ , __ , __ , and Francesco Passarelli, "Promises, expectations & causation," *Games and Economic Behavior*, 2019, *113*, 137–146.

Bassey, Bassey Asuquo, Isaac Ubi, Effa German Anagbogu, and Valentine Joseph Owan, "Permutation of the UTME Multiple-Choice Test Items on Performance in Use of English and Mathematics among Prospective Higher Education Students," *The Journal of Social Sciences Research, ISSN (e)*, 2022, pp. 2411–9458.

Beltagy, Iz, Matthew E Peters, and Arman Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

Bender, Emily M, "On achieving and evaluating language-independence in NLP," *Linguistic Issues in Language Technology*, 2011, *6.*

__ , Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?ï¿½ï¿½," in

"Proceedings of the 2021 ACM conference on fairness, accountability, and transparency" 2021, pp. 610–623.

Berger, Benedikt, Martin Adam, Alexander Rühr, and Alexander Benlian, "Watch me improve–algorithm aversion and demonstrating the ability to learn," *Business & Information Systems Engineering*, 2021, *63* (1), 55–68.

Bhat, Savita and Vasudeva Varma, "Large Language Models As Annotators: A Preliminary Evaluation For Annotating Low-Resource Language Content," in "Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems" 2023, pp. 100–107.

Bicchieri, Cristina and Alex Chavez, "Behaving as expected: Public information and fairness norms," *Journal of Behavioral Decision Making*, 2010, *23* (2), 161–178.

Bjerva, Johannes, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein, "What do language representations really represent?," *Computational Linguistics*, 2019, *45* (2), 381–389.

Black, Michael J., "I asked #Galactica about some things I know about and I'm troubled. In all cases, it was wrong or biased but sounded right and authoritative.," Twitter Nov 2022. Tweet.

Blanco, Mariana, Bogaçhan Çelen, and Andrew Schotter, "On Blame-Freeness and Reciprocity: An Experimental Study," Technical Report, Department of Economics, New York University December 2009.

Bogert, Eric, Aaron Schecter, and Richard T Watson, "Humans rely more on algorithms than social influence as a task becomes more difficult," *Scientific reports*, 2021, *11* (1), 8028.

Bonaccio, Silvia and Reeshad S. Dalal, "Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences," *Organizational Behavior and Human Decision Processes*, 2006, *101* (2), 127–151.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al., "Language models are few-shot learners," *Advances in neural* *information processing systems*, 2020, *33*, 1877–1901.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock, "What can machines learn and what does it mean for occupations and the economy?," in "AEA papers and proceedings," Vol. 108 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2018, pp. 43–47.

Buolamwini, Joy and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in "Conference on fairness, accountability and transparency" PMLR 2018, pp. 77–91.

Burchardi, Konrad B. and Stefan P. Penczynski, "Out of your mind: Eliciting individual reasoning in one shot games," *Games and Economic Behavior*, 2014, *84* (1), 39 – 57.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *The Quarterly Journal of Economics*, August 2004, *119* (3), 861–898.

Capra, C Mónica, "Understanding decision processes in guessing games: a protocol analysis approach," *Journal of the Economic Science Association*, 2019, *5* (1), 123–135.

Castelli, Ilaria, Davide Massaro, Cristina Bicchieri, Alex Chavez, and Antonella Marchetti, "Fairness norms and theory of mind in an ultimatum game: judgments, offers, and decisions in school-aged children," *PloS one*, 2014, *9* (8), e105024.

Castelo, Noah, *Blurring the line between human and machine: marketing artificial intelligence*, Columbia University, 2019.

__ , Maarten W Bos, and Donald R Lehmann, "Task-dependent algorithm aversion," *Journal of Marketing Research*, 2019, *56* (5), 809–825.

Çelen, Bogaçhan, Shachar Kariv, and Andrew Schotter, "The advice puzzle: An experimental study of social learning where words speak louder than actions," *Mimeo Center for Experimental Social Science*, 2003.

Chae, Youngjin and Thomas Davidson, "Large language models for text classification: From zero-shot learning to fine-tuning," *Open Science Foundation*,

2023.

Chander, Ajay, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino, "Working with beliefs: AI transparency in the enterprise.," in "IUI Workshops," Vol. 1 2018.

Chang, Ting-Yun and Robin Jia, "Data curation alone can stabilize in-context learning," *arXiv preprint arXiv:2212.10378*, 2022.

Charness, Gary and Martin Dufwenberg, "Promises and partnership," *Econometrica*, 2006, *74* (6), 1579–1601.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann et al., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, 2023, *24* (240), 1–113.

Chugunova, Marina and Daniela Sele, "We and it: An interdisciplinary review of the experimental evidence on human-machine interaction," *Center for law & economics working paper series*, 2020, *12.*

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma et al., "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

Church, Barbara A, Eduardo Mercado III, Matthew G Wisniewski, and Estella H Liu, "Temporal dynamics in auditory perceptual learning: impact of sequencing and incidental learning.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2013, *39* (1), 270.

Clavié, Benjamin, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell, "Large language models in the workplace: A case study on prompt engineering for job type classification," in "International Conference on Applications of Natural Language to Information Systems" Springer 2023, pp. 3–17.

Conneau, Alexis and Guillaume Lample, "Cross-lingual language model pretraining," *Advances in neural information processing systems*, 2019, *32.*

Cooper, David J and John H Kagel, "Are two heads better than one? Team versus individual play in signaling games," *American Economic Review*, 2005, *95* (3), 477–509.

__ , **Ian Krajbich, and Charles N Noussair**, "Choice-process data in experimental economics," 2019.

Cooper, Russell Stevens, "Information processing in the judge-adviser system of group decision making." PhD dissertation, University of Illinois at Urbana-Champaign 1991.

Costa-Gomes, Miguel A. and Vincent P. Crawford, "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, December 2006, *96* (5), 1737–1768.

Coughlan, Peter J, "In defense of unanimous jury verdicts: Mistrials, communication, and strategic voting," *American Political Science Review*, 2000, *94* (2), 375–393.

Crawford, Vincent P. and Nagore Iriberri, "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?," *Econometrica*, November 2007, *75* (6), 1721–1770.

__ , **Miguel A. Costa-Gomes, and Nagore Iriberri**, "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, September 2013, *51* (1), 5–62.

__ , **Uri Gneezy, and Yuval Rottenstreich**, "The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yields Large Coordination Failures," *American Economic Review*, July 2008, *98* (4), 1443–1458.

Cross, Rob, Stephen P Borgatti, and Andrew Parker, "Beyond answers: Dimensions of the advice network," *Social networks*, 2001, *23* (3), 215–235.

Dakhel, Arghavan Moradi, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C Desmarais, and Zhen Ming Jack Jiang, "Github copilot ai pair programmer: Asset or liability?," *Journal of Systems and Software*, 2023, *203*,

111734.

d'Alessandro, Brian, Cathy O'Neil, and Tom LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big data*, 2017, *5* (2), 120–134.

Danks, David and Alex John London, "Algorithmic Bias in Autonomous Systems.," in "Ijcai," Vol. 17 2017, pp. 4691–4697.

Daschner, Stefan and Robert Obermaier, "Algorithm aversion? On the influence of advice accuracy on trust in algorithmic advice," *Journal of Decision Systems*, 2022, *31* (sup1), 77–97.

Deelstra, Janna T, Maria CW Peeters, Wilmar B Schaufeli, Wolfgang Stroebe, Fred RH Zijlstra, and Lorenz P Van Doornen, "Receiving instrumental support at work: when help is not welcome.," *Journal of applied psychology*, 2003, *88* (2), 324.

den Assem, Martijn J Van, Dennie Van Dolder, and Richard H Thaler, "Split or steal? Cooperative behavior when the stakes are large," *Management Science*, 2012, *58* (1), 2–20.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Dietvorst, Berkeley J and Soaham Bharti, "People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error," *Psychological science*, 2020, *31* (10), 1302–1314.

__ , Joseph P Simmons, and Cade Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *Journal of experimental psychology: General*, 2015, *144* (1), 114.

Dijkstra, Jaap J, "User agreement with incorrect expert system advice," *Behaviour & Information Technology*, 1999, *18* (6), 399–411.

Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

Dongen, Kees Van and Peter-Paul Van Maanen, "A framework for explaining reliance on decision aids," *International Journal of Human-Computer Studies*, 2013, *71* (4), 410–424.

Dvijotham, Krishnamurthy, Jim Winkens, Melih Barsbey, Sumedh Ghaisas, Robert Stanforth, Nick Pawlowski, Patricia Strachan, Zahra Ahmed, Shekoofeh Azizi, Yoram Bachrach et al., "Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians," *Nature Medicine*, 2023, *29* (7), 1814–1820.

Dzindolet, Mary T, Linda G Pierce, Hall P Beck, and Lloyd A Dawe, "The perceived utility of human and automated aids in a visual detection task," *Human factors*, 2002, *44* (1), 79–94.

Ederer, Florian and Alexander Stremitzer, "Promises and expectations," *Games and Economic Behavior*, 2017, *106*, 161–178.

Edwards, Kari and Edward E. Smith, "A disconfirmation bias in the evaluation of arguments," *Journal of Personality and Social Psychology*, 1996, *71* (1), 5.

Efrat, Avia and Omer Levy, "The turking test: Can language models understand instructions?," *arXiv preprint arXiv:2010.11982*, 2020.

Eich, Theresa and Stefan P. Penczynski, "On the replicability of intra-team communication classification," Working Paper, University of Mannheim 2016.

El-Gamal, Mahmoud A and David M Grether, "Are people Bayesian? Uncovering behavioral strategies," *Journal of the American statistical Association*, 1995, *90* (432), 1137–1145.

Enarsson, Therese, Lena Enqvist, and Markus Naarttijärvi, "Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts," *Information & Communications Technology Law*, 2022, *31* (1), 123–153.

Eyster, Erik and Matthew Rabin, "Cursed Equilibrium," *Econometrica*, September 2005, *73* (5), 1623–1672.

**Faghel-Soubeyrand, Simon, Meike Ramon, Eva Bamps, Matteo Zoia, Jessica

Woodhams, Arjen Alink, Frédéric Gosselin, and Ian Charest, "Decoding real-world visual recognition abilities in the human brain," *Journal of Vision*, 2021, *21* (9), 2604–2604.

Fanta, Nicolas and Roman Horvath, "Artificial intelligence and central bank communication: the case of the ECB," *Applied Economics Letters*, 2024, pp. 1–8.

Feddersen, Timothy and Wolfgang Pesendorfer, "Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting," *American Political Science Review*, 1998, pp. 23–35.

Feng, Xinran and A Mike Burton, "Identity documents bias face matching," *Perception*, 2019, *48* (12), 1163–1174.

Fenneman, Achiel, Joern Sickmann, Thomas Pitz, and Alan G Sanfey, "Two distinct and separable processes underlie individual differences in algorithm adherence: Differences in predictions and differences in trust thresholds," *Plos one*, 2021, *16* (2), e0247084.

Festinger, Leon, "A Theory of Cognitive Dissonance (Evanston: Row Peterson, 1957).'Cognitive Dissonance'," *Scientific American*, 1962, *207.*

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, June 2007, *10* (2), 171–178.

Fu, Yao, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot, "Complexity-based prompting for multi-step reasoning," in "The Eleventh International Conference on Learning Representations" 2022.

Fysh, Matthew C and Markus Bindemann, "Human–computer interaction in face matching," *Cognitive science*, 2018, *42* (5), 1714–1732.

Gardner, P. H. and D. C. Berry, "The effect of different forms of advice on the control of a simulated complex system," *Applied Cognitive Psychology*, 1995, *9*, S55–S79.

Gefen, David and Detmar W Straub, "Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services," *Omega*, 2004, *32* (6), 407–424.

Gekhman, Zorik, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig, "Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?," *arXiv preprint arXiv:2405.05904*, 2024.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Gibbons, A. M., J. A. Sniezek, and R. S. Dalal, "Antecedents and consequences of unsolicited versus explicitly solicited advice," in D. Budescu, ed., *Symposium in Honor of Janet Sniezek*, Vancouver, BC 2003. Symposium presented at the annual meeting of the society for judgment and decision making.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proceedings of the National Academy of Sciences*, 2023, *120* (30), e2305016120.

Gino, Francesca, "Do we listen to advice just because we paid for it? The impact of advice cost on its use," *Organizational behavior and human decision processes*, 2008, *107* (2), 234–245.

Glasserman, Paul and Caden Lin, "Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis," *arXiv preprint arXiv:2309.17322*, 2023.

Godbole, Shantanu and Sunita Sarawagi, "Discriminative methods for multi-labeled classification," in "Pacific-Asia conference on knowledge discovery and data mining" Springer 2004, pp. 22–30.

Goldsmith, Daena J, "Soliciting advice: The role of sequential placement in mitigating face threat," *Communications Monographs*, 2000, *67* (1), 1–19.

__ **and Kristine Fitch**, "The normative context of advice as social support," *Human communication research*, 1997, *23* (4), 454–476.

Gönül, M Sinan, Dilek Önkal, and Michael Lawrence, "The effects of structural characteristics of explanations on use of a DSS," *Decision support systems*, 2006, *42* (3), 1481–1493.

Gonzalez, Manuel F, Weiwei Liu, Lei Shirase, David L Tomczak, Carmen E

Lobbe, Richard Justenhoven, and Nicholas R Martin, "Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes," *Computers in Human Behavior*, 2022, *130*, 107179.

Goodwin, Paul, M Sinan Gönül, and Dilek Önkal, "Antecedents and effects of trust in forecasting advice," *International Journal of Forecasting*, 2013, *29* (2), 354–366.

Goodyear, Kimberly, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, and Frank Krueger, "An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents," *Social neuroscience*, 2017, *12* (5), 570–581.

Green, Ben and Yiling Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments," in "Proceedings of the conference on fairness, accountability, and transparency" 2019, pp. 90–99.

__ and __ , "The principles and limits of algorithm-in-the-loop decision making," *Proceedings of the ACM on Human-Computer Interaction*, 2019, *3* (CSCW), 1–24.

Grimm, Veronika and Friederike Mengel, "Experiments on belief formation in networks," *Journal of the European Economic Association*, 2020, *18* (1), 49–82.

Grother, Patrick, Mei Ngan, and Kayee Hanaoka, *Face recognition vendor test (fvrt): Part 3, demographic effects*, National Institute of Standards and Technology Gaithersburg, MD, 2019.

Gu, Jiasheng, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin, "Robustness of learning from task instructions," *arXiv preprint arXiv:2212.03813*, 2022.

Guarnaschelli, Serena, Richard D McKelvey, and Thomas R Palfrey, "An experimental study of jury decision rules," *American Political Science Review*, 2000, pp. 407–423.

Hambleton, Ronald K and Ross E Traub, "The effects of item order on test performance and stress," *The Journal of Experimental Education*, 1974, *43* (1), 40–46.

Hansen, Anne Lundgaard and Sophia Kazinnik, "Can chatgpt decipher fedspeak," *Available at SSRN*, 2023.

Harber, Kent D, Joanne Kraenzle Schneider, Kelly M Everard, and Edwin B Fisher, "Directive support, nondirective support, and morale," *Journal of Social and Clinical Psychology*, 2005, *24* (5), 691–722.

Harvey, N. and I. Fischer, "Taking advice: accepting help, improving judgment, and sharing responsibility," *Organizational Behavior and Human Decision Processes*, 1997, *70*, 117–133.

Hauck, Kendall B, Maya A Mingo, and Robert L Williams, "A review of relationships between item sequence and performance on multiple-choice exams.," *Scholarship of Teaching and Learning in Psychology*, 2017, *3* (1), 58.

Haviv, Adi, Jonathan Berant, and Amir Globerson, "BERTese: Learning to speak to BERT," *arXiv preprint arXiv:2103.05327*, 2021.

He, Zeyu, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao'Kenneth' Huang, "If in a Crowdsourced Data Annotation Pipeline, a GPT-4," *arXiv preprint arXiv:2402.16795*, 2024.

Heath, C. and R. Gonzalez, "Interaction with others increases decision confidence but not decision quality: evidence against information collection views of interactive decision-making," *Organizational Behavior and Human Decision Processes*, 1995, *61*, 305–326.

Herm, Lukas-Valentin, Kai Heinrich, Jonas Wanner, and Christian Janiesch, "Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability," *International Journal of Information Management*, 2023, *69*, 102538.

Heseltine, Michael and Bernhard Clemm von Hohenberg, "Large language models as a substitute for human experts in annotating political text," *Research & Politics*, 2024, *11* (1), 20531680241236239.

Hicks, Michael Townsen, James Humphries, and Joe Slater, "ChatGPT is bullshit," *Ethics and Information Technology*, 2024, *26* (2), 38.

Hofmann, David A., Zhike Lei, and Adam M. Grant, "Seeking help in the shadow of doubt: The sensemaking processes underlying how nurses decide whom to ask for advice," *Journal of Applied Psychology*, 2009, *94* (5), 1261.

Horstmann, Aike C and Nicole C Krämer, "Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment," *Frontiers in psychology*, 2019, *10*, 939.

Houser, Daniel and Erte Xiao, "Classification of natural language messages using a coordination game," *Experimental Economics*, 2011, *14*, 1–14.

Howard, John J, Laura R Rabbitt, and Yevgeniy B Sirotin, "Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making," *Plos one*, 2020, *15* (8), e0237855.

Huang, Allen H, Hui Wang, and Yi Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, 2023, *40* (2), 806–841.

Huang, Jie and Kevin Chen-Chuan Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.

Hüning, Hendrik, Lydia Mechtenberg, and Stephanie Wang, "Detecting arguments and their positions in experimental communication data," *Available at SSRN 4052402*, 2022.

__ , __ , and __ , "Using Arguments to Persuade: Experimental Evidence," *Available at SSRN 4244989*, 2022.

Ismayilov, Huseyn and Jan Potters, "Why do promises affect trustworthiness, or do they?," *Experimental Economics*, 2016, *19*, 382–393.

IV, Robert L Logan, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel, "Cutting down on prompts and parameters: Simple few-shot learning with language models," *arXiv preprint arXiv:2106.13353*, 2021.

Jacobs, Maia, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos, "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection," *Translational psychiatry*, 2021, *11* (1), 108.

Jago, Arthur S, "Algorithms and authenticity," *Academy of Management Discoveries*, 2019, *5* (1), 38–56.

Jha, Manish, Jialin Qian, Michael Weber, and Baozhong Yang, "ChatGPT and corporate policies," Technical Report, National Bureau of Economic Research 2024.

Jiang, Zhengbao, Frank F Xu, Jun Araki, and Graham Neubig, "How can we know what language models know?," *Transactions of the Association for Computational Linguistics*, 2020, *8*, 423–438.

Jungermann, H., "Advice-giving and taking," in "Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS-32)" Institute of Electrical and Electronics Engineers, Inc. (IEEE) Maui, HI 1999. [CD-ROM].

__ **and K. Fischer**, "Using expertise and experience for giving and taking advice," in T. Betsch and S. Haberstroh, eds., *The routines of decision making*, Mahwah, NJ: Lawrence Erlbaum, 2005, pp. 157–173.

Kaddour, Jean, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

Kahneman, Daniel, Barbara L Fredrickson, Charles A Schreiber, and Donald A Redelmeier, "When more pain is preferred to less: Adding a better end," *Psychological science*, 1993, *4* (6), 401–405.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

Kawaguchi, Kohei, "When will workers follow an algorithm? A field experiment with a retail business," *Management Science*, 2021, *67* (3), 1670–1695.

Kayande, Ujwal, Arnaud De Bruyn, Gary L Lilien, Arvind Rangaswamy, and Gerrit H Van Bruggen, "How incorporating feedback mechanisms in a DSS affects DSS evaluations," *Information Systems Research*, 2009, *20* (4), 527–546.

Kim, Alex, Maximilian Muhn, and Valeri Nikolaev, "From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI," *arXiv preprint arXiv:2310.17721*,

2023.

Klare, Brendan F, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on information forensics and security*, 2012, *7* (6), 1789–1801.

Klimko, Ivan P, "Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance," *The Journal of experimental education*, 1984, *52* (4), 214–219.

Koch, Christian and Stefan P Penczynski, "The winner's curse: Conditional reasoning and belief formation," *Journal of Economic Theory*, 2018, *174*, 57–102.

Kocher, Martin, Matthias Sutter, and Florian Wakolbinger, "The impact of naive advice and observational learning in beauty-contest games," 2007.

Koestner, Richard, Isabelle Gingras, Rania Abutaa, Gaëtan F Losier, Lina DiDio, and Marylene Gagné, "To follow expert advice when making a decision: An examination of reactive versus reflective autonomy," *Journal of personality*, 1999, *67* (5), 851–872.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, 2022, *35*, 22199–22213.

Kong, Aobo, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou, "Better zero-shot reasoning with role-play prompting," *arXiv preprint arXiv:2308.07702*, 2023.

Koskinen, Kari, Antti Lyyra, Niina Mallat, and Virpi Tuunainen, "Trust and risky technologies: Aligning and coping with Tesla Autopilot," in "Annual Hawaii International Conference on System Sciences" Hawaii International Conference on System Sciences 2019.

Krueger, Joachim I, "Return of the ego–Self-referent information as a filter for social prediction: Comment on Karniol (2003).," 2003.

Kumar, Sawan and Partha Talukdar, "Reordering examples helps during priming-based few-shot learning," *arXiv preprint arXiv:2106.01751*, 2021.

__ **and** __ , “Reordering examples helps during priming-based few-shot learning,” *arXiv preprint arXiv:2106.01751*, 2021.

Kuzman, Taja, Igor Mozetic, and Nikola Ljubešic, “ChatGPT: beginning of an end of manual linguistic data annotation,” *Use case of automatic genre identification*, 2023.

Kuznia, Kirby, Swaroop Mishra, Mihir Parmar, and Chitta Baral, “Less is more: Summary of long instructions is better for program synthesis,” *arXiv preprint arXiv:2203.08597*, 2022.

Lai, Viet Dac, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen, “Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning,” *arXiv preprint arXiv:2304.05613*, 2023.

Lester, Brian, Rami Al-Rfou, and Noah Constant, “The power of scale for parameter-efficient prompt tuning,” *arXiv preprint arXiv:2104.08691*, 2021.

Li, Lingyao, Lizhou Fan, Shubham Atreja, and Libby Hemphill, ““HOT” ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media,” *ACM Transactions on the Web*, 2024, *18* (2), 1–36.

Li, Lixiang, Xiaohui Mu, Siying Li, and Haipeng Peng, “A review of face recognition technology,” *IEEE access*, 2020, *8*, 139110–139120.

Li, Shengwu, “Obviously strategy-proof mechanisms,” *American Economic Review*, 2017, *107* (11), 3257–3287.

Li, Xiang Lisa and Percy Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” *arXiv preprint arXiv:2101.00190*, 2021.

Li, Xiaonan and Xipeng Qiu, “Finding support examples for in-context learning,” *arXiv preprint arXiv:2302.13539*, 2023.

__ **and** __ , “Finding support examples for in-context learning,” *arXiv preprint arXiv:2302.13539*, 2023.

Li, Ziyang, Pei-Luen Patrick Rau, and Dinglong Huang, “Who should provide clothing recommendation services: Artificial Intelligence or Human Experts?," *Journal of Information Technology Research (JITR)*, 2020, *13* (3), 113–125.

Liel, Yotam and Lior Zalmanson, "What if an AI told you that 2+ 2 is 5? Conformity to algorithmic recommendations," 2020.

Litterscheidt, Rouven and David J Streich, "Financial education and digital asset management: What's in the black box?," *Journal of Behavioral and Experimental Economics*, 2020, *87*, 101573.

Liu, Estella H, Eduardo Mercado III, Barbara A Church, and Itzel Orduña, "The easy-to-hard effect in human (Homo sapiens) and rat (Rattus norvegicus) auditory identification.," *Journal of Comparative Psychology*, 2008, *122* (2), 132.

Liu, Jiachang, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen, "What Makes Good In-Context Examples for GPT-3?," *arXiv preprint arXiv:2101.06804*, 2021.

Liu, Peng and Zhizhong Li, "Task complexity: A review and conceptualization framework," *International Journal of Industrial Ergonomics*, 2012, *42* (6), 553–568.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

Logg, J.M., J.A. Minson, and D.A. Moore, "Algorithm appreciation: people prefer algorithmic to human judgment," *Organ. Behav. Hum. Decis. Process.*, 2019, *151*, 90–103.

Longoni, Chiara, Andrea Bonezzi, and Carey K Morewedge, "Resistance to medical artificial intelligence," *Journal of Consumer Research*, 2019, *46* (4), 629–650.

Lopez-Lira, Alejandro and Yuehua Tang, "Can chatgpt forecast stock price movements? return predictability and large language models," *arXiv preprint arXiv:2304.07619*, 2023.

Lou, Renze and Wenpeng Yin, "Toward zero-shot instruction following," in "Proceedings of the 18th Conference of the European Chapter of the Association for

Computational Linguistics: Student Research Workshop” 2024, pp. 50–60.

__ **, Kai Zhang, and Wenpeng Yin**, “Is prompt all you need? no. a comprehensive and broader view of instruction learning,” *arXiv preprint arXiv:2303.10475*, 2023.

Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp, “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity,” *arXiv preprint arXiv:2104.08786*, 2021.

Lundberg, Scott M, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim et al., “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” *Nature biomedical engineering*, 2018, *2* (10), 749–760.

Luo, Man, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao, “Dr. icl: Demonstration-retrieved in-context learning,” *arXiv preprint arXiv:2305.14128*, 2023.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu, “Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases,” *Marketing Science*, 2019, *38* (6), 937–947.

Madaan, Aman and Amir Yazdanbakhsh, “Text and patterns: For effective chain of thought, it takes two to tango,” *arXiv preprint arXiv:2209.07686*, 2022.

Madhavan, Poornima and Douglas A Wiegmann, “Effects of information source, pedigree, and reliability on operator interaction with decision support systems,” *Human factors*, 2007, *49* (5), 773–785.

Mahmud, H., A. N. Islam, S. I. Ahmed, and K. Smolander, “What influences algorithmic decision-making? A systematic literature review on algorithm aversion,” *Technological Forecasting and Social Change*, 2022, *175*, 121390.

Mantelero, Alessandro, “The social and ethical component in AI systems design and management,” in “Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI,” Springer, 2022, pp. 93–137.

Matter, Daniel, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer, “Close to

Human-Level Agreement: Tracing Journeys of Violent Speech in Incel Posts with GPT-4-Enhanced Annotations," *arXiv preprint arXiv:2401.02001*, 2024.

McCaffery, Jennifer M and A Mike Burton, "Passport checks: Interactions between matching faces and biographical details," *Applied Cognitive Psychology*, 2016, *30* (6), 925–933.

McKelvey, Richard D. and Thomas R. Palfrey, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, July 1995, *10* (1), 6–38.

Mcknight, D Harrison, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on management information systems (TMIS)*, 2011, *2* (2), 1–25.

Merrienboer, Jeroen JG Van and John Sweller, "Cognitive load theory and complex learning: Recent developments and future directions," *Educational psychology review*, 2005, *17*, 147–177.

Meshi, Dar, Guido Biele, Christoph W Korn, and Hauke R Heekeren, "How expert advice influences decision making," *PLoS One*, 2012, *7* (11), e49748.

Methnani, Leila, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou, "Let me take over: Variable autonomy for meaningful human control," *Frontiers in Artificial Intelligence*, 2021, *4*, 737072.

Mikolon, Sven, Anika Kolberg, Till Haumann, and Jan Wieseke, "The complex role of complexity: How service providers can mitigate negative effects of perceived service complexity when selling professional services," *Journal of Service Research*, 2015, *18* (4), 513–528.

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?," *arXiv preprint arXiv:2202.12837*, 2022.

Mishra, Swaroop, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," *arXiv preprint arXiv:2104.08773*, 2021.

__ , __ , __ , Yejin Choi, and Hannaneh Hajishirzi, “Reframing Instructional Prompts to GPTk’s Language,” *arXiv preprint arXiv:2109.07830*, 2021.

Moghaddam, Shima Rahimi and Christopher J Honey, “Boosting theory-of-mind performance in large language models via prompting,” *arXiv preprint arXiv:2304.11490*, 2023.

Mozannar, Hussein, Gagan Bansal, Adam Fourney, and Eric Horvitz, “When to show a suggestion? Integrating human feedback in AI-assisted programming,” in “Proceedings of the AAAI Conference on Artificial Intelligence,” Vol. 38 2024, pp. 10137–10144.

Mucha, H., S. Robert, R. Breitschwerdt, and M. Fellmann, “Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach.,” *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-6. 2021.*, 2021.

Müller, Dominik, Iñaki Soto-Rey, and Frank Kramer, “Towards a guideline for evaluation metrics in medical image segmentation,” *BMC Research Notes*, 2022, *15* (1), 210.

Murray, Alex, JEN Rhymer, and David G Sirmon, “Humans and technology: Forms of conjoined agency in organizations,” *Academy of Management Review*, 2021, *46* (3), 552–571.

Nagel, Rosemarie, “Unraveling in Guessing Games: An Experimental Study,” *American Economic Review*, December 1995, *85* (5), 1313–1326.

Newman, Dianna L, Deborah K Kundert, David S Lane Jr, and Kay Sather Bull, “Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty,” *Applied Measurement in education*, 1988, *1* (1), 89–97.

Ngan, Mei, Patrick J Grother, and Mei Ngan, *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*, US Department of Commerce, National Institute of Standards and Technology, 2015.

Nickerson, Raymond S., “Confirmation bias: A ubiquitous phenomenon in many guises,” *Review of General Psychology*, 1998, *2* (2), 175–220.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira et al., "Universal dependencies v1: A multilingual treebank collection," in "Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)" 2016, pp. 1659–1666.

Nyarko, Yaw, Andrew Schotter, and Barry Sopher, "On the informational content of advice: A theoretical and experimental study," *Economic Theory*, 2006, *29* (2), 433–452.

Obaid, Khaled and Kuntara Pukthuanthong, "Distortions in Financial Narratives: A ChatGPT Approach," 2024.

Önkal, Dilek, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock, "The relative influence of advice from human experts and statistical methods on forecast adjustments," *Journal of Behavioral Decision Making*, 2009, *22* (4), 390–409.

OpenAI, "Fine-tuned Classification," `https://github.com/openai/openai-cookbook/blob/main/examples/Fine-tuned_classification.ipynb` 2023. Accessed: 2023-04-20.

__ , "Prompt design," `https://platform.openai.com/docs/guides/completion/prompt-design` 2023.

__ , "Using Logprobs," `https://cookbook.openai.com/examples/using_logprobs` 2023. Accessed: 2023-04-21.

O'Toole, Alice J, P Jonathon Phillips, Xiaobo An, and Joseph Dunlop, "Demographic effects on estimates of automatic face recognition performance," *Image and Vision Computing*, 2012, *30* (3), 169–176.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray et al., "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, 2022, *35*, 27730–27744.

Paik, Jihyun Esther, “The contextual effects of advice solicitation on advice outcomes: The role of perceived face threats and psychological reactance,” *Communication Monographs*, 2020, *87* (1), 70–91.

Pálfi, Bence, Kavleen Arora, and Olga Kostopoulou, “Algorithm-based advice taking and clinical judgement: impact of advice distance and algorithm information,” *Cognitive research: principles and implications*, 2022, *7* (1), 70.

Pan, Jane, Tianyu Gao, Howard Chen, and Danqi Chen, “What In-Context Learning” Learns” In-Context: Disentangling Task Recognition and Task Learning,” *arXiv preprint arXiv:2305.09731*, 2023.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching, “Automated annotation with generative ai requires validation,” *arXiv preprint arXiv:2306.00176*, 2023.

Patel, Bhavik N, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi et al., “Human–machine partnership with artificial intelligence for chest radiograph diagnosis,” *NPJ digital medicine*, 2019, *2* (1), 111.

Patt, Anthony G., Hannah R. Bowles, and David W. Cash, “Mechanisms for enhancing the credibility of an adviser: Prepayment and aligned incentives,” *Journal of Behavioral Decision Making*, 2006, *19* (4), 347–359.

Penczynski, Stefan P, “Persuasion: An experimental study of team decision making,” *Journal of Economic Psychology*, 2016, *56*, 244–261.

Penczynski, Stefan P., “Strategic thinking: The influence of the game,” *Journal of Economic Behavior & Organization*, 2016, *128*, 72–84.

Penczynski, Stefan P, “The nature of social learning: Experimental evidence,” *European Economic Review*, 2017, *94*, 148–165.

__ , “Using machine learning for communication classification,” *Experimental Economics*, 2019, *22* (4), 1002–1029.

Pescetelli, Niccolò, Anna-Katharina Hauperich, and Nick Yeung, “Confidence, advice seeking and changes of mind in decision making,” *Cognition*, 2021, *215*, 104810.

Peskine, Youri, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso, "Definitions Matter: Guiding GPT for Multi-label Classification," in "Findings of the Association for Computational Linguistics: EMNLP 2023" 2023, pp. 4054–4063.

Peskoff, Denis, Adam Visokay, Sander Schulhoff, Benjamin Wachspress, Alan Blinder, and Brandon M Stewart, "GPT Deciphering Fedspeak: Quantifying Dissent Among Hawks and Doves," in "Findings of the Association for Computational Linguistics: EMNLP 2023" 2023, pp. 6529–6539.

Phillips, P Jonathon, Harry Wechsler, Jeffery Huang, and Patrick J Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, 1998, *16* (5), 295–306.

Plake, Barbara S, Charles J Ansorge, Claire S Parker, and Steven R Lowry, "Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance," *Journal of Educational Measurement*, 1982, pp. 49–57.

Puri, Ravsehaj Singh, Swaroop Mishra, Mihir Parmar, and Chitta Baral, "How many data samples is an additional instruction worth?," *arXiv preprint arXiv:2203.09161*, 2022.

Rabin, Matthew, "Inference by Believers in the Law of Small Numbers," *The Quarterly Journal of Economics*, 2002, *117* (3), 775–816.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language Models are Unsupervised Multitask Learners," *OpenAI Blog*, 2019.

Rapoport, Amnon, Darryl A. Seale, Ido Erev, and James A. Sundali, "Equilibrium Play in Large Group Market Entry Games," *Management Science*, 1998, *44* (1), 119–141.

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J Van Bavel, "GPT is an effective tool for multilingual psychological text analysis," 2023.

Razeghi, Yasaman, Robert L Logan IV, Matt Gardner, and Sameer Singh, "Impact of pretraining term frequencies on few-shot reasoning," *arXiv preprint arXiv:2202.07206*, 2022.

Reinhardt, Joann P, Kathrin Boerner, and Amy Horowitz, "Good to have but not to use: Differential impact of perceived and received support on well-being," *Journal of Social and Personal Relationships*, 2006, *23* (1), 117–129.

Reiss, Michael V, "Testing the reliability of chatgpt for text annotation and classification: A cautionary remark," *arXiv preprint arXiv:2304.11085*, 2023.

Renier, Laetitia A, Marianne Schmid Mast, and Anely Bekbergenova, "To err is human, not algorithmic–Robust reactions to erring algorithms," *Computers in Human Behavior*, 2021, *124*, 106879.

Reynolds, Laria and Kyle McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in "Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems" 2021, pp. 1–7.

Robalo, Pedro and Rei Sayag, "Paying is believing: The effect of costly information on Bayesian updating," *Journal of Economic Behavior & Organization*, 2018, *156*, 114–125.

Rudolph, Jürgen, Samson Tan, and Shannon Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?," *Journal of applied learning and teaching*, 2023, *6* (1), 342–363.

Rytting, Christopher Michael, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate, "Towards coding social science datasets with language models," *arXiv preprint arXiv:2306.02177*, 2023.

Salewski, Leonard, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata, "In-Context Impersonation Reveals Large Language Models' Strengths and Biases," *Advances in Neural Information Processing Systems*, 2024, *36.*

Savadori, L., L. M. Van Swol, and J. A. Sniezek, "Information sampling and confidence within groups and judge advisor systems," *Communication Research*, 2001, *28*, 737–771.

Savelka, Jaromir, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and

Huihui Xu, "Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise?," *arXiv preprint arXiv:2306.13906*, 2023.

Scao, Teven Le and Alexander M Rush, "How many data points is a prompt worth?," *arXiv preprint arXiv:2103.08493*, 2021.

Schick, Timo and Hinrich Schütze, "It's not just size that matters: Small language models are also few-shot learners," *arXiv preprint arXiv:2009.07118*, 2020.

Schotter, Andrew, "Decision making with naive advice," *American Economic Review*, 2003, *93* (2), 196–201.

__ , "Decision making with naive advice," *American Economic Review*, 2003, *93* (2), 196–201.

__ **and Barry Sopher**, "Social learning and coordination conventions in intergenerational games: An experimental study," *Journal of political economy*, 2003, *111* (3), 498–529.

__ **and** __ , "Trust and trustworthiness in games: An experimental study of intergenerational advice," *Experimental economics*, 2006, *9*, 123–145.

Schrah, G. E., R. S. Dalal, and J. A. Sniezek, "The adaptive decision-maker is not an island: integrative expert advice with information search," *Journal of Behavioral Decision-Making*, 2006, *19*, 43–60.

Schultze, Thomas, Tanja M Gerlach, and Jacob C Rittich, "Some people heed advice less than others: Agency (but not communion) predicts advice taking," *Journal of Behavioral Decision Making*, 2018, *31* (3), 430–445.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr, "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting," *arXiv preprint arXiv:2310.11324*, 2023.

Shaffer, V.A., C.A. Probst, E.C. Merkle, H.R. Arkes, and M.A. Medow, "Why Do Patients Derogate Physicians Who Use a Computer-Based Diagnostic Support System?," *Medical Decision Making*, 2012, *33*, 108–118.

Shen, Xinyue, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang, "" do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," *arXiv preprint arXiv:2308.03825*, 2023.

Shrestha, Yash Raj, Shiko M Ben-Menahem, and Georg Von Krogh, "Organizational decision-making structures in the age of artificial intelligence," *California management review*, 2019, *61* (4), 66–83.

Siddiqi, Naeem, *Credit risk scorecards: developing and implementing intelligent credit scoring*, Vol. 3, John Wiley & Sons, 2012.

Smales, Lee A, "Classification of RBA monetary policy announcements using ChatGPT," *Finance Research Letters*, 2023, *58*, 104514.

Sniezek, J. A. and L. M. Van Swol, "Trust, confidence, and expertise in a judge-advisor system," *Organizational Behavior and Human Decision Processes*, 2001, *84*, 288–307.

Sniezek, Janet A and Timothy Buckley, "Cueing and cognitive conflict in judge-advisor decision making," *Organizational behavior and human decision processes*, 1995, *62* (2), 159–174.

Sniezek, Janet A., Gunnar E. Schrah, and Reeshad S. Dalal, "Improving judgement with prepaid expert advice," *Journal of Behavioral Decision Making*, 2004, *17* (3), 173–190.

Soll, Jack B., Asa B. Palley, and Christina A. Rader, "The bad thing about good advice: Understanding when and how advice exacerbates overconfidence," *Management Science*, 2022, *68* (4), 2949–2969.

Sorower, Mohammad S, "A literature survey on algorithms for multi-label learning," *Oregon State University, Corvallis*, 2010, *18* (1), 25.

Stahl, Dale O. and Paul W. Wilson, "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, July 1995, *10* (1), 218–254.

Su, Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith et al., "Selective annotation makes language models better few-shot learners," *arXiv preprint* *arXiv:2209.01975*, 2022.

Sutherland, Steven C, Casper Harteveld, and Michael E Young, "Effects of the advisor and environment on requesting and complying with automated advice," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2016, *6* (4), 1–36.

Sweller, John, Jeroen JG Van Merrienboer, and Fred Paas, "Cognitive architecture and instructional design: 20 years later," *Educational psychology review*, 2019, *31*, 261–292.

Swim, Janet K, Kathryn J Aikin, Wayne S Hall, and Barbara A Hunter, "Sexism and racism: Old-fashioned and modern prejudices.," *Journal of personality and social psychology*, 1995, *68* (2), 199.

Swol, L. M. Van and J. A. Sniezek, "Factors affecting the acceptance of expert advice," *British Journal of Social Psychology*, 2005, *44*, 443–461.

Takagishi, Haruto, Shinya Kameshima, Joanna Schug, Michiko Koizumi, and Toshio Yamagishi, "Theory of mind enhances preference for fairness," *Journal of experimental child psychology*, 2010, *105* (1-2), 130–137.

Talamadupula, Kartik, "A Guide to LLM Inference Performance Monitoring," `https://symbl.ai/developers/blog/a-guide-to-llm-inference-performance-monitoring/` 2024. Accessed: 2024-05-29.

Tardif, Jessica, Xavier Morin Duchesne, Sarah Cohan, Jessica Royer, Caroline Blais, Daniel Fiset, Brad Duchaine, and Frédéric Gosselin, "Use of face information varies systematically from developmental prosopagnosics to super-recognizers," *Psychological science*, 2019, *30* (2), 300–308.

Taylor, Ross, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085*, 2022.

Törnberg, Petter, "Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning," *arXiv preprint arXiv:2304.06588*,

2023.

Towler, Alice, James D Dunn, Sergio Castro Martínez, Reuben Moreton, Fredrick Eklöf, Arnout Ruifrok, Richard I Kemp, and David White, "Diverse types of expertise in facial recognition," *Scientific reports*, 2023, *13* (1), 11396.

Tversky, Amos and Daniel Kahneman, "Belief in the law of small numbers.," *Psychological bulletin*, 1971, *76* (2), 105.

__ **and** __ , "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, 1974, *185* (4157), 1124–1131.

Tzioti, Stefanie, *Let me give you a piece of advice: empirical papers about advice taking in marketing* number EPS-2010-211-MKT 2010.

Uhlmann, Eric Luis, Victoria L Brescoll, and Edouard Machery, "The motives underlying stereotype-based discrimination against members of stigmatized groups," *Social Justice Research*, 2010, *23* (1), 1–16.

van Elten, Jonas and Stefan P Penczynski, "Coordination games with asymmetric payoffs: An experimental study with intra-group communication," *Journal of Economic Behavior & Organization*, 2020, *169*, 158–188.

Vanberg, Christoph, "Why do people keep their promises? An experimental test of two explanations 1," *Econometrica*, 2008, *76* (6), 1467–1480.

von Walter, Benjamin, Dietmar Kremmel, and Bruno Jäger, "The impact of lay beliefs about AI on adoption of algorithmic advice," *Marketing Letters*, 2022, *33* (1), 143–155.

Vossen, Rainer and Gerrit J Dimmendaal, *The Oxford handbook of African languages*, Oxford University Press, USA, 2020.

Wang, Dayong, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

Wang, Lei, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim, "Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models," *arXiv preprint arXiv:2305.04091*, 2023.

Wang, Xiuxin and Xiufang Du, "Why does advice discounting occur? The combined roles of confidence and trust," *Frontiers in Psychology*, 2018, *9*, 2381.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

Webson, Albert and Ellie Pavlick, "Do prompt-based models really understand the meaning of their prompts?," *arXiv preprint arXiv:2109.01247*, 2021.

Wei, Jason, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

__ , Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, 2022, *35*, 24824–24837.

__ , Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

Wisniewski, Matthew G, Milen L Radell, Barbara A Church, and Eduardo Mercado III, "Benefits of fading in perceptual learning are driven by more than dimensional attention," *Plos one*, 2017, *12* (7), e0180959.

Workman, M., "Expert decision support system use, disuse, and misuse: a study using the theory of planned behavior," *Computers in Human Behavior*, 2005, *21*, 211–231.

Wu, Tongshuang, Michael Terry, and Carrie Jun Cai, "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts," in "Proceedings of the 2022 CHI conference on human factors in computing systems" 2022, pp. 1–22.

Wu, Xingjiao, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He, "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, 2022, *135*, 364–381.

Wu, Zhaofeng, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim, "Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks," *arXiv preprint arXiv:2307.02477*, 2023.

Wulf, Gabriele and Charles H Shea, "Principles derived from the study of simple skills do not generalize to complex skill learning," *Psychonomic bulletin & review*, 2002, *9* (2), 185–211.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

Yang, Jie, Judith Redi, Gianluca Demartini, and Alessandro Bozzon, "Modeling task complexity in crowdsourcing," in "Proceedings of the AAAI Conference on Human Computation and Crowdsourcing," Vol. 4 2016, pp. 249–258.

Yaniv, I. and E. Kleinberger, "Advice taking in decision making: egocentric discounting and reputation formation," *Organizational Behavior and Human Decision Processes*, 2000, *83*, 260–281.

Yaniv, Ilan, "Receiving other people's advice: Influence and benefit," *Organizational behavior and human decision processes*, 2004, *93* (1), 1–13.

Yarger, Lynette, Fay Cobb Payton, and Bikalpa Neupane, "Algorithmic equity in the hiring of underrepresented IT job candidates," *Online information review*, 2020, *44* (2), 383–395.

Ye, Junjie, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen et al., "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg, "Making sense of recommendations," *Journal of Behavioral Decision Making*, 2019, *32* (4), 403–414.

Yoo, Kang Min, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang goo Lee, and Taeuk Kim, "Ground-truth labels matter: A deeper look into input-label demonstrations," *arXiv preprint arXiv:2205.12685*, 2022.

You, Sangseok, Cathy Liu Yang, and Xitong Li, "Algorithmic versus human advice: does presenting prediction performance matter for algorithm appreciation?," *Journal of Management Information Systems*, 2022, *39* (2), 336–365.

Yu, Kun, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen, "Do I trust my machine teammate? an investigation from perception to decision," in "Proceedings of the 24th International Conference on Intelligent User Interfaces" 2019, pp. 460–468.

Yuan, J, P Bao, Z Chen, M Yuan, J Zhao, J Pan, Y Xie, Y Cao, Y Wang, Z Wang et al., "Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers," *The Innovation*, 2023.

Yuan, Weizhe, Graham Neubig, and Pengfei Liu, "Bartscore: Evaluating generated text as text generation," *Advances in Neural Information Processing Systems*, 2021, *34*, 27263–27277.

Zaleskiewicz, Tomasz, Agata Gasiorowska, Katarzyna Stasiuk, Renata Maksymiuk, and Yoram Bar-Tal, "Lay evaluation of financial experts: The action advice effect and confirmation bias," *Frontiers in psychology*, 2016, *7*, 1476.

__ **and** __ , "Tell me what I wanted to hear: Confirmation effect in lay evaluations of financial expert authority," *Applied Psychology*, 2018, *67* (4), 686–722.

Zanzotto, Fabio Massimo, "Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, 2019, *64*, 243–252.

Zhang, Bowen, Daijun Ding, and Liwen Jing, "How would stance detection techniques evolve after the launch of chatgpt?," *arXiv preprint arXiv:2212.14548*, 2022.

Zhang, Lixuan, Iryna Pentina, and Yuhong Fan, "Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services," *Journal of Services Marketing*, 2021, *35* (5), 634–646.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin et al., "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong et al., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

Zheng, Huaixiu Steven, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou, "Take a step back: Evoking reasoning via abstraction in large language models," *arXiv preprint arXiv:2310.06117*, 2023.

Zhong, Qihuang, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," *arXiv preprint arXiv:2302.10198*, 2023.

Zhu, Qingxiaoyang and Hao-Chuan Wang, "Leveraging Large Language Model as Support for Human Problem Solving: An Exploration of Its Appropriation and Impact," in "Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing" 2023, pp. 333–337.

Zhu, Yiming, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson, "Can chatgpt reproduce human-generated labels? a study of social computing tasks," *arXiv preprint arXiv:2304.10145*, 2023.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang, "Can large language models transform computational social science?," *Computational Linguistics*, 2024, pp. 1–55.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this dissertation I used GPT-4 in order to improve readability and language of the text. After using this tool, I reviewed and edited the content as needed.

This page is intentionally left blank

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig angefertigt und die benutzten Hilfsmittel vollständig und deutlich angegeben habe.

Datum:

Can Çelebi

This page is intentionally left blank

Curriculum Vitae

2014-2024	Ph.D. Studies in Economics, University of Mannheim
2012-2013	M.Sc. in Economics, University of Nottingham
2005-2011	B.Sc. in Economics, University of Istanbul