Proceedings of the American Society for Engineering Management 2020 International Annual Conference
H. Keathley, J. Enos and M. Parrish eds.

COMMUNICATING UNCERTAIN INFORMATION FROM DEEP LEARNING MODELS IN HUMAN MACHINE TEAMS

Harishankar V. Subramanian
Casey Canfield, Ph.D.*
Daniel B. Shank, Ph.D.
Luke Andrews
Cihan Dagli, Ph.D.
Missouri University of Science and Technology

***canfieldci@mst.edu**

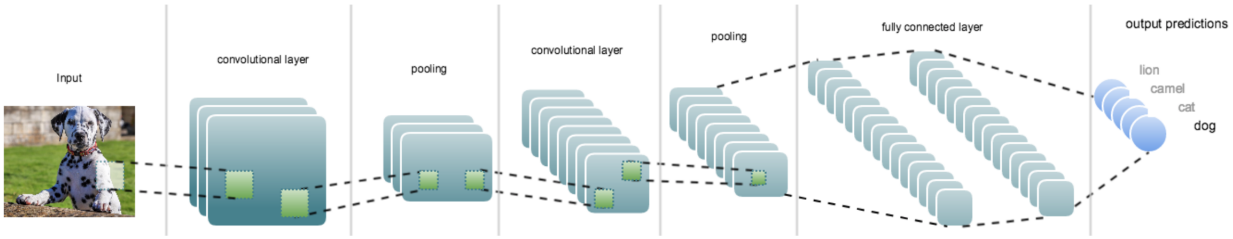
Abstract
The role of human-machine teams in society is increasing, as big data and computing power explode. One popular approach to AI is deep learning, which is useful for classification, feature identification, and predictive modeling. However, deep learning models often suffer from inadequate transparency and poor explainability. One aspect of human systems integration is the design of interfaces that support human decision-making. AI models have multiple types of uncertainty embedded, which may be difficult for users to understand. Humans that use these tools need to understand how much they should trust the AI. This study evaluates one simple approach for communicating uncertainty, a visual confidence bar ranging from 0-100%. We perform a human-subject online experiment using an existing image recognition deep learning model to test the effect of (1) providing single vs. multiple recommendations from the AI and (2) including uncertainty information. For each image, participants described the subject in an open textbox and rated their confidence in their answers. Performance was evaluated at four levels of accuracy ranging from the same as the image label to the correct category of the image. The results suggest that AI recommendations increase accuracy, even if the human and AI have different definitions of accuracy. In addition, providing multiple ranked recommendations, with or without the confidence bar, increases operator confidence and reduces perceived task difficulty. More research is needed to determine how people approach uncertain information from an AI system and develop effective visualizations for communicating uncertainty.

Keywords
Human Systems Integration, Recommendation System, Artificial Intelligence, Uncertainty, Human Machine Team

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 Exhibit 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.

Exhibit 1. Example of a Deep Learning Model with Artificial Neural Networks for Image Recognition.



This study provides human participants with recommendations from an image recognition deep learning model to answer two primary research questions:

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

This research draws on insights from the literature on communicating AI recommendations and communicating uncertainty.

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.

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).

Method

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 is an alternative to Amazon mTurk created by a group of researchers from Oxford and Sheffield universities. The data quality of Prolific is comparable to Amazon mTurk, but Prolific offers a more diverse group of English-speaking participants in terms of geographical location and ethnicity (Peer et al., 2017).

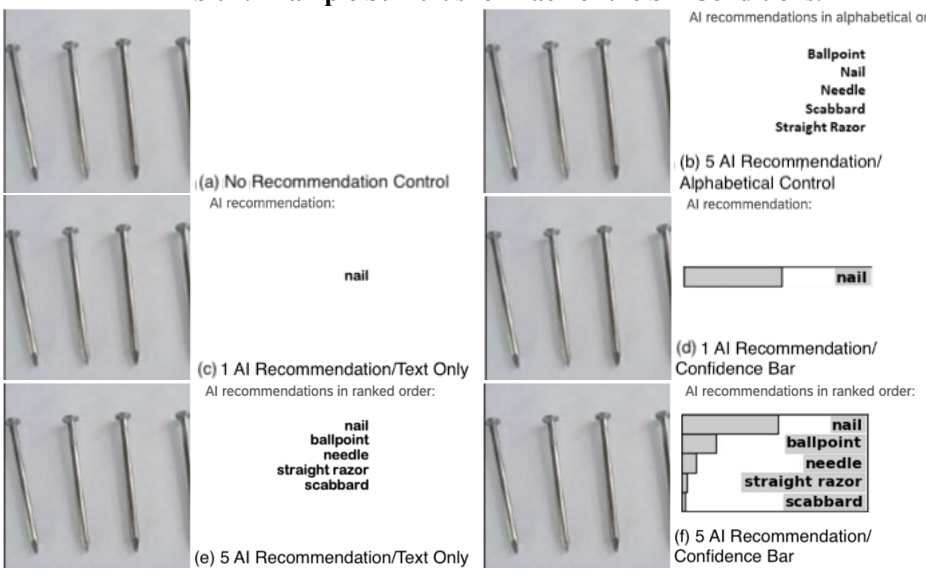
Participants performed an image recognition task. In a between-subjects design, participants were randomly assigned to one of six conditions:

a) *No Recommendation Control* – no AI recommendation or confidence bar provided,
b) *5 AI Recommendation/Alphabetical Control* – top five recommendations by the AI in alphabetical order,
c) *1 AI Recommendation/Text Only* – top recommendation by AI,
d) *1 AI Recommendation/Confidence Bar* – top recommendation by AI with confidence bar,
e) *5 AI Recommendation/Text Only* – top five recommendations by the AI in ranked order, and

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

The figures below show an example experimental stimulus for each condition. Within each condition, each participant identified 24 images and answered additional survey questions.

Exhibit 2. Example Stimulus for Each of the Six Conditions.



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.

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.

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. Exhibit 3 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, $r(284) = 0.27$, $p < .001$. In addition, participants that were more confident tended to spend more time per image, $r(284) = 0.15$, $p = .01$, and perceive the task as more difficult, $r(284) = 0.26$, $p < .001$.

Exhibit 3. 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 Exhibit 4, 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 Exhibits 3 and 5). 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 Exhibit 3). 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 Exhibit 3). 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.

Exhibit 4. 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
Experimental Condition	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
log(Age)	0.06	0.00	1.20	0.00	0.69	0.00	0.08	0.00

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

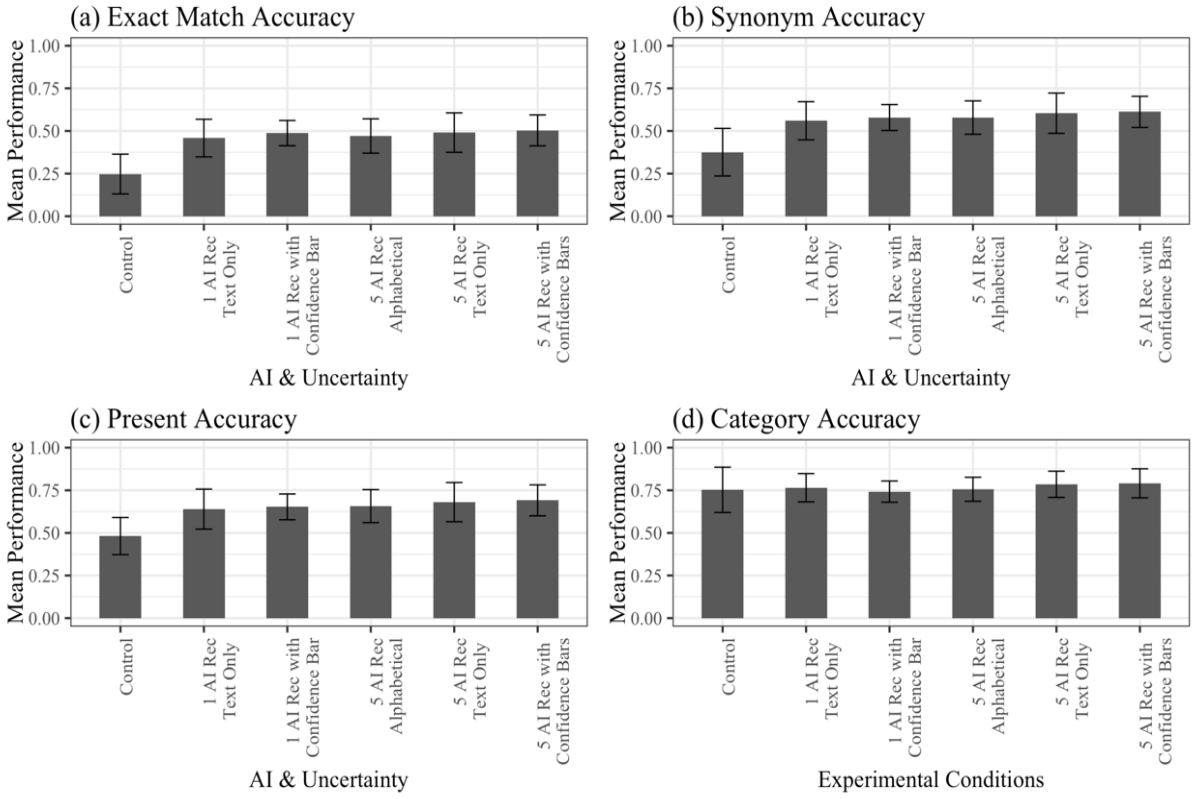
When excluding the Control conditions, it is possible to examine the potential interaction of the number of AI recommendations and the use of the confidence bar. As shown in Exhibit 5, there is a significant difference due to the number of AI recommendations for all definitions of accuracy. However, 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.

Exhibit 5. 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 ***$p < .001$

Exhibit 6. 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.



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 when 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.

Recommendations

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.

Acknowledgements

The authors would like to thank an anonymous reviewer for helpful comments that were partially addressed in this analysis and will be further investigated in future work.

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

About the Authors

Harishankar V. Subramanian has a bachelor's degree in Mechanical Engineering from Missouri University of Science and Technology (Missouri S&T). He is currently a master's student in the Engineering Management department at Missouri S&T working under Dr. Casey Canfield with research focusing on communicating uncertainty for Artificial Intelligence recommendations.

Dr. Casey Canfield is an Assistant Professor in Engineering Management & Systems Engineering at Missouri University of Science & Technology. Her research is focused on improving data-driven decision-making in the context of infrastructure, governance, and healthcare. She has a PhD in Engineering & Public Policy from Carnegie Mellon University and a BS in Engineering: Systems from Franklin W. Olin College of Engineering.

Dr. Daniel B. Shank is an assistant professor in the Department of Psychological Science at Missouri University of Science & Technology with a PhD in Sociology from the University of Georgia. His research interests involve social psychological approaches to interaction with artificial intelligence agents. He currently has grants from the Army Research Office and the Leonard Wood Institute to study affective and moral perceptions and interactions with artificial agents.

Luke Andrews is a senior at the Missouri University of Science and Technology pursuing a B.S in Engineering Management with an Industrial Engineering emphasis. Luke will also earn a minor in Industrial/Organizational Psychology when he graduates in December of 2020.

Dr. Cihan Dagli is a Fellow of INCOSE and IISE; Life member of IEEE and a member of NDIA; and a Fellow of International Foundation for Production Research. He is Professor of Engineering Management and Systems

Engineering at the Missouri University of Science and Technology. Dr. Dagli is the founder and Director of the Missouri S&T's System Engineering graduate program. He is the director of Smart Engineering Systems Laboratory and a Senior Investigator in DoD Systems Engineering Research Center-URAC. His current research interests are in the areas of Architecting Cyber Physical Systems, Complex Adaptive Systems, System of Systems, Data Analytics and Machine Learning. He is currently working on systems architectures and meta- architectures for self-organizing systems ensembles and deep neural networks in creating adaptive behavior.