Summer 2021

Communicating uncertain information from deep learning models to users

Harishankar Vasudevanallur Subramanian

Department:

COMMUNICATING UNCERTAIN INFORMATION FROM DEEP LEARNING MODELS TO USERS

by

HARISHANKAR VASUDEVANALLUR SUBRAMANIAN

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN ENGINEERING MANAGEMENT

2021

Approved by:

Casey Inez Canfield
Daniel B. Shank
Cihan B. Dagli

PUBLICATION THESIS OPTION

This thesis consists of the following articles, formatted in the style used by the Missouri University of Science and Technology:

Paper I: Pages 3-18 has been accepted by ASEM 41st International Annual Conference Proceedings "Leading Organizations through Uncertain Times."

Paper II: Pages 19-53 are intended for submission to Journal of Risk Research.

ABSTRACT

The use of Artificial Intelligence (AI) decision support systems is increasing in high-stakes contexts, such as healthcare, defense, and finance. Uncertainty information may help users better leverage AI predictions, especially when combined with domain knowledge. I conducted two human-subject experiments to examine the effects of uncertainty information with AI recommendations. The experimental stimuli are from an existing image recognition deep learning model, one popular approach to AI. In Paper I, I evaluated the effect of the number of AI recommendations and provision of uncertainty information. For a series of images, participants identified the subject and rated their confidence level. Results suggest that AI recommendations, especially multiple, increased accuracy and confidence. However, uncertainty information, which was represented visually with bars, did not significantly improve participants' performance. In Paper II, I tested the effect of AI recommendations in a within-subject comparison and the effect of more salient uncertainty information in a between-subject comparison in the context of varying domain knowledge. The uncertainty information combined both numerical (percent) and visual (color-coded bar) formats to make the information easier to interpret and more noticeable. Consistent with Paper I, results suggest that AI recommendations improved participants' accuracy and confidence. In addition, the more salient uncertainty information significantly increased accuracy, but not confidence. Based on a subjective measure of domain knowledge, participants had higher domain knowledge for animals. In general, AI recommendations and uncertainty information had less of an effect as domain knowledge increased. Results suggest that uncertainty information, can improve accuracy and potentially decrease over-confidence.

ACKNOWLEDGEMENTS

My sincere thanks to Dr. Casey Inez Canfield, without whom this research will not be possible. Dr. Canfield's mentorship and leadership qualities enabled me to complete this project with much confidence. I gained much knowledge working under her, and I am more than excited to learn more in the future as a Ph.D. student. I would also like to thank Dr. Daniel B. Shank and Dr. Cihan B. Dagli for their invaluable contributions to this project. My sincere thanks to fellow Canfield Lab peers Hannah Elder-Felske, Ankit Agarwal, Matthew Kinnison, Casey Hines, and Luke Andrews for helping me with this project. Last but most importantly, I would like to thank my friends and family for all the love and support!

We would like to thank the anonymous reviewer through the American Society of Engineering Management (ASEM) body for helpful comments on Communicating Uncertain Information from Deep Learning Models in Human Machine Teams paper that were partially addressed in section 2 analysis.

We would also like to thank the National Science Foundation (NSF) for partially funding this project.

TABLE OF CONTENTS

Page

SECTION

LIST OF ILLUSTRATIONS

PAPER I Page

PAPER II

LIST OF TABLES

1. INTRODUCTION

The use of Artificial Intelligence (AI) has exploded in high-stakes contexts. Studies have tested both fully automated systems (Rauschecker et al., 2020) and recommender systems (Bien et al., 2018; Lakhani & Sundaram, 2017; Patel et al., 2019) in high-stakes scenarios like medical diagnosis. In most cases, recommender systems (i.e., AI decision support systems), are desired so that human experts can use their domain knowledge along with decision support system recommendations to ensure a successful outcome (Zhang et al., 2020). As experimental evidence has shown, human-AI teams, especially for lay people, are sometimes less accurate than the decision support system alone (Green & Chen, 2019; Grgic-Hlaca et al., 2019; Lin et al., 2020). Uncertainty information may help users better leverage AI predictions, especially when combined with their own domain knowledge. However, empirical research on the effects of communicating uncertainty with AI recommendations is limited (Bhatt et al., 2020). This thesis includes two studies that use an existing image recognition deep learning model to examine the effects of an AI decision support system on users' accuracy and confidence. We also measure the interaction effects of decision support system and users' self-reported domain knowledge on their accuracy and confidence.

In Paper I, we evaluated the effect of the number of AI recommendations and provision of uncertainty information. For a series of images, participants identified the subject and rated their confidence level. Results suggest that providing AI recommendations increased accuracy and confidence, especially when multiple AI recommendations were present. However, uncertainty information, which was

represented visually with bars, did not significantly improve participants' accuracy or confidence. In Paper II, we tested the effect of AI recommendations in a within-subject comparison and the effect of more salient uncertainty information in a between-subject comparison in the context of varying domain knowledge. The uncertainty information combined both numerical (percent) and visual (color-coded bar) formats to make the information easier to interpret and more noticeable. Consistent with Paper I, results suggest that AI recommendations improved participants' accuracy and confidence. In addition, the more salient uncertainty information significantly increased accuracy, but not confidence. Based on a subjective measure of domain knowledge, participants had higher domain knowledge for animals. In general, AI recommendations and uncertainty information had less of an effect as domain knowledge increased. Future work will further investigate the role of domain knowledge in the use and interpretation of AI predictions.

PAPER

I. COMMUNICATING UNCERTAIN INFORMATION FROM DEEP LEARNING MODELS IN HUMAN MACHINE TEAMS

1. INTRODUCTION

Artificial intelligence (AI) recommendations are not only found in online shopping, streaming services, and smart home devices. Increasingly, there are efforts to embed AI recommendations in high-risk work contexts such as the military, healthcare, and manufacturing (Ashiku & Dagli, 2019; Gottapu & Dagli, 2018). Consequently, it is critical to understand how people use AI recommendations in situations with varying uncertainty and potential impacts.

One popular approach to AI is deep learning. In the context of image recognition, deep learning models use neural networks to find similarities in each image and categorize them accordingly (see Figure 1). Neural networks are essentially rows of computational cells in layers that process information individually and pass information on to the next layer. The network learns and thus improves the more it is used. These networks start to recognize patterns between examples, which helps classify future examples or information. While neural networks excel at specific tasks as they learn from data, they are poor at extrapolation. It is possible to give prediction probabilities for different choices in clustering problems for deep learning models that use "softmax" functions in the last layer of the network. This probability is valuable for AI systems that

interact with humans as a representation of uncertainty or confidence for each recommendation.

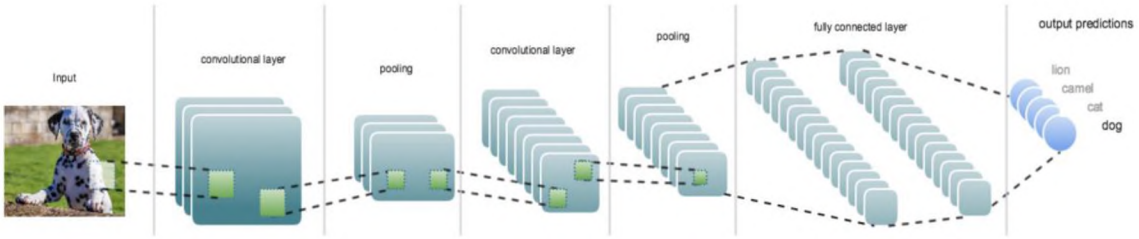

Figure 1. Example of a deep learning model with artificial neural networks for image recognition.

This research draws on insights from the literature on communicating AI recommendations and communicating uncertainty. This study provides human participants with recommendations from an image recognition deep learning model to answer two primary research questions:

- Does human performance improve when participants receive multiple recommendations instead of a single recommendation? Do multiple recommendations need to be ranked?
- Does providing a confidence bar for each recommendation improve performance?

1.1. COMMUNICATING AI RECOMMENDATIONS

It is important for human users to understand both the capabilities and limitations of AI when used for decision-making. Experimental evidence suggests that a detailed example of how the AI will help the user in the activity may provide a better

understanding for the users (Amershi et al., 2019). Raising awareness of mistakes made by the AI can increase acceptance of AI assistance. This "expectation-setting intervention" helps users understand how the AI works and be more accepting of mistakes (Kocielnik et al., 2019). People are also sensitive to how AI recommendations are communicated. For example, when performing a 2D task (such as on a computer screen), people are more influenced by a 2D on-screen agent. However, when performing a 3D task (such as operating a machine), people are more influenced by the recommendations of a 3D robot interface (Shinozawa et al., 2005). This suggests that the AI recommendations need to be presented in a way that is consistent with the task.

1.2. COMMUNICATING UNCERTAINTY INFORMATION.

One strategy for communicating the limitations of AI is to include uncertainty or confidence information with the recommendations. However, one of the challenges is that there may be different types of uncertainty associated with the training and test data vs. the model (van der Bles et al., 2019). In addition, visual communications of risk (or uncertainty) that improve quantitative understanding differ from the types of visualizations that encourage behavior change. Being able to make comparisons between categories (e.g., part vs. whole) is effective for increasing understanding. Without the ability to make comparisons, it is much more challenging to interpret the information (Ancher et al., 2006). In a review of the health communication literature, Lipkus & Hollands (1999) find that providing numerical and written information in addition to visualizations improves the perception of risk and perceived helpfulness. The visual

representation of risk (or uncertainty) is more effective for helping people make decisions that affect them positively (Lipkus & Hollands, 1999; Lipkus, 2007).

2. METHOD

2.1. DESIGN

We recruited 286 participants from Prolific, an online participant pool platform. In order to participate, participants had to be over 18 and speak English. Prolific offers a more diverse group of English-speaking participants in terms of geographical location and ethnicity (Peer et al., 2017). Participants performed an image recognition task. Each participant was randomly assigned to one of six conditions:

a) *No Recommendation Control* – no AI recommendation or confidence bar provided,

b) *1 AI Recommendation/Text Only* – top recommendation by AI,

c) *1 AI Recommendation/Confidence Bar* – top recommendation by AI with confidence bar,

d) *5 AI Recommendations/Alphabetical Control* – top five recommendations by the AI in alphabetical order,

e) *5 AI Recommendations/Text Only* – top five recommendations by the AI in ranked order, and

f) *5 AI Recommendations/Confidence Bar* – top five recommendations by the AI in ranked order with confidence bar for each recommendation.

Figure 2 below shows examples of experimental stimulus for each condition. Within each condition, each participant identified 24 images and answered additional survey questions.

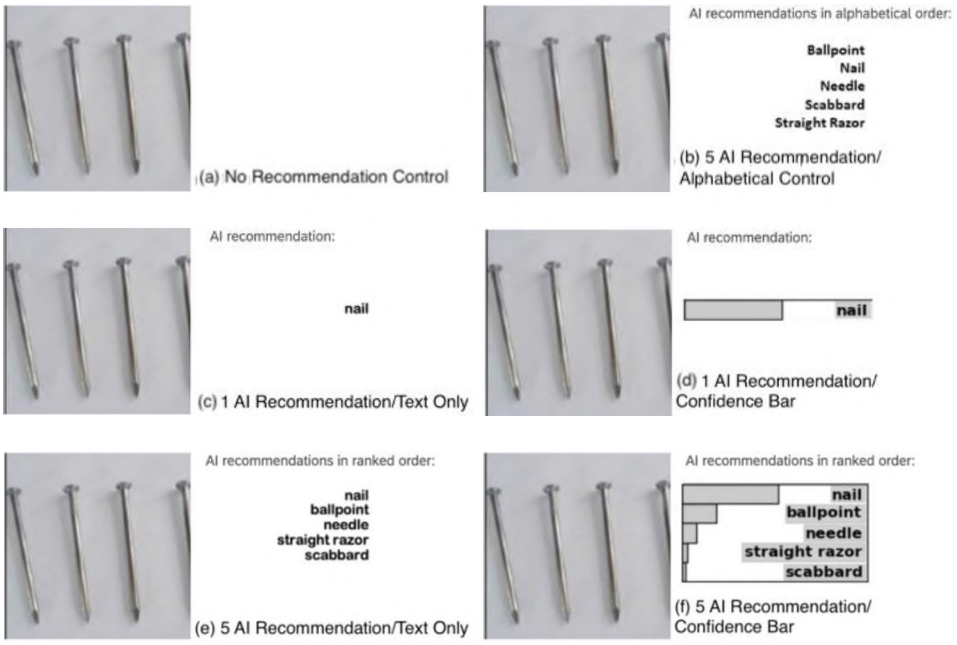

Figure 2. Example stimulus for each of the six conditions.

2.2. STIMULI

The images, AI recommendations, and confidence bars were drawn from the supplementary materials of Krizhevsky et al. (2012), which leverages the ImageNet database (Deng et al., 2010). The ImageNet database is made up of 12 subsets consisting of 3.2 million images in 5,247 categories. Deng et al. (2010) used participants from Amazon mTurk to label these images. The Krizhevsky et al. (2012) model used in this study was trained on 1.2 million images in 1,000 categories. To avoid overfitting,

Krizhevsky et al. (2012) augmented the model by scaling all the input images to 256 x 256 resolution and by altering the RGB scales of all the images. From the 88 images provided in the supplementary materials by Krizhevsky et al., (2012), we selected 24 to use in this study where the image label was clearly a focus of the image and there was a mix of correct and incorrect AI recommendations.

2.3. MEASURES

Before viewing the images, participants completed two attention check questions: "In the instructions, an example image was given along with the correct label for that image. What was the correct answer for the example image?" (answer: "howler monkey") and "How did the instructions say to describe the picture?" (answer: "be specific"). In addition, there was one attention check embedded in the images where participants were asked to identify the image that was explained in the instructions. These items were combined into an attention indicator, where 1 indicates that the participant passed all three of the attention checks and 0 indicates that they failed at least one. In addition, we measured the average time spent per image.

For each of the 24 images, participants identified the subject of the image ("What is this a picture of?") in an open textbox. The responses were manually categorized into the following types of accuracy:

(1) Exact Match – answer matched the image label,

(2) Synonym – answer was an alternate or similar name to the image label (e.g., Metal Nails instead of Nail),

(3) Present – the answer was present in the image but not the image label (e.g. White Wall instead of Nail),

(4) Category – the answer was a broader category, rather than specific (e.g. Hardware instead of Nail),

where each level includes the previous level. In other words, if the response was "Category correct", then it was also considered correct for the other levels. After each image, participants indicated their confidence on a 6-point scale that ranged from 0-100% confident ("How confident are you in your answer?").

Following the series of images, participants rated the difficulty of the task ("How difficult was this task?") on a 5-point Likert scale that ranged from "extremely difficult" to "extremely easy". We also measured demographics including gender, education, and age. Four participants did not report their education level. Age was highly skewed, so a log transformation was used to normalize the measure. A separate ANOVA was run for each definition of accuracy, where the outcome (or dependent) variable was the performance of an individual participant across 24 images. Due to the high number of statistical tests, we focus on interpreting effects with $p < 0.01$ to reduce false positives.

3. RESULTS AND DISCUSSION

Participants were predominantly female (67%) and approximately half had at least a 4-year college degree. The average age was 33 years old and ranged from 18 to 67 years old. Table 1 summarizes measures across experimental conditions. The demographics and attention measures did not significantly vary across the experimental conditions. This suggests that the random assignment was successful and there are no

systematic differences between the experimental groups. Older participants tended to spend more time per image, $r(284) = .27$, $p < .001$. In addition, participants that were more confident tended to spend more time per image, $r(284) = .15$, $p = .01$, and perceive the task as more difficult, $r(284) = .26$, $p < .001$.

Table 1. Mean and standard deviation for each experimental condition. Accuracy, confidence, and task difficulty differed across experimental conditions.

		Controls		1 AI Recommendation		5 AI Recommendations	
	Total	**No Rec**	**Alphabetical Recs**	**Text Only**	**Confidence Bar**	**Ranked Text**	**Confidence Bar**
Participants	286	46	45	49	49	48	49
Exact Match Accuracy	45% (31%)	25% (27%)	47% (35%)	46% (35%)	49% (39%)	49% (40%)	50% (37%)
Synonym Accuracy	55% (32%)	38% (31%)	58% (33%)	56% (36%)	58% (38%)	60% (33%)	61% (35%)
Present Accuracy	64% (30%)	48% (33%)	66% (31%)	64% (34%)	65% (37%)	68% (31%)	69% (32%)
Category Accuracy	77% (27%)	75% (20%)	76% (28%)	77% (31%)	74% (34%)	78% (27%)	79% (30%)
Confidence	69% (14%)	63% (19%)	67% (10%)	71% (15%)	66% (14%)	73% (11%)	73% (11%)
Task Difficulty	3.2 (1.1)	3.8 (1.0)	4.3 (1.0)	3.3 (1.1)	3.2 (1.1)	2.7 (1.0)	2.9 (1.1)
% Passed Attention	75% (43%)	78% (42%)	76% (43%)	80% (41%)	80% (41%)	69% (47%)	69% (47%)
Time per image (secs)	22 (15)	21 (12)	27 (14)	23 (18)	18 (14)	23 (16)	22 (12)
% Male	33%	33%	33%	29 %	33%	33%	33%
% College	50%	59%	36%	59%	45%	50%	52%
Age	33 (11)	33 (10)	33 (11)	34 (12)	31 (9)	34 (13)	31 (10)

As shown in Table 2, separate ANOVAs were conducted for each definition of accuracy. Performance differed across experimental conditions and confidence. In addition, there were weakly significant effects at the $p < .05$ level for attention and task difficulty. Tukey HSD post hoc tests indicated that when compared to the control condition, accuracy was higher in all of the AI conditions ($p < .01$), but there was no significant difference between the AI conditions (see Table 1 and Table 3). This was true across all definitions of accuracy except Category accuracy, which uses the most lenient definition. In this case, there was no significant difference between the control and AI conditions (although post hoc tests indicated that a few comparisons approached, but did not achieve, statistical significance).

Participants that were more confident tended to have higher Synonym and Category accuracy. From a metacognition perspective, the Category accuracy effect suggests that participants knew when they did or did not have a vague sense (i.e. the category) of an image. More investigation is needed to determine the mechanism for Synonym accuracy. A one-way ANOVA indicates that the average confidence varied across experimental conditions, $F(5, 280) = 4.41$, $p < .001$. Post hoc comparisons using the Tukey HSD test suggest that participants in the 1 AI Recommendation/Text Only, 5 AI Recommendation/Text Only, and 5 AI Recommendation/Confidence Bar conditions were significantly more confident that the No Recommendation Control group (see Table 1). This suggests that the confidence bar increased confidence (compared to the No Recommendation Control condition) when there were 5 AI recommendations, but not when there was only 1 AI recommendation. The confidence bar may help sort among

multiple recommendations, but simply serves to decrease confidence if there are no alternative recommendations.

A one-way ANOVA showed that the perceived task difficulty varied across the experimental conditions, $F(5, 280) = 6.28$, $p < .001$. Tukey HSD post-hoc tests indicate that the 5 AI Recommendations/Alphabetical Control condition was perceived as significantly more difficult than the 5 AI Recommendations/Text Only condition (see Table 1). In addition, the 5 AI Recommendations/Text Only and 5 AI Recommendations/Confidence Bar conditions were perceived as significantly less difficult than the No Recommendation Control condition. This suggests that providing multiple recommendations made the task less difficult, as long as the recommendations were ranked.

Table 2. Separate ANOVA for each accuracy definition. Accuracy differed across experimental conditions ($p < .01$).

	Exact Match		**Synonym**		**Present**		**Category**	
	F	η^2	**F**	η^2	**F**	η^2	**F**	η^2
AI Recommendation	45.11***	0.44	34.46***	0.38	29.00***	0.34	2.88*	0.05
Confidence	6.31*	0.001	5.70*	0.01	1.64	0.00	6.37*	0.02
Task Difficulty	3.40	0.01	1.936	0.00	0.65	0.00	0.08	0.00
Attention	5.96*	0.01	5.10*	0.01	4.58*	0.01	2.97	0.01
Time per Question	2.62	0.01	1.57	0.00	2.13	0.01	0.16	0.00
% Male	2.50	0.00	1.16	0.00	0.80	0.00	0.54	0.00
% College	0.94	0.00	0.07	0.00	0.03	0.00	0.00	0.00
Age (logged)	0.06	0.00	1.20	0.00	0.69	0.00	0.08	0.00

Note: *$p < .05$, **$p < .01$, and ***$p < .001$

When excluding the Control conditions, it is possible to examine the potential interaction of the number of AI recommendations and the use of the confidence bar. As shown in Table 3, there is a significant difference due to the number of AI recommendations for all definitions of accuracy. However, the difference is weakly significant for the Exact Match accuracy ($p < .05$), which is the most restrictive definition of accuracy. Providing 5 recommendations rather than 1 recommendation increased performance for exact match (50% vs. 47%), synonym (61% vs. 57%), present (68% vs. 64%), and category (79% vs. 75%) accuracy. However, the use of confidence bars was not associated with any significant differences, suggesting that this information did not improve participant accuracy.

Table 3. Two-way ANOVA for each accuracy definition. Accuracy differed for the number of AI recommendations, but not use of confidence bar ($p < .01$).

	Exact Match		**Synonym**		**Present**		**Category**	
	F	η^2	**F**	η^2	**F**	η^2	**F**	η^2
Number of AI Recs	4.06*	0.02	8.59**	0.04	7.89**	0.04	11.47***	0.06
Bar	3.36	0.02	1.63	0.01	1.23	0.01	0.02	0.00
Number of AI Recs * Bar	0.98	0.00	0.38	0.00	0.05	0.00	0.62	0.00
Confidence	0.00	0.00	0.51	0.00	1.34	0.01	0.24	0.00
Task Difficulty	2.60	0.01	1.51	0.01	0.48	0.00	0.01	0.00
Attention	4.49*	0.02	2.82	0.00	3.14	0.02	3.61	0.02
Time per Question	0.44	0.00	0.11	0.00	0.73	0.00	0.06	0.00
% Male	2.55	0.01	1.31	0.01	0.88	0.00	0.34	0.00
% College	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00
log (Age)	0.12	0.00	0.65	0.00	0.29	0.00	0.43	0.00

Note: *$p < .05$, **$p < .01$, and ***$p < .001$

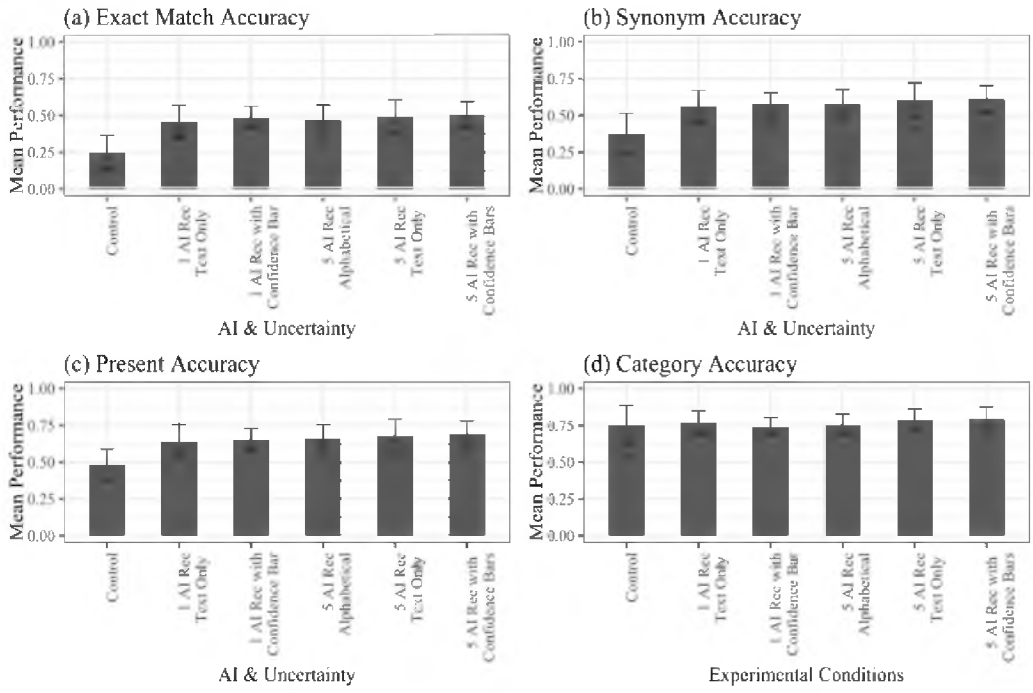

Figure 3. Mean performance of the participants in each experimental condition across all accuracy definitions. The AI conditions improved performance for exact match, synonym, and present accuracy.

4. CONCLUSION

The results suggest that AI recommendations improve accuracy for human-led image recognition tasks across multiple definitions of accuracy. In addition, providing additional recommendations (5 vs. 1) improves accuracy, but the use of confidence bars was not associated with any significant differences. For Category accuracy, the broadest definition of accuracy, there was a weak difference between the experimental and control conditions. This suggests that there were some images that did not benefit from AI recommendations, when using the most generous definition of accuracy. In addition, when examining the effect of the number of AI recommendations, there was a weak

effect for Exact Match accuracy, suggesting that additional recommendations may not help for narrow definitions of accuracy. This work suggests that AI recommendations are generally helpful even when the human and machine or AI components of a system have different definitions of accuracy. In this experiment, the Exact Match accuracy is the only case where the human and AI definitions match. For Synonym and Present accuracy, the human is recognizing more aspects of the image than the AI, yet the AI recommendations are still improving accuracy.

The AI recommendation conditions differ in how they influenced confidence. Participants in the 1 AI Recommendation/Text Only, 5 AI Recommendations/Text Only, and 5 AI Recommendations/Confidence Bar conditions were significantly more confident that the No Recommendations Control group. This suggests that ranked AI recommendations are associated with higher confidence. In addition, the confidence bars are more helpful for increasing confidence when sorting through multiple recommendations. In terms of metacognition or people's ability to "know what they know", participants were able to distinguish between Category accuracy and wrong answers. However, they did not know whether they were focusing on the same aspect of the image as the AI. More investigation is needed to determine the mechanism for Synonym accuracy.

Providing multiple recommendations made the task seem less difficult, as long as the recommendations were ranked. The 5 AI Recommendations/Alphabetical Control condition was perceived as the most difficult while the 5 AI Recommendations/Text Only and 5 AI Recommendations/Confidence Bar conditions were perceived as the least difficult. This suggests that providing multiple ranked recommendations with confidence

bars from an AI system may increase human operator confidence and reduce the perceived difficulty of the task.

Future research efforts will further investigate principles for designing AI recommendation communications. The research team will explore stimuli-level effects, the impact of AI recommendations that are not correct, and the role of attention. This work is based on a laboratory experiment and does not represent an ecologically valid task. As a result, these findings may not be directly generalizable to workplaces or specific applications. Further research is needed to determine if there are any differences based on domain or application.

5. IMPLICATIONS

AI recommendations are increasingly being integrated into a variety of engineering management contexts (e.g., healthcare, military, manufacturing, supply chain). However, to date, there is insufficient research on integrating uncertainty or confidence information into AI recommendation communications. The results of this study suggest that it may be valuable for AI systems to provide multiple ranked recommendations, particularly if the AI is trained on a narrower task than the human operators are performing. In the context of image recognition, the AI may be focused on specific features while a human analyst is examining the broader context and may focus on different features or levels of precision. Engineering managers must consider the task characteristics to determine the appropriate strategy for communicating AI recommendations and the impacts on human performance.

More research is needed on designing communications of uncertainty for AI outputs. This study found no evidence of a performance benefit associated with including uncertainty or confidence bars for each recommendation. However, there are many types of uncertainty. For example, temporal uncertainty refers to uncertainty about future events. Structural uncertainty refers to uncertainty that is introduced as a function of the model. Measurement uncertainty refers to uncertainty associated with measuring specific values and translational uncertainty refers to the uncertainty introduced in the communication process (Rowe, 1994). This work focuses on developing communications for a measure that incorporates multiple types (e.g., structural and metrical). Future work should explore strategies for designing communications that differentiate between kinds of uncertainty. In addition, future work should investigate combining visual and numerical uncertainty information. Ultimately, this research effort aims to develop communications that improve the performance of human-machine teams.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-gil, R., & Horvitz, E. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-13.

Ashiku, L. & Dagli, C. J. (2019). System of Systems (SoS) Architecture for Digital Manufacturing Cybersecurity. *Procedia Manufacturing*, 39, 132-140. https://doi.org/10.1016/j.promfg.2020.01.248.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248-255. https://doi.org/10.1109/cvpr.2009.5206848

Gottapu, R. D. & Dagli, C. H. (2018). DenseNet for Anatomical Brain Segmentation. *Procedia Computer Science*, 140, 179-185, https://doi.org/10.1016/j.procs.2018.10.327

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceeding of Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300641

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097-1105.

Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute. Monographs*, *27701*(25), 149–163. https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, 27(5), 696–713. https://doi.org/10.1177/0272989X07307271

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Rowe, W. D. (1994). Understanding Uncertainty. *Risk Analysis*, *14*(5), 743–750.

Shinozawa, K., Naya, F., Yamato, J., & Kogure, K. (2005). Differences in effect of robot and screen agent recommendations on human decision-making. *International Journal of Human Computer Studies*, 62(2), 267–279. https://doi.org/10.1016/j.ijhcs.2004.11.003

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(181870), 1-42. https://doi.org/10.1098/rsos.181870

II. ROLE OF UNCERTAINTY INFORMATION AND DOMAIN KNOWLEDGE IN USE OF AI RECOMMENDATIONS

1. INTRODUCTION

Artificial Intelligence (AI) decision support systems are increasingly common across sectors, especially for high-risk scenarios in healthcare, manufacturing, and the military (Ashiku & Dagli, 2019; Gottapu & Dagli, 2018). Peng (2018) summarized key AI system failures in 2018, highlighting the need for further understanding AI and its limitations. IBM's Watson AI Health was geared to help doctors in cancer treatment. Doctors stopped using it as they found it provided unsafe recommendations that could have had dire or fatal consequences. Similarly, Amazon stopped using its AI software for screening resumes after finding the AI's gender bias. The AI was trained on benchmark engineering applicant resumes, which predominantly belonged to white men. As a result, the AI predicted males would be a better fit for engineering jobs (Blier, 2020). Consequently, it is not sufficient to simply communicate an AI prediction and assume human users will know how to use the information.

People are more likely to accept an AI prediction when given a choice, particularly in a high-stakes scenario. In general, participants tend to weigh AI recommendations similar to an expert's recommendation when making decisions (Wang et al., 2020). However, there is mixed evidence of which they prefer. Ashktorab et al. (2020) found evidence that participants preferred human experts, but decision performance was not affected by whether participants perceived their partner as a human expert or an AI. In contrast, Logg et al. (2019) found that participants chose to side with

AI recommendations more than experts' when given a choice. In three different experiments, participants received the same advice either from an algorithm or other people. Results suggested that lay people relied more on advice when they thought it came from an algorithm (Logg et al., 2018). In other words, participants tended to accept the recommendation more when it came from a recommender system than humans. This mixed preference to recommender system could be tied to user's perception on the accuracy of a decision support system.

AI decision support systems can improve human decision-making if they can compensate for each other's errors, and people can discern when to follow or not follow the AI. Experimental evidence suggests that human-AI teams tend to perform better than either alone (Bansal et al., 2020; Rosenberg & Willcox, 2019), even in high-stake medical situations (Bien et al., 2018; Lakhani & Sundaram, 2017; Patel et al., 2019; Xiong et al., 2020). However, humans-AI teams tend to be less accurate than AI alone in high-stake prediction tasks like recidivism (when participants decide to grant or not grant bail for defendants) prediction (Green & Chen, 2019; Grgic-Hlaca et al., 2019; Lin et al., 2020). Recidivism studies used laypeople in their experiments, whereas the medical studies involved domain experts. This suggests that laypeople may inappropriately rely on AI predictions when it is not warranted. So, it may be helpful to provide prediction-specific guidance, such as uncertainty information. In addition, people with domain knowledge may be better able to leverage the uncertainty information.

1.1. PROVIDING UNCERTAINTY INFORMATION WITH AI PREDICTIONS

Uncertainty measures the "lack of knowledge" about an outcome. In tasks such as image recognition, uncertainty of a recommender system is the predicted probability to match the ground truth (Bhatt et al., 2020). Communicating this uncertainty to users is rather complex as almost 30% of the participants in a study could not differentiate the levels of risk between 1 in 10, 1 in 100, or 1 in 1000 (Galesic, 2010). Another study by Zikmund-Fisher et al. (2007) found that participants' risk comprehension abilities are significantly affected by their numeracy skills. Hence, it may be necessary to communicate uncertainty in more than one manner that is easy to understand for the task at hand.

Uncertainty can be communicated via text, numbers, and visuals. However, the best representation may vary based on the task and individual characteristics. For example, a study conducted by Budescu et al. (2012) used an Intergovernmental Panel on Climate Change (IPCC) report and asked the participants to translate the verbal uncertainties mentioned in the report, numerically. They found that participants interpreted "very likely" to mean 60% probability as opposed to 90% probability as intended by the IPCC. Bhatt et al. (2020) reviewed literature on communicating uncertainty and recommended using categorical or numerical methods to overcome this misinterpretation. Another common method to represent uncertainty is through graphs. Graphical representation of data enables users to identify patterns, and trends (Lipkus & Hollands, 1999). Different graphs could be used to communicate uncertainty like pie charts for proportions, bar charts for comparisons, or line charts for time-series data. Choosing the appropriate visual tool depends on the task at hand. Additionally, Gkatzia et

al. (2016) found providing uncertainty in text, numerical, and visual formats together significantly improved users' accuracy and confidence when compared to providing just visual information.

There is mixed evidence on the effectiveness of providing uncertainty information with AI predictions to improve decision-making. In some studies, providing uncertainty or confidence information increases accuracy (Bansal et al., 2020; Fernandes et al., 2018; Gkatzia et al., 2016), often because users trust the AI more (Antifakos et al., 2005). However, studies have also found no effect or limitations to the use of uncertainty information. In contrast to the studies above, Subramanian et al. (2020) found no effect of uncertainty information when represented in terms of confidence bars. Similarly, providing an explanation for why the AI is providing a recommendation, has not improved accuracy (Bansal et al., 2020). This suggests that although visual representations of uncertainty tend to be most effective, there are exceptions likely related to the saliency of the uncertainty information. In Antifakos et al. (2005), users needed more information when the AI's confidence level was below 50%. This suggests that users tend to agree with the AI when the probabilities are above 50% and choose their own answer when it is not. When AI recommendations' uncertainty is low, it requires the users to expertly navigate among the choices. As a result, domain knowledge may play a more important role when the AI is uncertain.

1.2. EFFECT OF DOMAIN KNOWLEDGE.

In general, experts are able to perform a task much better than novices. In a study by Snow et al. (2008), they found that they needed four novices to label an item to the

same accuracy of one expert. Brand-Gruwel et al. (2017) found that domain experts performed better than novices in an experimental setting of gathering reliable information on topics relating to psychology using the Internet. Another study found that with AI recommendations, experts performed better under time pressure whereas there was no accuracy differences between experts and novices without a time constraint (Dane et al., 2012). After reviewing literature on human-AI teams, Maadi et al. (2021) found that as the task difficulty or complexity increased, the need for a higher domain knowledge human expert to be in the loop also increased. This suggests that experts are better able to navigate and integrate AI information than novices.

Experts also benefit more from a complex recommender system since it gives them more control whereas novices preferred the opposite (Knijnenburg et al., 2011). However, there is some evidence that AI systems can hurt experts' accuracy if they under-rely on it. Logg et al. (2019) found experts tended to rely less on algorithmic recommendations than human recommendations, which hurt their accuracy. In contrast, novices who had low domain knowledge, tended to rely on recommendation systems more (Wang & Benbasat, 2013), especially when uncertainty information was provided (Bussone et al., 2015). This suggests that novices may agree with the AI even if the recommendations are incorrect. Additionally, an empirical study by Feng & Boyd-Graber (2019) found that experts are better able to navigate AI recommendations, suggesting that it may be extremely valuable to adapt AI interfaces to match users' skillsets.

2. AIM OF STUDY

Even though the accuracy of the users is highly studied in AI-assisted high-stake scenarios, the research on uncertainty information is limited to the effect on users' confidence (Greis et al., 2017; Zhou et al., 2015). Communicating uncertainty has been highly studied in the risk communication literature (Spiegelhalter, 2017) but, literature on communicating uncertainty and its effects on task performance in human-AI teams is also minimal (Arshad et al., 2015; Bhatt et al., 2020; Gkatzia et al., 2016). Lastly, many of the high-stakes decision support system research are in the medical field and only use experts in their studies. One exception is Huang et al. (2020) who found that the AI performed better with experts (doctors) than novices (interns). They only tested the accuracy of the decision support system when experts and interns provided training data for the system and not vice-versa. Using breast cancer patient cases from April 2017 to August 2018, McNamara et al. (2019) used IBM's AI Watson for Oncology to find that breast cancer experts' accuracy did not vary with and without the AI, whereas novices (tumor and hematologic focused oncologists) accuracy improved significantly after AI recommendations. However, Peng (2018) reported that IBM's AI Watson was stopped from use as it provided unreliable recommendations. Consequently, it is difficult to accept the findings of McNamara et al. (2019). As a result, effects of different levels of domain knowledge (experts vs novices) interacting with the decision support system recommendations and uncertainty information in high-stakes scenarios is not well documented.

To address these gaps, this study evaluates the effect on accuracy and confidence of providing AI predictions with uncertainty information and how this effect varies based on the participant's domain knowledge. Based on findings from the risk communication literature, the uncertainty information includes both numerical and visual representations as well as color to increase saliency. We use an image recognition task for pictures of plants and animals with an existing deep learning model to provide the AI recommendations and uncertainties. We test three hypotheses in a mixed-subject design:

H1. In a within-subjects comparison, ranked AI recommendations increase image recognition accuracy and confidence compared to random order AI recommendations.

H2. In a between-subjects comparison, ranked AI recommendations with uncertainty information increase accuracy and confidence more than without uncertainty information.

H3. Participants with higher domain knowledge have higher accuracy and confidence when provided ranked AI recommendations (Domain Knowledge X AI) and uncertainty information (Domain Knowledge X Uncertainty).

3. METHODS

3.1. PARTICIPANTS

We recruited 201 participants from Prolific, an online participant recruitment platform. All participants were over 18 years old and spoke English. Prolific offers comparable data and a more diverse group of participants than Amazon mTurk (Peer et

al., 2017). All participants provided informed consent and were compensated $5. This study was approved by the University of Missouri Institutional Review Board.

3.2. DESIGN

In a mixed-subject design, participants performed an image recognition task with and without AI recommendations (within-subjects). With the AI recommendations, participants were randomly assigned to receive or not receive uncertainty information (between-subjects). To evaluate the effect of domain knowledge, we used a self-reported subjective measure of knowledge related to plants and animals and analyzed the different types of stimuli separately.

In the image identification task, the image stimuli were of plants and animals. For the AI predictions and associated uncertainties, we use an existing image recognition deep learning model (Krizhevsky et al., 2012) that was trained on the ImageNet database. The ImageNet database is made up of 12 subsets consisting of 3.2 million images in 5,247 categories. Deng et al. (2010) used participants from Amazon mTurk to label these images. The Krizhevsky et al. (2012) model used in this study was trained on 1.2 million images in 1,000 categories. In their supplementary materials, Krizhevsky et al. (2012) provide 88 images with 5 label predictions and associated uncertainties displayed as bars.

We selected 33 images to use in this study where the image label was clearly a focus of the image and there was a mix of correct and incorrect AI recommendations. For each image, we provided 6 potential labels to ensure that the correct label was always included. In addition to the 5 labels provided by Krizhevsky et al. (2012), we either added the correct label or another similar but incorrect label. In addition, we redesigned the

uncertainty information to improve ease-of-interpretation compared to Subramanian et al. (2020). In the original presentation, the label text overlapped the uncertainty bars, the bars were monochrome, and no numerical information was provided. In this study, the uncertainty bars were color coded to represent the AI's confidence in its recommendation, where green is 100% - 76%, yellow is 75% - 51%, orange is 50% - 26%, and red is 25% - 0%. The uncertainty bars were separated from the label text, and we added a percentage value indicating the AI's confidence in its recommendations based on measuring the bars.

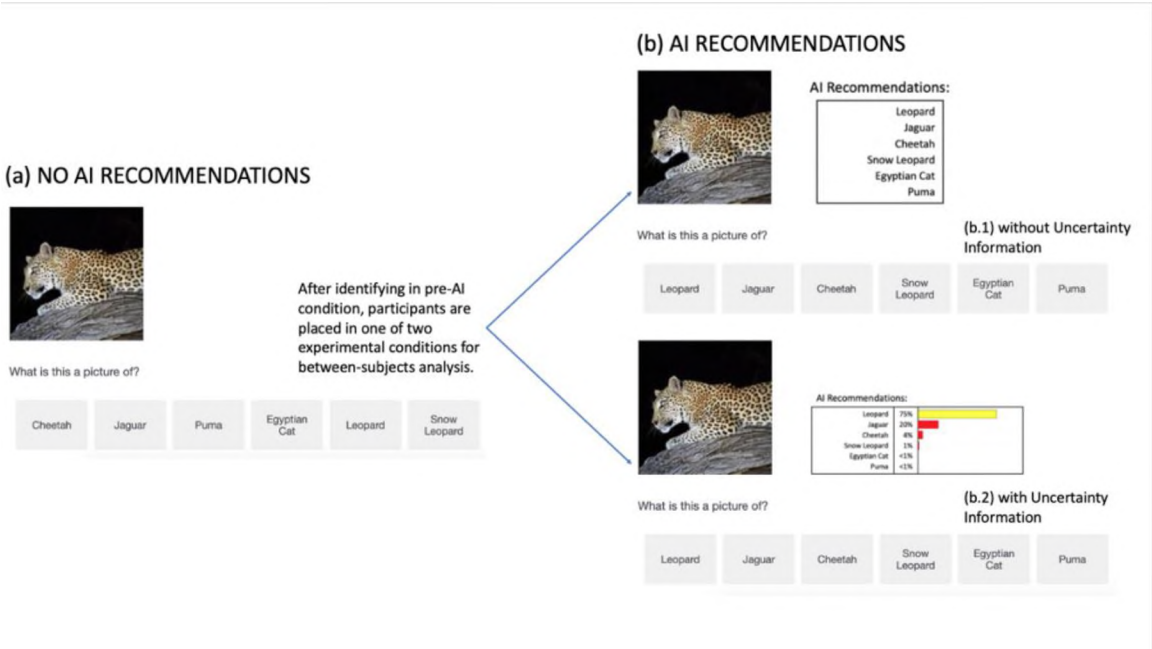

Figure 1. Example stimulus for each experimental condition. Within-subjects comparison between (a) No-AI recommendations & (b) AI recommendations. Between-subjects comparison between (b.1) AI recommendations without Uncertainty information & (b.2) AI recommendations with uncertainty information.

3.3. PROCEDURE

After providing informed consent and reading instructions, participants performed the image identification task. Participants identified 32 images, of which 19 were plants and 13 were animals via a multiple-choice question, "what is this a picture of?" First, participants selected one from six random ordered options, where the options were the AI recommendations. This response was scored as correct or incorrect to determine *pre-AI accuracy*. For each image, participants rated their *pre-AI confidence* via "How confident are you in your answer?" [1 = not confident at all (0-20%), 5 = extremely confident (80-99%), 6 = absolutely confident (100%)].

Each image was viewed twice for the within-subject comparison. However, in the second viewing, the multiple-choice options were provided in rank order according to the AI recommendations. To measure *post-AI accuracy*, participants again answered, "what is this a picture of?" For each image, participants rated (a) *post-AI confidence,* "How confident are you in your answer?" [1 = not confident at all (0-20%), 5 = extremely confident (80-99%), 6 = absolutely confident (100%)] and (b) *perceived usefulness,* "How useful was the AI in recognizing the image?" (1 = not useful at all, 6 = extremely useful). We also measured the *time taken/image* by the participant.

In order to evaluate data quality, we used three attention checks. After the instructions, we asked, "what was mentioned as the correct answer to the image provided in the instructions?" (multiple choice; answer: Howler Monkey). After the image identification task, we asked, "how many AI recommendations did you get for each image?" (multiple choice; answer: 6). We also embedded an attention check in the images where participants responded to "What is this a picture of?" (answer: Howler

Monkey), which was used in the instructions example. All three attention checks were combined into an *attention* score.

After the image identification task, participants rated their *domain knowledge* via two questions "How well can you identify plants?" and "How well can you identify animals?" (1 = not well at all, 6 = extremely well). In addition, participants rated (a) *perceived difficulty*, "How difficult was this task?" (1 = extremely easy, 6 = extremely difficult) and (b) *perceived trustworthiness*, "How trustworthy was the AI?" (1 = very untrustworthy, 6 = very trustworthy). In both cases, an "I do not know" option was treated as missing. Participants also completed cognitive measures relating AI usefulness (4 items) (Viswanath,V. & Fred D., 2000) and AI reliability (5 items) (Madsen & Gregor, 2000) on a 7-point Likert scale (1 = Strongly disagree, 4 = Neither agree nor disagree, and 7 = Strongly agree). Detailed information on the scales is provided in the Appendix. Lastly, we measured demographics, including age, gender, and education. A log transformation was used to normalize age.

3.4. ANALYSIS

We performed linear regressions to evaluate the effect of (1) AI recommendations, (2) uncertainty information, and (3) domain knowledge on (a) accuracy and (b) confidence. Each person was observed for both within and between-subject analysis. Accuracy, confidence ratings, time taken per image, and perceived AI usefulness ratings were averaged for every person. Dummy variables were used to denote AI recommendations and Uncertainty Information. Each person appears twice in the dataset, once representing their accuracy and other measures without AI

recommendations and again representing their measures with AI recommendations. Linear mixed-effects regressions evaluated the within-subject effects of AI recommendations using this data. Then, participants were identified separating for receiving and not receiving Uncertainty information used for between-subjects analysis using frequentist linear regression. To examine the effects of domain knowledge, the data was separated into plant and animal stimuli.

Due to the large number of planned tests, we use $\alpha < 0.01$ for interpretation to reduce false positives. We preregistered the analyses at Open Science Framework (https://osf.io/bjgu9/). This analysis deviates from the preregistration in one way. Initial approach included logistic mixed-effects model for analyzing accuracy to account for repeated measures in within-subjects design. Instead, we averaged the participants' accuracy across images and performed linear mixed-effects regression to account for the same repeated measures. All the data, R code, and survey materials are available on Open Science Framework (https://osf.io/bjgu9/).

4. RESULTS AND DISCUSSION

Of the 201 participants, 49% were male and 48% had completed a 4-year college degree. The average age was 33 years old, ranging from 18 to 64 years old. Overall, 74% of the participants passed all three attention checks. More than 90% of the participants responded correctly to two of the three attention checks. Participants mostly failed the last attention check in the experiment. Only 78% of the participants passed "how many AI recommendations did you get for each image?" (Answer: 6) which was asked at the

end of the survey. As summarized in Table 1, the mean time to identify an image was 13 seconds (*Median* = 10, *SD* = 14). Before the AI information, the mean time per image was 16 seconds (*Med = 12, SD =19)*. With the AI information, the mean time per image was 10 seconds (*Med = 9, SD = 6*). In a paired t-test, participants spent significantly less time on the images when viewing for a second time with the AI information, $t(200) = 4.19$, $p < .001$.

Accuracy increased in the post-AI and uncertainty conditions, but confidence did not. In a paired t-test, accuracy was higher for participants after AI information *(M = .54, Med = .56, SD = .08)* than before *(M = .36, Med = .34, SD = .08), t(200) = 30.43, p < .001; d = 2.36.* As reported in Table 2, participants average pre-AI accuracy and post-AI condition accuracy were moderately correlated, *r(201) = .39, p <.001.* In a two-sample t-test, accuracy was higher for participants who received uncertainty information *(M = .57, Med = .56, SD = .05)* than participants who did not *(M = .52, Med = .53, SD = .09), t(147.6) = 5.40, p < .001; d = .77.* The human-AI team was less accurate than the AI alone as the AI's top recommendation accuracy was 59% which is in disagreement with (Bien et al., 2018; Lakhani & Sundaram, 2017; Patel et al., 2019; Xiong et al., 2020) findings and similar to (Green & Chen, 2019; Grgic-Hlaca et al., 2019; Lin et al., 2020).

As summarized in Table 1, in a paired-test, confidence was higher for participants after AI information *(M = .57, Med = .58, SD = .16)* than before *(M = .43, Med = .41, SD = .15), t(200) = -18.07, p < .001; d = .95*. In a two-sample test, confidence was not significantly higher for participants who received uncertainty information *(M =.58, Med = .60, SD = .15)* than participants who did not *(M = .56, Med = .57, SD = .16), t(196) =*

1.07, p = .29; d = .15. Participants pre-AI condition confidence and post-AI condition confidence ratings were also highly correlated, *r(201) = .72, p <.001.*

Overconfidence was calculated by calculating the difference between participants average accuracy and average confidence ratings. In a paired-test, participants were more overconfident before AI recommendations *(M = .07, Med = .04, SD = .17)* than after AI recommendations *(M= .03, Med = .03, SD = .17), t(200) = 4.90, p < .001; d = .26.* Participants' overconfidence was not significantly different in the between-subjects part (with & without uncertainty information). Lastly, a two-sample t-test for cognitive measures (scales from (Madsen & Gregor, 2000; Viswanath, V. & Fred D., 2000)) show participants found the AI information more useful (*M = 5.20, Med = 5.25, SD = 1.21)* than reliable *(M = 4.14, Med = 4.20, SD = 1.23)* with a large effect, *t(400) = 8.69, p <.001; d = .87.* AI recommendations' uncertainty information showed over 50% predicted probability of success for the correct label only for 15 out of the 32 images. So, this suggests that participants were able to realize when the AI provided reliable information and when it did not. Individual item statistics for the cognitive measures are provided in the Appendix.

As summarized in Table 2, few of the measured covariates were significantly correlated with the outcome variables, accuracy and confidence. Participants with high animal domain knowledge tended to also report high plant domain knowledge, *r(201) = .40, p < .001.* Participants with high pre-AI confidence tended to have high animal domain knowledge, *r(201) = .46, p < .001.* Participants who perceived the AI as more useful tended to also perceive the AI as more trustworthy, *r(200) = .41, p < .001.*

Table 1. Summary of measures by conditions. *

	Within-Subjects		Between-Subjects	
	No AI	AI	AI w/o Uncertainty	AI w/ Uncertainty
	M (SD)	M (SD)	M (SD)	M (SD)
Participants	201		99	102
Accuracy	36% (8%)	54% (8%)	52% (9%)	57% (5%)
Confidence	43% (15%)	57% (16%)	56% (16%)	58% (15%)
Overconfidence	7% (17%)	3% (17%)	5% (18%)	1% (15%)
Time per Image (s)	16 (19)	10 (6)	10 (6)	10 (5)
AI Usefulness		56% (33%)	55% (33%)	57% (33%)
Task Difficulty		3.2 (1.0)	3.2 (1.1)	3.1 (1.0)
AI Trustworthiness		3.5 (1.0)	3.5 (1.0)	3.5 (1.0)

***AI's top recommendation accuracy is 59%**

4.1. EFFECT OF AI RECOMMENDATIONS

The results support H1, suggesting that providing rank ordered AI recommendations increased accuracy and confidence (see Figure 2). As shown in Table 3.3, separate linear mixed effects regressions were conducted to test the within-subject effects of the AI recommendations on accuracy and confidence. For each outcome variable, we estimated two models to measure the effect of additional covariates. Results for the effect of AI were consistent across both models.

Across participants, ranked AI recommendations increased accuracy by 0.19 ($t = 30.43$, $p < .001$; $d = 2.36$), representing a large effect consistent with the paired t-tests. None of the additional covariates were significant, suggesting that this effect was

independent of perceptions of the AI or task (α = .01). Low variance for the random effects suggests that the variability between individual participants was not substantial. Similarly, ranked AI recommendations increased confidence by 0.14 ($t = 18.07$, $p < .001$; $d = .95$), representing a large effect consistent with the paired t-tests. None of the

Table 2. Pearson correlation matrix. Bolded coefficients are significant at α=.05.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Pre-AI Accuracy		**.39**	.05	.11	.06	.11	.05	.04	.02	**.15**	.14	-.11	-.04
2. Post-AI Accuracy			**-.16**	.10	**.36**	.02	**-.03**	**.24**	-.04	.03	.04	.13	.00
3. Pre-AI Confidence				**.72**	**-.14**	**.24**	**.15**	**.24**	**-.15**	**.39**	**.46**	**-.21**	.07
4. Post-AI Confidence					.08	**.19**	0.06	**.62**	**-.15**	**.29**	**.25**	**-.16**	**.16**
5. Uncertainty						.02	.00	.07	**-.24**	-.08	-.06	-.07	.01
6. Pre-AI Avg Time taken							**.29**	**.16**	-.01	.12	.10	**-.16**	.10
7. Post-AI Avg Time taken								.03	-.07	**.16**	.03	.03	.12
8. AI usefulness									-.08	**.25**	.04	-.11	**.41**
9. Attention score										**-.18**	**-.17**	.06	.00
10. Plant knowledge											**.40**	**-.24**	**.15**
11. Animal knowledge												-.12	-.02
12. Task difficulty													-.11
13. AI trustworthiness													

additional covariates were significant, suggesting that this effect was independent of perceptions of the AI or task ($\alpha = .01$). Low variance for the random effects suggests that the variability between individual participants was not substantial.

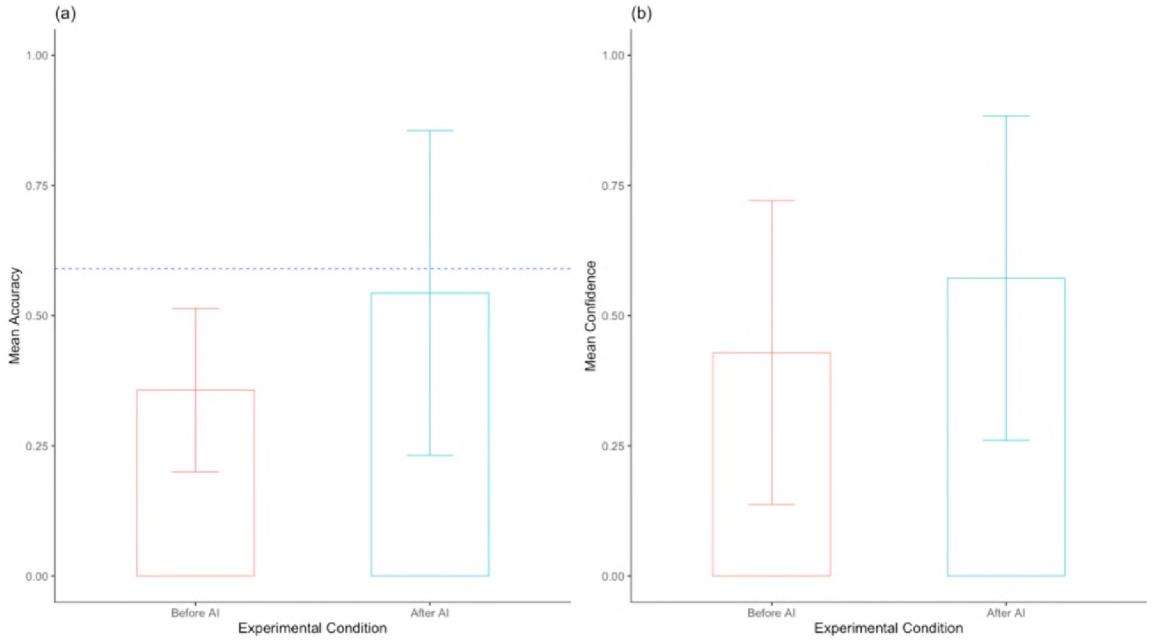

Figure 2. Accuracy and confidence is significantly improved by AI recommendations. Dotted blue line represents the accuracy of AI (59%)

4.2. EFFECT OF UNCERTAINTY INFORMATION

As shown in Figure 3, providing uncertainty information with AI recommendations increased accuracy, but not confidence, partially supporting H2. As shown in Table 4, separate linear regressions were conducted to test the between-subject effects of uncertainty information on accuracy and confidence. For each outcome variable, we estimated two models to measure the effect of additional covariates related

Table 3. Linear mixed effects regression models suggest AI recommendations improve accuracy and confidence.

	Accuracy		Confidence	
Fixed Effects	**Model 1 B (SE)**	**Model 2 B (SE)**	**Model 3 B (SE)**	**Model 4 B (SE)**
Intercept	.36 (.01) ***	.41 (.06) ***	.43 (.01) ***	.61 (.10) ***
AI	.19 (.01) ***	.19 (.01) ***	.14 (.01) ***	.15 (.01) ***
Avg. Time/Image (s)		.00 (.00)		.00 (.00)
Attention Score		-.00 (.01)		-.05 (.02) *
Task Difficulty		.00 (.00)		-.02 (.01) *
AI Trustworthiness		-.00 (.00)		.01 (.01)
Age (logged)		-.02 (.02)		-.04 (.03)
Male		-.01 (.01)		.01 (.02)
College		.01 (.01)		.02 (.02)
Random Effects	**V (σ)**	**V (σ)**	**V (σ)**	**V (σ)**
Individual	.002 (.05)	.002 (.05)	.02 (.13)	.02 (.12)
Residual	.004 (.06)	.004 (.06)	.01 (.08)	.01 (.08)
N	402	398	402	398

*Note: *p < .05, **p < .01, and ***p < .001.*
See Table A.1 in appendix for detailed results of each model.
B = Estimate, SE = Standard Error, V = Variance, σ = Standard Deviation, N = Number of Observation

to attention, perceptions of the AI and task, and demographics. Results for the effect of uncertainty were consistent across both models.

Across participants, uncertainty information increased accuracy by 0.06 *(d = .77)*, representing a medium effect consistent with the paired t-tests. This suggests that participants were able to leverage additional information from the uncertainty information to improve their accuracy. In Model 2 (Table 4), participants who perceived

the AI as more useful had higher accuracy, with medium effect *(d = .43)*. Participants who perceived the AI as more useful may have been more likely to rely on the AI. None of the other covariates were significant ($\alpha = .01$).

Across participants, uncertainty information did not significantly increase confidence ($\alpha = .01$). Similarly, in Model 4 (Table 4), participants who perceived the AI as more useful had higher confidence with a very large effect *(d = 1.54)*, likely because participants who perceived the AI as more useful were more likely to rely on it. None of the other covariates were significant ($\alpha = .01$).

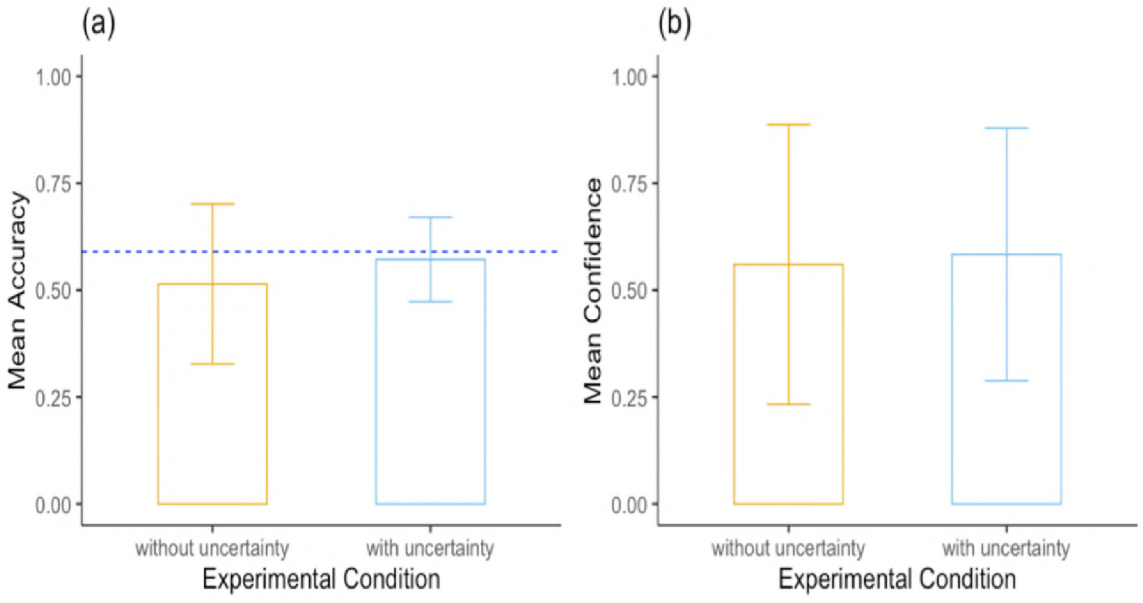

Figure 3. Accuracy is significantly improved by uncertainty information, but confidence is not. Dotted blue line represents the accuracy of AI (59%).

Table 4. Linear regression models suggest that uncertainty information improves accuracy, but not confidence.

	Post-AI Accuracy		Post-AI Confidence	
	Model 1 B (SE)	**Model 2 B (SE)**	**Model 3 B (SE)**	**Model 4 B (SE)**
Intercept	.51 (.01) ***	.43 (.07) ***	.56 (.02) ***	.52 (.11) ***
Uncertainty Information	.06 (.01) ***	.06 (.01) ***	.02 (.02)	-.00 (.02)
Avg. Time/Image (s)		-.00 (.00)		.00 (.00)
Attention Score		.01 (.01)		-.04 (.02)
Avg. AI Usefulness		.13 (.03) ***		.60 (.06) ***
Task Difficulty		.01 (.01) *		-.02 (.01)
AI Trustworthiness		-.01 (.01)		-.02 (.01)
Age (logged)		-.00 (.02)		-.05 (.03)
Male		-.02 (.01)		.01 (.02)
College		.01 (.01)		.02 (.02)
N	201	199	201	199
Adjusted R^2	.13	.20	.00	.40
F	29.67***	6.39***	1.14	15.32***

*Note: *p < .05, **p < .01, and ***p < .001*
See Table A.1 in appendix for detailed results of each model.
B = Estimate, SE = Standard Error, N = Number of Observation

4.3. EFFECT OF DOMAIN KNOWLEDGE

As shown in Figure 4 and Figure 5, the interaction of domain knowledge with AI recommendations and uncertainty information were situational, partially supporting H2. To evaluate the effect of domain knowledge, we calculated accuracy and confidence for the plant and animal images separately. Thus, separate regression models are conducted for plants and animals to measure the interaction between domain knowledge and ranked AI recommendations (Figure 4, Table 6) or uncertainty information (Figure 5, Table 7).

Table 6 uses linear mixed effects regression to capture the sampling hierarchy from repeated measures. In contrast, Table 7 uses linear regression to measure the effect of between-subject effects and is limited to the post-AI performance.

In general, participants were better at identifying animals than plants. Participants reported higher domain knowledge for animals *(M = .47, Med = .40 SD = .23)* than plants *(M = .19, Med = .20, SD = .21), t (395.41) = 13, p < .001.* Overall, participants were more accurate when identifying animals *(M = .61, Med = .62, SD = .09),* rather than, plants *(M = .50, Med = .53, SD = 10), t (395.96) = 11.89, p < .001.* In addition, participants were more confident when identifying animals *(M = .63, Med = .65, SD = .15)* rather than plants (*M* = .53, *SD* = .17), *t* (396.11) = 6.07, *p* < .001. Similar to the participants, the AI was better at identifying animals than plants. The AI's first recommendation was correct 62% of the time for animals and 58% for plants.

Table 5. Summary of measures by plants and animals.

	Plants M% (SD%)				Animals M% (SD%)			
	Pre-AI	**Post-AI**	**No Uncertainty**	**Uncertainty**	**Pre-AI**	**Post-AI**	**No Uncertainty**	**Uncertainty**
AI Accuracy		58				62		
Accuracy	27 (9)	50 (10)	46 (12)	53 (7)	48 (12)	61 (9)	51 (9)	57 (5)
Confidence	39 (14)	53 (17)	55 (16)	58 (15)	49 (17)	63 (15)	56 (16)	58 (15)
Overconfidence	1 (21)	2 (17)	3 (18)	1 (16)	1 (21)	2 (17)	5 (18)	1 (15)

Consistent with Table 3, AI recommendations significantly increased participants' accuracy and confidence even when separated by image topic. However, the effect on

accuracy is larger for plant images (β = .24, t = 21.69, p < .001, d = .15) than animal images (β = .16, t = 8.60, p < .001, d = .23). This is likely because participants relied on the AI more heavily for plants, where they had lower domain knowledge. Conversely, the effect on confidence is larger for animal images (β = .23, t = 12.09, p < .001, d = .16) than plant images (β = .16, t = 13.07, p < .001, d = .09), likely due to higher domain knowledge for animals.

When separated by plants and animals, domain knowledge significantly increased participants' accuracy for animals (β = .10, t = 3.16, p = .002) but not plants (p = .05) in Models 1 and 2 in Table 6. This suggests that participants effectively employed their domain knowledge to identify animals. However, domain knowledge significantly increased participants' confidence for both plants (β = .24, t = 4.71, p < .001) and animals (β = .36, t = 8.16, p < .001) in Models 3 and 4 in Table 6. This suggests that domain knowledge more consistently increased confidence.

In most of the models, the interaction between domain knowledge and AI recommendations was not significant (see Models 1-3, Table 6). However, for animal images, the effect of AI recommendations on confidence decreased as domain knowledge increased (see Figure 4). In Model 4 (Table 6), the interaction of animal domain knowledge and AI recommendations decreased confidence (β = -.19, t = -5.32, p < .001). As animal domain knowledge increased, the additive effect of domain knowledge decreased when AI recommendations were provided. This suggests that the AI recommendations were more effective for increasing confidence when domain knowledge was low. When domain knowledge was high, the AI recommendations had little effect on confidence.

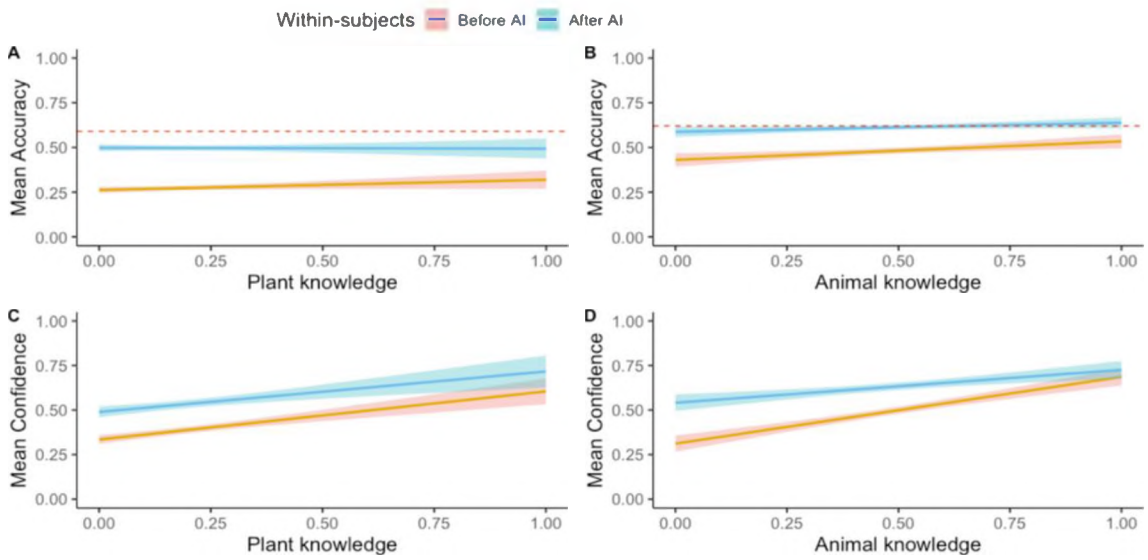

Figure 4. Effect of animal (b, d) and plant (a, c) domain knowledge on accuracy (a, b) and confidence (c, d). For animal domain knowledge, the effect of ranked AI recommendations on confidence (d) decreases as domain knowledge increases. Dotted red line represents AI accuracy in plants (58%) and animals (62%).

In contrast with Table 4, when separated by plants and animals, uncertainty information significantly increased participants' accuracy for plant images and confidence for animal images (see Table 7). This suggests the results in Table 4 are largely driven by plants, rather than animals.

Participants had higher accuracy when provided uncertainty information for plants (β = .08, t = 4.33, p < .001, d = .13), but there was no effect for animals (p = .21). This suggests that participants may have relied on the uncertainty information more when they had lower domain knowledge. In addition, perceived AI usefulness increased accuracy for animals (β = .14, t = 3.67, p < .001, d = .04). This suggests that participants who perceived usefulness may be a better predictor of AI reliance when domain knowledge is higher. Perceived task difficulty increased accuracy for plants (β = .02, t =

2.76, p = .006, d = .05), suggesting that participants who perceived the task as more difficult may have relied on the AI more.

Table 6. Linear mixed effects regression model suggests the interaction of animal domain knowledge and AI recommendations decreases confidence.

	Accuracy		**Confidence**	
Fixed Effects	**Model 1: Plants B (SE)**	**Model 2: Animals B (SE)**	**Model 3: Plants B (SE)**	**Model 4: Animals B (SE)**
Intercept	.38 (.07) ***	.37 (.08) ***	.55 (.12) ***	.44 (.12) ***
AI Recommendations	.24 (.01) ***	.16 (.02) ***	.16 (.01) ***	.23 (.02) ***
Domain Knowledge	.07 (.03) *	.10 (.03) **	.24 (.05) ***	.36 (.04) ***
Knowledge*AI recommendations	-.06 (.04)	-.06 (.03)	-.04 (.04)	-.19 (.04) ***
Avg. time taken	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
Attention check	.00 (.01)	.00 (.01)	-.04 (.02)	-.02 (.02)
Task difficulty	.00 (.01)	-.00 (.01)	-.01 (.01)	-.02 (.01) *
AI trustworthiness	.00 (.01)	-.00 (.01)	.01 (.01)	.01 (.01)
Age (logged)	-.05 (.02) *	.03 (.02)	-.06 (.03)	-.03 (.03)
Male	-.01 (.01)	-.02 (.01)	.01 (.02)	.03 (.02)
College	.03 (.01) *	-.01 (.01)	.03 (.02)	.02 (.02)
Random Effects	**V (σ)**	**V (σ)**	**V (σ)**	**V (σ)**
Individual	.00 (.05)	.00 (.07)	.01 (.12)	.01 (.11)
Residuals	.01 (.08)	.01 (.08)	.01 (.09)	.01 (.08)
Number of Images	19	13	19	13
N	398	398	398	398

Note: *p < .05, **p < .01, and ***p < .001.
See Table A.1 in appendix for detailed results of each model.
B = Estimate, SE = Standard Error, V = Variance, σ = Standard Deviation, N = Number of Observation

Participants had higher confidence when provided uncertainty information for animals (β = .11, t = 2.81, p = .005, d = .004), but there was no effect for plants (p = .07). This suggests that participants may have used the AI recommendations for confirmation, increasing their confidence. However, this did not work for plants, when they had low domain knowledge. In addition, perceived AI usefulness increased confidence for plants (β = .55, t = 8.90, p < .001, d = .29) and animals (β = .48, t = 9.17, p < .001, d = 1.35), consistent with Table 3.

When separated by plants and animals, domain knowledge significantly increased confidence, but not accuracy in Table 7. For both plants and animals, domain knowledge did not significantly increase post-AI accuracy (p >.01). This suggests that domain knowledge did not allow participants to better leverage the uncertainty information. However, domain knowledge significantly increased post-AI confidence for both plants (β = .19, t = 3.20, p = .002, d = .08) and animals (β = .28, t = 5.54, p < .001, d = .08). This suggests that participants with higher domain knowledge tended to be more confident in the post-AI assessment, although the reasons may be different between plants and animals.

In most of the models, the interaction between domain knowledge and uncertainty information was not significant (see Models 1-3, Table 7). However, for animal images, post-AI confidence increased when no uncertainty information was provided and stayed constant when AI information was provided (see Figure 5d). In Model 4 (Table 7), the interaction of animal domain knowledge and uncertainty information decreased confidence (β = -0.21, t = -2.95, p = 0.004, d = 0.03). The interaction estimate effectively cancels out the effect of domain knowledge (0.28 – 0.21). This suggests that participants

who received uncertainty information tended to not have different confidence levels despite varying in terms of domain knowledge.

It is important to note that there were more images of plants than animals. As a result, participants were penalized less for each mistake on the plant images (18/19 = 95%) than the animal images (12/13 = 92%). Some, but not all, of the differences in results for plants and animals can be attributed to this difference. Additionally, no participant who were randomly placed in receiving the uncertainty information reported plant knowledge higher than 0.6.

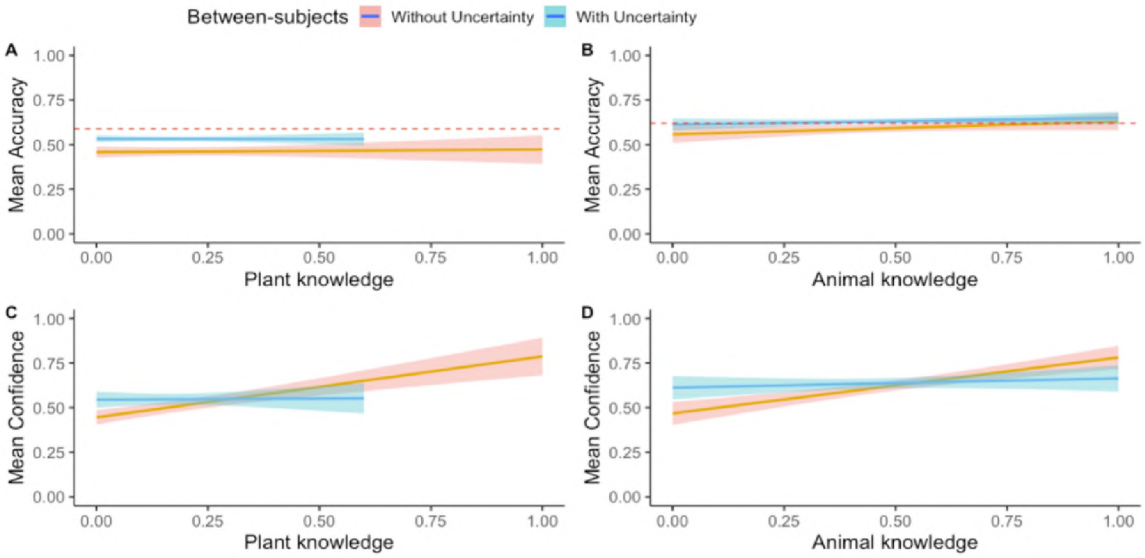

Figure 5. Effect of animal (b, d) and plant (a, c) domain knowledge on accuracy (a, b) and confidence (c, d). For animal domain knowledge, the effect of uncertainty information on confidence (d) decreases as domain knowledge increases. Dotted red line represents AI accuracy in plants (58%) and animals (62%).

Table 7. Linear regression model suggests the interaction of animal domain knowledge and uncertainty information decreases confidence.

	Accuracy		Confidence	
	Model 1: Plants B (SE)	**Model 2: Animals B (SE)**	**Model 3: Plants B (SE)**	**Model 4: Animals B (SE)**
Intercept	.40 (.08) ***	.42 (.08) ***	.51 (.12) ***	.46 (.11) ***
Uncertainty information	.08 (.02) ***	.04 (.03)	.05 (.03)	.11 (.04) **
Domain Knowledge	.04 (.04)	.05 (.04)	.19 (.06) **	.28 (.05) ***
Knowledge*Uncertainty information	.01 (.07)	.00 (.05)	-.23 (.10) *	-.21 (.07) **
Avg. time taken	-.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
Attention check	.02 (.02)	.00 (.02)	-.04 (.02)	-.01 (.02)
Avg. AI usefulness	.11 (.04) *	.14 (.04) ***	.55 (.06) ***	.48 (.05) ***
Task difficulty	.02 (.01) **	.01 (.01)	-.01 (.01)	-.02 (.01) *
AI trustworthiness	-.00 (.01)	-.01 (.01)	-.02 (.01)	-.01 (.01)
Log(age)	-.02 (.02)	.02 (.02)	-.07 (.03) *	-.05 (.03)
Gender (male = 1)	-.01 (.01)	-.02 (.01)	.01 (.02)	.02 (.02)
Education (college = 1)	.03 (.01) *	-.02 (.01)	.02 (.02)	.04 (.02) *
N	199	199	199	199
Adjusted R^2	.19	.10	.40	.42
F	5.25	2.89	13.13	14.07

Note: *$p < .05$, **$p < .01$, and ***$p < .001$

5. CONCLUSION

In a mixed-subject design, participants performed an image recognition task with and without AI recommendations (within-subjects). With the AI recommendations, participants were randomly assigned to receive or not receive uncertainty information

(between-subjects). We also used a self-reported measure of domain knowledge related to plants and animals to evaluate the effect of domain knowledge. We hypothesized that (1) post-AI accuracy and confidence will be higher, (2) accuracy and confidence of participants receiving uncertainty information will be higher than participants who did not receive it, and (3) the interaction between AI recommendations and uncertainty information with domain knowledge will increase accuracy.

Based on the results, we have 3 primary findings, (1) AI recommendations increased accuracy and confidence overall, (2) uncertainty information increased accuracy but did not affect users' confidence, and (3) domain knowledge interacting with AI recommendations and uncertainty information decreased confidence, particularly for animals. In our study, AI recommendations improved users' accuracy and confidence across all the models, even in different domains. None of the other covariates were significant, suggesting the increase in accuracy and confidence was mainly due to AI recommendations. In general, the human-AI team did not perform better than the AI alone, similar to studies on recidivism prediction (Green & Chen, 2019; Grgic-Hlaca et al., 2019; Lin et al., 2020) and unlike high-stakes healthcare studies (Bien et al., 2018; Lakhani & Sundaram, 2017; Patel et al., 2019; Xiong et al., 2020). Similar to this study, the recidivism studies used laypeople from online survey platform like Amazon mTurk, whereas the healthcare studies used specific domain experts.

Uncertainty information improved accuracy but did not affect confidence. Even though the overall accuracy improved, uncertainty information had significant effects on plant images but not animal images which could mean the effects of uncertainty may depend on domain knowledge as well. Results also indicate that participants found the AI

recommendations and uncertainty information useful, which suggests they may have been relying on the AI more which agrees with the findings of Antifakos et al. (2005). However, unlike the results from our study, Gkatzia et al. (2016) found uncertainty information to increase users' confidence as well. Providing numbers, text, and visual together could be reason Gkatzia et al. (2016) found significant results in their study. In addition, complexity of this study is significantly higher than Gkatzia et al. (2016).

The interaction effects between domain knowledge and AI recommendations or between domain knowledge and uncertainty information is situational since the interaction effects are significant in decreasing the confidence of the participants for images of animals only. Overall, participants reported higher domain knowledge for animals than plants. Results also suggest that participants were more accurate in identifying animal images. Results indicate domain knowledge increased users' confidence consistently across both domains. Effect of AI recommendations on accuracy was larger in plant images and on confidence was larger in animal images suggesting that users relied on AI more when domain knowledge was low (plants) and used AI's recommendations as confirmation when domain knowledge was high (animals). Results of uncertainty information on accuracy and confidence across both domains are similar to the effects of AI recommendations. The interaction effect between domain knowledge & AI recommendations and domain knowledge & uncertainty information significantly decreased users' confidence in animal domain. This suggests that AI recommendations had little effect on confidence when domain knowledge was high. Results also suggests that participants who received uncertainty information tended to not have different confidence levels despite varying in terms of domain knowledge.

This study also has its limitations. Participants were gathered from Prolific, as it would be expensive and difficult to gather experts for this study. The domain knowledge scores were self-reported and may not reflect the true level of knowledge as people might not be a good judge and may have been influenced by their perception of task performance. As Greis et al. (2017) and Zhou et al. (2015) mention, research on users' confidence due to AI recommendations and uncertainty information is limited. So, future work should continue to examine this outcome. Research on expert vs novice interacting with decision support systems in high-stake situations were also limited.

In the future, studies can focus on domain knowledge interacting with recommendation systems for high-stake decisions to determine the effects on users' metacognition. Since domain knowledge effects carry less weight in this study due to subjective interpretation, in the future, it is recommended to design an experiment where domain knowledge interaction with recommendation systems is examined in terms of expert vs. novices.

Decision support systems are integrated into various fields at an increasing rate. However, effects of providing uncertainty information on users' accuracy and confidence is still being researched. Results of this study indicate that user's performance will improve with AI recommendations and uncertainty information. So, it will be valuable for decision support systems to provide uncertainty information in more than one format along with the AI recommendations.

REFERENCES

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. *ACM International Conference Proceeding Series*, *111*, 9–14. https://doi.org/10.1145/1085777.1085780

Arshad, S. Z., Zhou, J., Bridon, C., Chen, F., & Wang, Y. (2015). Investigating User Confidence for Uncertainty Presentation in Predictive Decision Making. *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, 352–360. https://doi.org/10.1145/2838739.2838753

Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2020). Human-AI Collaboration in a Cooperative Game Setting. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–20. https://doi.org/10.1145/3415167

Bansal, G., Tongshuang, W. U., Zhou, J., Raymond, F. O. K., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *ArXiv*.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2020). *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*. http://arxiv.org/abs/2011.07586

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D. F., Beaulieu, C. F., Riley, G. M., Stewart, R. J., Blankenberg, F. G., Larson, D. B., … Lungren, M. P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, *15*(11), 1–19. https://doi.org/10.1371/journal.pmed.1002699

Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., & van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *Journal of Computer Assisted Learning*, *33*(3), 234–251. https://doi.org/10.1111/jcal.12162

Budescu, D. V., Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, *113*(2), 181–200. https://doi.org/10.1007/s10584-011-0330-3

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, *October*, 160–169. htts://doi.org/10.1109/ICHI.2015.26

Dane, E., Rockmann, K. W., & Pratt, M. G. (2012). When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness. *Organizational Behavior and Human Decision Processes*, *119*(2), 187–194. https://doi.org/10.1016/j.obhdp.2012.07.009

Feng, S., & Boyd-Graber, J. (2019). *What can AI do for me?* 229–239. https://doi.org/10.1145/3301275.3302265

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*, 1–12. https://doi.org/10.1145/3173574.3173718

Galesic, M. (2010). Statistical Numeracy for Health. *Archives of Internal Medicine*, *170*(5), 462. https://doi.org/10.1001/archinternmed.2009.481

Gkatzia, D., Lemon, O., & Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 264–268. https://doi.org/10.18653/v1/p16-2043

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Greis, M., Avci, E., Schmidt, A., & Machulla, T. (2017). Increasing users' confidence in uncertain data by aggregating data from multiple sources. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 828–840. https://doi.org/10.1145/3025453.3025998

Grgic-Hlaca, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine advice: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359280

Huang, Q., Chen, Y., Liu, L., Tao, D., & Li, X. (2020). On Combining Biclustering Mining and AdaBoost for Breast Tumor Classification. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 728–738. https://doi.org/10.1109/TKDE.2019.2891622

Knijnenburg, B. P., Reijmer, N. J. M., & Willemsen, M. C. (2011). Each to his own: How different users call for different interaction methods in recommender systems. *RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems*, 141–148. https://doi.org/10.1145/2043932.2043960

Lakhani, P., & Sundaram, B. (2017). THORACIC IMAGING: Deep Learning at Chest Radiography Lakhani and Sundaram. *Radiology*, *284*(2), 574–582. http://pubs.rsna.org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1148/radiol.2017162326

Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*(7), 1–9. https://doi.org/10.1126/sciadv.aaz0652

Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute. Monographs*, *27701*(25), 149–163. https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191

Logg, J. M., Minson, J. A., & Moore, D. A. (2018). 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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*(December 2018), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Maadi, M., Khorshidi, H. A., & Aickelin, U. (2021). A review on human–ai interaction in machine learning and insights for medical applications. *International Journal of Environmental Research and Public Health*, *18*(4), 1–21. https://doi.org/10.3390/ijerph18042121

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. *Proceedings of Eleventh Australasian Conference on Information Systems*, 6–8. http://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+Big+Five+Trait+Taxonomy:+History,+measurement,+and+Theoretical+Perspectives&ots=758BNaTvOi&sig=L52e79TS6r0Fp2m6xQVESnGt8mw%5Cn http://citeseerx.ist.psu.edu/viewdoc/download?doi=

McNamara, D. M., Goldberg, S. L., Latts, L., Atieh Graham, D. M., Waintraub, S. E., Norden, A. D., Landstrom, C., Pecora, A. L., Hervey, J., Schultz, E. V., Wang, C. K., Jungbluth, N., Francis, P. M., & Snowdon, J. L. (2019). Differential impact of cognitive computing augmented by real world evidence on novice and expert oncologists. *Cancer Medicine*, *8*(15), 6578–6584. https://doi.org/10.1002/cam4.2548

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarappallil, J., Mariano, A. J., Riley, G., Seekins, J., Shen, L., Zucker, E., & Lungren, M. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *Npj Digital Medicine*, *2*(1). https://doi.org/10.1038/s41746-019-0189-7

Rauschecker, A. M., Rudie, J. D., Xie, L., Wang, J., & Gee, J. C. (2020). Neuroradiologist-level Differential Diagnosis Accuracy at Brain MRI. *Radiology*, *00*, 1–12.

Rosenberg, L., & Willcox, G. (2019). Artificial Swarm Intelligence The technology of Artificial Swarm Intelligence (ASI) has been shown to amplify. *IntelliSys*, *September*, 1–18.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast - But is it good? Evaluating non-expert annotations for natural language tasks. *EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL*, *October*, 254–263.

Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, *4*, 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

Subramanian, H. V., Canfield, C., Shank, D. B., Andrews, L., & Dagli, C. (2020). Communicating uncertain information from deep learning models in human machine teams. *ASEM 41st International Annual Conference Proceedings "Leading Organizations through Uncertain Times."*

Viswanath, Venkatesh, & Fred D., Davis. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46 (2)* (May 2014), 186–204.

Wang, J., Molina, M. D., & Sundar, S. S. (2020). When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior*, *107*(July 2019), 106278. https://doi.org/10.1016/j.chb.2020.106278

Wang, W., & Benbasat, I. (2013). A Contingency approach to investigating the effects of user-system interaction modes of online decision aids. *Information Systems Research*, *24*(3), 861–876. https://doi.org/10.1287/isre.1120.0445

Xiong, Z., Wang, R., Bai, H. X., Halsey, K., Mei, J., Li, Y. H., Atalay, M. K., Jiang, X. L., Fu, F. X., Thi, L. T., Huang, R. Y., Liao, W. H., Pan, I., Choi, J. W., Zeng, Q. H., Hsieh, B., CuiWang, D., Sebro, R., Hu, P. F., … Qi, Z. Y. (2020). Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology*, *296*(3), E156–E165. https://doi.org/10.1148/radiol.2020201491

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *ArXiv*.

Zhou, J., Bridon, C., Chen, F., Khawaji, A., & Wang, Y. (2015). *Be Informed and Be Involved*. 923–928. https://doi.org/10.1145/2702613.2732769

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making*, *27*(5), 663–671. https://doi.org/10.1177/0272989X07303824

SECTION

2. CONCLUSION

The results of both studies suggest that AI recommendations significantly improve the accuracy of the participants. In Paper I, results indicated that ranked multiple AI recommendations improves accuracy rather than single AI recommendations. Paper II confirms that ranked multiple AI recommendations improve the accuracy of the participants across all domains as well. Paper II also suggests that ranked multiple AI recommendations significantly improve participants' confidence. Results of both these studies agree with the literature reviewed on recidivism prediction with AI recommendations (Green & Chen, 2019; Grgic-Hlaca et al., 2019; Lin et al., 2020). However, studies in this thesis and in recidivism studies, human-AI teams performed better than humans but not the AI alone. All these studies included laypeople making the decisions so, future studies should design experiments in terms of experts vs novices to compare results with studies that suggests Human-AI teams perform better than humans and AI alone (Bansal et al., 2020; Bien et al., 2018; Lakhani & Sundaram, 2017; Patel et al., 2019; Rosenberg & Willcox, 2019; Xiong et al., 2020).

The literature reviewed also support that accuracy increased when AI recommendations uncertainty information is provided (Bansal et al., 2020; Fernandes et al., 2018; Gkatzia et al., 2016). The effects of providing AI recommendations uncertainty information on accuracy were unclear in Paper I as it was presented in terms of confidence bars that were monochrome and were overlapped by texts. Limitations of Paper I helped design the representation of uncertainty information in Paper II better

which confirmed uncertainty information will significantly improve participants accuracy. In Paper II, uncertainty information was provided both numerically and visually in terms of percentage values and bars. The bars were color coded to signify varying uncertainty and were separated from texts as well. As a result, the saliency of representing uncertainty information improved in Paper II which may have been the reason participants accuracy improved in Paper II compared to Paper I.

Gkatzia et al. (2016) found providing uncertainty information in terms of texts, numbers, and graphs improved users' confidence however, research on AI recommendations uncertainty information effect on users is limited (Greis et al., 2017; Zhou et al., 2015). Effects of uncertainty information on participants' confidence is unclear in Paper II. Future studies should examine the effects of uncertainty on users' confidence. One method would be improving the AI's accuracy on its top recommendation and involving domain experts.

Lastly, our findings agree with the literature reviewed – high domain knowledge increases accuracy (Brand-Gruwel et al., 2017; Snow et al., 2008). Participants reported higher domain knowledge for animals than plants and results show that confidence increased across both domains suggesting that domain knowledge more consistently increased confidence. Results of Paper II indicates that AI recommendations improved accuracy across both domains however, the effect was larger for plant images suggesting that participants relied on AI more when domain knowledge was low which is similar to findings of Wang & Benbasat (2013).

Providing uncertainty information significantly improved accuracy in plants but not for animals, again indicating that participants relied on AI more when domain

knowledge was low. This finding agrees with Bussone et al. (2015) who also found participants rely on AI uncertainty information more when domain knowledge is low. The interaction between AI recommendation and knowledge or uncertainty information and knowledge was situational as it only showed significant effects on animal images domain by decreasing users' confidence only. This suggests that AI recommendations had little effect on confidence when domain knowledge was high.

Results also suggests that participants who received uncertainty information tended to not have different confidence levels despite varying in terms of domain knowledge. Research on different levels of knowledge interacting AI recommendations or AI recommendations uncertainty information is limited so future work can include investigating this further. Domain knowledge was self-rated by the participants at the end of the experiment of Paper II. In the future it will be beneficial if the experiment is designed in terms of expert vs novice. It would be ideal if the experiment setting involves a high-stakes scenario. In addition, one could examine the effects of AI recommendations and uncertainty on accuracy and confidence when there is a time limit like kidney organ transplant process. In a kidney organ allocation process several stakeholders make decisions from Organ Procurement Organizations (OPOs), Transplant Centers (TCs), and organ recipients typically in a very short time frame. A National Science Foundation (NSF) planning grant funded project focuses on reducing the kidney discard rate in the kidney allocation process with the use of an AI decision support system. Results of this thesis can be applied when designing the AI for users by providing multiple AI recommendations in ranked order instead of single AI recommendations. Providing AI

recommendation's uncertainty information both numerically and visually through bar plots and percentage value will also aid transplant center workers compare between the multiple recommendations.

APPENDIX

Table A.1. Mean of individual cognitive measures responses show that participants found the AI information more useful than reliable.

	AI recommendation w/o uncertainty information	**AI recommendations w/ uncertainty information**
AI Reliability Scales	M (SD)	M (SD)
The AI always provided the advice I required to make a decision.	3.73 (1.73)	3.85 (1.59)
The AI performed reliably.	4.16 (1.54)	4.44 (1.45)
The AI responded the same way under the same conditions at different times.	4.32 (1.48)	3.94 (1.49)
I could rely on the AI to function properly.	4.20(1.44)	4.20 (1.39)
The AI evaluated the images consistently.	4.32 (1.56)	4.28 (1.36)
AI Usefulness scales		
Using the AI recommendations improved my performance.	4.82 (1.40)	5.42 (0.99)
Using AI recommendations in the task increased my productivity	4.94 (1.56)	5.38 (1.08)
Using the AI recommendations in the task enhanced my effectiveness.	4.88 (1.64)	5.51 (1.02)
I found the AI to be useful in completing the task	4.98 (1.49)	5.64 (1.05)

Table A.2. Linear mixed effects regression model suggests the interaction of animal domain knowledge and AI recommendations decreases confidence.

	Accuracy		Confidence	
Fixed Effects	**Model 1: Plants B (SE)**	**Model 2: Animals B (SE)**	**Model 3: Plants B (SE)**	**Model 4: Animals B (SE)**
Intercept	0.26 (0.01) ***	0.43 (0.02) ***	0.33 (0.01) ***	0.31 (0.02) ***
AI Recommendations	0.24 (0.01) ***	0.16 (0.02) ***	0.15 (0.01) ***	0.23 (0.02) ***
Domain Knowledge	0.06 (0.03)	0.10 (0.03) **	0.27 (0.05) ***	0.37 (0.04) ***
Knowledge*AI recommendations	-0.06 (0.04)	-0.05 (0.03)	-0.04 (0.04)	-0.19 (0.04) ***
Random Effects	**V (σ)**	**V (σ)**	**V (σ)**	**V (σ)**
Individual	0.00 (0.05)	0.00 (0.07)	0.01 (0.12)	0.01 (0.11)
Residuals	0.01 (0.08)	0.01 (0.08)	0.01 (0.09)	0.01 (0.08)
Number of Images	19	13	19	13
N	402	402	402	402

Table A.3. Linear regression model suggests the interaction of animal domain knowledge and uncertainty information decreases confidence

	Accuracy		Confidence	
	Model 1: Plants B (SE)	**Model 2: Animals B (SE)**	**Model 3: Plants B (SE)**	**Model 4: Animals B (SE)**
Intercept	0.46 (0.01) ***	0.55 (0.02) ***	0.44 (0.02) ***	0.46 (0.03) ***
Uncertainty information	0.07 (0.02) ***	0.06 (0.03)	0.10 (0.03) **	0.14 (0.05) **
Domain Knowledge	0.01 (0.04)	0.07 (0.04)	0.34 (0.06) ***	0.31 (0.06) ***
Knowledge*Uncertainty information	-0.02 (0.07)	-0.04 (0.05)	-0.33 (0.11) **	-0.26 (0.09) **
N	201	201	201	201
Adjusted R^2	0.11	0.05	0.11	0.11
F	9.30***	4.39**	9.93***	9.42***

Note: *$p < .05$, **$p < .01$, and ***$p < .001$

Table A.4. Linear mixed effects regression shows that domain knowledge significantly increased participants overconfidence.

	Over confidence 1	**Over confidence 2**	**Over confidence 3 Plants**	**Over confidence 4 Animals**
	B (SE)	**B (SE)**	**B (SE)**	**B (SE)**
Intercept	0.07 (0.01) ***	0.20 (0.13)	0.22 (0.14)	0.07 (0.15)
AI Recommendations	- 0.04 (0.01) ***	- 0.04 (0.01) ***	0.01 (0.02)	0.07 (0.03) **
Domain Knowledge			0.21 (0.07) **	0.25 (0.06) ***
Knowledge*Uncertainty information			- 0.06 (0.05)	- 0.13 (0.05) **
Avg. time taken		0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Task difficulty		- 0.02 (0.01) *	- 0.02 (0.01)	0.02 (0.01)
AI trustworthiness		0.01 (0.01)	0.01 (0.01)	0.02 (0.01)
Attention check		- 0.05 (0.02) *	- 0.03 (0.03)	0.02 (0.03)
log (Age)		- 0.03 (0.04)	0.07 (0.04)	- 0.06 (0.04)
Male		0.03 (0.02)	0.04 (0.02)	0.05 (0.02)
College		0.01 (0.02)	0.03 (0.03)	0.03 (0.03)
Random Effects	**V (σ)**	**V (σ)**	**V (σ)**	**V (σ)**
Individual	0.02 (0.14)	0.02 (0.14)	0.02 (0.14)	0.02 (0.14)
Residuals	0.01 (0.09)	0.01 (0.09)	0.01 (0.11)	0.01 (0.11)
N	402	398	398	398

Note: *$p < .05$, **$p < .01$, and ***$p < .001$

Table A.5. Effects of uncertainty information on overconfidence is situational however, participants who perceived AI as more useful were significantly overconfident.

	Over confidence 1 **B (SE)**	**Over confidence 2** **B (SE)**	**Over confidence 3 Plants** **B (SE)**	**Over confidence 4 Animals** **B (SE)**
Intercept	0.05 (0.02) **	0.09 (0.13)	0.10 (0.14)	0.04 (0.14)
Uncertainty	- 0.03 (0.02)	- 0.06 (0.02) **	- 0.01 (0.03)	0.07 (0.05)
Information				
Domain Knowledge			0.12 (0.07)	0.22 (0.06) ***
Knowledge*Uncerta			- 0.16 (0.11)	- 0.22 (0.09) *
inty information				
Avg. AI use		0.47 (0.07) ***	0.40 (0.07) ***	0.33 (0.07) ***
Avg. time taken		0.00 (0.00)	0.00 (0.00)	0.00
Task difficulty		- 0.03 (0.01) **	- 0.02 (0.01) *	- 0.03 (0.01) *
AI trustworthiness		- 0.01 (0.01)	- 0.01 (0.01)	- 0.00 (0.01)
Attention check		- 0.05 (0.02)	- 0.02 (0.03)	- 0.01 (0.03)
Log(age)		- 0.05 (0.04)	0.07 (0.04) *	- 0.08 (0.04) *
Male		0.02 (0.02)	0.05 (0.02) *	0.04 (0.02)
College		0.01 (0.02)	0.06 (0.02) *	0.06 (0.04) *
N	201	199	199	199
Adj-R^2	0.01	0.26	0.24	0.22
F	2.03	8.59 ***	6.84 ***	6.03 ***

Note: *$p < .05$, **$p < .01$, and ***$p < .001$

BIBLIOGRAPHY

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). Towards improving trust in context-aware systems by displaying system confidence. *ACM International Conference Proceeding Series*, *111*, 9–14. https://doi.org/10.1145/1085777.1085780

Arshad, S. Z., Zhou, J., Bridon, C., Chen, F., & Wang, Y. (2015). Investigating User Confidence for Uncertainty Presentation in Predictive Decision Making. *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, 352–360. https://doi.org/10.1145/2838739.2838753

Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2020). Human-AI Collaboration in a Cooperative Game Setting. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–20. https://doi.org/10.1145/3415167

Bansal, G., Tongshuang, W. U., Zhou, J., Raymond, F. O. K., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *ArXiv*.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2020). *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*. http://arxiv.org/abs/2011.07586

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D. F., Beaulieu, C. F., Riley, G. M., Stewart, R. J., Blankenberg, F. G., Larson, D. B., … Lungren, M. P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, *15*(11), 1–19. https://doi.org/10.1371/journal.pmed.1002699

Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., & van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *Journal of Computer Assisted Learning*, *33*(3), 234–251. https://doi.org/10.1111/jcal.12162

Budescu, D. V., Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, *113*(2), 181–200. https://doi.org/10.1007/s10584-011-0330-3

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, *October*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Dane, E., Rockmann, K. W., & Pratt, M. G. (2012). When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness. *Organizational Behavior and Human Decision Processes*, *119*(2), 187–194. https://doi.org/10.1016/j.obhdp.2012.07.009

Feng, S., & Boyd-Graber, J. (2019). *What can AI do for me?* 229–239. https://doi.org/10.1145/3301275.3302265

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*, 1–12. https://doi.org/10.1145/3173574.3173718

Galesic, M. (2010). Statistical Numeracy for Health. *Archives of Internal Medicine*, *170*(5), 462. https://doi.org/10.1001/archinternmed.2009.481

Gkatzia, D., Lemon, O., & Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers*, 264–268. https://doi.org/10.18653/v1/p16-2043

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Greis, M., Avci, E., Schmidt, A., & Machulla, T. (2017). Increasing users' confidence in uncertain data by aggregating data from multiple sources. *Conference on Human Factors in Computing Systems - Proceedings*, *2017-May*, 828–840. https://doi.org/10.1145/3025453.3025998

Grgic-Hlaca, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine advice: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359280

Huang, Q., Chen, Y., Liu, L., Tao, D., & Li, X. (2020). On Combining Biclustering Mining and AdaBoost for Breast Tumor Classification. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 728–738. https://doi.org/10.1109/TKDE.2019.2891622

Knijnenburg, B. P., Reijmer, N. J. M., & Willemsen, M. C. (2011). Each to his own: How different users call for different interaction methods in recommender systems. *RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems*, 141–148. https://doi.org/10.1145/2043932.2043960

Lakhani, P., & Sundaram, B. (2017). THORACIC IMAGING: Deep Learning at Chest Radiography Lakhani and Sundaram. *Radiology*, *284*(2), 574–582. http://pubs.rsna.org.ezp-prod1.hul.harvard.edu/doi/pdf/10.1148/radiol.2017162326

Lin, Z. J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*(7), 1–9. https://doi.org/10.1126/sciadv.aaz0652

Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute. Monographs*, *27701*(25), 149–163. https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191

Logg, J. M., Minson, J. A., & Moore, D. A. (2018). 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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*(December 2018), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Maadi, M., Khorshidi, H. A., & Aickelin, U. (2021). A review on human–ai interaction in machine learning and insights for medical applications. *International Journal of Environmental Research and Public Health*, *18*(4), 1–21. https://doi.org/10.3390/ijerph18042121

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. *Proceedings of Eleventh Australasian Conference on Information Systems*, 6–8. http://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+Big+Five+Trait+Taxonomy:+History,+measurement,+and+Theoretical+Perspectives&ots=758BNaTvOi&sig=L52e79TS6r0Fp2m6xQVESnGt8mw%5Cn http://citeseerx.ist.psu.edu/viewdoc/download?doi=

McNamara, D. M., Goldberg, S. L., Latts, L., Atieh Graham, D. M., Waintraub, S. E., Norden, A. D., Landstrom, C., Pecora, A. L., Hervey, J., Schultz, E. V., Wang, C. K., Jungbluth, N., Francis, P. M., & Snowdon, J. L. (2019). Differential impact of cognitive computing augmented by real world evidence on novice and expert oncologists. *Cancer Medicine*, *8*(15), 6578–6584. https://doi.org/10.1002/cam4.2548

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarappallil, J., Mariano, A. J., Riley, G., Seekins, J., Shen, L., Zucker, E., & Lungren, M. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *Npj Digital Medicine*, *2*(1). https://doi.org/10.1038/s41746-019-0189-7

Rauschecker, A. M., Rudie, J. D., Xie, L., Wang, J., & Gee, J. C. (2020). Neuroradiologist-level Differential Diagnosis Accuracy at Brain MRI. *Radiology*, *00*, 1–12.

Rosenberg, L., & Willcox, G. (2019). Artificial Swarm Intelligence The technology of Artificial Swarm Intelligence (ASI) has been shown to amplify. *IntelliSys*, *September*, 1–18.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast - But is it good? Evaluating non-expert annotations for natural language tasks. *EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL*, *October*, 254–263.

Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, *4*, 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

Subramanian, H. V., Canfield, C., Shank, D. B., Andrews, L., & Dagli, C. (2020). Communicating uncertain information from deep learning models in human machine teams. *ASEM 41st International Annual Conference Proceedings "Leading Organizations through Uncertain Times."*

Viswanath, Venkatesh, & Fred D., Davis. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46 (2)* (May 2014), 186–204.

Wang, J., Molina, M. D., & Sundar, S. S. (2020). When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior*, *107*(July 2019), 106278. https://doi.org/10.1016/j.chb.2020.106278

Wang, W., & Benbasat, I. (2013). A Contingency approach to investigating the effects of user-system interaction modes of online decision aids. *Information Systems Research*, *24*(3), 861–876. https://doi.org/10.1287/isre.1120.0445

Xiong, Z., Wang, R., Bai, H. X., Halsey, K., Mei, J., Li, Y. H., Atalay, M. K., Jiang, X. L., Fu, F. X., Thi, L. T., Huang, R. Y., Liao, W. H., Pan, I., Choi, J. W., Zeng, Q. H., Hsieh, B., CuiWang, D., Sebro, R., Hu, P. F., … Qi, Z. Y. (2020). Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology*, *296*(3), E156–E165. https://doi.org/10.1148/radiol.2020201491

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *ArXiv*.

Zhou, J., Bridon, C., Chen, F., Khawaji, A., & Wang, Y. (2015). *Be Informed and Be Involved*. 923–928. https://doi.org/10.1145/2702613.2732769

Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making*, *27*(5), 663–671. https://doi.org/10.1177/0272989X07303824

VITA

Harishankar Vasudevanallur Subramanian was born in Chennai, India. He received his bachelor's degree in Mechanical Engineering in December 2019 from Missouri University of Science and Technology. After graduation, he joined Dr. Casey Canfield's Lab at Missouri University of Science and Technology in January 2020 and received his master's degree in Engineering Management in July 2021. He continued with Dr. Casey Canfield to pursue a doctoral degree in Engineering Management.