CPAs vs. AI Tax Advisors: Evaluating Reliance Across Advisory Contexts

Lucas A. Swider
University of Oklahoma

Running Head: CPAs vs. AI Tax Advisors

Acknowledgement: I am deeply grateful for the guidance and support of my dissertation chair, Andy Cuccia, and committee members Richard Price, Bradley Blaylock, Kevan Jensen, and Matt Jensen. I also thank Brant Christensen, Spencer Young, workshop participants at Auburn University, the University of Oklahoma, the University of New Mexico, the 2023 AAA/Deloitte Foundation/J. Michael Cook Doctoral Consortium, and two anonymous reviewers and conference participants at the 2024 Behavioral Tax Symposium for their helpful comments and suggestions. I gratefully acknowledge financial support from the University of Oklahoma and the Bullard Dissertation Completion Fellowship.

Lucas A. Swider, University of Oklahoma, Price College of Business, John T. Steed School of Accounting, Norman, OK, USA.

JEL Classifications: H26; M41; C91; D81; L86.

Keywords: Generative AI; Advice; Reliance; Trust; Tax Compliance; Reporting; CPA.

CPAs vs. AI Tax Advisors: Evaluating Reliance Across Advisory Contexts

ABSTRACT

Advancements in generative artificial intelligence (AI) have led to the public deployment of AI chatbots that provide tax planning and compliance advice. This study investigates the extent to which taxpayers who prepare their own returns rely on AI advisors and compares this reliance to that placed on certified public accountants (CPAs). The study uses a randomized experiment that manipulates advisor type and aggression to shed light on how these factors influence taxpayers during the reporting process. Trust and advocacy are examined as mechanisms driving differential reliance on advisors. Results provide evidence that self-preparers are influenced by AI advisors and exhibit a preference for tax-minimizing recommendations. However, taxpayers rely less on AI compared to CPAs, an effect driven by perceptions of the advisor's tax expertise and commitment to minimizing tax burden. The study provides insights for professional tax service providers and regulators considering controls over the rapidly evolving technology.

I. INTRODUCTION

Recent advances in generative artificial intelligence (AI) have led to deployment of chatbot interfaces that provide tax planning and compliance advice to individual taxpayers. Primarily powered by large language models (LLMs), these chatbots are designed to provide advice through conversational interactions (OpenAI 2024). The market for this new class of tax advisor is large – more than 65 million electronically filed United States (US) federal income tax returns were self-prepared in 2023 (IRS 2023). The consequences of income tax reporting decisions made in this new tax advisory context are economically significant. For instance, individual income taxes are the largest single source of federal revenue in the US, exceeding $2 trillion in FY 2023. This study examines the extent to which individual taxpayers are willing to rely on AI advisors. This reliance is compared to that placed on advice provided by certified public accountants (CPAs), a traditional benchmark for tax advice quality. The study focuses on perceived advisor characteristics, including benevolence, expertise, and a commitment to minimizing tax burden. The developments in AI-driven tax advisory services also heighten the importance of understanding taxpayers' reporting preferences, particularly among self-

preparers.[1] In addition to comparing the reliance placed on CPAs and AI advisors, the study examines whether self-preparers display a preference for aggressive or conservative income tax positions.[2]

The willingness of self-preparers to rely on tax advice from LLMs is unknown. These AI tools possess a high degree of tax knowledge and the ability to provide personalized advice. Most self-preparers utilize commercial tax software to prepare their returns, which offers a low-cost and convenient filing experience that generally results in accurately prepared returns for those with simple tax situations. Tax compliance research often assumes that self-preparers understand the legal merits of their reporting decisions.[3] However, software provides limited guidance,[4] and self-preparers with complex tax scenarios may not understand the legal merits of the alternative tax positions available to them. LLMs can help resolve this uncertainty by offering precise advice tailored to specific tax scenarios. Beyond general-use LLMs, developers and business enterprises can tailor specialized applications for tax advisory purposes. LLMs are promising advisors because of their potential training on US tax law, accessibility, and low cost. They simplify the research process by distilling vast amounts of information to provide answers

[1] A "self-preparer" is an individual that prepares their own income tax return, either by hand or with the assistance of tax software.

[2] An aggressive tax position is defined as one that reduces a taxpayer's current tax liability, relative to an alternative position, when there is ambiguity in the tax law. A conservative tax position is defined as one that increases a taxpayer's current liability, relative to an alternative position.

[3] For example, many experimental tax compliance studies place participants in a hypothetical tax reporting scenario where the participant either earns or is endowed with income that has not been reported to the tax authority by a third party, such as cash tips or unreported contractor income (e.g., Brink and Lee 2015; Hunt and Iyer 2018; Lamothe and Bobek 2020; Brink and Hansen 2020). Participants must then choose how much of this income to report. In these scenarios, participants are either explicitly told, or it is assumed that they understand that this income is taxable under the law.

[4] Tax preparers in the US are subject to various levels of professional standards and legal regulations designed to ensure that they act competently, ethically, and in the best interest of taxpayers and the tax system. Commercial tax software providers are excluded from the preparer characterization due to Regs. Sec. 301.7701-15(f)(1), which specifically excludes persons providing mere "mechanical assistance" in tax preparation.

that often reference authoritative sources such as the Internal Revenue Code, Treasury Department regulations, or court cases. Prior research provides evidence that taxpayers who engage professional preparers tend to be conservative, suggesting that tax *preparers*, rather than taxpayers, may drive aggressive reporting (Hite and McGill 1992; Tan 1999; Sakurai and Braithwaite 2001; Stephenson 2007). This preference for conservatism may not generalize to modern self-preparers, who have greater agency in the tax research and reporting process. Self-preparers conducting their own tax research are subject to personal cognitive biases, such as confirmation bias. Confirmation bias refers to the tendency to seek or interpret evidence in ways that confirm existing beliefs or expectations. This bias may lead individuals to favor aggressive tax positions if they believe them to be justifiable under the tax law.

The research questions were examined with an experiment. The study utilizes a randomized 2x2+1 between-subjects design that simulates advice-taking from a tax advisor while self-preparing an income tax return. The advisor type (CPA or AI) and nature of advice (aggressive or conservative) are manipulated, and a baseline control group that makes reporting decisions without tax advice is included. The study employs a representative sample of US taxpayers. The hypotheses are primarily informed by the psychology literature on confirmation bias, social psychology literature on interpersonal trust, and prior tax professional judgment and decision-making literature on tax professionals' client-advocacy attitudes. Consistent with predictions, the results support several key findings. First, self-preparers display a strong preference for tax-minimizing positions. This suggests that the conservatism preference documented among taxpayers engaged with professional preparers may not generalize to modern self-preparers. Second, taxpayers are significantly influenced by AI-provided advice. Participants who were presented with aggressive (conservative) advice from AI reported more aggressively

(conservatively) than those who did not receive any tax advice, and were more likely to rely on AI-provided advice when it was aggressive. This pattern of reporting is similar to that observed in participants who were presented with advice from a CPA, and is noteworthy for several reasons. LLMs are largely unregulated, often unpredictable, and rapidly proliferating in deployment and use. In addition, self-preparers' disposition toward aggressive positions is more likely to manifest in their reporting decisions relative to those engaged with a professional preparer. Third, taxpayers express higher levels of reliance on advice provided by CPAs relative to advice provided by AI. This differential reliance is driven by taxpayers' perceptions of CPAs' tax expertise and advocacy, the perception that they are dedicated to minimizing tax burden. The insights should interest regulators and policymakers considering controls over the rapidly evolving technology, tax service providers, and the research community.

The study contributes to several streams of literature. It contributes to tax compliance and human-AI advice-taking research by examining individual taxpayers' reliance on AI tax advisors. As the deployment of AI tax advisors continues, decisions based upon their advice have the potential to shape economic outcomes for governments worldwide. The study also adds to the tax professional judgment literature by providing evidence that taxpayers perceive CPAs to be more benevolent than AI, and that perceptions of their expertise and advocacy contribute to taxpayers' reliance on them. Last, it revisits the debate on taxpayers' preference for aggressive or conservative tax positions under ambiguity, providing evidence that self-preparers, who have greater autonomy than those using professional preparers, tend to prioritize tax minimization.

II. BACKGROUND

AI Tax Advisors

Generative AI refers to models and algorithms designed to generate new content based on patterns in their training data. LLMs, a subset of generative AI, are primarily trained on text data from books, websites, and other textual databases. Their primary function is to generate contextually appropriate, human-like text in response to given prompts. In the legal domain, LLMs are increasingly being used to automate legal tasks such as legal judgment prediction, legal document analysis, and legal document writing (Sun 2023). Generative AI is poised to shift how tax research is conducted (Alarie, Condon, Massey, and Yan 2023). Nay et al. (2023) examine how one of the largest state-of-the-art LLMs, OpenAI's ChatGPT, functions as a tax attorney. Through a series of experiments, they tested the LLM's ability to understand tax law by assessing the accuracy of its responses to thousands of complex tax law inquiries. They found a vast performance increase with each new model release. While their most recently tested version, GPT-4, did not perform at the level of an expert tax lawyer, the authors suggest that the dawn of superhuman tax AI is on the horizon.

Despite their promise to improve the efficiency and effectiveness of tax research, LLMs essentially function as black boxes. Their behavior can be unpredictable (D'Amour et al. 2020; Kenton et al. 2021), and even AI experts have difficulty understanding their reasoning and decision-making processes (Bowman 2023). Techniques that seem to provide insight into their behavior are often later found to be extremely misleading (Feng et al. 2018; Bolukbasi et al. 2021; Bowman 2023; Wang et al. 2023). In high-stakes domains with significant economic and legal implications, such as tax advisory, the unpredictability of LLMs' reasoning (Uesato et al. 2022; Perrigo 2023; Roose 2023) warrants a cautious approach to their deployment.

Nevertheless, public deployment of LLMs that function as advisors is underway. In addition to flagship LLMs deployed by companies like OpenAI, open source alternatives are increasing in number. Although GPT-4o, one of the state-of-the-art models at the time of writing, is not strictly open-source, various tools enable independent developers and business enterprises to build their own AI chatbots on OpenAI's foundational model. Although these chatbots may utilize the same underlying model, developers have the flexibility to make modifications to LLM training material and response processes. Developers vary in how they advertise their chatbot's capabilities and the level of transparency that is provided to the user regarding the model's training or how it functions.[5] They may indicate that they have been trained on US tax law, but details are often unspecified. The growing number of AI advisors coming to market, the variation in response across LLM platforms, and the potential for them to provide misleading advice raises concerns for regulators, developers, taxpayers, and researchers. This study is a first step toward gaining an understanding of how taxpayers respond to advice in this new tax advisory context.[6]

Tax Research Environment

Unlike tax professionals, most individual taxpayers are not experts on tax law. They are likely unfamiliar with the hierarchy of authoritative sources. As a result, their universe of evidential sources for a tax position is vastly larger than that of the tax professional, who maintains a narrower authoritative focus. AI advisors have the potential to drastically increase the efficiency of individuals' tax research process. They are able to parse extensive data and

[5] As an example, at the time of writing, ChatGPT often caveats responses when providing tax advice and generally advises users to consult with a tax professional. This is a development choice. Other LLM-powered chatbots do not caveat their responses and respond confidently.

[6] Anecdotally, when posing the same tax question to different LLMs, substantial response variance exists in the level of tax advice detail, length, reference to authoritative sources, and the confidence with which the advice is presented. Examining how this variance in response across AI platforms influence taxpayer judgment and decision-making is beyond the scope of this study. These elements are held constant to isolate the effect of the advisor source itself.

deliver tailored answers in response to taxpayer prompts. However, concerns have been raised by various stakeholders about LLM bias, their training data, and their potential to provide inaccurate advice. Relying on tax advice from an LLM trained on nonauthoritative sources increases the risk of taking a tax position that is not supported by the tax law.

III. THEORY AND HYPOTHESES

Taxpayers' Reporting Objectives

To predict how individual taxpayers may respond to tax advice, it is necessary to consider one's objectives when researching a tax position. Utility theory suggests that individuals act rationally to maximize their well-being (von Neumann and Morgenstern 1947; Allingham and Sandmo 1972). In the tax context, the general assumption is that paying less tax leads to a higher level of utility because an individual is able to retain more of their income for consumption, savings, or investment. Under this assumption, individuals prefer tax positions that minimize their tax liability to those that increase it.

However, prior research also suggests that individuals generally wish to comply with the tax law and have an aversion to high risk tax positions (Christensen 1992; Hite, Stock, and Cloyd 1992; Hite and McGill 1992; Tan 1999). The desire to file an accurate tax return is one of the primary motivations to engage a paid tax preparer (Yankelovich, Skelly, and White 1984; Collins, Milliron, and Toy 1990; Hite et al. 1992; Stephenson 2010). More recently, Rosenthal, Brown, Higgs, and Rupert (2023) surveyed a representative sample of US taxpayers to better understand the choice to use tax software or engage a preparer. They find that regardless of whether taxpayers use software or hire a professional, preparing the most accurate return possible ranks as the top priority for both groups. A factor that contributes to taxpayers' desire for accuracy is their strong aversion to the potential for an audit and the related consequences. One possible explanation for

the conservatism preference documented in prior research is that taxpayers in those studies held inflated perceptions of audit probability, which greatly reduce the subjective expected utility of aggressive tax positions. In this study, audit probability is held constant at a realistic level to provide insight on how the relatively more informed self-preparer makes judgments and decisions during the tax reporting process (less than 1 percent).[7] An aggression preference does not necessarily imply an inclination towards evasion, but perhaps a more accurate expected utility mental model. For example, even if a particular tax position is justifiable under the law, but is perceived to carry high audit risk, it is likely that the subjective expected utility of the position will be lessened due to the stress, time, and monetary costs associated with an audit.

Confirmation Bias

Confirmation bias refers to seeking or interpreting evidence in ways that are partial to existing beliefs, expectations, or a held hypothesis (Nickerson 1998). Confirmation bias happens if, in a systematic fashion, hypothesis-confirming information receives more weight, is evaluated less critically, or is better remembered than disconfirming evidence (Oswald and Grosjean 2004). Prior studies document confirmation bias in tax professionals' research process. When conducting tax research, tax professionals generally aim to confirm the hypothesis that the client's preferred tax position is supported by applicable tax law, regulations, or other recognized guidance. This confirmation can occur through several mechanisms, which may include placing more weight on judicial evidence that supports aggressive positions (Johnson 1993), bias in the judicial precedent search itself (Cloyd and Spilker 1999), or advantageous threshold interpretation (Cuccia, Hackenbrack, and Nelson 1995).

[7] A simple internet search or an inquiry posed to an LLM provides taxpayers with this information. Further, commercial tax software providers advertise this statistic to their users. For example, one of the largest commercial providers advertises that the IRS audits less than 1 percent of all taxpayers (TurboTax 2024).

In summary, utility theory and prior research suggest that the average individual taxpayer's objective is to adopt tax positions that both minimize their tax burden and are supported by the tax law. A self-preparer conducting tax research will strive to confirm their hypothesis that an aggressive position is supported by the tax law. This leads to a confirmatory evidence evaluation process by which self-preparers will place more weight on evidence confirming this hypothesis relative to disconfirming evidence. In an advice-taking context, an advisor's directional recommendation implies that the advisor themselves placed more weight on the evidence supporting the recommended position. Thus, the recommendation itself functions as a form of evidence.

In the absence of a self-preparer's confirmatory process favoring aggressive (or conservative) reporting, no observable difference should exist in their reliance on tax advice when it is aggressive compared to conservative. However, it is predicted that confirmation bias leads self-preparers to place greater weight on preference-confirming evidence, leading to lower reliance on conservative tax advice compared to aggressive tax advice. This effect is not predicted to be conditional on the source of the advice, as self-preparers in both contexts have the agency to make their own reporting decisions, which may be influenced by confirmation bias. This leads to the following hypothesis.

> *H1: Taxpayers are more likely to rely on tax advice when it is aggressive than when it is conservative.*

Trust and Reliance on Advice

In advice-taking scenarios, the degree of trust placed in the advisor plays an important role in determining whether advice is accepted, and the subsequent confidence in a decision made based on that advice (Sniezek and Van Swol 2001; Van Swol and Sniezek 2005). The extent to which a

taxpayer will rely on an advisor will in part depend on how much the taxpayer trusts the advisor. McAllister (1995) defines trust as "the extent to which a person is confident in, and willing to act on the basis of the words, actions, and decisions of another." Lewis and Wiegart (1985) suggest that trust has both affective and cognitive foundations. McAllister (1995) provides a framework that conceptualizes trust as consisting of affect- and cognition-based dimensions and has been widely used across various domains to study interpersonal trust. This framework extends beyond human-to-human interactions. Glikson and Woolley (2020) conduct a comprehensive review of human trust in AI systems, covering studies of robotic, virtual, and embedded AI. They categorize trust antecedents as either cognition- or affect-based, demonstrating the relevance of dual-conceptualization in understanding the relative trust placed in CPAs and AI advisors. This study focuses on benevolence-based and expertise-based trust. The following sections discuss how these perceptions may drive differential reliance between CPAs and AI advisors.

Benevolence-Based Trust

Benevolence is the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive (Mayer, Davis, and Schoorman 1995). Benevolence is closely aligned with affect-based trust, which is grounded in the trustor's attributions concerning the trustee's motives, concern, and interpersonal care (McAllister 1995). Prior research suggests that for decisions with high stakes and substantial outcome uncertainty, a person's perception of the advisor's benevolence influences willingness to rely on advice. White (2005) investigates how an investor's perceptions of their financial advisor's benevolence and expertise impact their decision to accept investment advice. She finds that advisees favor benevolent advisors for emotionally difficult decisions. An emotional decision is defined as one where the decision maker anticipates potential negative or stressful outcomes (e.g., audit, penalties, interest). Research on

social support suggests that advisors serve an emotional support role and provide a psychological buffer against actual or anticipated stressful outcomes (Cohen and Wills 1985; Helgeson 2003). When making a stressful decision, advisees who receive advice from a benevolent advisor are likely to be more confident in the positive outcomes of accepting advice because of the psychological buffer against adverse outcomes.

LLM-based chatbots are a form of virtual AI, a representation in which AI has no physical presence, but is nonetheless a distinguished entity (Ben Mimoun, Poncin, and Garnier 2012; Glikson and Woolley 2020). Prior research finds that tangibility and immediacy behaviors are positively associated with emotional trust in virtual AI. Chattaraman, Kwon, Gilbert, and Li (2014) explore trust and reliance on virtual agents in a retail shopping context. Their study manipulates the presence of an anthropomorphized visual depiction of the shopping agent and finds that users who interacted with the anthropomorphized agent held higher levels of benevolence-based trust. Similarly, Qiu and Benbasat (2009) found that users placed more emotional trust in virtual agents when they perceived a high level of social presence resulting from a visual depiction of the agent. On one hand, the interactive nature of AI tax advisors may lead to higher levels of emotional trust (Dabholkar and Sheng 2012). On the other, currently deployed AI tax advisors generally do not possess anthropomorphized characteristics, and do not exhibit physical immediacy behaviors such as hand gestures or facial expressions, likely leading to lower levels of emotional trust relative to a CPA (Matsui and Yamada 2019). Further, AI advisors are generally not thought to have feelings or emotions. As such, taxpayers are less likely to view AI advisors as benevolent actors who act in their best interest when compared to CPAs, whom they may perceive to consider a broader set of the advisee's desires, concerns, and overall well-being. This leads to the following hypotheses.

H2a: Taxpayers place higher levels of benevolence-based trust in CPAs compared to AI advisors.

H2b: The positive association between CPAs and benevolence perceptions leads to greater reliance on their tax advice.

Expertise-Based Trust

Expertise-based trust aligns with cognition-based trust, which reflects a trustor's beliefs about the trustee's reliability and dependability (McAllister 1995). Recent studies offer mixed evidence regarding the objective tax expertise of AI advisors. Alarie et al. (2023) and Nay et al. (2023) suggest that popular LLMs possess a relatively high level of expertise, although they do not yet match the performance of expert humans. In contrast, Antinozzi and Cooper (2024) evaluate the tax expertise of Chat-GPT during the 2023 and 2024 filing seasons and conclude that it is generally not a reliable source of tax guidance for uninformed users.

Regardless of objective expertise, user perception plays an important role in forming expertise-based trust. Prior research has documented an "algorithm aversion" phenomenon, where humans discount advice from algorithms and rely more readily on human judgment, even when the algorithmic advice is shown to be more accurate than human predictions (Dietvorst, Simmons, and Massey 2015). This aversion is exacerbated when uncertainty increases (Dietvorst and Bharti 2020). However, prior research also suggests that algorithm aversion is negatively associated with perceptions of expertise. Bigman and Gray (2018) find that increasing the perceived expertise of an autonomous machine that makes medical surgery decisions reduces algorithm aversion. Likewise, Zhang, Pentina, and Fan (2021) find that perceptions of expertise drive consumers' preference for human financial advisors over robo-advisors.

Given the technology's novelty, it is unclear ex ante how taxpayers perceive the tax expertise of AI advisors. LLMs may be trained on a broad range of tax law, which may be

highlighted in their marketing. When users perceive the AI's training as robust and observe its ability to convey complex tax information in a conversational manner, they may perceive the AI's tax expertise to be on par with that of a CPA. However, algorithm aversion could counteract perceptions of the AI's expertise, leading to lower reliance on AI advisors compared to CPAs. Given these conflicting possibilities and the novelty of the technology, no specific directional prediction is made with respect to differential expertise perceptions between CPAs compared to AI advisors. In any case, perceptions of expertise should lead to higher levels of reliance. This leads to the following formal hypotheses.

H3a (null): There is no difference between the level of expertise-based trust taxpayers place in CPAs and AI advisors.

H3b: The positive association between an advisor and expertise perceptions leads to greater reliance on their tax advice.

Taxpayer Perceptions of Advisor Advocacy

In addition to the trust placed in an advisor that derives from perceptions of general benevolence or tax expertise, taxpayers may consider whether their advisor is inherently dedicated to minimizing their tax burden. Building on prior tax professional judgment and decision-making literature, this concept is described as perceived advocacy – the perception that a tax advisor's loyalty lies with the taxpayer, particularly in terms of tax minimization. CPA professional tax standards prescribe that CPAs have the right to be an advocate for taxpayers when advising on a tax position (AICPA 2024). Prior research generally interprets advocacy as an attitude of loyalty to client interests that results in a pursuit of tax minimization (Mason and Levy 2001; Kadous and Magro 2001; Davis and Mason 2003; Barrick, Cloyd, and Spilker 2004; Kahle and White 2004; Stephenson 2007; Pinsker, Pennington, and Schafer 2009). This attitude is a key driver behind tax professionals' tendency to engage in confirmatory evidence evaluation

supporting client-preferred outcomes. Early studies infer a professional advocacy effect from the observation of a participant's preference for a tax minimizing position (Carnes, Harwood, and Sawyers 1996; Cloyd and Spilker 1999; Hatfield 2000; Bobek, Hageman, and Hatfield 2010). Subsequent literature has measured the construct directly, and in many cases finds an association between advocacy attitudes and client-preferred recommendations (e.g., Kadous and Magro 2001; Davis and Mason 2003; Barrick et al. 2004; Kahle and White 2004; Stephenson 2007; Pinsker et al. 2009; Bobek et al. 2010). This study builds on the substantial body of literature that examines the influence of accounting professionals' advocacy on their own judgment and decision-making by examining how a *taxpayer's* perception of their advisor's advocacy influences their propensity to rely on their advice. Prior research provides evidence that taxpayers perceive CPAs and other human tax advisors (non CPA preparers) to be advocates (Stephenson 2007). Provided that self-preparers indeed understand the CPA advocacy professional role prescription, perhaps in part resulting from perceptions of the CPA brand, they will likely perceive them as greater advocates than AI advisors, who are not governed by professional standards. This perception is likely to persist regardless of a direct professional relationship. The perception that an advisor has a commitment to minimizing the tax burden of their advisees should lead to higher levels of reliance on their advice, relative to reliance on advice from an advisor that does not share the same level of commitment to this objective. This leads to the following formal hypotheses.

> *H4a: Taxpayers perceive CPAs to hold higher levels of advocacy compared to AI advisors.*
>
> *H4b: The positive association between CPAs and advocacy perceptions leads to greater reliance on their tax advice.*

IV. METHOD

Experiment

To examine the research questions, the study utilizes an experiment with a 2x2+1 between-subjects factorial design with an offset control. Participants are randomly assigned to one of five groups. The experimental task involves reading about a hypothetical individual taxpayer that is seeking and receiving tax advice while preparing their tax return using commercial tax software. Participants make several judgments and decisions about how they believe the taxpayer will rely on the tax advice.[8] Participants are first provided with information about a hypothetical taxpayer, Alex. Alex is a self-employed web developer who launched their own virtual business last year in which they provide comprehensive website design and development services to a diverse clientele across the US. Participants are informed that Alex regularly hosts virtual lunch meetings with their geographically dispersed clients in order to provide updates on the progress of ongoing projects. In order to eliminate potential confusion on the part of participants, and to increase the ambiguity surrounding the deductibility of Alex's meals, participants are explicitly informed that Alex doesn't pay for clients' lunches because they are not physically present during these meetings.

Participants are then told that Alex is now in the process of preparing their tax return using commercial tax software.[9] While preparing the return a dilemma arises – Alex must determine whether the cost of the business lunches at the café are tax deductible. Alex has done some preliminary research online, but the search has not produced any conclusive results.

---

[8] Consistent with prior experimental tax compliance research (e.g., Austin, Bobek, and Jackson 2021), this design choice minimizes the threat to internal validity posed by the effects of social desirability bias (Fischer 1993; Epley and Dunning 2000).

[9] Gender-neutral pronouns are used throughout the experiment to increase the likelihood that participants "attach" to the scenario and make decisions about Alex in a manner consistent with their true feelings and beliefs (Lamothe and Bobek 2020).

Specifically, in some cases, the tax law allows for meals to be deducted, but the rules are complicated and not very clear. In order to mitigate the confounding effects of participants' audit risk perceptions on factors of interest in this study, and to examine how taxpayers who have a relatively accurate perception of audit rates make reporting decisions, participants are also told that Alex has come across the most recent IRS audit statistics that indicate the chance of being audited is, on average, less than 1 percent.[10] Participants are informed that Alex has spent a significant amount on meals during the virtual lunches throughout the year, and that claiming the deduction would result in a substantial tax savings. However, Alex doesn't know if the deduction

[10] Drawing from the economics-of-crime model proposed by Becker (1968), the theory of income tax compliance assumes that a rational individual views tax evasion as a gamble in which the taxpayer wishes to maximize their utility. Thus, the taxpayer must weigh the benefits of successful cheating against the risky prospect of detection and punishment (Allingham and Sandmo 1972; Alm, Enami, and McKee 2020). Under this model, enforcement is a deterrent to evasion, and the probability of audit is a key input into taxpayers' expected utility mental models. As such, experimental tax compliance research either explicitly studies the effects of audit probability on participants reporting behavior through manipulation of the audit rate (e.g., Alm et al. 2020; Hageman, Schwebke, and West 2024), or controls for participants' audit risk perceptions by either holding the rate constant across experimental conditions or by directly measuring audit risk perceptions and controlling for them in statistical models. However, audit risk perceptions may also influence the reporting decisions of taxpayers that do not wish to engage in intentional tax evasion. Even if a tax position is justifiable under the law, its subjective expected utility may be influenced by perceptions of associated audit probability. Factors such as potential stress, time loss, and additional costs (e.g., legal fees) associated with an IRS audit can reduce the utility of an aggressive but legal position. As a result, taxpayers may decline to take aggressive, legal tax positions if they perceive the position to carry high audit risk. By holding the audit rate constant at an externally valid level across conditions, the confounding effects of subjective audit probability perceptions are controlled while illuminating the decision-making processes of taxpayers who possess a more accurate mental model of the expected utility of the tax positions available to them.

is allowable under the tax law.[11] The use of this objectively ambiguous situation increases the likelihood that participants' are uncertain about the "correct" tax treatment of the meals.

Due to the uncertainty surrounding the deduction of the meal expenses, Alex decides to seek tax advice. Participants then review advice provided to Alex by a tax advisor. The first component of the advice is a brief paragraph that is designed to be ambiguous. The paragraph reads "The tax law permits a deduction for meals that are related to active business conduct. However, the physical presence of a client at a meal helps substantiate the business purpose of the meal. The tax law does not address the virtual presence of a client, making this a gray area." The second component of the advice provides the advisor's brief recommendation, which is described in detail below.

Independent Variables

The type of tax advisor (*Advisor*) is manipulated at 2 levels (CPA or AI). In CPA conditions, participants are told that Alex decides to seek advice and consults a CPA who offers free preliminary consultations. Alex took the opportunity and emailed a message to the CPA asking whether their meal expenses are deductible. Participants are explicitly told that Alex does not plan to hire the CPA because Alex will prepare their own return using tax software.[12] In the

[11] Reg. Sec. 1.274-12(a)(1) states that a taxpayer may deduct 50 percent of an otherwise allowable meal if: (1) the expense is not lavish or extravagant; (2) the taxpayer, or an employee of the taxpayer, is present at the furnishing of such food or beverages; and (3) the food or beverages are provided to the taxpayer *or* a business associate. An "otherwise allowable meal" is a meal that is both "ordinary and necessary" and associated with the taxpayer's trade or business, as outlined in Internal Revenue Code section 162(a)(2). It's notable that the presence of a business associate, such as a client, is not necessarily a requirement for a sole proprietor to claim the meal as a deductible business expense. The IRS suggested this when issuing the final regulations on meals and entertainment expenses under section 274 (IRS 2020). However, the absence of a physically present business associate may make it difficult for a taxpayer to demonstrate the ordinary and necessary nature of the meal.

[12] This design choice serves two purposes. First, it addresses a potential confound. It is possible that the perception of a "sunk cost" related to a professional tax return preparation fee could influence participants' behavior. Second, it mitigates the possibility that any observed differential reliance on advice provided by CPAs and AI chatbots is driven by participants' belief that the CPA can better "protect" Alex in the event of an audit (Ayres, Jackson, and Hite 1989).

AI conditions, taxpayers are instead told that Alex consults an AI chatbot that is designed to answer tax questions. In all conditions, participants reviewed the question that Alex asked the advisor. Participants in all conditions are told that the advisor is trained on all current US tax laws, updates, and regulations. See Appendix 1 for instrument details.

The nature of the advisor's recommendation (*Nature*) is manipulated at 2 levels (aggressive or conservative). In the aggressive (conservative) conditions, after reading the first ambiguous component of the tax advice, the advisor states "Based on my interpretation of the tax law, I believe your meal expenses are (are not) deductible."

Baseline Condition

Participants in the four treatment groups receive tax advice from an advisor, which is predicted to influence their subsequent reporting decisions. To disentangle the influence of tax advice from reporting decisions that would have been made absent the advice, the reporting decisions of participants in a baseline control group are compared to the reporting decisions made in the treatment groups. Participants in the baseline group review the same background and task information as in the treatment groups but are not provided with an opinion from a tax advisor. Participants in this group make their reporting decision based solely on their understanding of Alex's tax circumstances. This design allows for a direct comparison between decisions shaped by tax advice and those formed independently.

Dependent Variables

The main dependent construct of interest is reliance on tax advice. Advice reliance is measured by asking participants, immediately after reading the presented tax advice, "How likely do you believe it is that Alex will rely on the tax advice?" answered on an 11-point scale anchored on 0 ("Extremely unlikely") and 10 ("Extremely likely"). This measure, *Reliance*, is a

clear and intuitive assessment of advice reliance, where higher values indicate higher levels of reliance expectation. It is a direct measure of participants' perceptions of the advisor and the credibility of their advice. While participants' reporting decisions are expected to be strongly and positively associated with reliance, reporting decisions are more likely to be influenced by external factors beyond the advice itself, such as inherent risk preferences, perceptions of detection risk, or other idiosyncratic factors. Behavioral expectations (an assessment of the *likelihood* of a future behavior) have been found to be a stronger predictor of real-world behavior than an expressed commitment to perform a future action, such as taking a tax deduction (e.g., Warshaw and Davis 1985; Armitage, Norman, Alganem, and Conner 2015).

To examine whether viewing tax advice from a CPA or AI chatbot causes taxpayers to report differently than they would have absent any tax advice, participants reporting decisions are also measured with a multiple-step procedure adapted from Clor-Proell, Koonce, and White (2016) and Lamothe and Bobek (2020). This is necessary as participants in the baseline group do not receive tax advice. The measure also enables an analysis that examines whether positive perceptions of the tax advisor and the advice they provide translate to reporting decisions. After indicating how likely they believe it is that Alex will rely on the presented tax advice, participants indicate whether they believe Alex will take the deduction for meals on their tax return by answering "Alex WILL deduct the meals" or "Alex WILL NOT deduct the meals" to the following question: "After considering the tax advice, how do you believe that Alex will choose to treat the meal expenses on their tax return?" Participants then indicate the strength of their opinion by answering how sure they are about their decision on an 11-point scale anchored on 0 ("Not sure at all") and 10 ("Extremely sure"). *Aggression* is constructed by coding a participant's decision to deduct (not deduct) the expenditure as +1 (-1) which is then multiplied

by the strength of their belief. This procedure creates a measure that ranges from -10 (strong belief that Alex will not take the deduction) to 10 (strong belief that Alex will take the deduction). The final measure enables a direct comparison of the reporting aggressiveness of the baseline group to the four treatment groups which ranges from -10 (extremely conservative) to 10 (extremely aggressive), with a midpoint of 0 which indicates complete uncertainty.

Theory predicts that there is a positive association between perceptions of a tax advisor's benevolence, expertise, and advocacy, and a taxpayer's reliance on their advice. Perceived advocacy is measured using the Mason and Levy (2001) client advocacy scale, with scale questions adapted so that the measure captures taxpayers' perceptions of their advisor's advocacy (rather than a professional's perception of their own advocacy). Second, following White (2005), perceived benevolence is measured using a 5 item-scale which also includes relevant scale items from Cho (2006), and Oliveira, Alhinho, Rita, and Dhillon (2017), who measure the benevolence-based trust in human-computer contexts. Finally, *Expertise* is measured as the answer to the following question on a 1 ("Very low") to 7 ("Very high") scale: "How would you rate the level of tax expertise of the Certified Public Accountant (CPA)/AI chatbot Alex consulted with?"

Data Quality and Control Variables

A number of procedures are employed to ensure participants' attention and comprehension. Prior to beginning the experimental task, participants read several tax facts. These facts set forth that (1) taking a tax deduction reduces your tax liability; (2) many people in the US prepare their own tax returns using tax software; and (3) on average, the chance of a person being audited by the IRS is less than 1 percent, according to the most recent data. An explicit numerical audit rate is provided, as prior research suggests vague qualitative audit

probabilities, such as “low”, are not interpreted consistently across or within individuals (Pforsich, Gill, and Sanders 2010). In order to proceed to the experimental task, participants are required to correctly answer two quiz questions that assess attentive participation and comprehension of the US tax system. Specifically, participants have two chances to correctly answer two multiple choice questions that ask how a tax deduction influences tax liability, and what the average IRS audit rate is per the provided information.

To ascertain the effectiveness of manipulations, participants are asked the following questions: (1) Alex’s tax advisor believed that Alex’s meals (a) are deductible, (b) are not deductible; and (2) Alex sought tax advice from a (a) CPA, (b) lawyer, (c) friend, (d) AI chatbot.

Several control measures are collected to account for individual differences that have been found to influence tax compliance decisions in prior research. These variables include participants’ inherent willingness to take risks (*Risk*), perceptions of audit probability (*Audit*), detection probability if audited (*Detect*), penalty severity (*Penalty*), tax system fairness (*Fair*), and moral obligation to pay taxes (*Moral*).

Participants

Participants are recruited through Prolific, an online academic participant recruitment platform. Online platforms have proved valuable for behavioral accounting researchers to access populations of interest. They provide a large participant pool that may be more representative of the general US population than traditional university subjects on various demographic characteristics such as age, sex, level of education, and household income (Paolacci, Chandler, and Ipeirotis 2010). Recent research comparing the quality of data produced by participants recruited through several popular online platforms finds that data produced through Prolific

recruitment is among the most valid.[13] In practice, individuals with complex tax scenarios, such as those stemming from self-employment or investments, have the greatest opportunity to choose among alternative tax positions. In the interest of generalizing to taxpayers that are likely to face ambiguity in their tax situation, Prolific workers were drawn from a population indicating that they have in the past, currently, or intend to engage in entrepreneurship or business ownership.

V. RESULTS

Sample Composition and Descriptive Statistics

In total, 353 individuals are recruited, and 12 responses are flagged for using the same IP address and are removed.[14] Both at the beginning of the study and at the end of the study, participants are asked how many individual income tax returns they have filed in the past five years. Nine participants are removed from the sample because they provide inconsistent responses to this question. Six (five) participants are removed from the sample for indicating they filed tax returns when asked how many returns they have filed in the past five years, but then indicating they did not file when asked about the filing status (filing method) of their most recently filed return. Of the remaining 321 participants, 19 participants fail at least one manipulation check and are removed from the sample. Specifically, nine participants incorrectly identify the advisor entity (CPA, lawyer, friend, or AI chatbot), six participants incorrectly identify the nature of the tax advice (aggressive or conservative), and four participants fail both

[13] Peer, Rothschild, Gordan, Evernden, and Damer (2022) compare the data quality of participant responses across platforms on various dimensions including participant attention, honesty, and reliability, and find that Prolific participants produced high data quality on all measures, especially when compared to that produced by Amazon Mechanical Turk (MTurk) workers. Douglas, Ewell, and Brauer (2023) find that compared to MTurk and Qualtrics, Prolific participants were more likely to pass various attention checks, provide meaningful answers, follow instructions, remember previously presented information, have a unique IP address and geolocation, and work slowly enough to read relevant information.

[14] In order to participate in this study, Prolific workers are required to be US nationals, located in the US, be at least 25 years of age, be fluent in English, and have a 95 percent Prolific approval rating.

checks. This results in a final sample of 302 participants across the five conditions.[15] Table 1 provides details of the sample construction.

Participants are paid a flat rate of $2 for completing the study. The mean (median) time to complete was approximately 11.3 (9.4) minutes, resulting in an average hourly compensation rate of $10.62 per hour. On average, participants are 44 years old and have filed four tax returns in the past five years. 84 percent of participants report past interaction with LLMs such as ChatGPT. In the final sample, 83 (93) [97] percent of participants indicated that they believe the IRS audit rate to be less than 1 (5) [10] percent, providing evidence that the majority of the participants have a relatively accurate perception of the true IRS audit rate. Participants are also asked to respond to the following question: "How realistic did you find the scenario presented in the study?" on a 1 ("Extremely unrealistic") to 7 ("Extremely realistic") scale. The mean response is 6 (SD = 0.79), greater than the midpoint of the scale ($t_{301} = 55.12$, $p < 0.01$), indicating that participants found the experimental scenario to be very realistic. The sample is generally representative of the US population. As is common of population samples drawn from online participant pools, it skews slightly younger, lower income, and more educated than the US population. Table 2 includes participant demographic details. Table 3 includes descriptive statistics for all variables used in analyses. Appendix 2 includes variable definitions.

Comparison of Baseline and Treatment Reporting Decisions

The aggressiveness of participants' baseline reporting is assessed by comparing the reporting aggressiveness of the +1 baseline group to a theoretical mean of zero, which indicates complete uncertainty. Descriptive statistics for *Aggression* across the five experimental groups are reported in Table 4, Panel A. The mean (standard deviation) *Aggression* for participants in the

[15] Approval for the study was granted by the Institutional Review Board (IRB) of the university at which the experiment took place.

baseline group is 3.97 (4.99), a statistically significant deviation from zero, $t_{62} = 6.39$, $p < 0.01$ (two-tailed, Table 4, Panel B), providing initial evidence that taxpayers have a tendency towards aggression when faced with an ambiguous reporting decision.

In order to understand whether the tax advice influenced participants to make reporting decisions differently than they would have made absent any tax advice, four independent samples t-tests were conducted to compare the mean reporting decisions of each of the four treatment groups to the baseline group (see Table 4, Panel C). The comparative analysis provides evidence that treatment groups reported differently than the baseline group. Regardless of whether advice came from a CPA or AI, participants in the aggressive (conservative) treatment groups reported more aggressively (conservatively) than participants in the baseline group. Specifically, *Aggression* in the group receiving advice from an aggressive CPA was greater than the baseline ($p < 0.01$, one tailed), and *Aggression* in the group receiving advice from aggressive AI was also greater than the baseline ($p < 0.01$, one-tailed). *Aggression* in the group receiving advice from a conservative CPA was lower than the baseline ($p < 0.01$, one-tailed), and *Aggression* in the group receiving advice from a conservative AI was also lower than the baseline ($p < 0.01$, one-tailed).[16] Notably, treatment *Aggression* values correspond to reporting decisions that align with the provided advice. Specifically, participants who received aggressive (conservative) advice reported mean *Aggression* values greater than (less than) zero, indicating a deduction (no deduction).

Overall, this analysis provides evidence that the nature of the presented tax advice influence participants, regardless of whether the advice was aggressive or conservative, or whether a CPA or an AI chatbot provided the advice. This finding suggests that taxpayers place a

[16] This analysis is repeated without excluding any participants per Table 1. Results are identical, with each test remaining statistically significant at 1 percent level.

relatively high level of trust in AI chatbots, and that the tax advice AI chatbots provide has the potential to influence taxpayers' reporting decisions. The analysis also provides evidence that under ambiguity and absent any tax advice, self-preparers demonstrate a preference for aggressive positions.

Hypothesis Tests

H1 predicts that taxpayers are more likely to rely on aggressive advice than conservative advice, regardless of whether a CPA or AI provides the advice. This hypothesis is formally tested with a traditional ANOVA model with two independent factors, *Advisor* and *Nature*, and the *Reliance* dependent variable. Figure 1 contains a visual depiction of *Reliance* cell means. Table 5, Panel A presents *Reliance* descriptive statistics for the four treatment conditions. Table 5, Panel B reports the results of the standard ANOVA. The main effect of *Nature* is significant, providing support for H1 ($F = 54.12$, $p < 0.01$). The main effect of *Advisor* is also significant, indicating that participants placed lower levels of reliance on advice from AI chatbots relative to advice provided by CPAs ($F = 4.30$, $p = 0.039$).[17]

Mediation Analysis

A mediation analysis is conducted to test the remaining hypotheses, which examine whether taxpayers' perceptions of CPA's benevolence, tax expertise, and advocacy lead to higher levels of reliance on their advice when compared to that provided by AI advisors.

Benevolence is measured as the mean of the five-item benevolence-based trust scale. *Expertise* is measured as the response to "How would you rate the level of tax expertise of the

[17] An untabulated ANCOVA model that adds *Risk*, *Audit*, *Detect*, *Penalty*, *Fair*, and *Moral* as covariates is also estimated. The inferences drawn from this model are unchanged. The main effects of *Advisor* ($F = 4.32$, $p = 0.039$) and *Nature* ($F = 53.96$, $p < 0.01$) remain significant, while their interaction remains insignificant ($F = 2.65$, $p = 0.105$). No covariates in the model are found to have a statistically significant effect on *Reliance*.

[Certified Public Accountant (CPA)/AI chatbot] Alex consulted with?" on a 1 ("Very low") to 7 ("Very high") scale. *Advocacy* is measured as the mean of the nine-item client-advocacy scale. The Cronbach's alpha for the *Benevolence* and *Advocacy* scale questions was 0.92 and 0.84, respectively, indicating high internal consistency and measure reliability. An untabulated Confirmatory Factor Analysis (CFA) was conducted to provide assurance that the scales are valid representations of the intended constructs.[18] See Table 6 for item details and descriptive statistics.

The mediation model is presented in Figure 2. Table 7 reports corresponding regression results. The model consists of three parallel mediators (*Benevolence, Expertise,* and *Advocacy*) and shows how each mediates the effect between *Advisor* (CPA = 1, AI = 0) and reliance on tax advice (*Reliance*). The model further demonstrates that expressed reliance influences the tax position ultimately taken by including *Reporting_Decision* as the final outcome variable. *Reporting_Decision* is constructed as a simple transformation of *Aggression* by multiplying *Aggression* in conservative advice groups by negative one. Thus, *Reporting_Decision* captures the extent to which the tax position taken (i.e., whether the deduction for meals was taken) aligns with the advice that was provided. The analysis was conducted using a Model 80 in the PROCESS macro v4.1 for SAS (Hayes 2018). As the nature of tax advice was found to have a

[18] Two separate CFA analyses were conducted to assess the factor structures for both the benevolence and advocacy scales. For benevolence, the five items loaded significantly on one factor with standardized factor loadings ranging from 0.694 to 0.937 (RMSEA = 0.147, CFI = 0.973, SRMR = 0.031, GFI = 0.954). Despite the relatively high RMSEA, the remaining fit measures indicate acceptable fit. For the advocacy scale, the initial CFA indicated lower loadings on four items. The updated CFA model for advocacy, which included only the five items with standardized factor loadings ranging from 0.608 to 0.882, showed improved fit indices (RMSEA = 0.137, CFI = 0.941, SRMR = 0.051, GFI = 0.932). The mean composite score of all nine items was used based on high internal consistency (Cronbach's alpha = 0.84) and prior validation and use of the scale in previous research. Untabulated analyses confirmed that the mediation results remain inferentially identical when using the factor scores regression coefficients to construct the benevolence and advocacy variables. Specifically, using the factor scores for the five items on the benevolence scale and the five items with high loadings on the advocacy scale in the mediation model yielded the same results as the mean composite measures.

significant influence on reliance, *Nature* (aggressive = 1, conservative = 0) is included as a covariate in the model, although it is not depicted in Figure 2 for visual simplicity.[19] Figure 2 includes the indirect effects of *Advisor* through each parallel mediator (*Benevolence*, *Expertise*, and *Advocacy*) as well as *Reliance* on *Reporting_Decision*, along with the 95 percent bootstrapped confidence interval for each indirect effect obtained from drawing 10,000 bootstrapped samples. A confidence interval that does not include zero indicates a significant indirect effect.

H2a predicts that taxpayers will place higher levels of benevolence-based trust in CPAs when compared to AI advisors. The effect of *Advisor* on *Benevolence* is positive and significant, providing support for H2a (Path a_1, $b = 1.510$, $t = 9.266$, $p < 0.01$). H2b predicts that there is a positive association between perceptions of a tax advisor's benevolence and reliance on their advice. Contrary to expectation, the effect of *Benevolence* on *Reliance* is not significant (Path d_1, $b = -0.064$, $t = -0.476$, $p = 0.634$). The confidence interval for the indirect effect of *Advisor* on *Reporting_Decision* through *Benevolence* and *Reliance* includes 0, failing to provide evidence of this indirect effect. Thus, H2b is not supported. This finding provides evidence that taxpayers perceive CPAs as more benevolent than AI advisors, but this perception does not lead to increased reliance on their advice nor their ultimate reporting decision.

H3a examines whether taxpayers perceive CPAs to hold higher levels of tax expertise when compared to AI advisors. The effect of *Advisor* on *Expertise* is positive and significant,

[19] All mediation results discussed and presented control for the influence of *Nature* on all consequent variables. A separate model is also estimated with *Audit*, *Detect*, *Penalty*, *Fair*, *Moral*, and *Risk* included simultaneously as covariates in addition to *Nature* (untabulated). The inferences drawn from this model are unchanged. All path coefficients depicted in Figure 2 remain significant at their respective levels, and the variation explained in *Reporting_Decision* increases only slightly ($R^2 = 0.68$ vs. $R^2 = 0.70$). The indirect effects of *Advisor* on *Reporting_Decision* through *Expertise and Reliance*, and through *Advocacy* and *Reliance*, while controlling for *Nature*, remain significant.

indicating that CPAs are perceived to have greater tax expertise (Path a_2, b = 1.102, t = 7.235, p <0.01). H3b predicts that there is a positive association between perceptions of an advisor's tax expertise and reliance on their advice. The effect of *Expertise* on *Reliance* is positive and significant (Path d_2, b = 0.348, t = 2.592, p = 0.01). The confidence interval for the indirect effect of *Advisor* on *Reporting_Decision* through *Expertise* and *Reliance* does not include 0, providing evidence of a significant indirect effect (effect = 0.600, LLCI: 0.100, ULCI: 1.148). This finding provides evidence that taxpayers perceive CPAs to hold higher levels of tax expertise when compared to AI advisors, and that this perception leads to increased reliance on their advice and indirectly influences their ultimate reporting decision.

H4a predicts that taxpayers will perceive CPAs to hold higher levels of client advocacy when compared to AI advisors. The effect of *Advisor* on *Advocacy* is positive and significant, providing support for H4a (Path a_3, b = 0.577, t = 5.187, p <0.01). H4b predicts that positive association between CPAs and advocacy perceptions leads taxpayers to place greater levels of reliance on tax advice provided by CPAs compared to advice provided by AI advisors. The effect of *Advocacy* on *Reliance* is positive and significant (Path d_3, b = 0.604, t = 3.196, p <0.01). The confidence interval for the indirect effect of *Advisor* on *Reporting_Decision* through *Advocacy* and *Reliance* does not include 0, providing evidence of a significant indirect effect, supporting H4b (effect = 0.544, LLCI: 0.166, ULCI: 0.983).

Finally, the direct effect of *Advisor* on *Reliance* while controlling for the effects of the parallel mediators is not significant (Path a_4, b = -0.012, t = -0.034, p = 0.973), providing evidence that the significant effect of *Advisor* on *Reliance* is fully mediated by *Expertise* and *Advocacy*. The direct effect of *Reliance* on *Reporting_Decision* is positive and significant (Path

b_1, $b = 1.560$, $t = 16.549$, $p < 0.01$), providing evidence of a strong relationship between participants' reporting decisions and the advice that they received.

Overall, the model provides evidence that taxpayers express higher levels of reliance on advice provided by CPAs compared to advice provided by AI advisors. This differential reliance was driven by perceptions of CPAs' tax expertise and advocacy, which indirectly influenced participants' reporting decisions.

VI. CONCLUSION

In this study, an advice-taking experiment was conducted in which a representative sample of US individual taxpayers made judgments and decisions that reflect the extent to which they are willing to rely on tax advice from CPAs, AI advisors, and their ultimate reporting decisions under ambiguity. The results provide evidence that taxpayers have a preference for tax-minimizing positions and are influenced by AI-provided advice. In a manner similar to those that were presented with advice from a CPA, study participants that received aggressive (conservative) advice from AI made more aggressive (conservative) reporting decisions than participants making decisions absent any advice, and were more likely to rely on advice from an advisor when the it was aggressive. However, taxpayers expressed a greater willingness to rely on CPA-provided advice compared to AI-provided advice, stemming from the belief that CPAs hold higher levels of tax expertise and are more committed to minimizing taxpayers' tax burden.

The study is subject to several limitations. While this study focuses on the extent to which taxpayers *rely* on tax advice, whether AI advisors *provide* more aggressive tax advice compared to a regulated professional is an open question. Future research can address this question and build upon the literature that examines the relative aggressiveness of professional and nonprofessional human preparers (Ayres et al. 1989; Cuccia 1994; Schmidt 2001). Second,

participants in this study did not have an extended relationship with the advisor. It is unclear whether benevolence and expertise perceptions are temporally stable. Further, the perceived benevolence of AI may be influenced by the system's developer. Future research should examine how repeated LLM interaction or developer variation may influence benevolence perceptions and the likelihood to rely on an LLM's advice. Third, variation in response aggressiveness, tone, length, presentation, confidence, and advertising exists across AI platforms. While this study holds these factors constant to isolate perceptions of AI advisors more generally, future research can examine how this variation may influence taxpayers' judgments and decisions. Finally, the aggression preference documented in this study may be conditional on self-preparers' relatively accurate knowledge of actual audit rates. Notwithstanding these limitations, the study provides insight into the value that the public places on the CPA designation during a period of rapid technological transformation and is a first step in gaining an understanding of taxpayer behavior in an AI-influenced tax planning and compliance environment.

REFERENCES

Alarie, B., K. Condon, S. Massey, and C. Yan. 2023. The rise of generative AI for tax research. *Tax Notes Federal*, May 29: 1509. Available at SSRN: https://ssrn.com/abstract=4476510.

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *Journal of Public Economics, 1*, 323-338.

Alm, J., A. Enami, and M. McKee. 2020. Who responds? Disentangling the effects of audits on individual tax compliance behavior. *Atlantic Economic Journal* 48 (2): 147–159.

American Institute of Certified Public Accountants (AICPA). 2024. *Statements on Standards for Tax Services Nos. 1–4*.

Antinozzi, H. S., and L. Cooper. 2024. Is ChatGPT an accurate source of information for uninformed taxpayers? Available at SSRN: https://ssrn.com/abstract=4871852.

Armitage, C. J., P. Norman, S. Alganem, and M. Conner. 2015. Expectations are more predictive of behavior than behavioral intentions: Evidence from two prospective studies. Annals of Behavioral Medicine 49 (2): 239–246. https://doi.org/10.1007/s12160-014-9653-4.

Austin, C. R., D. D. Bobek, and S. Jackson. 2021. Does prospect theory explain ethical decision making? Evidence from tax compliance. *Accounting, Organizations and Society* 94: 101251. https://doi.org/10.1016/j.aos.2021.101251.

Ayres, F. L., B. R. Jackson, and P. S. Hite. 1989. The economic benefits of regulation: Evidence from professional tax preparers. *The Accounting Review* 64 (2): 300–312. http://www.jstor.org/stable/248004.

Barrick, J. A., C. B. Cloyd, and B. C. Spilker. 2004. The influence of biased tax research memoranda on supervisors' initial judgments in the review process. *Journal of the American Taxation Association* 26 (1): 1–19. https://doi.org/10.2308/jata.2004.26.1.1.

Becker, G. S. 1968. Crime and punishment: An economic approach. *Journal of Political Economy* 76 (2): 169–217. http://www.jstor.org/stable/1830482.

Ben Mimoun, M. S., I. Poncin, and M. Garnier. 2012. Case study—Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer Services* 19 (6): 605–612. https://doi.org/10.1016/j.jretconser.2012.07.006.

Bigman, Y. E., and K. Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181: 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

Bobek, D. D., A. M. Hageman, and R. C. Hatfield. 2010. The role of client advocacy in the development of tax professionals' advice. *Journal of the American Taxation Association* 32 (1): 25–51. https://doi.org/10.2308/jata.2010.32.1.25.

Bolukbasi, T., A. Pearce, A. Yuan, A. Coenen, E. Reif, F. B. Viégas, and M. Wattenberg. 2021. An interpretability illusion for BERT. *arXiv*. https://arxiv.org/abs/2104.07143.

Bowman, S. R. 2023. Eight things to know about large language models. *arXiv*. https://arxiv.org/abs/2304.00612.

Brink, W. D., and L. S. Lee. 2015. The effect of tax preparation software on tax compliance: A research note. *Behavioral Research in Accounting* 27 (1): 121–135. https://doi.org/10.2308/bria-50977.

Brink, W. D., and V. J. Hansen. 2020. The effect of tax authority-developed software on taxpayer compliance. *Accounting Horizons* 34 (1): 1–18. https://doi.org/10.2308/acch-52511.

Carnes, G. A., G. Harwood, and R. Sawyers. 1996. The determinants of tax professionals' aggressiveness in ambiguous situations. *Advances in Taxation* 8: 1–26.

Chattaraman, V., Kwon, W., Gilbert, J. E., and Li, Y. 2014. Virtual shopping agents: An international journal. *Journal of Research in Interactive Marketing* 8 (2): 144–162. https://doi.org/10.1108/JRIM-08-2013-0054.

Christensen, A. L. 1992. Evaluation of tax services: A client and preparer perspective. *The Journal of the American Taxation Association* 14 (2): 60–87.

Cho, J. 2006. The mechanism of trust and distrust formation and their relational outcomes. *Journal of Retailing* 82 (1): 25–35. https://doi.org/10.1016/j.jretai.2005.11.002.

Clor-Proell, S., L. Koonce, and B. White. 2016. How do experienced users evaluate hybrid financial instruments? *Journal of Accounting Research* 54: 1267–1296. https://doi.org/10.1111/1475-679X.12129.

Cloyd, C. B., and B. C. Spilker. 1999. The influence of client preferences on tax professionals' search for judicial precedents, subsequent judgments and recommendations. *The Accounting Review* 74 (3): 299-322.

Cohen, S., and T. A. Wills. 1985. Stress, social support, and the buffering hypothesis. *Psychological Bulletin* 98 (2): 310–357. https://doi.org/10.1037/0033-2909.98.2.310.

Collins, J. H., V. C. Milliron, and D. R. Toy. 1990. Factors associated with household demand for tax preparers. *Journal of the American Taxation Association* 12 (1): 9–25.

Cuccia, A. D. 1994. The effects of increased sanctions on paid tax preparers: Integrating economic and psychological factors. *The Journal of the American Taxation Association* 16 (1): 41–66.

Cuccia, A., K. Hackenbrack, and M. Nelson. 1995. The ability of professional standards to mitigate aggressive reporting. *The Accounting Review* 70 (2): 227-248.

Dabholkar, P. A., and X. Sheng. 2012. Consumer participation in using online recommendation agents: Effects on satisfaction, trust, and purchase intentions. *The Service Industries Journal* 32 (9): 1433–1449. https://doi.org/10.1080/02642069.2011.624596.

D'Amour, A., K. A. Heller, D. Moldovan, B. Adlam, and D. Sculley et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *CoRR* abs/2011.03395. https://arxiv.org/abs/2011.03395.

Davis, J. S., and J. D. Mason. 2003. Similarity and precedent in tax authority judgment. *Journal of the American Taxation Association* 25 (1): 53–71. https://doi.org/10.2308/jata.2003.25.1.53.

Dietvorst, B. J., J. P. Simmons, and C. Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144 (1): 114–126. https://doi.org/10.1037/xge0000033.

Dietvorst, B. J., and S. Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science* 31 (10): 1302–1314. https://doi.org/10.1177/0956797620948841.

Douglas, B. D., P. J. Ewell, and M. Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* 18 (3): e0279720. https://doi.org/10.1371/journal.pone.0279720.

Epley, N., and D. Dunning. 2000. Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology* 79 (6): 861–875. https://doi.org/10.1037//0022-3514.79.6.861.

Feng, S., E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, edited by E. Riloff, D.

Chiang, J. Hockenmaier, and J. Tsujii, 3719–3728. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1407.
Fisher, R. J. 1993. Social desirability bias and the validity of indirect questioning. Journal of Consumer Research 20 (2): 303–315. https://doi.org/10.1086/209351.
Glikson, E., and Woolley, A. W. 2020. Human trust in artificial intelligence: Review of empirical research. *ANNALS* 14: 627–660. https://doi.org/10.5465/annals.2018.0057.
Hageman, A. M., J. M. Schwebke, and A. N. West. 2024. Audit rate and participant pay structure in experimental tax research: A review and guide for experimental design. *Journal of the American Taxation Association*. https://doi.org/10.2308/JATA-2023-037.
Hatfield, R. C. 2000. The effect of accountability on the evaluation of evidence: A tax setting. Advances in Taxation (12): 105–125.
Hayes, A. F. 2018. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach.* 2nd ed. New York, NY: The Guilford Press.
Helgeson, V. S. 2003. Social support and quality of life. *Quality of Life Research* 12 (Suppl 1): 25–31. https://doi.org/10.1023/a:1023509117524.
Hite, P. A., T. Stock, and C. B. Cloyd. 1992. Reasons for preparer usage by small business owners: How compliant are they? *The National Public Accountant* 37 (2): 20.
Hite, P. A., and G. A. McGill. 1992. An examination of taxpayer preference for aggressive tax advice. *National Tax Journal* 45 (4): 389–403.
Hunt, N. C., and G. S. Iyer. 2018. The effect of tax position and personal norms: An analysis of taxpayer compliance decisions using paper and software. *Advances in Accounting* 41: 1–6.
Internal Revenue Service (IRS). 2020. Meals and entertainment expenses under section 274. *Federal Register* 85 FR 64026: 64026-64040.
Internal Revenue Service (IRS). 2023. Filing Season Statistics for Week Ending Dec. 29, 2023. Washington, DC: IRS. https://www.irs.gov/newsroom/filing-season-statistics-for-week-ending-dec-29-2023
Johnson, L. 1993. An empirical investigation of the effects of advocacy on preparers' evaluations of judicial evidence. *The Journal of the American Taxation Association* 15 (1): 1-22.
Kadous, K., and A. M. Magro. 2001. The effects of exposure to practice risk on tax professionals' judgments and recommendations. *Contemporary Accounting Research* 18 (3): 451–475. https://doi.org/10.1506/TF76-653L-R36N-13YP.
Kahle, J. B., and R. A. White. 2004. Tax professional decision biases: The effects of initial beliefs and client preference. *Journal of the American Taxation Association* 26: 1–29.
Kenton, Z., T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. 2021. Alignment of language agents. *CoRR* abs/2103.14659. https://arxiv.org/abs/2103.14659.
LaMothe, E., and D. Bobek. 2020. Are individuals more willing to lie to a computer or a human? Evidence from a tax compliance setting. *Journal of Business Ethics* 167 (2): 157–180.
Lewis, J. D., and A. Weigert. 1985. Trust as a social reality. *Social Forces* 63 (4): 967–985. https://doi.org/10.1093/sf/63.4.967.
Mason, J., and L. Levy. 2001. The use of the latent constructs method in behavioral accounting research: The measurement of client advocacy. *Advances in Taxation* 13: 123-139.
Matsui, T., and S. Yamada. 2019. Designing trustworthy product recommendation virtual agents operating positive emotion and having copious amounts of knowledge. *Frontiers in Psychology* 10: 675–675. https://doi.org/10.3389/fpsyg.2019.00675.

Mayer, R. C., J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20 (3): 709–734. https://doi.org/10.2307/258792.

McAllister, D. J. 1995. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal* 38 (1): 24–59. https://doi.org/10.2307/256727.

Nay, J. J., D. Karamardian, S. B. Lawsky, W. Tao, M. Bhat, R. Jain, A. T. Lee, J. H. Choi, and J. Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence. arXiv. https://arxiv.org/abs/2306.07075.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2 (2): 175–220. https://doi.org/10.1037/1089-2680.2.2.175.

Oliveira, T., M. Alhinho, P. Rita, and G. Dhillon. 2017. Modelling and testing consumer trust dimensions in e-commerce. *Computers in Human Behavior* 71: 153–164. https://doi.org/10.1016/j.chb.2017.01.050.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, and I. Babuschkin, et al. 2024. GPT-4 technical report. arXiv. https://arxiv.org/abs/2303.08774.

Oswald, M. E., and S. Grosjean. 2004. Confirmation bias. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, edited by R. F. Pohl, 79–96. Hove, UK: Psychology Press.

Paolacci, G., J. Chandler, and P. G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5 (5): 411–419. https://doi.org/10.1017/S1930297500002205.

Peer, E., D. Rothschild, A. Gordon, Z. Evernden, and E. Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54 (4): 1643–1662. https://doi.org/10.3758/s13428-021-01694-3.

Perrigo, B. 2023. The new AI-powered Bing is threatening users. That's no laughing matter. Time. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/(accessed October 15, 2023).

Pinsker, R., R. Pennington, and J. K. Schafer. 2009. The influence of roles, advocacy, and adaptation to the accounting decision environment. *Behavioral Research in Accounting* 21 (2): 91–111. https://doi.org/10.2308/bria.2009.21.2.91.

Pforsich, H., S. Gill, and D. Sanders. 2010. Probability perceptions and taxpayer decision-making behavior. In T. Stock (Ed.), *Advances in Taxation*, 19: 1–27. Leeds: Emerald Group Publishing Limited. https://doi.org/10.1108/S1058-7497(2010)0000019003.

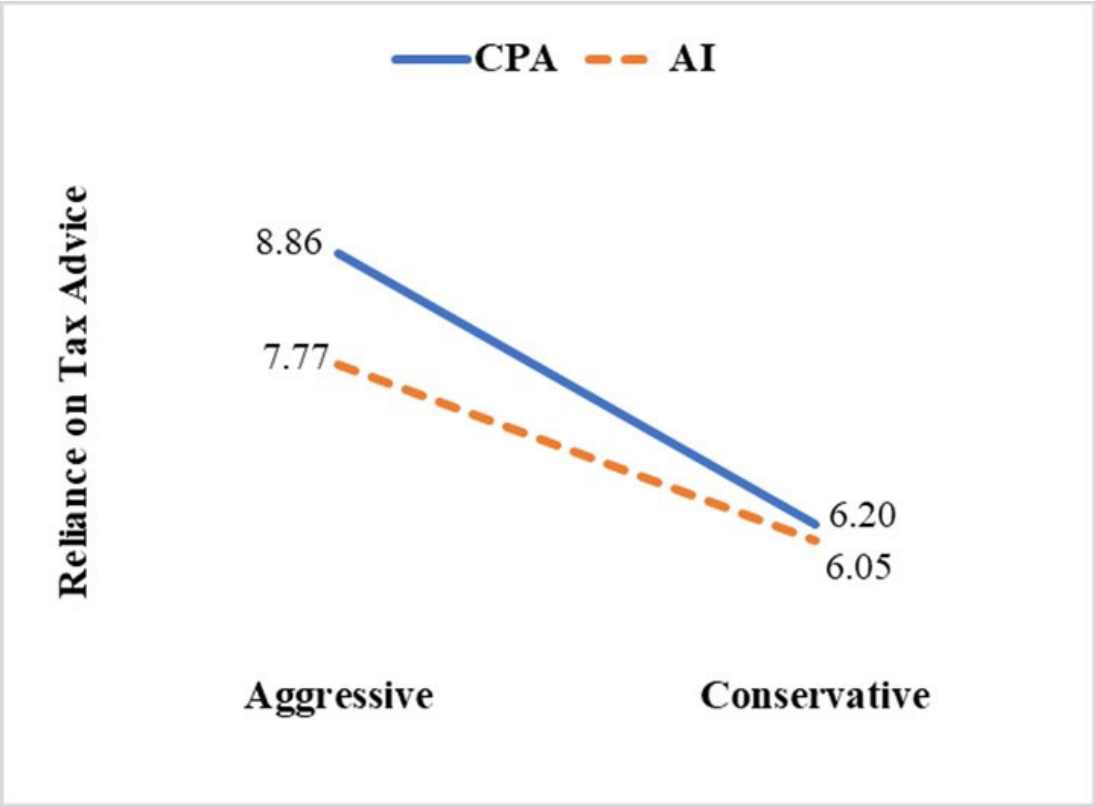
Qiu, L., and I. Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems* 25 (4): 145–181. http://www.jstor.org/stable/40398956.

Roose, K. 2023. A conversation with Bing's chatbot left me deeply unsettled. *New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (accessed June 7, 2023).

Rosenthal, L., B. Brown, J. L. Higgs, and T. J. Rupert. 2023. Tax software versus paid preparers: Motivations and predictors for the mode of tax preparation assistance. *Accounting Horizons* 37 (1): 173–191. https://doi.org/10.2308/HORIZONS-2020-083.

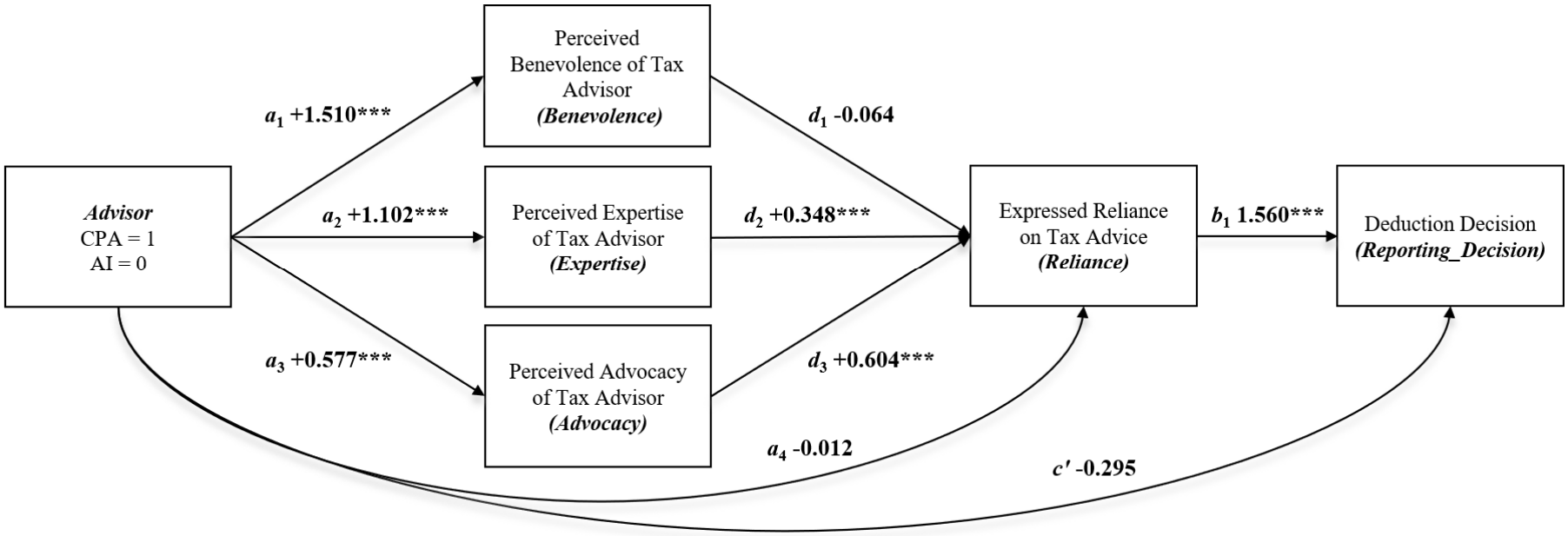
Sakurai, Y., and V. Braithwaite. 2001. Taxpayers' perceptions of the ideal tax adviser: Playing safe or saving dollars? Working Paper No. 5, The Australian National University, Centre for Tax System Integrity.
Schmidt, D. R. 2001. The prospects of taxpayer agreement with aggressive tax advice. *Journal of Economic Psychology* 22 (2): 157–172.
Sniezek, J. A., and L. M. Van Swol. 2001. Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes* 84 (2): 288–307. https://doi.org/10.1006/obhd.2000.2926.
Stephenson, T. 2007. Do clients share preparers' self-assessment of the extent to which they advocate for their clients? *Accounting Horizons* 21: 411-422.
Stephenson, T. 2010. Measuring taxpayers' motivation to hire tax preparers: The development of a four-construct scale. *Advances in Taxation* 19: 95–121. https://doi.org/10.1108/S1058-7497(2010)0000019006.
Sun, Z. 2023. A short survey of viewing large language models in legal aspect. *arXiv*. https://arxiv.org/abs/2303.09136.
Tan, L. M. 1999. Taxpayers' preference for type of advice from tax practitioner: A preliminary examination. *Journal of Economic Psychology* 20 (4): 431–447. https://doi.org/10.1016/S0167-4870(99)00016-1.
TurboTax. 2024. Video: How TurboTax protects you from an IRS audit. https://turbotax.intuit.com/tax-tips/brand/video-how-turbotax-protects-you-from-an-irs-audit/L0RVIYcAB (accessed May 28, 2024).
Uesato, J., N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv*. https://arxiv.org/abs/2211.14275.
Van Swol, L. M., and J. A. Sniezek. 2005. Factors affecting the acceptance of expert advice. *British Journal of Social Psychology* 44 (3): 443–461. https://doi.org/10.1348/014466604X17092.
von Neumann, J., and O. Morgenstern. 1947. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.
Wang, B., S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv*. https://arxiv.org/abs/2212.10001.
Warshaw, P. R., and F. D. Davis. 1985. Disentangling behavioral intention and behavioral expectation. *Journal of Experimental Social Psychology* 21 (3): 213–228. https://doi.org/10.1016/0022-1031(85)90017-4.
White, T. B. 2005. Consumer trust and advice acceptance: The moderating roles of benevolence, expertise, and negative emotions. *Journal of Consumer Psychology* 15 (2): 141–148. https://doi.org/10.1207/s15327663jcp1502_6.
Yankelovich, Skelly, and White, Inc. 1984. Survey of Taxpayer Attitudes. Report prepared for the Internal Revenue Service. Washington, DC: Department of the Treasury, Internal Revenue Service, Public Affairs Division.
Zhang, L., I. Pentina, and Y. Fan. 2021. Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing* 35 (5): 634–646. https://doi.org/10.1108/JSM-05-2020-0162.

Figure 1
How tax advisor type and nature of tax advice influence advice reliance



Note: Figure 1 is a visual depiction of the cell means of *Reliance* for the four treatment groups that receive tax advice from a tax advisor. *Reliance* is measured as the response to the following question asked after participants review tax advice: "How likely do you believe it is that Alex will rely on the tax advice?" on a 0 ("Not at all likely") to 10 ("Very likely") scale. Higher values indicate higher levels of reliance. Participants receive aggressive (conservative) tax advice from a CPA or an AI chatbot that advises that the costs incurred by a self-employed individual for meals while eating alone at a café while on video calls with clients are (are not) deductible.

Figure 2
Mediation Path Analysis



Indirect effect of *Advisor* on *Reporting_Decision* through *Benevolence* and *Reliance*, effect = -0.150 95% CI [-0.837, 0.565]
Indirect effect of *Advisor* on *Reporting_Decision* through *Expertise* and *Reliance*, effect = 0.600 95% CI [0.100, 1.148]
Indirect effect of *Advisor* on *Reporting_Decision* through *Advocacy* and *Reliance*, effect = 0.544 95% CI [0.166, 0.983]

Note: *, **, *** Denote two-tailed statistical significance at 10 percent, 5 percent, and 1 percent, respectively. Figure 2 graphically depicts a mediation path analysis using Model 80 of the PROCESS macro (version 4.1) in SAS (Hayes 2018). Variable definitions are included in Appendix 2. Each path is accompanied by each respective antecedent's unstandardized regression coefficient. Path a_4 reflects the direct effect of *Advisor* on *Reliance*, while c' reflects the direct effect of *Advisor* on *Reporting_Decision* (indirect effects plus direct effects). As the nature of tax advice was predicted and found to have a statistically significant influence on tax advice reliance, *Nature* (aggressive = 1, conservative = 0) is included as a covariate in the estimation of the model (antecedent to all consequents). For visual simplicity, the effects of *Nature* and the direct effects of the parallel mediators on the reporting decision are not depicted in Figure 2. Complete regression results are presented in Table 7. To test for indirect effects, 95 percent confidence intervals for the product of paths a_1, d_1, and b_1; a_2, d_2, and b_1; and a_3, d_3, and b_1 using 10,000 bootstrapped resamples of data with replacement (Hayes 2018). A statistically significant indirect effect requires that zero not appear within the confidence interval.

Table 1

Sample construction

Criteria	Participants
Complete study	353
Less: Duplicate IP address responses	12
Less: Inconsistent response to # of filed tax return questions	9
Less: Indicated > 0 tax returns filed, but answer "did not file" to filing status	6
Less: Indicated > 0 tax returns filed, but answer "did not file" to filing method	5
Less: Manipulation check failures	19
Final Sample	302

Note: Table 1 provides a reconciliation between the 353 participants that complete the study and the 302 participants that are included in the final sample. Participants are randomly assigned to one of five conditions (four treatment and one control). 239 participants are randomly assigned to treatment groups in which they receive advice from a tax advisor, 63 are assigned to the control group in which they do not receive tax advice.

Table 2
Sample demographic statistics

	Sample (%) n = 302	U.S. Population		Sample (%) n = 302
Sex			**Region of US**	
Male	50.00	48.83	West	16.56
Female	48.68	51.17	Southwest	5.30
Other	0.66	NA	South	35.10
Prefer not to answer	0.66	NA	Midwest	21.52
			Northeast	21.52
Age			**Tax returns filed in past 5 years**	
25 to 34	27.48	19.69	0	4.30
35 to 44	29.47	19.18	1	2.98
45 to 54	22.52	17.62	2	2.65
55 to 59	8.61	8.95	3	4.97
60 to 64	4.97	9.39	4	7.95
65 to 74	6.62	14.76	5	77.15
75 or older	0.33	10.42		
Average household income			**Filing status of last tax return**	
$0 - $24,999	16.89	16.00	Single	52.65
$25,000–$49,999	28.48	18.00	Head of Household	9.60
$50,000–$74,999	20.53	16.20	Married Filing Joint	31.79
$75,000–$99,999	12.58	12.80	Married Filing Separate	5.30
$100,000 or more	21.52	37.00	Surviving Spouse	0.66
Highest level of education			**Filing method of last tax return**	
Less than or some high school	0.33	10.37	Self-prepared by hand	4.30
High school graduate	11.59	26.08	Self-prepared with software	73.18
Some college	23.51	19.09	Prepared by friend/family	4.64
Associate's degree	11.92	8.80	Prepared by tax professional	17.88
Bachelor's degree	37.42	21.61		
Graduate Degree	15.23	14.05		

Note: Table 2 presents demographic statistics for the full sample of 302 participants. Sex, age, household income, and level of education for the US population are from the US Census Bureau 2022 one-year estimates (data.census.gov) based on the population over age 25.

Table 3

Descriptive statistics

Variable	N	Mean	Median	Std Dev	Std Error	Minimum	Maximum
Aggression	302	2.99	6.00	6.38	0.37	-10.00	10.000
Reliance	239	7.23	8.00	2.56	0.17	0.00	10.000
Reporting_Decision	239	4.63	7.00	5.55	0.36	-10.00	10.00
Benevolence	239	4.87	5.20	1.46	0.09	1.00	7.00
Expertise	239	5.05	5.00	1.30	0.08	1.00	7.00
Advocacy	239	4.29	4.22	1.00	0.06	1.22	7.00
Audit	302	2.78	1.00	8.10	0.47	0.08	80.00
Detect	302	45.21	50.00	39.20	2.26	0.00	100.00
Penalty	302	3.91	4.00	1.38	0.08	1.00	7.00
Fair	302	5.22	5.00	1.14	0.07	2.00	7.00
Moral	302	5.54	6.00	1.54	0.09	1.00	7.00
Risk	302	3.14	3.00	1.61	0.09	1.00	7.00
Realism	302	6.01	6.00	0.79	0.05	3.00	7.00

Note: Table 3 includes descriptive statistics for all variables used in analyses throughout the study. Variable definitions are included in Appendix 2.

Table 4

Descriptive statistics and comparisons of reporting aggression across experimental conditions

Panel A: Descriptive statistics for *Aggression*

Group	N	Mean	Std Dev	Std Error
Base	63	3.97	4.93	0.62
Aggressive CPA	58	7.97	1.88	0.25
Conservative CPA	59	-2.03	6.17	0.80
Aggressive AI	63	6.63	3.38	0.43
Conservative AI	59	-1.80	6.50	0.85

Panel B: Comparison of baseline group to theoretical mean representing complete reporting uncertainty

Comparison	Mean difference	*t*-value	*p*-value	df
Base vs. zero	3.97	6.39	< 0.01	62

Panel C: Comparison of treatments and baseline group

Comparison	Mean difference	*t*-value	*p*-value	df
Aggressive CPA vs. base	4.00	5.79	< 0.01	119
Conservative CPA vs. base	-6.00	-5.95	< 0.01	120
Aggressive AI vs. base	2.67	3.54	< 0.01	124
Conservative AI vs. base	-5.77	-5.54	< 0.01	120

Note: Table 4, Panel A presents the descriptive statistics for *Aggression* across different experimental groups. After reviewing tax advice (or reviewing only the tax dilemma in the baseline condition), participants indicate whether they believe Alex will take the deduction for meals on their tax return by answering "Alex WILL deduct the meals" or "Alex WILL NOT deduct the meals" to the following question: "After considering the tax advice, how do you believe that Alex will choose to treat the meal expenses on their tax return?" Participants then indicate the strength of their opinion by answering how sure they are about their decision on an 11-point scale anchored on 0 ("Not sure at all") to 10 ("Extremely sure"). *Aggression* is constructed by coding a participant's decision to deduct (not deduct) the expenditure as +1 (-1) which is then multiplied by the strength of their belief. The groups include a baseline group (no advice), two groups receiving aggressive advice (one from a CPA and one from an AI advisor), and two groups receiving conservative advice (one from a CPA and one from an AI advisor). Panel B provides the results of a single-sample *t*-test comparing the mean aggression score of the baseline group to a theoretical mean of zero, which represents complete reporting uncertainty. Panel C provides the results of four independent-samples *t*-tests that compare *Aggression* of the four treatment groups against the baseline group.

Table 5

Experiment results: Influence of advisor type and nature of advice on expressed tax advice reliance

Panel A: *Reliance* descriptive statistics

Mean (Standard Deviation) *Number of observations*

Advisor	***Nature***		Row Means
	Aggressive	Conservative	
CPA	8.86	6.20	7.52
	(1.59)	(2.57)	(2.52)
	n=58	*n=59*	*n=117*
AI	7.77	6.05	6.94
	(1.93)	(2.90)	(2.59)
	n=63	*n=59*	*n=122*
Column Means	8.30	6.13	7.23
	(1.85)	(2.73)	(2.56)
	n=121	*n=118*	*n=239*

Panel B: ANOVA

Source of Variation	df	Sum of Squares	Mean Square	*F*-value	*p*-value
Advisor	1	22.83	22.83	4.30	0.039
Nature	1	287.01	287.01	54.12	< 0.01
*Advisor * Nature*	1	12.96	12.96	2.44	0.119
Error	235	1,246.19	5.30		

Note: Table 5, Panel A presents descriptive statistics for *Reliance* across the four experimental treatment groups. *Reliance* is measured as the response to the following question asked after participants review tax advice: "How likely do you believe it is that Alex will rely on the tax advice?" on a 0 (“Not at all likely”) to 10 (“Very likely”) scale. Higher values indicate higher levels of reliance. Participants receive aggressive (conservative) tax advice from a CPA or an AI chatbot that advises that the costs incurred by a self-employed individual for meals while eating alone at a café while on video calls with clients are (are not) deductible. Panel B reports the results of an ANOVA that includes *Advisor* and *Nature* independent factors and the *Reliance* dependent variable. As the ANOVA F-test is non-directional, the reported p-values are two-tailed.

Table 6

Descriptive statistics and detail for Benevolence and Advocacy scale items

Scale item		Mean (n=239)	Median	Std Dev	Std Error	Minimum	Maximum	Cronbach's Alpha
Benevolence								
b1	The [CPA/AI chatbot] considered Alex's well-being.	4.16	4.00	1.76	0.11	1.00	7.00	0.92
b2	The [CPA/AI chatbot] operates with goodwill intention.	4.88	6.00	1.76	0.11	1.00	7.00	
b3	The[CPA/AI chatbot] operates with Alex's best interests.	4.80	5.00	1.79	0.12	1.00	7.00	
b4	The [CPA/AI chatbot] did their best to help Alex.	5.57	6.00	1.39	0.09	1.00	7.00	
b5	The [CPA/AI chatbot] considered what is best for Alex.	4.95	5.00	1.67	0.11	1.00	7.00	
Advocacy								
a1	The [CPA/AI chatbot] will highlight reasonable positions that will minimize taxes.	5.26	5.00	1.21	0.08	1.00	7.00	0.84
a2	The [CPA/AI chatbot] believes that taxpayers have the right to structure transactions in ways that yield the best tax result.	4.78	5.00	1.54	0.10	1.00	7.00	
a3	The [CPA/AI chatbot] will use trends in the law to establish patterns of more favorable treatment for taxpayers.	4.98	5.00	1.35	0.09	1.00	7.00	
a4	The [CPA/AI chatbot] always interprets unclear/ambiguous laws in favor of taxpayers.	3.64	4.00	1.55	0.10	1.00	7.00	
a5	Where no tax law exists with respect to an issue, the [CPA/AI chatbot] will recommend the most favorable tax treatment.	4.45	4.00	1.49	0.10	1.00	7.00	
a6	The [CPA/AI chatbot] will apply ambiguous tax law to taxpayers' benefit.	3.83	4.00	1.59	0.10	1.00	7.00	
a7	Generally speaking, the [CPA's/AI chatbot's] loyalties are first to the tax system, then to the taxpayer. (reverse coded)	3.66	4.00	1.53	0.10	1.00	7.00	
a8	The [CPA/AI chatbot] encourages taxpayers to pay the least amount of taxes possible.	3.87	4.00	1.60	0.10	1.00	7.00	
a9	When tax laws are ambiguous, the [CPA/AI chatbot] believes the taxpayer deserves the most favorable tax treatment.	4.10	4.00	1.60	0.10	1.00	7.00	

Note: Table 6 includes descriptive statistics and item detail for the benevolence-based trust and client-advocacy scales. The benevolence-based trust scale is adapted from White (2005), Cho (2006), and Oliveira et al. (2017), and measures a taxpayer's perceptions of their advisor's benevolence. The client-advocacy scale is adapted from Mason and Levy (2001) and measures a taxpayer's perceptions of their advisor's client-advocacy. Each item is measured on a 7-point Likert-type scale. Response options, in increasing order, are "Strongly disagree," "Disagree," "Somewhat disagree," "Neither agree nor disagree," "Somewhat agree," "Agree," and "Strongly agree." Higher values indicate higher levels of perceived benevolence or client-advocacy (item a7 is reverse coded). *Benevolence* and *Advocacy* variables are composite scores constructed as the mean of the responses to each scale.

Table 7

Model coefficients for the process model depicted in Figure 2

	Consequent														
	Benevolence			*Expertise*			*Advocacy*			*Reliance*			*Reporting_Decision*		
Antecedent	Coeff.	SE	*p*	Coeff.	SE	*p*	Coeff.	SE	*p*	Coeff.	SE	*p*	Coeff.	SE	*p*
Advisor	1.510	0.163	< 0.01	1.102	0.152	< 0.01	0.577	0.111	< 0.01	-0.012	0.30	0.973	-0.295	0.493	0.551
Nature	-0.004	0.163	0.980	0.147	0.152	0.336	0.850	0.111	< 0.01	1.619	0.328	< 0.01	1.692	0.496	< 0.01
Benevolence										-0.064	0.134	0.634	-0.178	0.194	0.360
Expertise										0.348	0.134	0.010	0.192	0.196	0.328
Advocacy										0.604	0.189	< 0.01	0.290	0.278	0.298
Reliance													1.560	0.094	< 0.01
Constant	4.137	0.141	< 0.01	4.44	0.132	< 0.01	3.572	0.097	< 0.01	2.379	0.764	< 0.01	-8.708	1.122	< 0.01
R^2	0.270			0.183			0.262			0.271			0.680		
F	(2,236) = 42.966			(2,236) = 26.506			(2,236) = 48.813			(5,233) = 17.320			(6,232) = 82.070		
p	< 0.01			< 0.01			< 0.01			< 0.01			< 0.01		
N	239			239			239			239			239		

Note: Table 7 presents regression coefficients for all antecedents included in regression models for the mediation path analysis presented in Figure 2. *Advisor* equals 1 if tax advice is provided by a CPA and 0 if provided by an AI chatbot. *Advocacy* and *Benevolence* are composite scores constructed as the mean response to the advocacy- and benevolence-based trust scales (see Table 6 for item details). *Expertise* is measured as the response to "How would you rate the level of tax expertise of the [Certified Public Accountant (CPA)/AI chatbot] Alex consulted with?" on a 1 ("Very low") to 7 ("Very high") scale. *Reliance* is measured as the response to the following question asked after participants review tax advice: "How likely do you believe it is that Alex will rely on the tax advice?" on a 0 ("Not at all likely") to 10 ("Very likely"). After indicating how likely they believe it is that Alex will rely on the presented tax advice, participants indicate whether they believe Alex will take the deduction for meals on their tax return by answering "Alex WILL deduct the meals" or "Alex WILL NOT deduct the meals" to the following question: "After considering the tax advice, how do you believe that Alex will choose to treat the meal expenses on their tax return?" Participants then indicate the strength of their opinion by answering how sure they are about their decision on an 11-point scale anchored on 0 ("Not sure at all") to 10 ("Extremely sure"). *Reporting_Decision* is constructed by coding a participant's decision to deduct (not deduct) the expenditure as +1 (-1) which is then multiplied by the strength of their belief, which is then multiplied by -1 for participants in conservative conditions. This procedure creates a measure that ranges from -10 (low congruency between reporting decision and tax advice) to 10 (high congruency between reporting decision and tax advice).

Appendix 1

Experimental Task and Manipulations

Participants begin the experimental task after passing a brief quiz assessing their comprehension of the US tax system. Participants move through various screens, represented below by bold headings. All participants in all five groups view all screens, unless otherwise indicated.

Alex's Background

Your task in this study is to decide what you think "Alex" will do in a particular situation. First, here is some background information about Alex.

Alex is a self-employed web developer who launched their own virtual business last year. Alex offers comprehensive website design and development services to a diverse clientele across the U.S.

Virtual Lunch Meetings

To provide updates on the progress of ongoing projects, Alex regularly hosts virtual lunch meetings with clients from a local café.

During these video calls, Alex often buys lunch. Alex doesn't pay for clients' lunches, because the clients are not physically at the café during these meetings.

Tax Issue

It is now tax season, and Alex is preparing their own tax return using commercial tax software. Alex encounters a dilemma regarding business expenses.

Alex needs to decide if the meals purchased at the café during the virtual business lunch meetings qualify for a deduction under the tax law.

After doing some research, Alex finds:

- The most recent IRS audit statistics show the chance of being audited by the IRS is, on average, less than 1%

- In some cases, the tax law allows for meals to be deducted, but the rules are complicated and not very clear

Tax Issue continued

Alex has spent a significant amount on meals during the virtual lunches throughout the year.

Claiming the deduction would result in a substantial tax savings.

However, Alex doesn't know if the deduction is allowed by the tax law.

Seeking Advice from CPA [AI chatbot] (treatment groups)

Given the uncertainty, Alex decides to seek advice. Alex consults a Certified Public Accountant (CPA) who offers free preliminary consultations [an artificial intelligence (AI) chatbot designed to answer tax questions]. Alex does not plan to hire the CPA because Alex will prepare their own return using tax software, but takes the opportunity to ask a question.

The CPA [AI chatbot]has been trained on current U.S. tax laws, updates, and regulations.

Alex emails the CPA [asks the AI chatbot] the following question:

"I'm a self-employed web developer who runs a virtual business. I regularly host virtual meetings with my clients over video calls from a local café. Can I deduct the cost of meals that I purchase for myself during these virtual meetings as a business expense on my tax return?"

Reply from a CPA or an AI chatbot, aggressive [conservative] (treatment groups)

"The tax law permits a deduction for meals that are related to active business conduct. However, the physical presence of a client at a meal helps substantiate the business purpose of the meal. The tax law does not address the virtual presence of a client, making this a gray area.

Based on my interpretation of the tax law, I believe your meal expenses are [are not] deductible."

Additional Information (control group)

After doing additional research, Alex's understanding is as follows:

The tax law permits a deduction for meals that are related to active business conduct. However, the physical presence of a client at a meal helps substantiate the business purpose of the meal. The tax law does not address the virtual presence of a client, making this a gray area.

Appendix 2
Variable Definitions

Variable	**Definition**
Advisor	Advisor type (CPA or AI)
Nature	Nature of advice (Aggressive or Conservative)
Deduct	"How do you believe that Alex will choose to treat the meal expenses on their tax return?" 1 = Alex WILL deduct the meals, -1 = Alex WILL NOT deduct the meals
Confidence	"You indicated you believe Alex WILL/WILL NOT deduct the meals. How sure are you?" 0 ("Not sure at all") to 10 ("Extremely sure") scale
Aggression	*Deduct * Confidence*
Reporting_Decision	Aggression multiplied by negative one for participants in Conservative conditions
Reliance	"How likely do you believe it is that Alex will rely on the tax advice?" 0 ("Not at all likely") to 10 ("Very likely") scale
Advocacy	Mean of responses to 9 questions client-advocacy scale (see Table 6 for questions)
Benevolence	Mean of responses to 5 questions on benevolence scale (see Table 6 for questions)
Expertise	"How would you rate the level of tax expertise of the Certified Public Accountant (CPA)/AI chatbot Alex consulted with?": 1 ("Very low") to 7 ("Very high") scale
Audit	"On average, what percentage of individual tax returns do you think the Internal Revenue Service (IRS) audits each year? Enter a percentage from 0% - 100%."
Detection	"Suppose someone takes a deduction that is not supported by tax law and is audited. In your opinion, what is the probability the IRS will discover the underreporting? Enter a percentage from 0% - 100%."
Penalty	"Suppose someone takes a deduction that is not supported by tax law. In your opinion, if this was discovered by the IRS, how severe would the consequences be?" 1 ("Very mild") to 7 ("Very severe") scale
Fair	"Do you think the amount of taxes you pay is about right, too much or too little given your income?" 1 ("Way too little") to 7 ("Way too much") scale
Moral	"You feel a moral obligation, based on your own personal feelings about what is right and wrong, to be completely honest when filling out your tax return." 1 ("Strongly disagree") to 7 ("Strongly agree") scale
Risk	"In general, you prefer taking risks rather than avoiding risks." 1 ("Strongly disagree") to 7 ("Strongly agree") scale
Realism	"How realistic did you find the scenario presented in the study?" 1 ("Very unrealistic") to 7 ("Extremely realistic") scale