The Impact of AI Risk Scores on Business Managers' Ethical Decision-Making

Heather Domin
IBM Corporation

Erin Klawitter
University of Notre Dame

Artificial intelligence (AI) presents significant benefits to the organizations and to the individuals, environments, and stakeholders they impact. However, AI systems can also pose a risk of harm. Development of an AI risk score representing the potential risk of the system may assist business managers with the ethical decision on whether to deploy an AI system. While the quantification of risks associated with AI has received attention from researchers, limited research exists analyzing summarized AI risk scores and their impact on decision-making in practice. Expanding on integrated and behavioral theories of ethical decision-making, this quantitative experimental study found that the presence of an AI risk score can reduce the likelihood of an unethical decision, and thus may positively influence business managers faced with an ethical decision. The study also explored the potential influence of the AI system's use case when an AI risk score is present; however, no significant influence was identified in the scenarios tested. This study has implications for practice for organizations developing, deploying, and using AI systems.

Keywords: artificial intelligence, risk, risk scoring, ethical decision-making, business managers

INTRODUCTION

Artificial intelligence (AI) risk management has emerged as a critical area of focus for policymakers, developers, and users of this technology. Driven by a rapid rate of adoption and an increase in the public awareness of potential risks associated with artificial intelligence, this focus presents a challenge to policymakers and industry actors given the current lack of standard risk management frameworks (Ezeani et al., 2021; Metcalf et al., 2021). Business managers within the industry must decide whether the benefits of using an AI system outweigh the potential risks. To holistically assess the risk of AI systems, qualitative and quantitative information is needed. One way of aiding business managers in weighing benefits and risks is to present a quantified risk score that can help them decide if they should proceed with the deployment of the AI system (Piorkowski et al., 2022).

Quantifying specific AI risks relating to bias, explainability, and robustness is an area of focus for researchers in the field (Islam et al., 2020; Szepannek and Lübke, 2021). However, while approaches for quantifying specific risks have been widely studied, additional research is needed to better understand their use in risk assessment processes, and the ethical implications of summarized risk scores (Piorkowski et al., 2022). Quantified AI risk scores can significantly impact human decision-making (Bucinca et al., 2022).

For example, AI risk scores have been shown to have a strong anchoring effect on human decision-making in a judicial system use case involving Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software (Vaccaro, 2019).

This quantitative experimental study tested whether the presence of an AI risk score influences the likelihood of a business manager making an unethical decision. It also tested whether the use case (education, employment, or government benefits) for the AI system plays a significant role in whether the presence of an AI risk score influences the likelihood of an unethical decision in business managers. Findings from this study will have important implications for scholars, policymakers, and practitioners seeking to understand how quantitative AI risk scoring methods impact the ethical decision-making of managers within organizations. Crucially, gaining such understanding will increase the likelihood of more ethical outcomes, positively impacting the lives and livelihoods of affected individuals.

Background

AI systems can present a variety of risks. The ethical literature concerning AI discusses the potential impact of AI systems on the lives of humans and other beings and the potential ethical concerns relating to the technology itself (e.g., the rights of AI systems) (Meek et al., 2016). Research regarding quantification of AI risk scores tends to focus on the category of ethical issues that focus on the potential impact on humans and other beings (Islam et al., 2020; Szepannek and Lübke, 2021). For example, categorical risk scores present a measure of fairness by quantifying bias metrics or provide a measure of robustness by quantifying how well a system may prevent adversarial attacks (Nicolae et al., 2019; Szepannek and Lübke, 2021).

The implications of summarized risk scoring for AI models require additional research because accurately summarizing an AI risk score often entails balancing conflicting objectives. For example, a mathematical model that optimizes for an objective of individual privacy might not optimize for an objective of fairness, given the reduced ability to include data points that might enable the assessment of potential unfair bias (Information Commissioner's Office, 2020). Optimizing for a single objective, such as fairness, may also result in a trade-off with the overall accuracy scores of the model (Szepannek and Lübke, 2021).

Conceptual Framework

Theories on ethical decision-making take a rational (i.e., based on logical reasoning) or non-rational (i.e., based on intuition and emotion) approach (Schwartz, 2015). Here, we use an integrated lens, assuming rational and non-rational factors influence the ethical decision-making process (Schwartz, 2015). This integrated lens also aligns with a behavioral model of ethical decision-making, where the decision process is impacted by environmental factors and individual attributes (Bommer et al., 1987). In both integrated and behavioral models of ethical decision-making, the ethical or unethical behavior is preceded by an intention or decision (Bommer et al., 2015; Schwartz, 2015). An integrated theoretical approach to ethical decision-making suggests that awareness of a quantified AI risk score would be a key factor in the ethical decision-making process (Schwartz, 2015). In addition, use case (e.g., education, employment, and government benefits) might also be a key factor in ethical decision-making, moderating the situational context of the decision (Schwartz, 2015). Similarly, a behavioral model for ethical decision-making, suggests that use case might be one of the government/legal environment that factors into the ethical decision-making process (Bommer et al., 2015). For example, use case is a key factor for determining risk under the proposed EU AI Act (European Commission, 2021).

The conceptual framework for this study draws its base from theories on ethical decision-making and normative ethics. Normative ethical theory helps define what can be considered an ethical or unethical decision (MacKinnon and Fiala, 2017). To guide normative decision-making, many organizations have proposed ethical principles for AI systems (Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019; Mittelstadt, 2019; OECD, 2019). Dual-theory processing models of persuasion theorize that when people are highly invested in a decision, they tend to engage in more intensive, "systematic" information processing. When they are less invested, they tend to engage in simple, rules-based, or "heuristic" decision-

making (Chaiken, 1980). We conceptualize that an AI risk score can serve as a heuristic for information processing, a rule of thumb to assist decision-making that might be influenced by the normative ethical concept to which it relates. For example, in a high-risk use case such as law enforcement, an AI risk score (heuristic) related to a normative ethical principle of fairness might influence the likelihood of an unethical decision more than an AI risk score related to transparency.

Given the complexity of the ethical dilemmas faced by business managers, who may find it challenging to gather and weigh all of the information necessary to make a decision (Meek et al., 2016), this study investigated whether AI risk scores influence business managers during the ethical decision-making process in such a way that these scores can help avoid unethical decisions. AI risk scores and other quantitative assessments of the internal workings of an AI system are only one part of effectively assessing the potential impact of an AI system (Metcalf et al., 2021). Decision-makers and those who support the decision-making process also need to understand the potential risks of over-reliance on risk scores and encourage strong engagement from the humans responsible for the decision (Bucinca et al., 2022). However, a quantified risk score may help reduce ambiguity in ethical decision-making, for example, through providing clarity during the creation of decision criteria and providing a basis for comparison against the defined decision criteria (Johansen and Rausand, 2015).

The theoretical contribution of this research study is that it extends what we already understand about the impact of AI on human decision-making to ethical decision-making in business managers. Ethics, sociology, and AI researchers have well laid the theoretical foundation for this study. The presence of AI has been shown to impact decision-making through different cognitive biases such as confirmation bias and anchoring bias (Rastogi et al., 2022). AI output can be in the form of a quantified score. Indeed, the presence of a quantified output from an algorithm can play a significant role in decision-making. This is true, even in use cases that may be higher risk, such as allocating resources for housing and school funding (Johnson & Zhang, 2022). It is, therefore, important to understand how humans use AI output and quantified scores. Towards this end, human-AI decision coordination researchers have explored how humans and AI work together to make decisions (Baudel et al., 2023; Bucina et al., 2022; Zhang et al., 2020). They have found that quantified scores from AI models can directly influence decision-making; for example, quantified confidence scores associated with AI output can influence reliance on the AI's output (Zhang et al., 2020).

Extending the individual decision-making context to a business and organizational setting can be viewed in the context of normative ethics theory. Normative ethics provides a basis for understanding right versus wrong in decision-making (MacKinnon & Fiala, 2017). A business manager's use of AI during decision-making is influenced by technical and organizational factors (Feuerriegel et al., 2022). For example, principles that are established for the organization. AI ethics principles that reflect normative ethical theories are intended to influence the ethical decisions made by the actors within those organizations towards right behaviors (Fjeld et al., 2020). While ethical principles do not represent business ethics theory, they do represent how theory might be translated into practice. Therefore, we use principles to reflect the theoretical underpinnings in a business context within this study. Specifically, we use principles from the Organisation for Economic Co-operation and Development (OECD) given they are a prominent example of how AI ethics principles have been translated into practice (Floridi & Cowls, 2019).

OECD Principles

The Organisation for Economic Co-operation and Development (OECD) has outlined five principles constituting a normative ethical framework for all AI systems. As the basis of the OECD Framework for the Classification of AI Systems, they are intended to inform risk management efforts, and contribute to the foundation of a framework that OECD plans to develop to empower organizations to implement risk assessments for AI systems (OECD, 2022). These principles include "inclusive growth, sustainable development and well-being," "human-centered values and fairness," "transparency and explainability," "robustness, security and safety," and "accountability" (OECD, 2019, paras. 18-22). The OECD principles provide an appropriate framework for categorizing ethical issues within this study given the relationship to AI risk assessments, and the potential for this work to directly inform industry actors as the OECD's work evolves to include more comprehensive guidance on implementing risk assessments.

Research Focus

This study had two primary areas of focus. The first area of focus was on the impact of AI risk score awareness on ethical decisions. The second area of focus was whether the AI system use case influenced ethical decisions when an AI risk score was present. The second area also included sub-questions related to the impact of OECD principles on these use cases.

To explore the first area, we sought to understand the answer to the following question:

RQ1: *Does an awareness of a system-generated quantified AI risk score (versus a non-quantified AI risk statement) influence the likelihood of a business management unethical decision?*

RQ 1 Hypothesis: *Awareness of a system-generated quantified AI risk score impacts the likelihood of a business management unethical decision.*

To explore the second area, we collected data related to the following question

RQ2: *When awareness of a system-generated quantified AI risk score exists, does use case influence the likelihood of a business management unethical decision?*

RQ 2 Hypothesis: *When awareness of a system-generated quantified AI risk score exists, the type of AI use case impacts the likelihood of a business management unethical decision.*

The sub-questions for the second area focused on the five OECD principles of "inclusive growth, sustainable development and well-being," "human-centred values and fairness," "transparency and explainability," "robustness, security and safety," and "accountability" (OECD, 2019). We asked each of these five principles: "For an ethical dilemma involving each principle, does the use case influence the likelihood of a business management unethical decision?"

RQ 2.1 Hypothesis*: When awareness of a system-generated quantified AI risk score relating to OECD principle 1 exists, the type of AI use case impacts the likelihood of a business management unethical decision.*

RQ 2.2 Hypothesis*: When awareness of a system-generated quantified AI risk score relating to OECD principle 2 exists, the type of AI use case impacts the likelihood of a business management unethical decision.*

RQ 2.3 Hypothesis*: When awareness of a system-generated quantified AI risk score relating to OECD principle 3 exists, the type of AI use case impacts the likelihood of a business management unethical decision.*

RQ 2.4 Hypothesis*: When awareness of a system-generated quantified AI risk score relating to OECD principle 4 exists, the type of AI use case impacts the likelihood of a business management unethical decision.*

RQ 2.5 Hypothesis*: When awareness of a system-generated quantified AI risk score relating to OECD principle 5 exists, the type of AI use case impacts the likelihood of a business management unethical decision.*

METHODOLOGY

To answer these research questions and test their associated hypotheses, this project used a quantitative experimental research methodology conducted via a Qualtrics questionnaire and tested via a pilot study.

The questionnaire consisted of ethical vignettes that manipulated the independent variables (the presence of an AI risk score, use case, and principles) to demonstrate their effect on the dependent variable (the participants' likelihood of choosing to deploy an AI system). See Appendix A for the ethical vignettes.

Experimental Design

This study used a true experimental design where participants were randomly assigned to either the control or experimental group for each research question. Experimental research is appropriate when the research questions require examining the potential influence of independent variables on dependent variables (Creswell and Creswell, 2018). The design for this study was similar to a prior study on ethical decision-making by Hassan et al. (2021), where participants were presented with ethical vignettes before being asked to make an ethical decision. After completing an IRB-approved consent form, each participant was presented with an ethical vignette that provided background information required to answer subsequent questions.

Variations in the ethical vignettes were minimized to limit the risk of introducing confounding variables. For example, in the ethical vignette for the first research question, in the sentence where the risk level was introduced, only the words relating to the independent variable for the presence of a risk score were manipulated. Participants in the control group were told the AI system has "too much ethical risk", while those in the experimental group were told the system has "an AI risk score of 25%".

After being presented with the background information, participants were asked to perform an ethical decision-making task and assess the likelihood of that decision on a 10-point Likert scale. Experimental group participants were then presented with a manipulation check question to confirm the participant was aware of the independent variable that had been manipulated.

The same experimental structure was repeated for the second research question, with variation in the vignette text limited to the use case-independent variable. Sub-questions for the second research question were then assessed in a similar way for each use case scenario, using questions that limited variation to the differences in the ethical principles. Following ethical decision-making questions, participants were also asked to answer a qualitative question regarding the primary reason for their decision.

Pilot Study

In order to refine the ethical vignettes, we recruited students from a medium-sized midwestern university to participate in a pilot study. Thirty-one students who possessed business management students participated in the pilot study, which was conducted in small groups during six 45-minute sessions. Participants were students enrolled in graduate business degree programs and a late-career fellows program for college graduates contemplating the next phase of their career. We compensated pilot student participants $10 in a Visa gift card. In the pilot study, participants completed the initial questionnaire in Qualtrics as well as a follow-up questionnaire inquiring about their experience of taking the study. Following completing both questionnaires, we held an open-ended qualitative debriefing session to gain insight into any difficulties the participants faced understanding the ethical vignettes, and to share with them the experimental design and manipulations of the questionnaire. Through the follow-up questionnaire and discussions, we learned that the ethical vignettes and experimental manipulations were clear as written and required only minor edits. We also learned that participants seemed to provide a variety of rationales for their decision-making, which led us to add question to the final study, inquiring about participants' reasoning for deciding to deploy the AI system.

Primary Study

We recruited 1,100 study participants via Prolific, an online research platform that solicits research participants worldwide. To ensure reading comprehension and generalizability to the population of interest –business managers– we restricted participation to adults possessing English fluency and business management experience. Prolific solicited participation from individuals meeting these characteristics three times in 24 hours to ensure convenient times for questionnaire completion in different parts of the world. After completing a consent form and the questionnaire, participants received $2. Participants completed an

informed consent question as approved by our institution's IRB, which also permitted us to obtain demographic information stored by Prolific.

Participant Descriptives

Of the 1,100 participants who completed the questionnaire, 145 were removed due to failing a manipulation check, leaving 955 participants' responses for analysis. Table 1 shows demographic characteristics of participants whose responses were analyzed. Of those who completed the questionnaire and passed all manipulation checks, nearly two-thirds (62.2%) were male while more than one-third (37.7%) were female. Three-quarters (75.8%) identified as White, nearly one-tenth as (9.4%) Black, and less than one-tenth as Asian (7.1%), Mixed (5.3%), or Other (1.2%). Nearly half were from the Americas region (45.9 %); nearly half were from Europe and Africa (44.7%), and less than 10% were from Asia and Australia (9.4 %). Roughly half the group (51%) were between 18 and 40 years of age.

Measures and Analyses

In order to test our hypotheses, we randomly assigned participants to control and experimental conditions, articulated in ethical vignettes in Appendix A. The independent variables (presence of an AI risk score, use case, and principles) were manipulated via narrative language in the ethical vignette. The dependent variable (likelihood of making an unethical decision) was measured via a Likert scale. The questionnaire included manipulation checks to ensure that participants recognize the presence of a risk score or of a use case. Statistical analyses were conducted via SPSS software, including two sample t-tests (independent samples) and analysis of variance (one-way ANOVA).

Scale for Evaluating Ethical Decision-Making

The scale for evaluating an ethical decision used a 10-point Likert scale, where 1 = Very unlikely and 10 = Very likely. As in the study by Hassan et al. (2021), this scale focused on the likelihood of an unethical decision, and not the likelihood of an ethical decision. Both ethical and unethical decisions can be difficult to measure; however, unethical decisions occur less frequently than ethical decisions (Trevino, 1992). Since an ethical decision is more likely to occur than an unethical decision, a relative difference in the likelihood of an ethical decision is less significant than a relative difference in the likelihood of an unethical decision. Therefore, framing the questions to focus on an unethical decision may provide a more sensitive scale.

Use Cases

The use cases referenced within the ethical vignettes were selected for consistency in risk-level and system purpose. All use cases meet the criteria for high-risk, as defined in the draft EU AI Act Annex II (European Commission, 2021). The number of use cases selected was limited to three identified use cases to minimize risks to internal validity (e.g., confounding variables, too few participants in a group). The use cases all have the same fundamental task (screening) but a different use case context. The use cases selected for this study are AI systems used in the context of education (college screening), employment (hiring screening), and essential services (government benefit screening).

RESEARCH FINDINGS

The study's objective was to determine the effect of an AI risk scoring on a business manager's ethical decision-making. This was explored through two primary research questions, and a qualitative question relating to the rationale for the decision. Our hypothesis that a system-generated quantified AI risk score influences ethical decision-making was supported in the finding for Research Question 1. Our hypothesis that use case influences this ethical decision-making was not supported in our findings for Research Question 2. Responses to our qualitative question indicated differences in the rationale for ethical decisions made by participants who were presented with an AI risk score, and those who were not.

Research Question 1

The research findings showed that awareness of a system-generated quantified AI risk score (versus a non-quantified AI risk statement) influences the likelihood of a business management unethical decision (See Table 2). The main effect was statistically significant [$F(1,913) = 20.44$, $p < 0.001$]. Compared to individuals in the control group, those in the experimental group who were presented with an AI risk score indicated they would be less likely to proceed with an unethical decision, which in the ethical vignette was the likelihood to deploy the AI system despite being presented with information that indicated the AI system contained too much risk to deploy per company guidance. The control group who was not presented with an AI risk score had mean average scores ($M = 4.65$, $SD = 2.73$) that were higher than those of the experimental group with an AI risk score ($M = 3.87$, $SD = 2.62$). The Likert scale used to test for likelihood of an unethical decision ranged from 1 (very unlikely to deploy) to 10 (very likely to deploy); therefore, these mean scores indicate participants in both groups were somewhat likely, on average, to proceed with an unethical decision. However, the likelihood of an unethical decision was lower for the experimental group that was presented with an AI risk score.

Exploring the impact of demographic characteristics can provide insight into the primary research findings. For example, sex and age may influence ethical decision-making (Loe et al., 2000) and risk-taking behavior (Charness and Gneezy, 2012; Vroom and Pahl, 1971) among business managers. In general, males and younger individuals may be more likely to engage in risk-taking behavior (Charness and Gneezy, 2012; Loe et al., 2000; Vroom and Pahl, 1971). While exploring the impact of these characteristics on the research findings using an ANOVA method, age and sex were found to have statistically significant effects when viewed by themselves; however, when age and sex were combined the main effect for age and sex disappeared. A main effect for sex was significant [$F(3,949)=10.2559$, $p < 0.001$]; where men indicated a higher likelihood to deploy than women. So too, was a main effect for age [$F(3,943) = 15.3772$, $p < 0.001$]; where younger individuals indicated a higher likelihood to deploy than older individuals. However, when age and sex were combined with AI risk score, the only main effect was for the risk score [$F(7,938)=8.8009$, $p < 0.001$]; where those individuals that were not presented with an AI risk score indicated a higher likelihood to deploy than those individuals that were presented with an AI risk score. The findings indicate that an AI risk score matters the most. Despite differences in individual demographic characteristics, the presence of an AI risk score is the best predictor of an individual's decision to deploy.

Research Question 2

In response to the second research question, which asked, "when awareness of a system-generated quantified AI risk score exists, does use case influence the likelihood of a business management unethical decision?" We hypothesized that when awareness of an AI risk score exists, the type of AI use case will impact the likelihood of a business management unethical decision. This would imply that use cases are an important factor in the potential influence of AI risk scores in business manager decision-making pointing to the relative importance of accuracy for certain types of use cases. To test this hypothesis, a one-way ANOVA was performed on the business managers' likelihood to deploy an AI system in the context of various use cases (no use case, education, employment, and government benefits). We found the main effect of the use case was not significant [$F(3, 950) = 1.49$, $p = 0.22$*)* did not find any statistically significant group differences. Thus, we failed to reject the null hypothesis. However, we did notice that while the mean likelihood to deploy the AI system was similar for the control ($M = 3.72$, $SD = 2.68$), education ($M = 3.79$, $SD = 2.56$), and employment conditions ($M = 3.74$, $SD = 2.52$) were similar, the mean likelihood of deploying a system pertaining to the distribution of government benefits ($M = 3.34$, $SD = 2.44$) was between .38 and .45 points lower than the other groups.

In response to the sub questions of research question 2, we found no statistically differences between use cases relative to the OECD principles. A one-way ANOVA was performed to measure the effect of use case for each of the OECD principles. In the scenario of OECD principle 1 on "inclusive growth, sustainable development, and well being," no main effect of use case was found [$F(3, 950.00) = 0.59$, $p = 0.62$]. In the context of OECD principle 2, concerning "human-centered values and fairness," no main effect of use case was found [$F(3, 950.00) = 1.42$, $p = 0.24$]. Regarding OECD principle 3, "transparency and explainability"

no main effect of use case was found [$F(3, 950) = 1.47$, $p = 0.22$]. Concerning OECD principle 4, regarding "robustness, security and safety," no main effect of use case was found [$F(3, 520.60) = 2.18$, $p = 0.09$]. And, for OECD principle 5, "accountability," no main effect of use case was found $F(3, 522.50) = 0.83$, $p = 0.48$].

We examined whether variations in the likelihood of an unethical decision occur at the OECD principle level based on use case, which if present could point to the relative importance of the principle concerning a use case. This type of information would be useful for informing future research on developing quantitative methods for AI risk scoring. For example, such research might highlight how best to calculate AI risk scores using input on principles and other use case factors. However, we did not find any significant differences between the likelihood of an unethical decision at the OECD principle level.

Qualitative Rationale

The qualitative components of our pilot study indicated that ethical decision-making may involve a complex array of factors in the scenarios we presented. To better understand these qualitative aspects, we included a qualitative question in our final study focused on soliciting the primary reason for the decision made by the participant. The question was presented to both experimental and control group participants; however, answers were modified for control-control participants who were not presented with an AI risk score. For participants in both control conditions, the primary reason they gave regarding their likelihood to deploy the AI system was that "The AI risk level of 'too much risk' was close enough to the risk threshold set by the company" (46.7%). Less than one-third said, "Additional context is needed to decide against deployment" (28.5%). Less than 10% each chose other options (See Table 3). In contrast, for participants presented with an experimental condition, the primary reason they gave regarding their likelihood to deploy the AI system was "Additional context is needed to decide again deployment" (32%), while the second most popular rationale was "The AI risk score was close enough to the risk threshold set by the company" (25.3%). The frequencies of other responses are shown in Table 3. These findings suggest that when participants were presented with a quantified AI risk score, the risk level of the system was viewed as less ambiguous, and therefore, participants were less likely to proceed with an unethical decision. Prior research explains that both individual and situational factors influence ethical decision-making, and quantified risk scores may contribute to providing a solid normative structure where decision-makers can more clearly understand the difference between right and wrong (Thiel et al., 2012; Trevino, 1986).

DISCUSSION

Strengths and Limitations

The study benefited from an experimental design in which participants are randomly assigned to control or experimental conditions, which helps strengthen the generalizability of the findings. Using an online platform to recruit participants allowed the study to recruit participants with diverse backgrounds across the characteristics of sex, ethnicity, and age. A potential limitation of using this type of platform is that it recruits participants skewed towards the backgrounds of the countries where the platform has the largest presence. The online platform used in this study, Prolific, does have this skew; however, the countries it operates in have significant diversity within them.

This study utilized ethical vignettes to convey the real-world scenarios in which business managers might encounter AI risk scores. The study benefited from using a pilot study to validate the vignettes used in the final study. However, the use of vignette-based research has limitations. To limit the risk of introducing confounding variables, we kept the ethical vignettes consistent, only changing the variables we sought to manipulate. The study was therefore limited, and future research on the areas where this study did not find significant differences is recommended. For example, differences between the likelihood of an unethical decision at the principle level.

We also sought consistency in the use cases selected, each with a similar system purpose and associated risk-level. The benefit of this approach was that it minimized risks to internal validity (e.g., confounding variables, too few participants in a group). However, the findings don't preclude that differences in use

cases may be important in the context of an AI risk score. Therefore, future studies exploring greater variation in use cases may also be warranted. Despite the limitations of the use cases selected for this study, the findings do have important implications for practice. The findings indicate that when use cases are consistent across purpose and risk-level, the impact of a risk score is consistent. This might be helpful information for organizations seeking to deploy risk management strategies that may differ across purpose and risk-level.

Contribution to Literature & Practice

This research contributes new knowledge to scoring risk within AI systems. While there is a significant body of research concerning quantitative methods for risk and ethical decision-making in general, limited research explores the intersection of these two areas. This gap is more pronounced regarding quantifying potential ethical risk scores for AI systems, and how such scores might impact ethical decision-making. This research helps address this limit to knowledge concerning the use of AI systems.

In practice, risk scores may positively influence ethical decision-making for business managers, for example, by reducing ambiguity. Therefore, organizations might better manage risks associated with AI systems by leveraging risk scores in their risk management process and tooling. To gain this benefit, organizations will need to invest in tooling that will produce AI risk scores as well as in training for staff to utilize risk scores appropriately. Organizations will also need to closely evaluate the tooling they adopt to ensure the highest level of accuracy and performance possible for the overall score and the underlying data and metrics that feed/comprise the score, given the potential significant influence on business manager decision-making. The potential for business managers to rely on risk scores, amplifies the obligation for organizations to ensure that the risk score is as representative of the true risk as possible. Our findings point to the broader significance of using AI risk scores as part of an AI governance framework.

ACKNOWLEDGEMENTS

The authors wish to acknowledge support from the Notre Dame - IBM Technology Ethics Lab, which aided in the completion of this project. They also wish to thank Anna Jang, Livia Johan, Nicole McAlee, Jing Tong, and Jungmin Lee for their research assistance and helpful contributions.

REFERENCES

Baudel, T., Colombet, G., & Hartmann, R. (2023, April). AI decision coordination: Easing the appropriation of decision automation for business users. *IHM'23 - 34e Conférence Internationale Francophone sur l'Interaction Humain-Machine, AFIHM*; Université de Technologie de Troyes, Troyes, France. Retrieved from https://hal.science/hal-04046408

Bommer, M., Gratto, C., Gravander, J., & Tuttle, M. (1987). A behavioral model of ethical and unethical decision-making. *Journal of Business Ethics*, *6*(4), 265–280. https://doi.org/10.1007/BF00382936

Bucinca, Z., Chouldechova, A., Vaughan, J.W., & Gajos, K.Z. (2022). Beyond end predictions: Stop putting machine learning first and design human-centered AI for decision support. In *Virtual workshop on human-centered AI workshop at NeurIPS (HCAI@ NeurIPS'22)* (pp. 1–4). Virtual Event, USA.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 752–766.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, *83*(1), 50–58.

Creswell, J.W., & Creswell, J.D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th edition). *Kindle for PC version*.

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence*. Retrieved September from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Ezeani, G., Koene, A., Kumar, R., Santiago, N., & Wright, D. (2021). *A survey of artificial intelligence risk assessment methodologies: The global state of play and leading practices identified [White paper].* EY, Trilateral Research. Retrieved from https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf.

Feuerriegel, S., Shrestha, Y.R., von Krogh, G., & Zhang, C. (2022). Bringing Artificial Intelligence to Business Management. *Nature Machine Intelligence*, *4*(7), 611–613. https://doi.org/10.1038/s42256-022-00512-5

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1.1*. https://doi.org/10.1162/99608f92.8cd550d1

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120.

Hassan, S., Pandey, S., & Pandey, S.K. (2021). Should managers provide general or specific ethical guidelines to employees: Insights from a mixed methods study. *Journal of Business Ethics*, *172*(3), 563–580.

Information Commissioner's Office. (2020). *Guidance on the AI auditing framework: Draft guidance for consultation*. Retrieved from https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Islam, S.R., Eberle, W., & Ghafoor, S.K. (2020). *Towards quantification of explainability in explainable artificial intelligence methods.* The Thirty-Third International Flairs Conference, North Miami Beach, United States.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Johansen, I.L., & Rausand, M. (2015). Ambiguity in risk assessment. *Safety Science*, *80*, 243–251.

Johnson, R.A., & Zhang, S. (2022). What is the bureaucratic counterfactual? Categorical versus algorithmic prioritization in U.S. social policy. *2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533223

Loe, T.W., Ferrell, L., & Mansfield, P. (2000). A review of empirical studies assessing ethical decision-making in business. *Journal of Business Ethics*, *25*, 185–204.

MacKinnon, B., & Fiala, A. (2017). *Ethics: Theory and contemporary issues* (9th edition). [Kindle for PC version].

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016). *Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review.* Portland International Conference on Management of Engineering and Technology (PICMET), Hawaii, US.

Metcalf, J., Moss, E., Watkins, E.A., Singh, R., & Elish, M.C. (2021). Algorithmic impact assessments and accountability. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 735–746). Virtual Conference. ACM.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507.

Nicolae, M., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., . . . Edwards, B. (2018). Adversarial Robustness Toolbox V1.0.0. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1807.01069

OECD. (2019). *Recommendation of the council on artificial intelligence.* OECD, Paris, France.

OECD. (2022). *OECD framework for the classification of AI systems (OECD Digital Economy Papers No. 323)*. OECD, Paris, France.

Piorkowski, D., Hind, M., & Richards, J. (2022). *Quantitative AI risk assessments: Opportunities and challenges*. Retrieved from https://arxiv.org/abs/2209.06317

Rastogi, C., Zhang, Y., Wei, D., Varshney, K.R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–22. https://doi.org/10.1145/3512930

Schwartz, M.S. (2015). Ethical decision-making theory: An integrated approach. *Journal of Business Ethics*, *139*(4), 755–776. https://doi.org/10.1007/s10551-015-2886-8

Szepannek, G., & Lübke, K. (2021). Facing the challenges of developing fair risk scoring models. *Frontiers in Artificial Intelligence*, *4*, 681915.

Thiel, C.E., Bagdasarov, Z., Harkrider, L., Johnson, J.F., & Mumford, M.D. (2012). Leader ethical decision-making in organizations: Strategies for sensemaking. *Journal of Business Ethics*, *107*(1), 49–64.

Trevino, L.K. (1986). Ethical decision-making in organizations: A person-situation interactionist model. *Academy of Management Review*, *11*(3), 601–617.

Trevino, L.K. (1992). Experimental approaches to studying ethical-unethical behavior in organizations. *Business Ethics Quarterly*, *2*(2), 121–136.

Vaccaro, M.A. (2019). *Algorithms in human decision-making: A case study with the COMPAS risk assessment software*. Retrieved from https://dash.harvard.edu/handle/1/37364659

Vroom, V.H., & Pahl, B. (1971). Relationship between age and risk taking among managers. *Journal of Applied Psychology*, *55*(5), 399–405.

Zhang, Y., Liao, Q.V., & Bellamy, R.K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-Assisted decision-making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372852

APPENDIX

ETHICAL VIGNETTES AND SURVEY QUESTIONS

Ethical Vignettes for Research Question 1

Ethical Vignette (Control Group - No AI Risk Score)

As a business manager in a company that builds artificial intelligence (AI) systems for clients, you are responsible for deciding whether to allow AI systems that have been developed to be used by clients. You must decide whether to deploy an AI system that your technology team has informed you has too much ethical risk. The company you work for has specified that clients should use no system if it contains too much ethical risk. However, your company has already invested considerable time and money into the development of the AI system, and it is expected to provide significant benefits to the client.

Ethical Vignette (Experimental Condition - AI Risk Score)

As a business manager in a company that builds artificial intelligence (AI) systems for clients, you are responsible for deciding whether to allow AI systems that have been developed to be used by clients. You must decide whether to deploy an AI system that your technology team has informed you has an AI risk score of 25%. This AI risk score represents the potential ethical risk associated with the system. The company you work for has specified that clients should use no system if the AI risk score is above 20%. However, your company has already invested considerable time and money into the development of the AI system, and it is expected to provide significant benefits to the client.

Survey Questions for Research Question 1

Dependent Variable Question

How likely are you to allow the system to be used by the client? (ethical decision)

Manipulation Check

I am aware that I have been informed of an AI risk score.

Ethical Vignettes for Research Question 2

RQ2 Introduction for Those in RQ1 Control Condition

You must decide whether to deploy a different AI system that your technology team has informed you also has an AI risk score of 25%. As was the case in the first scenario, this AI risk score represents the potential ethical risk associated with the system. The company you work for still specifies that no system should be used by clients if the AI risk score is above 20%. Your company has also invested considerable time and money into the development of this AI system, and it is also expected to provide significant benefits to the client.

{Insert Experimental Conditions from below}

RQ2 Introduction for Those in RQ1 Experimental Condition

You must decide whether to deploy a different AI system that your technology team has informed you has an AI risk score of 25%. This AI risk score represents the potential ethical risk associated with the system. The company you work for specifies that clients should use no system if the AI risk score is above 20%. Your company has also invested considerable time and money into the development of this AI system, and it is also expected to provide significant benefits to the client.

{Insert Experimental Conditions from below}

Experimental Conditions

Control Group: *No use case is described.*

Education Condition: This AI system is used for screening applications for students who have applied to attend a university.

Employment Condition: This AI system is used for screening applications for people who have applied for an employment opportunity with the client.

Government Benefits Condition: This AI system is used for screening applications for citizens who have applied for government benefits.

Survey Questions for Research Question 2

Dependent Variable

Question (ethical decision): How likely are you to allow the system to be used by the client?

Ethical Principle Questions

Note: We used layman's terms instead of a direct quote of OECD principles, and chose to narrow some of the principles to help make them understandable for participants and fit the context of the vignette.

Question (Principle 1: "Inclusive Growth, Sustainable Development, and Well-Being")

You learn one of the ethical concerns is related to the impact to people and planet. Given the use case for this system, how likely are you to allow the system to be used by the client?

Question (Principle 2: "Human-Centred Values and Fairness")

You learn one of the ethical concerns is related to human rights and fairness. Given the use case for this system, how likely are you to allow the system to be used by the client?

Question (Principle 3: "Transparency and Explainability")

You learn one of the ethical concerns is related to transparency and explainability. Given the use case for this system, how likely are you to allow the system to be used by the client?

Question (Principle 4: "Robustness, Security and Safety")

You learn one of the ethical concerns is related to security and privacy of the AI system. Given the use case for this system, how likely are you to allow the system to be used by the client?

Question (Principle 5: "Accountability")

You learn one of the ethical concerns is related to human accountability for the AI system. Given the use case for this system, how likely are you to allow the system to be used by the client?

Manipulation Checks

Education Condition: I am aware that the use case for this system is screening applications for students who have applied to attend a university.

Employment Condition: This AI system is used for screening applications for people who have applied for an employment opportunity with the client.

Government Benefits Condition: I am aware that the use case for this system is screening applications for citizens who have applied for government benefits.

Qualitative Rationale Question

For Participants in Control Groups for Both Research Questions

Please indicate your primary reason for your decision:

- There could be significant benefits to deployment.
- The company has already invested considerable time and money in the system.
- The AI risk level of "too much risk" was close enough to the risk threshold set by the company.
- I don't trust that the assessment of "too much risk" is accurate.
- I would not be personally accountable for deciding against deployment.
- Additional context is needed to decide against deployment.

For Participants in at Least One Experimental Condition

Please indicate your primary reason for your decision:

- There could be significant benefits to deployment
- The company has already invested considerable time and money in the system
- The AI risk score was close enough to the risk threshold set by the company
- I don't trust that the AI risk score is accurate
- I would not be personally accountable for deciding against deployment
- Additional context is needed to decide against deployment

TABLE 1
PARTICIPANT DEMOGRAPHICS

		N = 955	%
Sex			
	Female	360	37.7
	Male	594	62.2
	Prefer not to say	1	< 0.1
Race			
	White	724	75.8
	Black	90	9.4
	Asian	68	7.1
	Mixed	51	5.3
	Other	19	1.2
	Missing	3	< 0.1
Regions			
	Americas	438	45.9
	EU/Africa	427	44.7
	Asia/Australia	90	9.4
Age			
	18-19	2	< 0.1
	20-29	172	18
	30-39	312	32.7
	40-49	199	20.8
	50-59	152	15.9
	60-69	84	8.8
	70-79	26	2.7
	80-89	1	< 0.1
	Missing	7	< 0.1

TABLE 2
DESCRIPTIVE STATISTICS

			M	(S.D.)	N
RQ1	How likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system when risk score is not present	**4.65**	**(2.73)**	**360**
		Likelihood of deploying system when risk score present	3.87	(2.62)	594
RQ2	How likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system in general use case	3.72	(2.68)	277
		This AI system is used for screening applications for students who have applied to attend a university.	3.79	(2.56)	238
		This AI system is used for screening applications for people who have applied for an employment opportunity with the client.	3.74	(2.52)	218
		This AI system is used for screening applications for citizens who have applied for government benefits.	3.34	(2.44)	221
RQ2.1	If this system had an ethical concern related to impact on people and planet, how likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system in general use case	2.9	(2.39)	277
		This AI system is used for screening applications for students who have applied to attend a university.	2.97	(2.46)	238

		This AI system is used for screening applications for people who have applied for an employment opportunity with the client.	2.97	(2.27)	218
		This AI system is used for screening applications for citizens who have applied for government benefits.	2.71	(2.33)	221
RQ2.2	If this system had an ethical concern related to human rights and fairness, how likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system in general use case	2.7	(2.36)	277
		This AI system is used for screening applications for students who have applied to attend a university.	2.7	(2.27)	238
		This AI system is used for screening applications for people who have applied for an employment opportunity with the client.	2.82	(2.26)	218
		This AI system is used for screening applications for citizens who have applied for government benefits.	2.4	(2.03)	221
RQ2.3	If this system had an ethical concern related to transparency and explainability, how likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system in general use case	3.49	(2.51)	277
		This AI system is used for screening applications for students who have applied to attend a university.	3.47	(2.49)	238
		This AI system is used for screening applications for people who have applied for an employment opportunity with the client.	3.65	(2.36)	218
		This AI system is used for screening applications for citizens who have applied for government benefits.	3.18	(2.24)	221

RQ2.4	If this system had an ethical concern related to security and privacy of the AI system, how likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system in general use case	2.81	(2.35)	277
		This AI system is used for screening applications for students who have applied to attend a university.	2.81	(2.37)	238
		This AI system is used for screening applications for people who have applied for an employment opportunity with the client.	2.85	(2.38)	218
		This AI system is used for screening applications for citizens who have applied for government benefits.	2.4	(2.07)	221
RQ2.5	If this system had an ethical concern related to human accountability for the AI system, how likely are you to allow the system to be used by the client? (1= Very Unlikely; 10 = Very Likely)				
		Likelihood of deploying system in general use case	3.26	(2.53)	277
		This AI system is used for screening applications for students who have applied to attend a university.	3.35	(2.44)	238
		This AI system is used for screening applications for people who have applied for an employment opportunity with the client.	3.33	(2.34)	218
		This AI system is used for screening applications for citizens who have applied for government benefits.	3.05	(2.23)	221
Bold indicates statistical significance ($p < 0.001$)					

TABLE 3
QUALITATIVE RATIONALE

Control/Control Group		N = 137	%
	The AI risk level of "too much risk" was close enough to the risk threshold set by the company.	64	46.7
	Additional context is needed to decide against deployment.	39	28.5
	The company has already invested considerable time and money in the system.	13	9.5
	There could be significant benefits to deployment.	12	8.8
	I don't trust that the assessment of "too much risk" is accurate.	6	4.4
	I would not be personally accountable for deciding against deployment.	3	2.2
Experimental Groups		**N = 818**	**%**
	Additional context is needed to decide against deployment.	262	32.0
	The AI risk score was close enough to the risk threshold set by the company.	207	25.3
	The company has already invested considerable time and money in the system.	105	12.8
	I don't trust that the assessment of "too much risk" is accurate.	103	12.6
	There could be significant benefits to deployment.	75	9.2
	I would not be personally accountable for deciding against deployment.	61	7.5
	Did not respond	5	< 0.1