Differences in Algorithm Reliance for Operating and Reporting Decisions: The Influence of Accuracy Rates on Weight of Advice

Sanaz S. Aghazadeh
Louisiana State University
sanaz@lsu.edu

Tamara A. Lambert
Lehigh University
tal413@lehigh.edu

Jenny W. Ulla
University of Illinois – Urbana-Champaign
ullaje1@illinois.edu

November 2023

Acknowledgements: We thank Ben Commerford and Owen Brown for helpful comments and Martin Persson for his help with this project. We thank workshop participants at Lehigh University. We gratefully acknowledge financial support provided by the College of Business at Lehigh University and the Francis L. Durand Professorship at the E.J. Ourso College of Business.

Differences in Algorithm Reliance for Operating and Reporting Decisions: The Influence of Accuracy Rates on Weight of Advice

ABSTRACT

We examine how managers' decision-making context (operating versus reporting decision) differentially affects reliance on an algorithmic versus human advice source. With more pressure to justify a judgment to an evaluator, we expect reporting decisions to lead to more critical analysis of the available information than operating decisions; we manipulate the available information by varying the presence or absence of an accuracy rate. We also expect the presence of a high but imperfect accuracy rate to increase people's reliance on the human advisor, but to have a more nuanced effect on reliance on an algorithmic advisor because people penalize algorithms for erring more than humans. Using a 2×2×2 design to manipulate decision-making context, advice source, and accuracy rate, we predict and find that the positive impact of a high but imperfect accuracy rate on people's reliance on human advisors will attenuate for algorithmic advisors more in reporting versus operating contexts.

Keywords: algorithm reliance, accountability, decision making, accuracy, advice reliance
Data: Available upon request

I. INTRODUCTION

Companies (e.g., manufacturers, financial service firms, audit firms) are investing significant resources in developing artificial intelligence (AI) systems capable of assisting professionals with complex business and financial decision-making (Deloitte 2018b; KPMG 2018; Bloomberg Tax 2020; CPA Canada and AICPA 2020; KPMG 2020a). For example, financial service companies have developed AI technology to provide managers with financial forecast estimates such as future cash flow (Deloitte 2019a). Managers will likely use these forecasted estimates to make both operating and reporting decisions. Operating decisions involve strategic performance choices (e.g., whether or not to purchase or sell an asset) while reporting decisions involve measuring and disclosing the results of previously made performance decisions (e.g., disclosing the value of held assets) to an outside party (i.e., users of financial statements, stakeholders, and regulators) (Bloomfield 2018; Bentley, Bloomfield, Bloomfield, and Lambert 2023; Buchanan, Griffith, Lambert, and Perreault 2023). Research finds that in some cases professionals are less willing to rely on an algorithm-based system as they are advice from a human, but that this effect is highly contextually dependent (Dietvorst, Simmons, and Massey 2015; Dietvorst 2016; Commerford, Dennis, Joe, Ulla 2023). Given that operating and reporting decisions capture many common business contexts, research is needed to examine whether and how the decision-making context differentially affects reliance on an algorithm relative to a human advisor.

A key distinction between operating and reporting decisions is the degree of accountability that is present. Accountability is typically defined in psychology literature as the pressure to justify a judgment or decision to an evaluator (Lerner and Tetlock 1999). Reporting involves collecting, gathering, and disclosing information to others that will be reviewed. As

such, this setting induces greater accountability in elevating concerns related to the perceptions of others (Dirsmith and Lewis 1982; Tetlock et al. 1989; Ashton 1990; Harker 2003). Prior literature on accountability shows that higher-accountability contexts lead people to conduct a more critical analysis of information (e.g., Dezoort, Harrison, and Taylor 2006).

Such critical analysis of information leads people to more carefully consider potentially relevant cues and to increase attention on the cues used to form their judgments and decisions (Lerner and Tetlock 1999). One of the characteristics likely to be provided about an advisor is its accuracy rate, as industries adopting AI systems are "narrowly focused on model accuracy" (Kesari 2021) and emphasize greater transparency in the performance of algorithmic systems (CPA Canada and AICPA 2020). Accuracy contributes to an advisor's credibility and in general, the advice-taking literature finds people place greater weight on and are more likely to rely on advice from more accurate advisors (Nowlis and Simonson 1997; Yaniv and Kleinberger 2000; Sniezek and Van Swol 2001; Mercer 2005; Patt, Bowles, and Cash 2006; Bonaccio and Dalal 2006; Sah, Moore, and MacCoun 2013). However, it is unclear whether knowledge of an algorithmic advisor's accuracy rate would have the same effect on advice reliance relative to human advisors, particularly in higher-accountability (e.g., reporting) contexts. While prior literature has found that under greater accountability people often exhibit greater effort and produce higher-quality decisions, the "benefits" of greater accountability may have certain limits, as (in some cases) anticipating the preferences of a third party may result in less optimal decision making (Messner 2009; Donnelly and Donnelly 2023). Thus, we examine whether and how knowledge of an important advisor characteristic—the advisor's accuracy rate—differentially affects reliance on advice from a human relative to an algorithm in a reporting versus operating decision context.

In a reporting context, managers may expect the party to whom they are accountable will value the algorithm's advice for its objectivity (e.g., Castelo et al. 2019), suggesting a high accuracy rate may be expected to further increase reliance on the algorithm. However, this may not be the case if the high but imperfect accuracy rate suggests the algorithm is prone to error. Research indicates algorithms are penalized more for errors and imperfections relative to humans (Highhouse 2008; Dietvorst et al. 2014; Bogert et al. 2021; Zhang, Chong, Kotovsky, and Cagan 2023).[1] In higher-accountability contexts, where people more critically-analyze information, the imperfections of the advisor will be highlighted by a less-than-perfect accuracy rate, which may counteract the positive boost a high accuracy rate might otherwise be expected to give to advice source reliance. Thus, we examine whether the positive impact of a high but imperfect accuracy rate on people's reliance on human advisors attenuates more for algorithmic advisors in reporting versus operating settings.

We conduct a 2×2×2 between-participants experiment that manipulates (1) advice source (human versus algorithm), (2) knowledge of a high but imperfect accuracy rate (present versus absent) and, (3) the decision context (operating versus reporting). Both contexts task participants with providing an estimate of the value of a house that we ask them to assume is theirs. We use a setting that should be familiar to general population participants and is designed to be able to generalize to the context of a manager making a fair value estimate, as our theory only requires a setting that evokes ecological validity (i.e., participant-specific realism). Participants estimate the house's value to decide whether or not to sell the house (operating decision) or to report on the

[1] Such research is supported by popular press articles and technology developers. For example, in an article on why ChatGPT has not "ignited the employment apocalypse," the authors note it, "still gets things wrong" (Kantrowitz and Gorman 2023); Steve Wozniak, co-founder of Apple recently warned attendees at the Digital X 2022 event hosted by Deutsche Telekom, "The trouble is it does good things for us, but it can make horrible mistakes" (Sauer 2023).

application for a small business loan (reporting decision). All participants are provided with photos and information on the house (i.e., number of beds and baths, house and lot size, house location) along with the sales price on comparable properties that were sold. Our dependent variable is based on the judgment-advisor system paradigm (Sniezek and Buckley 1995), which requires people to make a judgment, receive advice, and then finalize their judgment. Thus, participants first provide their initial estimate of the house's value. Next, participants are informed they will receive a house value estimate provided by either the eValEstate algorithm (i.e., web-based algorithmic valuation tool) or Val Jones (i.e., human real estate agent). The eValEstate algorithm is described as a leading real estate marketplace website that develops estimates based on proprietary statistical models and algorithms using historical data. Val Jones is described as a leading licensed real estate agent that develops estimates based on professional judgment-based experience. Next, participants are provided with the advisor's estimate of the house value. Lastly, participants provide their final estimate of the house's value. Our dependent variable, weight-of-advice, captures the degree to which participants adjust their estimate to be in accordance with the advisor's recommended value.

We find that for an operating decision context, the effect of providing a high but imperfect accuracy rate has a positive effect on advice source reliance relative to when an accuracy rate is not provided, regardless of whether the advisor is a human or an algorithm. For a reporting decision context, the accuracy rate continues to have a positive effect for human advisors, but that effect is attenuated for algorithmic advisors. This is consistent with our expectation that in higher-accountability settings, people are more likely to critically analyze diagnostic information and focus on the advisor's imperfection or likelihood of making an error when provided with a high but imperfect accuracy rate. In additional analyses, we find the

interactive effect of accuracy rate and decision context on advisor reliance is mediated by perceptions of credibility for a human advisor, but not for algorithmic advisors. Thus, while providing a relatively high but imperfect accuracy rate generally should affect credibility perceptions of the advisor, this is not the case with algorithmic advisors for people making reporting decisions.

Consistent with some of our theoretical arguments (e.g., Castelo et al. 2019), but not specifically hypothesized, results also reveal that when an accuracy rate for the advisor is not provided, people more heavily rely on an algorithmic advisor under a reporting relative to an operating decision-making context. We expect this preference is due to objectivity being highly valued for reporting tasks (e.g., Kadous, Koonce, and Towry 2005), and the preconceived notion that algorithms are less biased decision makers (e.g., Miller 2018; Castelo et al. 2019). While there is extensive research supporting objectivity being valued (e.g., Tetlock 1983; Kadous et al. 2005; Russell and Norvig 2010; Garvey, Kim, and Duhachek 2023), there is less evidence regarding the relative preference for algorithmic versus human advisors based on the decision-maker having an objectivity goal. To test whether the relative preference for algorithms increases when people have an objectivity goal, we run a supplemental experiment within the reporting decision-making context/accuracy rate absent condition, where we manipulate the decision maker's goal as objectivity or accuracy.[2] We find that when people have an objectivity goal, their relative preference for algorithms (versus humans) is higher than when they have an accuracy goal. Thus, we conclude that people seem to highly value or prefer algorithms when they are seeking to appear objective. The results of our main experiment suggest however, that

[2] An accuracy goal provides us with a benchmark, "optimal" goal to test the impact of an objectivity goal against, allowing us to hold constant that the participant is given some goal.

algorithms do not get a further "credibility boost" in such settings from information indicating a high but imperfect accuracy rate.

This study makes theoretical contributions by extending the literature related to reliance on algorithm-based evidence. We find that decision context (operating versus reporting) differentially affects how people perceive contextual cues to determine whether to rely on an algorithm relative to a human advisor. We contribute to contemporary research that has started differentiating operating versus reporting decisions (Bentley et al. 2023; Buchanan et al. 2023). The study also highlights that findings from prior literature on advice reliance—such as the notion that disclosure of a high accuracy rate should increase reliance— may not be uniformly applied to algorithmic advisors. Specifically, a high accuracy rate increases reliance on a human advisor, but not reliance on an algorithm for a reporting decision.

This study also has important implications for practice. As companies continue to encourage managers to seek and utilize advice to improve the quality of developing and evaluating judgment-based tasks, such as complex estimates (Ranzilla, Chevalier, Herrmann, Glover, and Prawitt 2011; PCAOB 2016; Deloitte 2019b), they may be interested to know that providing an algorithmic advisors' accuracy rate may not lead to increased reliance on advice; and that any efforts they make toward increasing reliance on algorithmic advice should consider the decision-making context. Overall, our findings should be useful to accounting and management researchers and practitioners, and others interested in understanding algorithm reliance in operating and reporting settings.

II. BACKGROUND AND HYPOTHESES DEVELOPMENT

Algorithm use

Advanced technology is being developed to perform business-related tasks historically

reserved for humans. Advanced algorithm-based technology now surpasses human capability in some financial reporting settings and areas of the audit (Antretter, Blohm, Siren, Grichnik, Malmström, and Wincent 2020; Leonard, Haarman, DeMelis, and Youngberg 2020; Ding, Lev, Peng, Sun, and Vasarhelyi 2020). For example, companies use artificial intelligence to develop estimates of financial statement items and to synthesize unstructured data for ESG reporting purposes (FRC 2019; EY 2020). Likewise, audit firms deploy artificial intelligence (AI) technology or "chains of algorithms" to assist auditors with activities such as detecting fraud and evaluating commercial loan grades (KPMG 2016, 4; Deloitte 2017). Although algorithm-based technology can significantly improve financial reporting quality, especially in complex areas and those involving uncertainty, these benefits will only materialize if accountants are willing to appropriately incorporate algorithm-developed input into their financial statement estimates (Ding et al. 2020).

Studies suggest people generally believe algorithms are designed to function in a predictable manner, in accordance with the constraints of their programming and without independent motives or intentions, and people are therefore likely willing to rely on algorithms when making objective decisions (Russell and Norvig 2010; Castelo et al. 2019; Kim and Duhachek 2020; Garvey, Kim, and Duhachek 2023). However, research also finds people are sometimes more hesitant to rely on algorithmic advice relative to human advice, which is often used as a baseline when studying algorithm reliance (Dietvorst et al. 2015; Dietvorst 2016; Commerford, Dennis, Joe, and Ulla 2022), particularly when the advice is low-quality or the algorithm errs. For example, one study manipulates the quality of advice provided to people estimating the number of individuals in a photograph (Bogert, Schecter, and Watson 2021); participants relied less on lower quality advice than higher quality advice, particularly when the

source of advice was an algorithm.[3] Another study using a team chess-game setting finds that less accurate teammates are perceived as more competent if people believe their teammate is a human rather than an algorithm (Zhang, Chong, Kotovsky, and Cagan 2023). An additional study finds that people reduce their reliance on erring algorithms more quickly than erring humans when forecasting the success of MBA applicants; this effect persists even when people are incentivized to make accurate predictions (i.e., given bonuses), and despite observing the algorithm make smaller average forecast errors relative to a human advisor (Dietvorst et al. 2014). In an audit setting, after specialist errors become more salient via second-hand information, auditors more heavily discount audit evidence from AI systems relative to human specialists (Commerford et al. 2023). Collectively, this research indicates people are highly sensitive to imperfect advice from algorithms (Castelo et al. 2019; Burton, Stein, and Jensen 2020); some research suggests this may be driven by the general belief that algorithmic errors are systematic (Highhouse 2008).

Notably, much of the research on algorithm reliance focuses on settings in which people are not expected to justify their choices to others (Dzindolet, Pierce, Beck, and Dawe 2002; Castelo et al. 2019; Yeomans et al. 2019; Dietvorst and Bharti 2020). These settings range from forecasting student performance in graduate school and estimating people's weights, to predicting the popularity of songs (Dietvorst et al. 2015; Logg, Minson, and Moore 2019). While algorithm reliance has been studied in contexts with lower and higher levels of justification pressure, no study of which we are aware manipulates the decision-making context to focus on how justification pressure affects algorithm reliance. Thus, we explore the impact of an operating

[3] The lower quality advice provided by the advisor was 100 percent too high and high quality advice provided the correct estimate.

versus reporting context on algorithm reliance, which should be helpful for companies, investors, and others in society interested in algorithm use within business contexts.

Operating versus Reporting Decision-Making Contexts

Operating decisions involve strategic performance choices (e.g., whether or not to purchase or sell an asset), while reporting decisions involve measuring and disclosing the results of previously made strategic performance choices (e.g., disclosing the value of held assets).[4] Operating decisions tend to focus on performance and achieving strategic goals; reporting decisions involve discretion in summarizing and reporting to "outside" stakeholders (e.g., creditors and investors) (Bloomfield 2018; Buchanan et al. 2023). Reporting settings tend to invoke higher potential for work to be questioned and reviewed, which induces greater accountability and elevates concerns related to the perceptions of others (Dirsmith and Lewis 1982; Tetlock et al. 1989; Ashton 1990; Harker 2003). Accountability represents an individual's expectation that they will have to justify a decision or position to another party who may evaluate the individual (Lerner and Tetlock 1999). For example, financial statement users often hold accountants accountable for inaccuracies or misrepresentations of the company's financial health in disclosures (Kadous 2000; Lowe, Reckers, and Whitecotton 2002; Backof 2015; Brasel, Doxey, Grenier, and Reffett 2016). Thus, decision makers in these types of reporting settings expect their judgments to be scrutinized. Accountability can encourage emphasis on the provision of socially acceptable, or audience-preferred information (Brown 1999). Research finds that people in higher-accountability settings tend to make decisions tailored to the

[4] While there can be overlap between an operating and reporting decision, Bloomfield (2018) distinguishes operating decisions as those that "generate the raw data" used in reporting, and reporting decisions as exercising discretion over "how raw data are transformed" and disclosed (64-65). Bentley et al. (2023) provide empirical support for the distinction; across four experiments the paper finds that the acceptability of distorting operating decisions (e.g., real earnings management) is perceived differently than the acceptability of distorting reporting decisions (e.g., accruals management).

audience's preferences (Frink and Ferris 1998; Koreff and Perreault 2023), engage in self-criticism, and try to anticipate the audience's objections to their conclusions (Tetlock 1983; Tetlock et al. 1989; Fehrenbacher, Kaplan, and Moulang 2023).[5] In a reporting context, given the discretion involved, managers may expect the party to whom they are accountable will value the algorithm's advice for its objectivity (e.g., Castelo et al. 2019).

Given greater levels of justification pressure in higher-accountability contexts (e.g., reporting decisions), people will likely exert more effort, engage in a more thorough analysis of information, and (importantly) conduct a more *critical* analysis of available information as they consider how much to rely on advice sources (Dezoort, Harrison, and Taylor 2006). Higher-accountability settings often lead people to increase attention on potentially relevant information (Lerner and Tetlock 1999), and to focus more on details they believe are diagnostic of the problem at hand (Brown 1999). While prior literature has documented many benefits of greater accountability (e.g., Koonce et al. 1995; Asare et al. 2000; Stefaniak et al. 2017), accountability pressure may not always improve decisions. When people anticipate preferences of a third party (i.e., "evaluative others") it may result in less optimal decision making in lieu of intuitively high-quality decisions (Messner 2009; Donnelly and Donnelly 2023, 2).[6] We contend that the way in which potentially relevant information is processed depends on people's beliefs about how third parties will perceive such information. In this study, we expect that people in a reporting decision context compared to an operating context will engage in a more thorough analysis of the advisor and their associated characteristics as people consider relying on advice sources.

[5] For example, auditors who feel greater accountability provide more thorough justifications for conclusions (when audience views are unknown) compared to auditors who are not held accountable by the review process (Koonce, Anderson, and Marchant 1995), and they put forth more effort in general (Asare et al. 2000; Stefaniak, Houston, and Brandon 2017; Bhaskar 2020).

[6] For example, when subordinates become aware of supervisor's preferences in a going-concern audit task, subordinates are more likely to provide evidence more consistent with the supervisors' views (Wilks 2002).

Accuracy Rates

One of the characteristics likely to be provided about an advice source is its reliability or accuracy rate, as industries adopting AI systems are "narrowly focused on model accuracy" (Kesari 2021). An advisor's accuracy rate is a highly influential characteristic that often signals the credibility or trustworthiness of the advisor (Nowlis and Simonson 1997; Yaniv and Kleinberger 2000). Providing information about the advice source's reliability, such as an accuracy rate, is perceived to be transparent and can increase reliance on the advice source (Harries, Yaniv, and Harvey 2004; Bonaccio and Dalal 2006; Sniezek and Van Swol 2001). For example, the AICPA has called for greater transparency in algorithmic systems so that auditors and accountants can assess their prediction accuracy (CPA Canada and AICPA 2020). Prior literature has established that communications from sources perceived to be higher in credibility are generally relied upon more and more heavily weighted by decision makers such as investors (Hirst, Koonce, & Venkataraman, 2008; Hodge, Hopkins, & Pratt, 2006; Mercer, 2005), and that revealing information on advisors, such as a high accuracy rate, has a positive effect on advice reliance (Patt et al.2006). Collectively, literature suggests that providing a high accuracy rate may signal high advisor credibility whether the advice source is a human or an algorithm.

People perceive algorithms as particularly well-suited for objective tasks (Russell and Norvig 2010; Castelo et al. 2019; Kim and Duhachek 2020; Garvey, Kim, and Duhachek 2023). In fact, people may prefer algorithms over humans for reporting contexts, where perceived objectivity can possibly alleviate audience's concerns that the reporter is using the discretion naturally involved in a reporting decision (e.g., Bloomfield 2018; Bentley 2023) to bias the estimate provided.[7]

[7] Indeed, research demonstrates that financial managers and accountants prefer to provide and receive information that is based on quantitative and objective information, rather than qualitative and ambiguous information (Dirsmith

While algorithms are perceived to be well suited for objective tasks (e.g., Castelo et al. 2019) and objectivity is highly valued for reporting tasks (e.g., Kadous et al. 2005), prior literature also suggests algorithms and algorithm use are penalized for errors and imperfections more than humans and human advice-source use (Highhouse 2008; Dietvorst et al. 2014; Bogert et al. 2021; Zhang, Chong, Kotovsky, and Cagan 2023). We expect that in a reporting (i.e., higher-accountability) context, peoples' increased attention on the accuracy rate will highlight the imperfections of the advisor more than in an operating (lower-accountability) context (Lerner and Tetlock 1999), as accuracy rates less than 100% suggest the advice source errs. Specifically, when provided a high but imperfect advisor accuracy rate in a reporting setting, people may focus on the less-than-perfect accuracy rate, which may attenuate any overall positive effect the accuracy rate may otherwise have on credibility perceptions. Thus, the algorithm may not get a further credibility boost from a high accuracy rate when the accuracy rate also implies the algorithm errs. We expect a high but imperfect accuracy rate will have a differential effect on advice reliance for an algorithmic advisor relative to a human advisor, particularly in a reporting setting compared to an operating setting.

Hypothesis: The positive impact of a high but imperfect accuracy rate on people's reliance on human advisors will attenuate for algorithmic advisors more in reporting versus operating decision-making contexts.

and Lewis 1982; Kadous, Koonce, and Mercer et al. 2005; Joe, Vandervelde, and Wu 2017). Decision-makers often rely on reporting methods that employ systematic, consistent processes (i.e., with less discretion and room for bias), as resulting judgments (such as estimates) from such processes are easier to document, review, and defend to others (Kunda 1990; Shankar and Tan 2006; Piercey 2011; Joe, Vandervelde, and Wu 2017). Finally, algorithm reliance studies set in reporting contexts generally find people discount advice provided by algorithms (relative to humans), especially with high inspection risk, when people feel a deficit in their digital knowledge and skills, and (importantly) when contradictory evidence appears more objective (Cao, Duh, Tan, and Xu 2022; Commerford et al. 2022; Commerford and Holman 2023; Peecher, Pietsch, Stirnkorb, and Yamoah 2023).

III. EXPERIMENTAL DESIGN AND METHOD

Design, Context, and Participants

We test our theory using a 2×2×2 between-participants design that manipulates (1) *advice source* (human versus algorithm), (2) knowledge of a high but imperfect *accuracy rate* (present versus absent) and, (3) the *decision-making context* (operating versus reporting). Both contexts involve the participant providing an estimate of the value of a house that we ask them to assume is theirs. For the *operating context*, participants are estimating the value to decide whether or not to sell the house, and for the *reporting context* they are estimating the value to report on a small business loan application.[8] We use a setting that should be familiar to general population participants, designed to be able to generalize to business contexts, to examine our theory. We use such a setting because our theory does not require participants to be placed in a specific type of task, just one that evokes ecological validity for our *decision-making context* manipulation; and because using general population participants preserves the valuable resource of experienced participants (such as managers or auditors). We obtained 410 participants from Amazon's Mechanical Turk (MTurk).[9] We chose MTurk workers as appropriate participants given the task is one that general population participants should be able to understand and perform.[10]

[8] In the *reporting context* condition, participants are told that the estimated value of their house will be reported on a loan application. This context captures greater accountability (relative to the operating context) given that loan applications are reviewed by a third-party (i.e., potential creditor) which places a responsibility on the individual to explain and justify the reported value of the house.

[9] The experiment in this paper were approved by the Institutional Review Board. In this experiment, we required participants to meet the following three criteria: (1) no participation in our three pilot tests; (2) overall Human Intelligence Task (HIT) approval rate of 99% or greater; (2) overall number of HITs approved is greater than 500 (3) located in the United States (Farrell, Grenier, and Leiby 2017; Dennis, Goodson, and Pearson 2020). Participants completed the experiment in 7.06 minutes on average and this does not vary across experimental conditions (p = 0.22, untabulated).

[10] Participants on average are 40 years old and report an average level of knowledge of pricing and selling houses (mean of 3.92 on a scale from 1 = "Very Poor" and 7 = "Very Good"). There are no significant differences across experimental conditions based on demographic measures (see Footnote 8 for demographic measures).

Participants received $1.00 to participate.[11]

We carefully designed our experiment to capture the tensions and incentives that affect accounting-related business decisions. The task parallels tensions relevant to managers and auditors who often develop estimates (i.e., financial forecast estimates such as future cash flows) with an incomplete set of information as they make decisions (Deloitte 2019a; Rowe 2019). Further, managers and auditors often seek advice from others and utilize professional judgment to reconcile conflicting information (Griffith 2018; Griffith 2020). The *operating context* resembles a situation in which a manager must make the strategic decision of whether and when to dispose of an asset, while the *reporting context* embeds features that closely resemble the pressures and incentives that a manager or auditor would face as they use their discretion to estimate values to report to others outside their organization. Similar to the type of scrutiny given to reported values managers disclose and auditors audit, participants in the *reporting context* should bring to the experiment the situational understanding that their estimate will come under scrutiny by a loan officer, and there could be negative consequences if the house value reported on the legal document is considered to be grossly inaccurate or biased. The post-experimental measures confirm the ecological validity of our *decision-making context* manipulation.

Task, Additional Manipulations, and Dependent Measure

We placed participants in either the *operating* or *reporting decision-making context*, each described above. Participants receive information about the house and on comparable sales.

[11] Following prior literature's recommendation on improving the reliability of research, we asked an attention check question to ensure participants were reading the case after participants read general instructions related to their role. Participants responded to the following question: What type of task will you complete in this case? With the choices (a) Estimate value of real estate to report on a small business loan application, (b) Estimate value of real estate to sell your home, (c) Rate TV shows, and (d) Estimate nutritional facts of food. Participants who responded incorrectly were redirected back to the screen of information providing detail on the type of task they will complete. Participants were prompted with the same question asking which type of task they will complete. If participants responded incorrectly again, we concluded these participants were not paying attention and they were directed to close the survey.

Next, participants provide their initial estimate of the house value. Afterward, they receive an assessment from either a real estate agent named Val Jones, described as a leading realtor who uses professional judgment based on experience (*human advice source*), or an online real estate marketplace called eValEstate that uses proprietary statistical models and algorithms based on historical data (*algorithm advice source*). Both advice sources use comparable sales to provide the participant with an assessment of the house's estimated value. The case emphasizes this assessment is not an official real estate appraisal. Instead, it is independent and unbiased advice regarding the house's market value based on publicly available data. Participants are then informed that on average, 90% of the advisor's (eValEstate or Val Jones) assessments are accurate (i.e., within $5,000 of the actual sales price) (*accuracy rate present*) or are not provided with an accuracy rate (*accuracy rate absent*).[12] Participants receive the advisor's estimate of their house value, which is 6% less than the participant's initial house value estimate.[13] Following this, participants provide a final assessment of the house value and then answer a series of case-related and demographic questions.

Following psychology research on advice taking, our primary dependent variable is advice utilization, which is calculated as weight of advice (WOA) (e.g., Yaniv 2004; Önkal et al. 2009; Kadous, Leiby, and Peecher 2013). Expressed mathematically:

$$WOA = \frac{(Initial\ estimate - Final\ estimate)}{(Initial\ estimate - Advisor's\ estimate)}$$

[12] In a pilot study where participants are provided with advice source accuracy rates of 90% and 99%, results show no significant differences on WOA based on the variation in accuracy rates ($p = 0.85$). Thus, we conclude that a 90% accuracy rate represents a relatively high accuracy rate, and our results should remain unchanged if we provided participants with a higher one.

[13] Prior advice-taking literature has found that weight placed on advice decreases as the distance between the advice and initial opinion increases (Yaniv 2004). Therefore, we keep the relative distance equal in all conditions, by setting the E-Val system's estimate as 6% less than the participant's initial fair value estimate.

WOA captures the extent to which an individual incorporates the advisor's recommendation into their final estimate. WOA values can range from 0 to 1.[14] If participants' final estimate is equal to their initial estimate, then WOA would equal 0 and represent a complete discounting of the advisor's recommendation. In contrast, if there is a complete shift of participant's initial estimate to the advisor's estimate, WOA would equal 1, which represents full weighting of the advisor's recommendation. Partial reliance on the advisor's recommendation results in values ranging between 0 and 1. Following measurement of our dependent variable, we ask participants post-experimental and demographic questions.

IV. RESULTS

Manipulation Checks and Ecological Validity

To assess the effectiveness of the *advice source* manipulation, we ask participants to answer the following question: "Among the two descriptions below, which one more accurately describes eValEstate (algorithm condition)/Val Jones (human condition)?" on a 7-point scale with endpoints 1 = "Definitely a Human" and 7 = "Definitely an Algorithm". Participants in the *algorithm* condition (mean = 6.29) responded higher on the scale, while participants in the *human* condition (mean = 2.76) responded lower on the scale ($t_{408} = 24.72$, $p < 0.01$, untabulated).[15] To assess the effectiveness of the *accuracy rate* manipulation, we ask participants to answer the following question: "How reliable was the advice provided by [eValEstate / Val Jones]?" on a 7-point scale with endpoints 1 = Not at all Reliable and 7 = Very Reliable. The mean ratings in the *present* condition (5.20) are higher than the *absent* condition (4.85), as expected given the accuracy rate provided was designed to be perceived to be high ($t_{408} = 2.76$ p

[14] Following previous research, we truncate the WOA value to 1 if the participant "overshoots" the advice (i.e., participants' final estimate is less than the E-Val system's estimate) (Gino and Moore 2007; Gino, Shang, and Croson 2009).

[15] Consistent with our directional prediction, all reported p-values are one-tailed equivalents, unless otherwise noted.

< 0.01, untabulated).[16] To assess the effectiveness of the *decision-making context* manipulation, we ask participants to answer the following question: "Among the two descriptions below, which one more accurately describes why you were estimating your house's value?" on a 7-point scale with endpoints 1= "Definitely to report the value on a loan application" and 7 = "Definitely to determine the value to decide whether to sell my house." The mean ratings in the *operating* condition (mean = 6.41) are lower than the *reporting* condition (mean = 2.10, $t_{408} = 26.92$, $p <$ 0.01, untabulated). Results indicate successful manipulations of *advice source*, *accuracy rate*, and *decision-making context*.

As discussed in Section III, because we are using general population participants, it is important that our *decision-making context* manipulation evokes similar incentives/pressures as practicing managers or auditors would experience in the *reporting* condition. We measured *justification pressure* and *verification pressure* to assess whether we successfully evoked accountability pressure in this condition. We began each question with, "As you determined how much to rely on the advice provided by [eValEstate/Val Jones]" and asked, "How concerned were you that you would be able to justify using the advice to others?" (*justification pressure*) and "How concerned were you that others could verify the advice you received?" (*verification pressure*). We measured responses on a 7-point scale with 1 = Not at all Concerned and 7 = Very Concerned. We find a significant main effect of *decision-making context* on *justification pressure*, where participants feel more pressure for the *reporting* versus the *operating context* (relative means = 3.82 vs. 3.44; $p = 0.02$), and a similar significant main effect on *verification pressure* (*reporting* mean = 4.00, *operating* mean = 3.62; p = 0.02). These analyses suggest we

[16] We similarly measure participants' perceptions of how consistent and trustworthy the advice provided was. We find a main effect of *accuracy rate* on both measures, where means for the *present* condition are higher than means for the *absent* condition (p's < 0.01).

successfully incorporated relatively higher accountability pressure into the *reporting context* versus the *operating context*.

Hypothesis Testing

Our hypothesis predicts that the positive impact of a high but imperfect accuracy rate on human advice source reliance will attenuate for algorithms more in reporting versus operating decision contexts. As such, our hypothesis predicts a three-way interaction where the effect of *accuracy rate* on advice reliance will be greater for a *human* than for an *algorithm*, particularly for *reporting* versus *operating* settings. Table 1, panel B shows the three-way interaction of *advice source* × *accuracy rate* × *decision-making context* on WOA is significant ($F_{1,402} = 3.96$; $p = 0.02$). To further probe this interaction, we split our sample across *human* and *algorithm* advisor conditions. In Figure 1, Panel A and Panel B, provide visual depictions of the cell means which suggest the presence of an *accuracy rate* increases reliance on both *operating* and *reporting context* conditions for a *human* advisor. For *algorithm* conditions, provision of the *accuracy rate* increases reliance for the *operating context* condition, but does not induce a corresponding increase of reliance in the *reporting* condition. We test the interaction of *accuracy rate* and *decision-making context* using separate 2×2 ANOVAs for each *advice source* condition (Table 1, panel C). Within the *human* condition, the interaction of *accuracy rate* and *decision-making context* is not significant ($F_{1,204} = 1.55$; $p = 0.21$, two-tailed), consistent with a high but imperfect accuracy rate increasing weight of advice for human advisors regardless of the *decision-making context*. Within the *algorithm* condition, the interaction of *accuracy rate* and *decision-making context* is marginally significant ($F_{1,204} = 2.48$; $p = 0.06$) which is consistent with our expectation that the positive impact of a high but imperfect accuracy rate on weight of advice for algorithmic advisors will be attenuated in under reporting contexts. To further

evaluate this interaction, we examine the simple effect of *accuracy rate* within the *algorithm* conditions (Table 1, panel D). The simple effect of *accuracy rate* is significant in the *operating context* (0.53 versus 0.65; $t_{1,101} = 1.96$, $p = 0.05$, two-tailed) where participants exhibited greater reliance on the algorithmic advisor when they were provided with an accuracy rate relative to no accuracy rate. However, the simple effect of *accuracy rate* is not significant in the *reporting context* (0.72 versus 0.70; $t_{1,96} = 0.30$, $p = 0.77$, two-tailed).

We further test the interaction of *advice source* and *accuracy rate* using separate 2×2 ANOVAs for each *decision-making context* condition. For the *operating* condition, the interaction of *advice source* and *accuracy rate* is not significant ($F_{1,204} = 0.11$, $p = 0.74$, two-tailed), and thus, we interpret the main effect of *accuracy rate* on WOA as a true main effect (0.66 versus 0.55; $F_{1,204} = 5.50$, $p = 0.01$, untabulated). That is, the *accuracy rate* increases reliance on both the *human* and *algorithm* and does not differentially affect reliance on the *human* more than the *algorithm* for the *operating context*. This is consistent with our arguments suggesting that in operating settings, people are less likely to conduct a more critical analyses of information (e.g., the advisor's accuracy rate) and as a result are less likely to focus on an algorithm's imperfections. For the *reporting context*, we find a significant interaction ($F_{1,194} = 6.24$; $p = 0.01$) where the presence of an *accuracy rate* increases reliance on the *human* (0.58 versus 0.79, $t_{1,102} = 3.28$, $p < 0.01$) but not the *algorithm* (0.72 versus 0.70, $t_{1,96} = 0.30$, $p = 0.77$, two-tailed).

[Insert Table 1 and Figure 1]

Collectively, these findings are consistent with the arguments that comprise our hypothesis such that the positive effect of a high but imperfect accuracy rate for human advice

source reliance will attenuate for algorithms more in reporting versus operating decision contexts. Next, we consider the role of credibility perceptions in our results.

Additional Analyses: Credibility Perceptions

In developing the hypothesis, we note that having knowledge of a high accuracy rate relative to not having such knowledge could affect weighting of an advisor's advice via perceptions of the advice source's credibility. To capture a holistic measure of credibility, we asked participants to assess "How credible do you believe the advice provided by [eValEstate/Val Jones] was?" on 7-point Likert scales with endpoints 1 = "Not at all Credible" and 7 = Very Credible. We also asked, "How credible do you believe others will think the advice provided by [eValEstate/Val Jones] was?" on 7-point Likert scales with endpoints 1 = "Not at all Credible" and 7 = Very Credible. We use the average of these two measures, and name the variable *credibility.*[17] Using *credibility* as the dependent variable in a 2×2×2 ANOVA we find a main effect of *advice source* ($F_{1,402} = 3.76$; $p = 0.05$, two-tailed, untabulated) where participants find the human (mean = 5.30) to be more credible than the algorithm (mean = 5.08), and a main effect of *accuracy rate* ($F_{1,402} = 8.09$; $p < 0.01$, one-tailed, untabulated) where participants find the advice source to be more credible in the *accuracy rate present* (mean = 5.35) than the *accuracy rate absent* (mean = 5.03) condition.

To explore how *accuracy rate* and *decision-making context* interact to affect *credibility* perceptions and WOA differently for *human* versus *algorithm* advice sources we perform separate mediation analyses for the two *advice source* conditions. Specifically, we conduct a moderated mediation analysis using the Hayes (2022) PROCESS Model 8 for 10,000 bootstrapped samples with *accuracy rate* as the independent variable, *decision-making context* as

[17] Inferences are unchanged if we use factor scores of the two items measuring *credibility.*

the moderator, *credibility* as the mediator, and WOA as the dependent variable. To test for indirect effects, we construct 90% confidence intervals with 10,000 bootstrapped resamples of data with replacement. Figure 2, Panel A presents results (coefficients) of the moderated mediation model for the *human* condition and Figure 2, Panel B presents results for the *algorithm* condition.

[Insert Figure 2]

The analysis presented in Panel A (*human* condition) shows a significant index of moderated mediation (90 percent confidence interval of 0.01 to 0.12, indicating a one-tailed *p*-value less than 0.05). Examining the coefficients reported in the path model provides additional insights into the indirect effect of *accuracy rate* on WOA through *credibility*. Specifically, we observe a positive interaction effect of *accuracy rate* and *decision-making context* on *credibility* ($p = 0.05$). We interpret this as presence (relative to absence) of an accuracy rate has a positive effect on credibility perceptions, especially in a reporting decision-making context versus an operating decision-making context. Lastly, higher perceived *credibility* exhibits a significantly positive effect on WOA ($p < 0.01$). We also find the indirect effect of *accuracy rate* on WOA through *Credibility* is significant for the *reporting* context (90 percent confidence interval of 0.03 to 0.11) but not for the *operating* context (90 percent confidence interval of -0.03 to 0.06). Overall, these results are consistent with our expectations that the effect of *accuracy rate* on credibility perceptions is positive for a human advisor in *reporting* decision-making contexts, and that greater perceived credibility increases the degree to which people incorporate human advice into their final estimate.

The analysis presented in Figure 2, Panel B (*algorithm* condition) does not show a significant index of moderated mediation (90 percent confidence interval of -0.07 to 0.06). Further,

in contrast to the Panel A results, Panel B reveals two nonsignificant indirect effects for the *operating context* (90 percent confidence interval of -0.00 to 0.07) and the *reporting context* (90 confidence interval of -0.02 to 0.07) for the *algorithm advice source*. In contrast with a human advisor's results presented in Panel A, for an algorithmic advisor we do not observe positive credibility perceptions when a high but imperfect accuracy rate is present relative to absent in a *reporting* context. Thus, this evidence suggests provision of an accuracy rate has a differing impact on credibility perceptions, as well as WOA, depending on whether the advisor is an algorithm or a human. Additionally, the moderated mediation analyses suggest it is not the interaction of *accuracy rate* and *decision-making context* that impacts *credibility* perceptions to determine WOA for an algorithmic advisor.

Supplemental Experiment: Objectivity Goal

Purpose, Design, and Participants

Though not specifically hypothesized, consistent with the notion that objective sources of information are preferred in reporting contexts (e.g., footnote 8), we find that people more heavily relied on an algorithmic advisor for reporting decisions relative to operating decisions when its accuracy rate is not disclosed (i.e., Cell G = 0.72 versus Cell E = 0.53; $t_{1,95} = 2.78$, $p < 0.01$, untabulated). In a supplemental experiment we test whether an objectivity goal increases people's relative preference for algorithms (versus humans), which—to the extent that people are more likely to have an objectivity versus an accuracy goal for reporting decisions (e.g., Kadous et al. 2005)—may help to explain this result. An accuracy goal provides us with a benchmark, "optimal" goal to test the impact of an objectivity goal against, allowing us to hold constant that the participant is given some goal.

We design a 1 × 2 between-participants experiment. Within the *reporting decision context/accuracy rate absent* condition, we manipulate whether participants are instructed that their goal is to report an accurate or objective estimate of their house value. Additionally, following prior accounting research that primes a relationship (e.g., Kadous, Leiby, and Peecher 2013) or mindset (e.g., Griffith, Hammersley, Kadous and Young 2015), we ask participants to explain, in their own words, what it means to produce an accurate or objective estimate.[18] We obtained 101 participants from Mturk. We paid each MTurk participant $2.00 in exchange for completing the experiment.[19,20]

Materials, Manipulations, Dependent Measure

We placed participants in a scenario in which they are asked to estimate the value of a home that they assume to be theirs to report on a small business loan application. The house's value was described as depending on the house condition, age, features, neighborhood, etc. Participants are told that their goal is to come up with an objective or accurate estimate of the house's value. Participants are then asked to describe in their own words what it means to provide either an objective or accurate estimate. On the next screen, participants review information about the home (including pictures) and information about comparable sales. Participants are provided with two options of advice sources: eValEstate, an online real estate marketplace website or Val Jones, a licensed real estate agent. After reading the case, participants are asked to indicate which advice source (i.e., human or algorithm) they will select

[18] Two authors reviewed the open-ended responses to ensure that in general, participants understand what it means to provide either an accurate or objective estimate.
[19] 66% of participants are female and on average are 36 years old. Participants report an average level of knowledge of pricing and selling houses (mean of 4.24 on a scale from 1 = "Very Poor" and 7 = "Very Good").
[20] We ask participants several demographic questions at the end of the experiment and find that there are no significant differences across experimental conditions based on these measures (i.e., age, gender, whether participants owned a house, knowledge of pricing and selling houses, and comfort level with advanced technologies).

to help develop their accurate or objective estimate and how strong their preference is for using the selected advice source. Finally, participants complete a post-experimental questionnaire.

Following Bentley, Lambert, and Wang (2021), we calculate our primary dependent variable by combining participants' selection of advisor and their preference strength. We code the selection of eValEstate (i.e., algorithm) as "+1" and the selection of Val Jones (i.e., human) as "-1". We then multiply the selection by the strength of preference to get a variable ranging from -7 = "Strong preference for human" to +7 = "Strong preference for algorithm". We also ask participants to indicate which of the two advice sources (i.e., human or algorithm) is more likely to be prone to bias on a scale ranging from 1 = "Definitely human" to 7 = "Definitely algorithm".

Results

We find that when people have an accuracy goal they prefer humans as an advice source (mean = -1.75), but that this preference disappears when they have an objectivity goal (mean = 0.22; $t_{99} = 1.78$, $p = 0.04$, untabulated). We next conduct a chi-square analysis with the binary choice of source as the dependent variable and find that 54% of participants prefer algorithms when they have an objectivity goal while 63% of participants prefer humans when they have an accuracy goal ($\chi^2 = 2.85$, $p = 0.07$, untabulated). Results further reveal that people view humans as more likely to be prone to *bias* (mean of 3.38 is significantly below the midpoint of 4; $t_{1,99} = 3.23$, $p < 0.01$, untabulated) while algorithms are more likely to be prone to error (mean of 4.27 is significantly above the midpoint of 4; $t_{1,99} = 1.66$, $p = 0.05$, untabulated). Collectively, these results suggest people expect algorithms to be objective advice sources.

V. CONCLUSION

This study provides experimental evidence on how operating versus reporting decision-making contexts, and the provision of accuracy rates, differentially affect managers' reliance on

a human versus algorithmic advice source. We predict and find that while provision of a high but imperfect accuracy rate increases reliance on human advisors, such an effect will attenuate for algorithms more in reporting versus operating contexts. Specifically, we find that in a reporting context, the presence of an accuracy rate increases reliance on a human advisor but not on an algorithmic advisor. Furthermore, we find that perceived credibility of the advice source mediates the joint effect of context and accuracy rate on advice utilization for humans but not for algorithms. We also report results from a supplementary experiment that provides evidence suggesting that people's relative preference for algorithms increases when they have an objectivity versus an accuracy goal.

We make theoretical contributions to accounting literature and extend the literature on algorithm reliance. We find that operating versus reporting decision contexts differentially affect how people perceive relevant information (such as an accuracy rate) and contribute to contemporary research that has started differentiating operating versus reporting decisions in terms of their distortion (Bentley et al. 2023), and in terms of how trait behaviors might manifest differently for both types of decisions within the accounting process (Buchanan et al. 2023). We highlight that prior findings on advice reliance may not be uniformly applied to algorithmic advice sources. Future research can explore whether the lack of a positive effect of an accuracy rate for reporting decisions (in our main experiment) is due to a ceiling effect on algorithm advice reliance for reporting decisions (i.e., due to participants' being more likely to have an objectivity goal for such decisions); however, our mediation results suggest the lack of an effect is more likely due to the accuracy rate not increasing credibility perceptions for the algorithm the way that it does for the human—perhaps because people are simply more biased against algorithms that err than humans that err. To the extent that such a bias is a System 1 (intuition)

versus a System 2 (reasoning) bias (Kahneman 2003), collecting direct evidence on its cause will be challenging.

The results are important for practitioners as companies continue moving towards implementing algorithms to assist with more subjective and complex decisions. Prior literature identified methods of increasing reliance on advisors, such as providing advisor accuracy feedback, increasing perceived confidence of advisors, and providing explanation of advisor procedures (Bonaccio and Dalal 2006). Importantly, we identify that provision of an accuracy rate of an advisor could have differing effects depending on whether the advisor is a human or an algorithm. Thus, prior interventions aimed at increasing reliance on human advisors may not always generalize to algorithmic advisors.

This study is subject to limitations that provide interesting opportunities for future research. First, we examine settings where the advisor provides an estimate that is counter to the individual's intrinsic incentive to report a higher value for a house they assume to be theirs (i.e., to sell it for a higher price or obtain a higher loan value). In our experiment, participants receive the advisor's estimate of the house value, which is 6% less than the participant's initial house value estimate. Future studies could examine whether our results hold in settings where the advisor's estimate aligns with participants' incentives. Second, we provide a succinct description of how the algorithm and the real estate agent develop the house's estimated value. Future research could examine how providing more detailed explanation of the advice source's decision-making process could affect advice reliance. For example, research could examine whether "explainable AI," a tool that allows for more transparency on the algorithm's models, can assist others in understanding the models' behavior and increase reliance on algorithmic advice. Our study only examines factors that affect people's reliance on human or algorithmic

advice. However, human advisors are already developing their own estimates with the assistance of technology. Future research could investigate whether the combination of human advisor and algorithmic advisor results in greater reliance on human advice relative to an only-human and only-algorithm condition regardless of decision-making context. Lastly, future research can examine whether these results generalize to more contextually-rich settings with more experienced business professionals as participants.

References

Antretter, T., I. Blohm, C. Siren, D. Grichnik, M. Malmström, and J. Wincent. 2020. Do algorithms make better-and fairer-investments than angel investors? *Harvard Business Review*. Available at:https://hbr.org/2020/11/do-algorithms-make-better-and-fairer-investments-than-angel-investors#:~:text=By%20predicting%20survival%20probabilities%2C%20the,of%20the%20255%20angel%20investors

Asare, S. K., Trompeter, G. M., & Wright, A. M. 2000. The effect of accountability and time budgets on auditors' testing strategies. *Contemporary Accounting Research*, *17*(4), 539-560.

Ashton, R. H. 1990. Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification. *Journal of Accounting Research* 28: 148-180. http://hdl.handle.net/10.2307/2491253

Backof, A. G. 2015. The impact of audit evidence documentation on jurors' negligence verdicts and damage awards. The Accounting Review, 90(6), 2177-2204.

Bentley, J. W., Lambert, T. A., and E. Wang. 2021. The effect of increased audit disclosure on managers' real operating decisions: Evidence from disclosing critical audit matters. The Accounting Review, 96(1), 23-40.

Bentley, J., Bloomfield, M., Bloomfield, R.,and Lambert, T. 2023. Drivers of Public Opinion on the Acceptability of Distorting Performance Measures. Working paper, University of Massachusetts – Amherst, University of Pennsylvania, Cornell University, and Lehigh University. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2823705

Bhaskar, L. S. 2020. How do risk-based inspections impact auditor behavior? Experimental evidence on the PCAOB's process. The Accounting Review 95 (4): 103–126. https://doi.org/10.2308/tar-2016-0007

Bloomfield, R. J. 2016. What counts and what gets counted. Working paper, Cornell University. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2427106

Bloomberg Tax. 2020. Big four invest billions on tech, reshaping their identities. *Bloombergtax.com* (January 2). Available at: https://news.bloombergtax.com/financial-accounting/big-four-invest-billions-in-tech-reshaping-their-identities`

Bogert, E., Schecter, A. and R.T. Watson.. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports* 11(1): 8028.

Bonaccio, S., and R. S. Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101 (2): 127-151. https://doi.org/10.1016/j.obhdp.2006.07.001

Brasel, K., Doxey, M. M., Grenier, J. H., and A. Reffett. 2016. Risk disclosure preceding negative outcomes: The effects of reporting critical audit matters on judgments of auditor liability. The Accounting Review, 91(5), 1345-1362.

Brown, C. L. 1999. "Do the Right Thing:" Diverging Effects of Accountability in a Managerial Context. *Marketing Science 18* (3): 230-246.

Buchanan, J, E. Griffith, T. Lambert, and S. Perreault. 2023. Dark Triad Traits and the Audited Financial Statements: An Operating, Reporting, and Assurance Framework. Working paper, Providence College, University of Wisconsin-Madison, and Lehigh University.

Burton, J. W., M. K. Stein, and T. B. Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33 (2): 220-239. https://doi.org/10.1002/bdm.2155

Cao, T., Duh, R. R., Tan, H. T., and T. Xu. 2022. Enhancing auditors' reliance on data analytics under inspection risk using fixed and growth mindsets. *The Accounting Review*, *97*(3), 131-153.

Castelo, N., M. W. Bos, and D. R. Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56 (5): 809-825. https://doi.org/10.1177/0022243719851788

Commerford, B. P., S. A. Dennis, J. R. Joe, and J. W. Ulla. 2023. "Alexa, audit loan grades!": Does humanizing artificial intelligence enhance auditor reliance? Working paper, University of Kentucky, University of Central Florida, University of Delaware, and University of Illinois at Urbana-Champaign.

Commerford, B. P., S. A. Dennis, J. R. Joe, and J. W. Ulla. 2022. Man versus machine: Complex estimates and auditor reliance on artificial intelligence. *Journal of Accounting Research* 60 (1): 171-201. https://doi.org/10.1111/1475-679X.12407

Commerford, B. P., and B. A. Holman. 2023. "Tell Me More? Explanation Detail and Auditor Reliance on Human and Non-Human Specialists." Working Paper, University of Kentucky and University of Central Florida.

CPA Canada and AICPA. 2020. The data-driven audit: how automation and ai are changing the audit and the role of the auditor. Available at: https://us.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/the-data-driven-audit.pdf

Deloitte 2017. Business impacts of machine learning. Available at: https://www2.deloitte.com/tr/en/pages/strategy-operations/articles/business-impacts-of-machine-learning.html

Deloitte. 2018. *State of AI in the Enterprise, 2nd Edition.* Available at: https://www2.deloitte.com/content/dam/insights/us/articles/4780_State-of-AI-in-the-enterprise/DI_State-of-AI-in-the-enterprise-2nd-ed.pdf

Deloitte. 2019a. *AI leaders in financial services.* Available at: https://www2.deloitte.com/us/en/insights/industry/financial-services/artificial-intelligence-ai-financial-services-frontrunners.html

Deloitte 2019b. Transparency and responsibility in artificial intelligence. A call for explainable AI. Available at: https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/Innovation/lu-bringing-transparency-ethics-ai.pdf

Dennis, S. A., B. M. Goodson, and C. A. Pearson. 2020. Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting* 32 (1): 119-134. https://doi.org/10.2308/bria-18-044

DeZoort, T., Harrison, P. and M. Taylor. 2006. Accountability and auditors' materiality judgments: The effects of differential pressure strength on conservatism, variability, and effort. *Accounting, Organizations and Society* 31(4-5): 373-390.

Dietvorst, B. J. 2016. People reject (superior) algorithms because they compare them to counter-normative reference points. Working paper, University of Chicago. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2881503

Dietvorst, B.J., Simmons, J., and C. Massey. 2014. Understanding algorithm aversion: forecasters erroneously avoid algorithms after seeing them err. In Academy of management proceedings (Vol. 2014, No. 1, p. 12227). Briarcliff Manor, NY 10510: Academy of Management.

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. http://dx.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.

Ding, K., B. Lev, X. Peng, T. Sun, and M. A. Vasarhelyi. 2020. Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies* 25 (3): 1098-1134. http://dx.doi.org/10.1007/s11142-020-09546-9

Dirsmith, M. W., and B. L. Lewis. 1982. The effect of external reporting on managerial decision making: Some antecedent conditions. *Accounting, Organizations and Society* 7 (4): 319-336. https://doi.org/10.1016/0361-3682(82)90008-3

Donnelly, A. M., and D. P. Donnelly. 2023. A Systematic Review of Experimental Research on Accountability in Auditing. *Behavioral Research in Accounting*, 1-36.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and L. A. Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human factors 44* (1): 79-94.

Ernst & Young (EY). 2020. How to make the most of ai incorporate reporting. Available at: https://www.ey.com/en_us/audit/how-to-make-the-most-of-ai-in-corporate-reporting

Farrell, A. M., J. H. Grenier, and J. Leiby. 2017. Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets. *The Accounting Review* 92 (1): 93-114. https://doi.org/10.2308/accr-51447

Fehrenbacher, D. D., S. E. Kaplan, and C. Moulang. 2020. The role of accountability in reducing the impact of affective reactions on capital budgeting decisions. *Management Accounting Research 47*.

Financial Reporting Council (FRC). 2019. Artificial intelligence and corporate reporting. How does it measure up? Financial Reporting Lab. Available at: https://www.frc.org.uk/getattachment/e213b335-927b-4750-90db-64139aee44f2/AI-and-Corporate-Reporting-Jan.pdf

Frink, D.D. and G. R. Ferris. 1998. Accountability, impression management, and goal setting in the performance evaluation process. *Human relations* 51(10): 1259-1283.

Garvey, A. M., Kim, T., and A. Duhachek. 2023. Bad news? Send an AI. Good news? Send a Human. *Journal of Marketing* 87 (1): 10-25. https://doi.org/10.1177/00222429211066972

Gino, F., and D. A. Moore. 2007. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making* 20 (1): 21-35. https://doi.org/10.1002/bdm.539

Gino, F., J. Shang, and R. Croson. 2009. The impact of information from similar or different advisors on judgment. *Organizational Behavior and Human Decision Processes* 108 (2): 287-302. https://doi.org/10.1016/j.obhdp.2008.08.002

Griffith, E. E. 2018. When do auditors use specialists' work to improve problem representations of and judgments about complex estimates? *The Accounting Review 93* (4): 177-202.

Griffith, E. E. 2020. Auditors, specialists, and professional jurisdiction in audits of fair values. *Contemporary Accounting Research 37* (1): 245-276.

Griffith, E. E., Hammersley, J. S., Kadous, K., and D. Young. 2015. Auditor mindsets and audits of complex estimates. Journal of Accounting Research, 53(1), 49-77.

Harker, A. L. 2003. *The impact of financial reporting on decision-making by boards of trustees: A case study at Stanford University*. University of Pennsylvania. Available at: https://repository.upenn.edu/dissertations/AAI3084858

Harries, C., I. Yaniv, and N. Harvey. 2004. Combining advice: The weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making* 17 (5): 333-348. https://doi.org/10.1002/bdm.474

Hayes, A. F. 2022. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (3rd edition). New York: The Guilford Press.

Highhouse, S., 2008. Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology* 1(3): 333-342.

Hirst, D. E., Koonce, L., and S. Venkataraman. 2008. Management earnings forecasts: A review and framework. Accounting horizons, 22(3), 315-338.

Hodge, F., Hopkins, P. E., and J. Pratt. 2006. Management reporting incentives and classification credibility: The effects of reporting discretion and reputation. Accounting, Organizations and Society, 31(7), 623-634.

Joe, J. R., Vandervelde, S. D., and Y. J. Wu. 2017. Use of high quantification evidence in fair value audits: Do auditors stay in their comfort zone?. The Accounting Review, 92(5), 89-116.

Kadous, K. 2000. The effects of audit quality and consequence severity on juror evaluations of auditor responsibility for plaintiff losses. The Accounting Review, 75(3), 327-341.

Kadous, K., Koonce, L., and K. L. Towry. 2005. Quantification and persuasion in managerial judgement. *Contemporary Accounting Research*, *22*(3), 643-686.

Kadous, K., J. Leiby, and M. E. Peecher. 2013. How do auditors weight informal contrary advice? The joint influence of advisor social bond and advice justifiability. *The Accounting Review* 88 (6): 2061-2087. https://doi.org/10.2308/accr-50529

Kahneman, D. 2003. A perspective on judgment and choice: Mapping bounded rationality. American Psychologist 58 (9): 697–720.

Kantrowitz, A., and D. Gorman. "Jobs are for humans." *Slate.com*. August 14, 2023. Available at: https://slate.com/technology/2023/08/chat-gpt-artificial-intelligence-jobs-economy-employment-labor.html

Kesari, G. "AI Accuracy is Overrated: How Even a "Wrong" Model can Transform Your Business." *Forbes*. January 21, 2021. Available at: https://www.forbes.com/sites/ganeskesari/2021/01/21/accuracy-isnt-everything-how-even-a-wrong-ai-model-can-transform-your-business/?sh=5ac215b47083

Kim, T. W., and A. Duhachek. 2020. Artificial intelligence and persuasion: A construal-level account. *Psychological science* 31 (4): 363-380. https://doi.org/10.1177/09567976209049

Koonce, L., U. Anderson, and G. Marchant. 1995. Justification of decisions in auditing. *Journal of Accounting Research 33* (2): 369-384.

Koreff, J., and S. Perreault. 2023. Is sophistication always better? Can perceived data analytic tool sophistication lead to biased judgments?. *Journal of Emerging Technologies in Accounting 20*(1): 91-110.

KPMG, LLP (KPMG). 2016. Harnessing the power of cognitive technology to transform the audit. Available at: https://assets.kpmg/content/dam/kpmg/us/pdf/2017/02/harnessing-the-power-of-cognitive-technology-to-transform-the-audit.pdf

KPMG, LLP. 2018. *Implementing the Expected Credit Loss model for receivables*. Wilmington, DE: KPMG. Available at: https://assets.kpmg/content/dam/kpmg/ch/pdf/treasury-news-26-en.pdf

KPMG, LLP. 2020a. *Business challenge 2020: keeping pace with the rising expectations for artificial intelligence*. Available at: https://home.kpmg/us/en/home/media/press-releases/2020/01/business-challenge-2020-keeping-pace-with-the-rising-expectations-for-artificial-intelligence.html

Kunda, Z. 1990. The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Lerner, J.S. and P. E. Tetlock. 1999. Accounting for the effects of accountability. *Psychological bulletin* 125(2): 255.

Leonard, A., B. Haarman, J. DeMelis, and T. Youngberg. 2020. "Innovative auditing using data and analytics-the role of academics." Panel at the AAA Auditing Section Midyear Meeting, Houston, TX.

Logg, J. M., J. A. Minson, and D. A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151: 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lowe, D. J., Reckers, P. M., and S. M. Whitecotton. 2002. The effects of decision-aid use and reliability on jurors' evaluations of auditor liability. The Accounting Review, 77(1), 185-202.

Mercer, M. 2005. The fleeting effects of disclosure forthcomingness on management's reporting credibility. *The Accounting Review* 80 (2): 723–744.

Messner, M. 2009. The limits of accountability. *Accounting, Organizations and Society*, *34*(8), 918-938.

Miller, A. P. 2018. "Want less-biased decisions? Use algorithms." July 26, 2018. *Harvard Business Review*. Available at: https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Nowlis, S. M., and I. Simonson. 1997. Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research* 34 (2): 205-218. https://doi.org/10.2307/3151859

Önkal, D., P. Goodwin, M. Thomson, S. Gönül, and A. Pollock. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22 (4): 390-409. https://doi.org/10.1002/bdm.637

Patt, A. G., H. R. Bowles, and D. W. Cash. 2006. Mechanisms for enhancing the credibility of an adviser: Prepayment and aligned incentives. *Journal of Behavioral Decision Making 19* (4): 347-359.

Peecher, M. E., C. Pietsch, S. Stirnkorb, and I. L. Yamoah. 2023. Upskilling Auditors in the Face of Changing Skills Requirements: Does Self-Affirmation Help Overcome Aversion to AI-Based Specialist Advice? Working Paper University of Illinois.

Piercey, M. D. 2011. Documentation requirements and quantified versus qualitative audit risk assessments. *Auditing: A Journal of Practice & Theory 30* (4): 223-248.

Public Company Accounting Oversight Board (PCAOB). 2016. *Using the Work of a Specialist. AS 1210*. Washington, DC: PCAOB.

Ranzilla, S., Chevalier, R. E., Herrmann, G., Glover, S. M., and Prawitt, D. F. 2011. *Elevating professional judgment in auditing: The KPMG professional judgment framework.* New York, NY: KPMG LLP.

Rowe, S. P. 2019. Auditors' comfort with uncertain estimates: More evidence is not always better. *Accounting, Organizations and Society* 76: 1-11.

Russell, S., and P. Norvig. 2010. Artificial intelligence: A modern approach. Upper Saddle River, New Jersey: Pearson Education, Inc.

Sah, S., Moore, D. A., and R. J. MacCoun. 2013. Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes* 121(2): 246-255.

Sauer, M. "Steve Wozniak's warning: No matter how 'useful' ChatGPT is, it can 'make horrible mistakes." *CNBC.com*. February 16, 2023. Available at: https://www.cnbc.com/2023/02/10/steve-wozniak-warns-about-ai-chatgpt-can-make-horrible-mistakes.html

Shankar, P. G., and H. T. Tan. 2006. Determinants of audit preparers' workpaper justifications. *The Accounting Review* 81(2): 473-495. https://doi.org/10.2308/accr.2006.81.2.473

Sniezek, J. A., and T. Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes* 62 (2): 159-174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J.A. and Van Swol, L.M., 2001. Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes* 84(2): 288-307.

Stefaniak, C. M., Houston, R. W., and D. M. Brandon. 2017. Investigating inspection risk: An analysis of PCAOB inspections and internal quality reviews. *Auditing: A Journal of Practice & Theory*, *36*(1), 151-168.

Tetlock, P.E., 1983. Accountability and complexity of thought. *Journal of personality and social psychology* 45(1): 74.

Tetlock, P. E., L. Skitka, and R. Boettger. 1989. Social and cognitive strategies for coping with accountability: conformity, complexity, and bolstering. *Journal of personality and social psychology* 57 (4): 632-640. https://doi.org/10.1037/0022-3514.57.4.632

Wilks, T. J. 2002. Predecisional distortion of evidence as a consequence of real-time audit review. *The Accounting Review 77* (1): 51-71.

Yaniv, I. 2004. Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes* 93 (1): 1-13. https://doi.org/10.1016/j.obhdp.2003.08.002

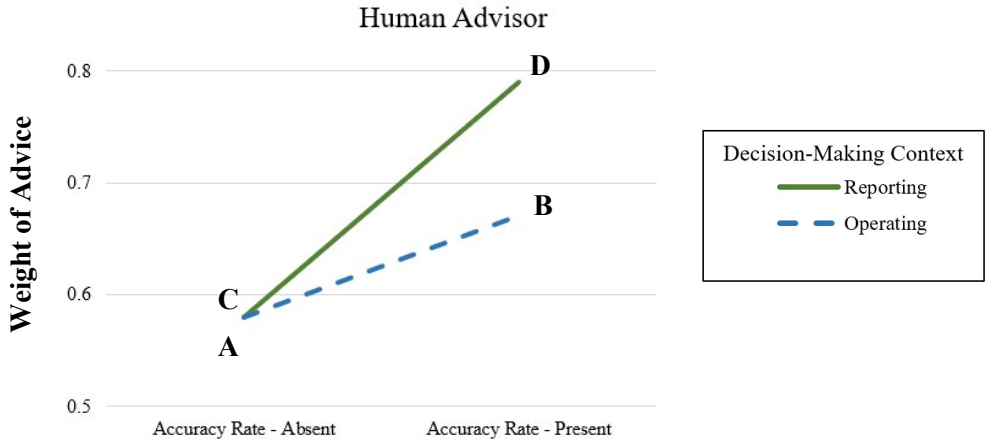
Yaniv, I., and E. Kleinberger. 2000. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes* 83 (2): 260-281. https://doi.org/10.1006/obhd.2000.2909

Yeomans, M., Shah, A., Mullainathan, S. and Kleinberg, J., 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32(4): 403-414.

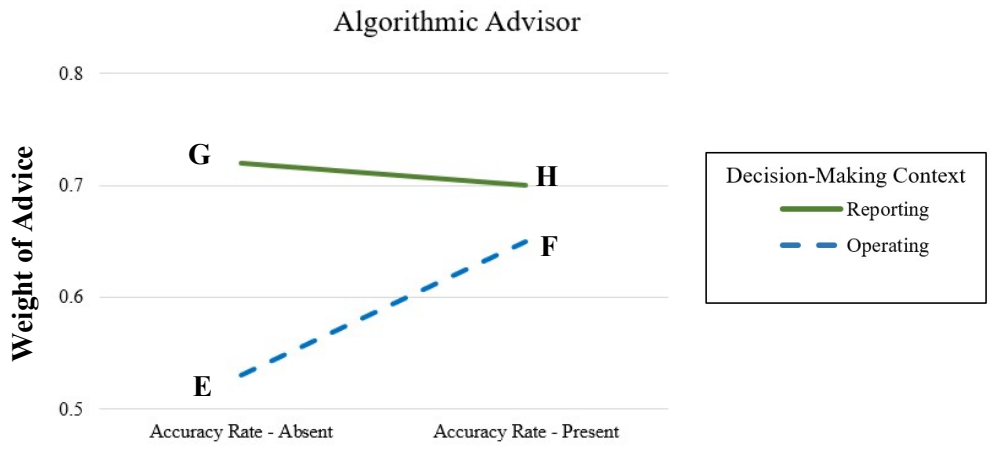
Zhang, G., Chong, L., Kotovsky, K., and J. Cagan.2023. Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior*, *139*, 107536.

FIGURE 1
Weight of Advice Results
All Conditions

Panel A: Human Advisor Conditions



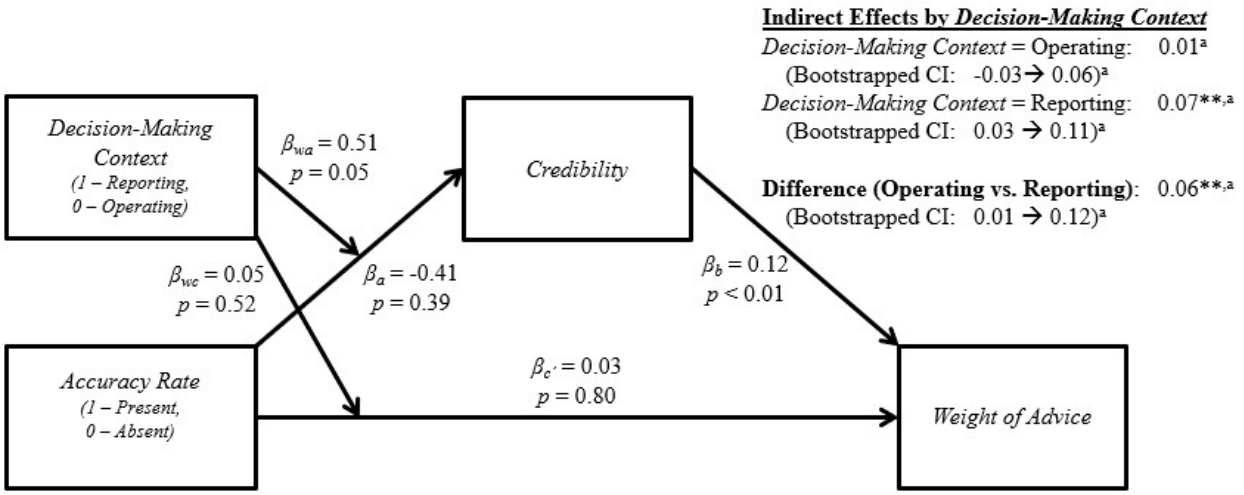
Panel B: Algorithm Advisor Conditions



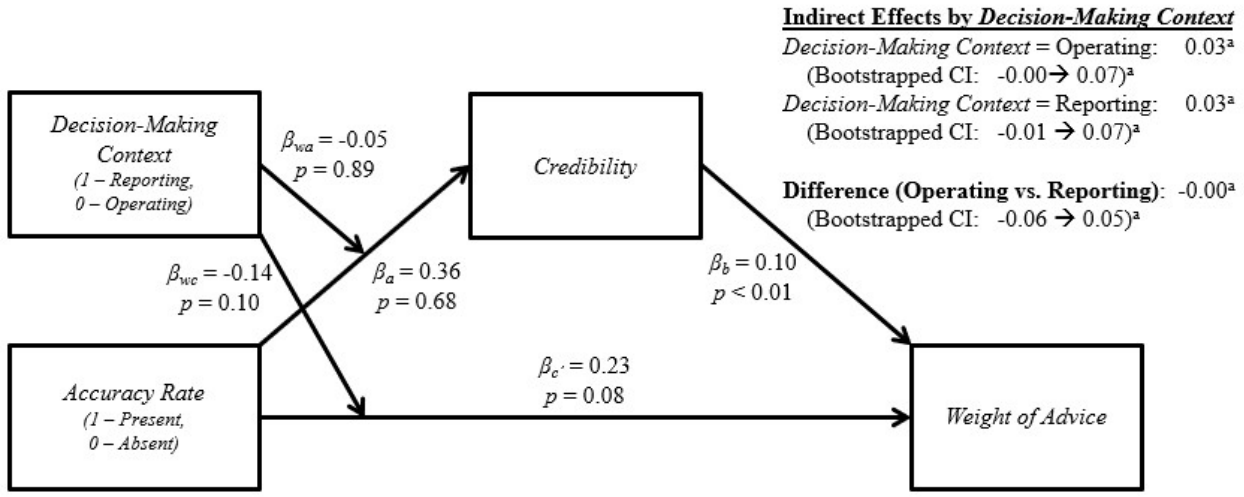
Note: Figure 1 Panel A and Panel B display results of participants' Weight of Advice (WOA) measure. The dependent variable is participants' advice utilization (WOA) which is calculated as (initial estimate – final estimate)/(initial estimate – advisor's estimate) and ranges from 0 to 1 where larger values of WOA indicate greater weighting of advice provided by the advisor (i.e., greater reliance on advice). We manipulate, between-participants, the advice source at two levels (human vs. algorithm). We also manipulate, between-participants, the decision-making context as a reporting context (i.e., reporting the value of a house for a loan) or operating context (i.e., estimating value of a house to decide whether to sell the house or not). Lastly, we manipulate, between-participants, whether a high accuracy rate is present (i.e., disclosed 90% accuracy) or absent (i.e., no disclosure of accuracy rate).

FIGURE 2
Moderated Mediation Analysis

Panel A: Human Conditions



Panel B: Algorithm Conditions



Note: The above diagram represents a moderated mediation model (Hayes 2022). We present results separately for the human and algorithm conditions. The model is calculated for human conditions and algorithm conditions separately using Model 8 in the PROCESS macro in SPSS. See notes in Figure 1 for descriptions of dependent variables, and independent factors. To construct the mediator, we ask participants to assess how credible they think the advice provided by the advisor (i.e., human or algorithm) was and how credible they believe others will think the advice provided by the advisor (i.e., human or algorithm) was on a 7-point Likert scales with endpoints 1 = "Not at all Credible" and 7 = "Very Credible". We then take the average of those two measures and construct a variable called *Credibility*.

[a] To test for indirect effects, we construct 90% confidence intervals for the product of paths *a* and *b*. We use 10,000 bootstrapped resamples of data with replacement (Hayes 2022). Reflecting our directional predictions, we use 90% confidence intervals (i.e., bounded at 0.05 and 0.95) to test whether one-tailed p-values are less than 0.05.

** denotes statistical significance equivalent to $p < 0.05$, one-tailed.

† one-tailed given directional prediction (all other *p*-values are two-tailed).

Table 1
Weight of Advice Results in all Conditions

Panel A: Descriptive Statistics
Mean, (standard deviation), [n]

Advice Source	Operating Decision-Making Context			Reporting Decision-Making Context			Total		
	Accuracy Rate - Absent	Accuracy Rate - Present	Total	Accuracy Rate - Absent	Accuracy Rate - Present	Total	Accuracy Rate - Absent	Accuracy Rate - Present	Total
Human	0.58 (0.35) [53] A	0.67 (0.33) [52] B	0.62 (0.34) [105]	0.58 (0.37) [51] C	0.79 (0.26) [53] D	0.68 (0.33) [104]	0.58 (0.36) [104]	0.73 (0.30) [105]	0.65 (0.34) [209]
Algorithm	0.53 (0.33) [52] E	0.65 (0.29) [51] F	0.59 (0.32) [103]	0.72 (0.32) [45] G	0.70 (0.31) [53] H	0.71 (0.32) [98]	0.62 (0.34) [97]	0.68 (0.30) [104]	0.65 (0.32) [201]
Total	0.55 (0.34) [105]	0.66 (0.31) [103]	0.61 (0.33) [208]	0.65 (0.35) [96]	0.74 (0.29) [106]	0.70 (0.33) [202]	0.60 (0.35) [201]	0.70 (0.30) [209]	0.65 (0.33) [410]

Table 1 (continued)
Results of Weight of Advice in all Conditions

Panel B: ANOVA (Tests of H2)

Source	Sum of Squares	*df*	*F*	*p*
Advice Source	0.00	1	0.00	0.98
Decision-Making Context	0.80	1	7.67	<0.01
Accuracy Rate	1.02	1	9.74	<0.01
Advice Source × Decision-Making Context	0.08	1	0.73	0.40
Advice Source × Accuracy Rate	0.24	1	2.31	0.06†
Advice Source × Decision-Making Context × Accuracy Rate	0.41	1	3.96	0.02†
Error	42.00	402	0.10	

Panel C: Two-way Interaction Tests

	Sum of Squares	*df*	*F*	*p*
Within the Human Conditions				
Accuracy Rate × Decision-Making Context	0.17	1	1.55	0.21
Within the Algorithm Conditions				
Accuracy Rate × Decision-Making Context	0.25	1	2.48	0.06†
Within the Reporting Conditions				
Advice Source × Accuracy Rate	0.63	1	6.24	<0.01†
Within the Operating Conditions				
Advice Source × Accuracy Rate	0.01	1	0.11	0.74

Panel D: Simple Effects Tests

	df	*t*	*p*
Within the Human Conditions			
Absent vs Present Accuracy Rate within Reporting conditions (C vs D)	102	3.28	< 0.01†
Absent vs Present Accuracy Rate within Operating conditions (A vs B)	103	1.38	0.09†
Within the Algorithm Conditions			
Present vs Absent Accuracy Rate within Reporting conditions (G vs H)	96	0.30	0.77
Present vs Absent Accuracy Rate within Operating conditions (E vs F)	101	1.96	0.05

Note: Panel A of Table 1 reports the descriptive statistics by experimental condition and Panel B reports the results of an ANOVA for the full factorial model. See notes to Figure 1 for descriptions of the dependent variable and independent factors.

† *p*-values are equivalent to a one-tailed test, consistent with our directional predictions.