

Spring 5-5-2019

Partnering People with Deep Learning Systems: Human Cognitive Effects of Explanations

Sean Dougherty

Follow this and additional works at: https://scholarworks.gsu.edu/bus_admin_diss

Recommended Citation

Dougherty, Sean, "Partnering People with Deep Learning Systems: Human Cognitive Effects of Explanations." Dissertation, Georgia State University, 2019.

https://scholarworks.gsu.edu/bus_admin_diss/113

This Dissertation is brought to you for free and open access by the Programs in Business Administration at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Business Administration Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

PERMISSION TO BORROW

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote from, copy from, or publish this dissertation may be granted by the author or, in her absence, the professor under whose direction it was written or, in his absence, by the Dean of the Robinson College of Business. Such quoting, copying, or publishing must be solely for scholarly purposes and must not involve potential financial gain. It is understood that any copying from or publication of this dissertation that involves potential gain will not be allowed without written permission of the author.

Sean E. Dougherty

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University Library must be used only in accordance with the stipulations prescribed by the author in the preceding statement.

The author of this dissertation is:

Sean E. Dougherty
J. Mack Robinson College of Business
Georgia State University
Atlanta, GA 30303

The director of this dissertation is:

Pamela Scholder Ellen
J. Mack Robinson College of Business
Georgia State University
Atlanta, GA 30303

Partnering People with Deep Learning Systems: Human Cognitive Effects of Explanations

by

Sean E. Dougherty

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree

Of

Executive Doctorate in Business

In the Robinson College of Business

Of

Georgia State University

GEORGIA STATE UNIVERSITY

ROBINSON COLLEGE OF BUSINESS

2019

Copyright by
Sean E. Dougherty
2019

ACCEPTANCE

This dissertation was prepared under the direction of the *SEAN E. DOUGHERTY* Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business Administration in the J. Mack Robinson College of Business of Georgia State University.

Richard Phillips, Dean

DISSERTATION COMMITTEE

Dr. Pamela Scholder Ellen

Dr. Balasubramaniam Ramesh

Dr. Mark Keil

ACKNOWLEDGEMENTS

My dissertation owes much to the guidance of my committee in translating my vague ideas and professional frustrations into a researchable topic. Dr. Ellen designs measures and experiments seemingly from intuition, and ensured the study resulted in interpretable and meaningful data. Dr. Keil's feedback not only helped me reach more coherent concepts, it kept the task lightweight but true to the real-world. Dr. Ramesh helped me focus not just into an area of promising technology, but also illuminated the "ah ha" that my interest was really in the human aspect of the problem. My committee was always patient and helped me improve my concepts between each proposal. The quality of the final product is far greater as a result.

This program required having both a lack of sense of what you could not do and the tenacity to learn whatever new research areas arose from the problem. Spending my formative years with a father that designed, built, and fixed whatever he wanted gave me the sense that I could do the same. And when I decided to start up a business selling on-line advertising space pre-internet both of my parents supported me at every step. These early experiences have clear echoes in my dissertation.

This program also required a source of nearly infinite daily support. It would have been a lot to ask my wife Kelly for tolerance of this crazy idea, much less for encouragement. My wife's support at every step of the path made this dissertation possible without having everything else in life fall apart.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	x
LIST OF FIGURES	xi
I INTRODUCTION	1
I.1 Explanations in Application.....	1
I.2 Computation in Disaster Assessment	4
I.3 Conceptual Framework.....	8
I.4 Theory and Previous Research	10
I.5 Research Questions.....	13
I.6 Motivation for the Study	14
I.7 Contribution and Significance.....	15
I.8 Dissertation Organization	15
II LITERATURE REVIEW	17
II.1 Computation in Disaster Assessment	17
II.1.1 <i>Background</i>.....	17
II.1.2 <i>Crowd Workers in Disaster Assessment</i>.....	18
II.1.3 <i>Automation in Disaster Assessment</i>	19
II.1.4 <i>Open Issues</i>	22
II.2 Social Cognition	23
II.2.1 <i>Scope</i>.....	23
II.2.2 <i>Intentional Systems Theory</i>	24
II.2.2.1 <i>Background</i>.....	24
II.2.2.2 <i>Dimensions of Mind</i>.....	25

II.2.2.3	<i>Attribution of Behavior of Technological Artifacts</i>	26
II.2.2.4	<i>Biases and Asymmetries in Attribution</i>	28
II.2.3	<i>Self-Efficacy</i>	29
II.2.3.1	<i>Background</i>	29
II.2.3.2	<i>Application to the Workplace</i>	30
II.2.3.3	<i>Application to Human-Computer Interaction</i>	30
II.2.3.4	<i>Illusory Understanding</i>	31
II.2.4	<i>Cognitive Load Theory</i>	32
II.2.4.1	<i>Background</i>	32
II.2.4.2	<i>Applications in Computing Systems</i>	33
II.2.4.3	<i>Expertise and Explanation</i>	34
II.2.5	<i>Open Issues</i>	34
II.3	Explanation by Intelligent Agents	35
II.3.1	<i>Scope</i>	35
II.3.2	<i>Background</i>	35
II.3.3	<i>Composition of Explanations</i>	36
II.3.4	<i>Explanation Engines in Deep Learning</i>	38
II.3.5	<i>Experimental Manipulation of Explanation</i>	40
II.3.6	<i>Open Issues</i>	44
III	RESEARCH MODEL AND HYPOTHESIS DEVELOPMENT	45
III.1	Research Model	45
III.2	Independent Variables	48
III.2.1	<i>Causal Explanation</i>	48
III.2.2	<i>Counterfactual Explanation</i>	49

III.2.3	<i>Hedging Explanation</i>	49
III.3	Dependent Variables	50
III.3.1	<i>Self-Efficacy</i>	50
III.3.2	<i>Cognitive Load</i>	51
III.3.3	<i>Attribution of Agent Intelligence</i>	51
III.4	Alternate Explanations	52
III.4.1	<i>Previous Task Experience</i>	52
III.4.2	<i>Trust in the Intelligent Agent</i>	53
III.4.3	<i>Dispositional and Learned Trust for AI</i>	53
III.4.4	<i>Perceived Interdependence</i>	54
III.4.5	<i>Perceived Level of Automation</i>	54
III.5	Scenario Measures	55
III.5.1	<i>Initial and Review Timing</i>	55
III.5.2	<i>Erroneous Agreement</i>	55
III.6	Manipulation Checks	56
III.7	Demographics and Feedback	56
III.8	Hypothesis Development	56
III.8.1	<i>Effect of Counterfactual Explanations</i>	56
III.8.2	<i>Effect of Hedging Explanations</i>	57
III.8.3	<i>Cognitive Load on Attribution of Agent Intelligence</i>	58
III.8.4	<i>Effect of Cognitive Load on Self-Efficacy</i>	59
III.8.5	<i>Effect of Attribution of Agent Intelligence on Self-Efficacy</i>	59
IV	RESEARCH METHODOLOGY	61
IV.1	Experiment Design	61

IV.2	Selection of Participants	63
IV.3	Experiment Process Flow	64
IV.4	Task Design.....	66
IV.5	Scenario Generation.....	69
IV.6	Data Analysis Plan	71
IV.7	Determination of Sample Size.....	74
IV.8	Development and Testing	75
V	DATA ANALYSIS AND RESULTS.....	76
V.1	Sample Description	76
V.2	Descriptive Statistics.....	80
V.3	Assessment Outcomes.....	83
V.4	Model Evaluation	85
V.5	Results	92
V.6	Manipulation Evaluation	97
V.6.1	<i>Interaction with Experiment Condition</i>	<i>98</i>
V.6.2	<i>Effect on Measures</i>	<i>99</i>
V.6.3	<i>Effect on Relationships.....</i>	<i>101</i>
V.7	Discussion.....	105
V.7.1	<i>Overview of the Study</i>	<i>105</i>
V.7.2	<i>Contributions to Theory.....</i>	<i>107</i>
V.7.3	<i>Practical Implications</i>	<i>110</i>
V.7.4	<i>Limitations.....</i>	<i>112</i>
V.7.5	<i>Future Research.....</i>	<i>114</i>
VI	CONCLUSIONS.....	116

APPENDICES	118
Appendix A: IRB Approval	118
Appendix B: Test Instrument	119
Appendix C: Study Instrument	120
Appendix D: Scenarios and Explanations by Condition	121
Appendix E: Pre-Test Interview Guide.....	122
Appendix F: Instrument Development and Testing	123
REFERENCES.....	124
VITA.....	138

LIST OF TABLES

Table 1 Concept Definitions.....	9
Table 2 Example Computer Vision Algorithms.....	21
Table 3 Theoretical Constructs.....	46
Table 4 English Proficiency Qualification Examples	64
Table 5 Selected Study Feedback	79
Table 6 Descriptive Statistics by Condition and Differences in Means	80
Table 7 Correlation Table.....	82
Table 8 Descriptive Results of Damage Assessments.....	85
Table 9 Erroneous Agreement and Incorrect Ratings	85
Table 10 Latent Construct Quality Metrics	90
Table 11 Item Outer Loading	91
Table 12 Cross Loading of Items Between Constructs.....	92
Table 13 Model Path Coefficients.....	96
Table 14 Multi-Group Analysis by Manipulation Outcome.....	104

LIST OF FIGURES

Figure 1 Example Incorrect Explanation	3
Figure 2 Example Damage Classification and Explanation.....	8
Figure 3 Conceptual Framework	10
Figure 4 Example Algorithmic Counterfactual Explanations	40
Figure 5 Research Model.....	48
Figure 6 Experiment Conditions	62
Figure 7 Image Provided for the English Proficiency Qualification.....	64
Figure 8 Experiment Process Flow.....	66
Figure 9 Scenario Interface Example.....	67
Figure 10 Microtask Process Loop.....	69
Figure 11 Simulated Output and Explanation Specification	70
Figure 12 Process to Generate Scenarios.....	71
Figure 13 Second-Order Cognitive Load Structural Model.....	72
Figure 14 Recruiting Flow.....	78
Figure 15 Revised Structural Model	87
Figure 16 Structural Model Path Analysis Results.....	95
Figure 17 Effects of Explanations Relative to the “Black Box” Condition	97
Figure 18 Review Times by Manipulation and Outcome.....	100
Figure 19 Attribution of Agent Intelligence by Manipulation and Outcome.....	101

LIST OF ABBREVIATIONS

Abbreviation	Definition
ADAM	Automated Damage Assessment Machine (the name of the simulated agent in the survey instrument)
AI	Artificial Intelligence
AVE	Average Variance Extracted
ECL	Extraneous Cognitive Load
GCL	Germane Cognitive Load
HAI	Human-Agent Interaction
HCI	Human-Computer Interaction
ICL	Intrinsic Cognitive Load
PLS-SEM	Partial Least Squares Structural Equation Modeling
UAV	Unmanned Aerial Vehicle (a.k.a. Drones)
VIF	Variance Inflation Factor
XAI	Explainable Artificial Intelligence

ABSTRACT

Partnering People with Deep Learning Systems: Human Cognitive Effects of Explanations

by

Sean E. Dougherty

May 2019

Chair: Pamela Scholder Ellen

Major Academic Unit: Executive Doctorate in Business

Advances in “deep learning” algorithms have led to intelligent systems that provide automated classifications of unstructured data. Until recently these systems could not provide the reasons behind a classification. This lack of “explainability” has led to resistance in applying these systems in some contexts. An intensive research and development effort to make such systems more transparent and interpretable has proposed and developed multiple types of explanation to address this challenge. Relatively little research has been conducted into how humans process these explanations. Theories and measures from areas of research in social cognition were selected to evaluate attribution of mental processes from intentional systems theory, measures of working memory demands from cognitive load theory, and self-efficacy from social cognition theory. Crowdsourced natural disaster damage assessment of aerial images was employed using a written assessment guideline as the task. The “Wizard of Oz” method was used to generate the damage assessment output of a simulated agent. The output and explanations contained errors consistent with transferring a deep learning system to a new disaster event. A between-subjects experiment was conducted where three types of natural language explanations were manipulated between conditions.

Counterfactual explanations increased intrinsic cognitive load and made participants more aware of the challenges of the task. Explanations that described boundary conditions and failure modes (“hedging explanations”) decreased agreement with erroneous agent ratings without a detectable effect on cognitive load. However, these effects were not large enough to counteract decreases in self-efficacy and increases in erroneous agreement as a result of providing a causal explanation. The extraneous cognitive load generated by explanations had the strongest influence on self-efficacy in the task. Presenting all of the explanation types at the same time maximized cognitive load and agreement with erroneous simulated output. Perceived interdependence with the simulated agent was also associated with increases in self-efficacy; however, trust in the agent was not associated with differences in self-efficacy. These findings identify effects related to research areas which have developed methods to design tasks that may increase the effectiveness of explanations.

Keywords: Interpretability, Human-Agent Interaction, XAI

I INTRODUCTION

I.1 Explanations in Application

Intelligent systems that utilize artificial intelligence (AI) are increasingly being paired with humans to perform tasks. Some of the earliest theorists in computer science anticipated that computers would augment human intelligence (Wiener, 1950). Today, humans are being introduced into work flows to augment intelligent systems and improve their performance (Kamar & Manikonda, 2017). Accenture (2017) surveyed 1,201 executives and senior managers and found that 61% reported an increase in the number of roles expected to collaborate with intelligent systems. Additionally, 46% reported that some job descriptions within their firms had become obsolete due to intelligent systems. A recent survey by Gartner (2018) found that of 460 executive and senior manager respondents, 40% believed AI will make a material impact on production or service operations, and 50% on the products or services themselves.

The benefits of partnering people with intelligent systems depends not just on the capacity of the technology, but also on the ability for people to understand the reasons behind the system's output and take proper action. The impact of failures in this has been significant. Knight Capital Group lost over \$400 million in 40 minutes after a coordination failure between its staff and automated trading software led to millions of erroneous trading orders. The system produced warnings prior to the opening of the market; however, those were not addressed by the staff (SEC, 2013). The MD Anderson Cancer Center's \$62 million effort to integrate IBM's Watson technology into the selection of treatment plans resulted in headlines about the system offering "unsafe and inappropriate" treatments, but users were not provided information about the limitations of the system's training data to be able to evaluate why they might disagree with the system's recommendations (Ross & Swetlitz, 2018). In the criminal justice context, some

jurisdictions have used intelligent systems to predict recidivism. These evaluations have influenced the bail and sentencing decisions for thousands of defendants. An investigative report claimed that Northpointe's COMPAS system exhibited racial biases in risk scoring; however, analysis of the system's predictions have not identified any distinctions in system rating performance by race (Flores, Bechtel, & Lowenkamp, 2016). On the other hand, there is evidence that biases do arise when people use the scores to make decisions (Green & Chen, 2019).

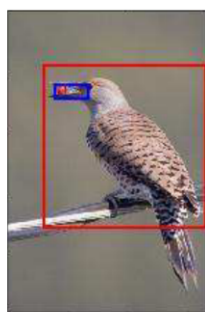
Compounding the challenge of understanding system output is a surge in the application of intelligent systems where the reasons behind system output are inherently difficult to understand (Došilović, Brčić, & Hlupić, 2018). Early algorithms in AI were designed to produce output that was intrinsically human intelligible, such as decision trees or lists of rules (Biran & Cotton, 2017). In contrast to these are "deep neural networks," which learn how to perform a task by estimating large numbers of numeric parameters in a multiple layer network. An example system of this type are "deep learning systems," which can process images and generate classifications based on objects the algorithm detects (Ball, Anderson, & Chan, 2017). A neural network in these systems has 5×10^7 learned parameters which define its function, and the system performs around 10^{10} mathematical operations to produce a single classification (Gilpin et al., 2018). Any explanation which describes the internal parameters or computation of the system would be unlikely to have meaning or be persuasive to a human user.

Not surprisingly, some stakeholders are uncomfortable deploying intelligent systems when even the creators of the system do not understand why it produces correct output (Holzinger et al., 2017). In addition, computer vision techniques do not yet have generalized understanding of the content of images and can produce implausible classifications with high

confidence (Akhtar & Mian, 2018). The lack of common sense in algorithms and fear over algorithmic bias has led to the European Union adopting broad legal requirements for AI systems to be able to explain their behavior (Doshi-Velez et al., 2017). While technologists had accepted AI systems that were unable to explain their output if the system was more capable than a more interpretable system, that is now changing (Adadi & Berrada, 2018).

Explainable Artificial Intelligence (XAI) is an area of research which seeks to maximize human interpretability of deep learning systems by developing means to expose logic or highlight elements of the data most salient to the output (Miller, 2019). These efforts have employed varying degrees of translation of data structures and interrogation of the deep learning system. Such a system is the automated bird classification system with explanations proposed by Hendricks, Hu, Darrell, and Akata (2018). This system presents the underlying reasons for a classification based on objects detected in the image by combining object recognition with a recurrent neural network to generate natural language explanations. An example output of this system is shown in Figure 1.

This is a **Northern Flicker** because ...



... this is a **black and white spotted bird** with **red beak**.

... this bird has a speckled belly and breast with a long pointy bill.

Figure 1 Example Incorrect Explanation

Note. Reprinted from Figure 4 (page 9) of Hendricks et al. (2018). Copyright 2018 by Springer Nature. Reprinted with permission.

The explanation in the example claims the bird is a “northern flicker” because it has a red beak. However, the bird in the image is in fact holding a red fruit in its beak. This does not mean the classification is necessarily incorrect. However, the basis of the justification provided by the explanation can be invalidated without needing domain knowledge about bird classification. The presence of the explanation may lead both to better understanding of the intelligent system and the requirements of the task.

Despite optimism in the literature and the apparent utility of explanations, there is reason to be cautious about the value of explanations in application. Adding “explainability” to deep learning systems may increase their acceptance, but there is conflicting evidence that explanations are processed by humans as expected in application. While the XAI area is relatively new and focused on algorithms that have been inherently uninterpretable, explanation has been examined within other types of intelligent systems such as expert systems, recommendation engines, and decision support systems (Nunes & Jannach, 2017). Past research has resulted in multiple proposed quality models for explanations with competing objectives, and has frequently identified challenges with humans utilizing explanations as intended (Miller, 2019). Nunes and Jannach (2017) called for explanations to be tailored and adapted to users; however, designers do not have clear guidelines for that effort.

I.2 Computation in Disaster Assessment

In the aftermath of a large natural disaster there is a need to rapidly assess vast amounts of unstructured data to make disaster relief and recovery decisions. Volunteers have been utilized extensively to monitor and assess information collected following natural disasters in the response and recovery phases of disaster management, using a method known as “crowdsourcing” (Yu, Yang, & Li, 2018). Chamales (2013) summarized crowdsourcing

technology as “bringing together a distributed workforce of individuals in order to collect resources, process information, or create new content” (p. 4). Many platforms have been designed to leverage crowdsourcing in numerous contexts to engage human evaluation as a central aspect of data processing systems, which has been termed “human computation” (Michelucci, 2016). Also driving this is an increase in the availability of high-resolution aerial imagery through the use of remote-sensing technologies, such as unmanned aerial vehicles (UAVs) (Kiatpanont, Tanlamai, & Chongstitvatana, 2016). As with any labor force, the quality of crowd workers must be monitored and controlled. This has led to the development of sophisticated architectures and methods to evaluate crowdsourcing products and individual workers (Daniel, Kucherbaev, Cappiello, Benatallah, & Allahbakhsh, 2018); however, full automation remains attractive.

A large area of literature explores deep learning algorithms which can automatically detect and rate damage to structures in aerial imagery. The goal of this research has been to fully automate damage assessment by utilizing successful systems trained on previous events to process new events, without adapting or retraining (known as “full transferability”). However, there are limits and challenges due to unique geographic and damage properties of each event (Vetrivel, Gerke, Kerle, & Vosselman, 2015). Efforts have improved transferability over time under conditions with ideal image quality (Vetrivel, Gerke, Kerle, Nex, & Vosselman, 2018), but in order to evaluate the performance of a transferred system on a new event at scale, images must be classified by hand to have a basis for performance evaluation.

State-of-the-art methods to achieve transferability include humans into the processing work flow as “trainers” for the algorithm and a source of inter-rater agreement. An example of this architecture is the microblog classification system from Imran, Castillo, Lucas, Meier, and

Vieweg (2014) where both crowdsourced individuals and trusted trainers review classifications made by the system, with the resulting data fed back to the system for further training. Proposals have been made to combine human computation and AI in disaster assessment with the goals of reducing human workload requirements and training intelligent systems to produce more rapid assessments (Imran et al., 2014; Ofli et al., 2016; Ostermann, 2015; R. Q. Wang, Mao, Wang, Rae, & Shaw, 2018). However, many of the AI methods applied today for automated assessment of damage are “black box” and unable to describe how they determined their output beyond highlighting areas of detected damage (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018). This interpretability issue and lack of training data have been identified as open issues by the review by Ball et al. (2017), which evaluated algorithms in remote sensing.

At the highest level, the goal of disaster management is to take advantage of aerial images collected immediately after an event to make better disaster response and recovery decisions. At the mid-level, the goal of the data analysis team is to produce more rapid assessments that identify areas requiring assistance. At the lowest-level of the task, the goal of individual damage assessors is to judge damage levels on individual images accurately. When that lowest level task is a partnership between a human and intelligent system, adding explanations of the system’s judgment for each image provides an opportunity to identify erroneous system performance, as well as challenge incorrect human evaluations or understanding of the damage assessment guidelines.

A damage assessment task environment was created for this study to test the effect of explanations. Damage classifications and explanations were developed using guidelines and templates from the literature. These classifications and explanations were provided as if they had been produced by an intelligent system (the “simulated agent”). Images of structures damaged

during Hurricane Michael (NGS, 2018) were used for the damage assessments. The objects identified in images and referenced by the explanations are aspects of structures identified by the Harvard Humanitarian Initiative's Wind Damage Rating guideline (Achkar, Baker, & Raymond, 2016). The causal and counterfactual explanation types generated by the system from Hendricks et al. (2018) were adopted for this study. In addition, an explanation type was included that discloses boundary conditions and failure modes relevant to the image ("hedging explanation"). These explanations were based on specific failure modes of real-world systems disclosed in the deep learning damage assessment literature. While an example implementation of this explanation type was not identified, such an explanation has been advocated by Hoffman, Miller, Mueller, Klein, and Clancey (2018).

An example damage assessment task from this study appears in Figure 2. The first statement is the simulated output where "Heavy Significant" refers to the structure and damage classifications from the guideline. Following this statement are the causal, counterfactual, and hedging explanation types examined by this study, in that order. As in the classification of bird species in Figure 1, domain knowledge is required to determine whether the classifications are correct. But it is possible to evaluate the coherence of the explanation with the image even without such knowledge. Further, with access to the guidelines for classification, the simulated output and explanations can be evaluated against a standard without requiring domain knowledge.



Figure 2 Example Damage Classification and Explanation

The advantage of utilizing the disaster assessment context in this research is that it has an extensive literature which has evaluated both human and automation performance. This supports the ability to generate a realistic task and to benchmark the results against prior findings. In addition, participants can be readily recruited to take part in the study. While this task is specialized, it may serve as an analogue for tasks in other contexts where the output of intelligent systems has unknown validity and people are partnered with intelligent systems to make rapid judgments.

I.3 Conceptual Framework

A conceptual framework was assembled to focus the research and identify constructs useful to develop understanding human information processing of explanations. The Van de Ven

(2007) formulation of Abraham Kaplan’s levels of abstraction is used here to start with semantically defined “theoretical concepts.” These concepts are expanded into one or more middle-range “theoretical constructs” drawn from the literature, which are then operationalized using “observable variables” of those constructs. Those variables become the basis of the research model. The theoretical concepts appear in Table 1 and the conceptual framework which connects these concepts appears in Figure 3. Theoretical constructs are developed in Chapter 3 to refine the concepts and provide theory and empirical evidence to evaluate.

Table 1 Concept Definitions

Concept	Definition
Worker	The human interacting with the simulated agent.
Microtask	A single instance of the process cycle where the human produces a structure and damage assessment.
Simulated Output	A classification generated by a simulated object recognition process of the agent.
Explanation	Information provided to the human to increase the performance of the human-agent partnership.
Type(s) of Explanation	One or more forms of explanation with content distinction from other types of explanation.
Information Processing	The human’s processing of the information presented to them by the microtask (particularly, the simulated output and any explanation).

Judgment	The human’s conclusion of the correct classification for the microtask, which is informed by and feeds back to other attitudes.
Attitudes about Agent	The human’s mental attitudes toward and beliefs about the simulated agent.
Performance Attitudes	The human’s attitudes about their performance of damage assessments.

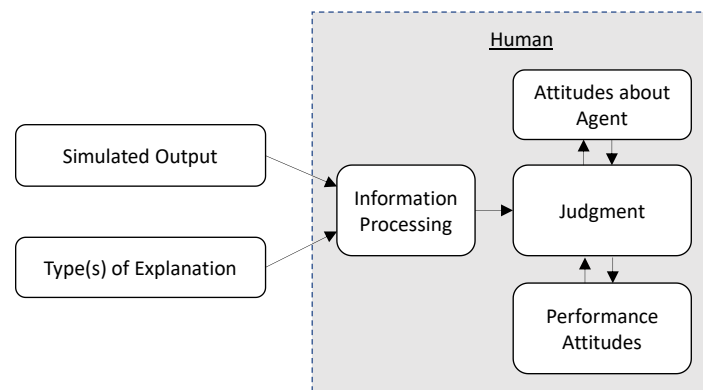


Figure 3 Conceptual Framework

I.4 Theory and Previous Research

Human interaction with intelligent agents has been extensively studied in the discipline of human-computer interaction (Smith, 2018). The earliest thinkers in this space considered natural communication the key to human-machine symbiosis (Licklider, 1960; Wiener, 1950). It was observed early that placing humans as supervisors of intelligent systems was a less-than-optimal configuration (Woods, 1985), but achieving “joint activity” between humans and intelligent systems instead of supervision by either participant is not trivial (Bradshaw et al., 2009). Several researchers within XAI research believe that the fitness for intelligent agent explanations is

based not just on the quality of the explanation logically but also how the explanation is perceived by humans (Doran, Schulz, & Besold, 2017; Miller, 2019; Tintarev & Masthoff, 2011). Despite human-computer interaction being strongly rooted in cognitive science and information theory from psychology, cognitive psychology has not strongly informed research in XAI (Abdul et al., 2018). Research of human perception of explanation in XAI has frequently focused on the extent to which specific explanation engines generate trust or rating preferences, but provide limited insight into how humans interpret those explanations (Miller, 2019). Understanding how explanations can improve human-agent interaction can aid not just the disaster assessment context, but many areas where intelligent agents have the potential to improve safety, quality, and cost.

Social cognition, rooted in psychology and cognitive science, seeks to understand how humans process information from their environment, perceive other intelligent behavior, and learn through observation (Fiske & Taylor, 2016). The theoretical framework of this study leverages three areas of theory from social cognition: intentional systems theory (Dennett, 1971), which examines the extent to which people perceive mental processes in observed behavior even in artifacts that do not have a mind; social cognitive theory (Bandura, 1986), which examines how social interactions inform attitudes about our capacity and motivation to engage and devote effort into an activity; and cognitive load theory (Sweller, 1988), which has developed understanding of the limits of human cognitive architecture.

Dennett (1971) proposed intentional systems theory, which states that humans will attribute an observed actor as having a mind, and predict its behavior as such, even when that is clearly impossible. When humans are interacting with other humans, empirical research in attribution theory has identified that differences in attribution of observed behavior are heavily

based on the information available to the observer and how well they can simulate the thinking of the observed (Malle, 2011). The perception of mind has been explored in an increasing number of dimensions (H. M. Gray, Gray, & Wegner, 2007), and these concepts have been specifically used in robotics to understand how humans interpret actions by robots that are normally conducted by humans (Terada & Yamada, 2017; Thellman, Silvervarg, & Ziemke, 2017). To the extent that workers perceive mental processes as taking place in the agent, making it predictable and rational, workers will be able to coordinate activity more appropriately and detect both the machine and their own inappropriate classifications.

Social cognitive theory (Bandura, 1986) and cognitive load theory (Sweller, 1988) were originally developed in instructional settings for the purpose of optimizing how learners build and manipulate mental schemas of a task, and to predict future behavior such as interest in and effort undertaken in a learning task. Central to social cognitive theory is the theoretical construct of “self-efficacy,” which is perhaps closest in meaning colloquially to “confidence,” and is distinct from the ability to perform a task (Bandura, 1977). It has been called the foundation of human performance (Peterson & Arnn, 2005), and strong correlations have been identified across many studies of both job performance and satisfaction in a workplace setting (Judge & Bono, 2001). Within the area of crowdsourcing damage assessment, the review by Dittus (2017) examined the factors of worker training and feedback through the lens of self-efficacy, citing its importance in having crowd workers return for future damage assessment efforts. Cognitive load theory has frequently been studied with self-efficacy and focuses on how limitations in working memory lead to challenges in manipulating and constructing mental schemas to understand a task (Sweller, 1988). Explanations of the intelligent agent’s logic for classification have the potential to focus damage assessment on relevant and differentiating content of the images. In

turn, this has the potential to direct the worker's attention to the guideline, and more rapidly and accurately construct mental schemas for damage assessment. However, high cognitive load has been found to produce less accurate attributions of observed behavior (Molden, Plaks, & Dweck, 2006), making it possible that the agent could create sufficient cognitive load to prevent its behavior from being perceived as intelligent.

The XAI literature has touted the benefits of counterfactual explanations, which offer contrasting reasons for not making other potential classifications (Wachter, Mittelstadt, & Russell, 2017). Hoffman et al. (2018) proposed that explanations which offer boundary conditions and known failure modes have the potential to increase trust. This type of explanation is termed as "hedging explanations" by this study. By allowing the worker to assess the role of any potential boundary conditions, hedging explanations may provide clarification for how an agent reached an otherwise seemingly irrational classification. The need for such explanations may rise to a legal requirement when the providers of a system are aware of failure modes, even if the explanations are challenging to accomplish (Doshi-Velez et al., 2017). While causal explanations that provide "why" information about output have been researched extensively since intelligent agents were first in use, no empirical evidence was identified regarding the effectiveness of counterfactual and hedging explanations in an interactive judgment. However, literature advocates for the utilization of such explanations (Hoffman et al., 2018; Miller, 2019; Wachter et al., 2017).

I.5 Research Questions

The explanation types appropriate in an application may be most determined by the "explainees" and whether they process the explanation types as expected. Counterfactual explanations allow damage assessment novices to observe the agent applying the guidelines and

demonstrate contrasting logic. Hedging explanations allow workers that are unaware of the limitations of the agent to better evaluate the appropriateness of the output based on the image contents. While the potential of both types of explanation is compelling, is not clear that they will achieve these objectives individually, nor that the types can be combined without exceeding cognitive limitations of the workers. As such, the following research questions are proposed:

- 1) How will people process counterfactual and hedging explanations of damage assessments by a deep learning system?
- 2) How does combining the explanation types affect their processing?

I.6 Motivation for the Study

I have spent nearly my entire professional career developing and deploying enterprise resource management systems and information systems which have included intelligent agents in their functionality. Those were expected to partner well with workers to improve outcomes, but the potential of these agents was rarely fully realized. This outcome is not uncommon despite high expectations for and excitement about the potential for intelligent agents. Over time I have come to see how intelligent agents do not make ideal partners and teammates. The workers have largely remained responsible for outcomes despite being paired up with intelligent agents, since these seem to operate mysteriously and carry little, if any, responsibility for outcomes. The ability for workers to understand how an intelligent agent has reached a decision is a major step towards opening areas that were previously impractical or impermissible. The context used to explore this challenge is also uniquely motivating compared to many other possible tasks because of the potential for humanitarian impact.

I.7 Contribution and Significance

This research contributes to the understanding of how three types of explanation are processed by humans by measuring cognitive load, attribution of agent intelligence, and self-efficacy in the task. This study found that the most common construct evaluated in the literature, trust in the intelligent agent, did not have a significant relationship with self-efficacy when these other constructs were considered. The study also identified that cognitive load and perceived interdependence had the greatest effect on self-efficacy. The instructional literature that developed cognitive load theory has identified a number of approaches that could be applied to explanation to operate within the limitations of human working memory. Perceived interdependence was also associated with increased self-efficacy, and is an existing area of research within human-robot interaction with the aim of increasing the effectiveness of joint activity. These two constructs were identified as having a similar or greater effect on self-efficacy as did previous task experience. Finally, this study indicates that when workers agree with erroneous output, it is most likely as a result of “going along,” rather than an illusion of explanatory understanding.

I.8 Dissertation Organization

This dissertation is organized as follows:

- **Chapter II:** A literature review is conducted in the areas of Computation in Disaster Assessment, Social Cognition, and Explanation by Intelligent Agents with the purpose of identifying current knowledge and theoretical constructs that have explanatory power.
- **Chapter III:** A research model is developed with hypotheses of relationships between the theoretical constructs, and a measurement model is developed to evaluate the research questions.

- **Chapter IV:** The methodology used to conduct the research and the analysis of the data is presented. A method to develop simulated output and explanations is described.
- **Chapter V:** The results of the analysis are reported with discussion of the results, contributions, practical implications, limitations, and future research directions.
- **Chapter VI:** A summary is offered of the study and its findings.

II LITERATURE REVIEW

II.1 Computation in Disaster Assessment

II.1.1 *Background*

On January 12, 2010 a magnitude 7.0 earthquake struck Haiti, creating one of the largest humanitarian crises of modern times, with enormous uncertainty of the impact and extent of damage (Kolbe et al., 2010). Large parts of the country were unmapped with tools common in developed countries (Zook, Graham, Shelton, & Gorman, 2010). It was one of the first major natural disasters where “crowdsourced” mapping and damage assessments using remotely sensed aerial imagery was utilized to make disaster relief decisions (Corbane et al., 2011). Information challenges and capabilities arising out of natural disasters has given rise to crisis informatics, which has been defined as the “interconnectedness of people, organizations, information, and technologies during crisis.” (Hagar, 2010).

The initial rapid assessment in the immediate aftermath of a disaster requires processing a large volume of unstructured data using a variety of approaches both to collect the data and to turn it into useful information (Poblet, García-Cuesta, & Casanovas, 2014). Crowdsourcing in disaster damage assessment has become highly organized, with groups such as Humanitarian OpenStreetMap Team emerging to develop training and pools of workers, as well as coordinating crowd efforts with international relief agencies (See et al., 2016). The results of crowdsourced disaster mapping efforts have been positive and the learnings from each disaster have been carried to the next in continuous improvement (GISCorps, 2013; Lallemand et al., 2017). While automated approaches perform well once trained, successful systems do not necessarily perform well on images from a new disaster event (Vetrivel et al., 2015). Complicating efforts across both crowdsourced and machine learning domains is an inherent

difficulty in establishing a “correct rating” for any given image, along with disagreement between aerial and ground-based surveys (GISCorps, 2013; Westrope, Banick, & Levine, 2014). As a result, inter-rater agreement is commonly used to determine the “correct” rating, even among groups of expert raters.

II.1.2 Crowd Workers in Disaster Assessment

Albuquerque, Herfort, and Eckerle (2016) examined volunteer crowdsourced humanitarian mapping in a non-disaster context in the Democratic Republic of Congo. Crowd workers reviewed images and reported whether a satellite image contained roads or a settlement. This provided an opportunity to compare crowdsourced interpretation with a reference map data set. The consensus of the crowdsourced data had high agreement and accuracy with the reference data (accuracy and precision of 89%, and sensitivity for feature detection of 73%). A low degree of rating consensus was strongly predictive of task difficulty on an image. Despite a large number of contributors, only a small number of volunteers completed the majority of the image classification effort. The average user classified 66 images with a median of just 21 images.

GISCorps (2013) compared damage ratings between volunteers and experts in rating aerial images from Hurricane Sandy. They found that volunteer classifications for images exceeded 95% inter-rater agreement confidence once they had been rated by five workers, with little practical difference made by the experience level of the volunteers. Expert consensus ratings were generated for the images with the lowest volunteer consensus (a total of 2% of images). Experts and volunteers produced identical consensus ratings in 64% of these most difficult images, and 11% of this subset was rated at the opposite end of the damage scale between the two groups. Images with greater volunteer agreement were not assessed by experts. GISCorps also identified several areas of improvement. Among these was a proposal to consider

allowing participants to decline to rate invalid images, and to provide specific classification definitions to raters. In a separate review, Westrope et al. (2014) found that low agreement between raters most often occurs due to image quality, there tends to be a focus on high impact areas such as major cities and high-profile incidents, and lack of pre-disaster imagery significantly hampers any classification effort. While some of these efforts are volunteer-driven, it has been recommended that crowdsourced damage assessments be a paid task supervised by expert raters (GIScorps, 2013).

Rather than being an on-call population of repeating workers, the review of 26 disaster events served by the Humanitarian OpenStreetMap Team by Dittus (2017) found that the 20,000 crowdsourced contributors were mostly first-time workers (50.2% overall, and as high as 84.7% for the Nepal Earthquake event). Collaborative mapping (including disaster assessment) was reviewed from a cognitive systems engineering perspective by Kerle and Hoffman (2013), with specific observations about how current processes in complex geographic information systems are designer-focused instead of work-focused, leading to overly complex task design unsuitable for the varied expertise levels of participants.

II.1.3 *Automation in Disaster Assessment*

Much of the literature on automated damage assessment has focused on fully automating assessments rather than combined human-AI approaches; however, humans frequently create the training information used by these fully-automated approaches. Techniques from computer vision, where algorithms interpret images, have been applied to disaster damage assessment for events such as tropical cyclones (Cao & Choe, 2018), earthquakes (Joshi, Tarte, Suresh, & Koolagudi, 2017), tsunamis (Fujita et al., 2017), and wildfires (Trekin, Novikov, Potapov, Ignatiev, & Burnaev, 2018). Some researchers have focused efforts on a single element of the

challenge such as understanding and evaluating the complex structure of residential rooftops (F. Wang, 2017). Computer vision techniques have been used to augment workers by highlighting areas of change and aligning images with maps (Trekin et al., 2018).

Cheng and Han (2016) reviewed 270 articles to identify a taxonomy of object detection methods in remote sensing, including damage assessment applications. That review divided algorithms into four types: template matching to detect change; knowledge-based, which uses geometric and context rules (such as shadows) to detect buildings; object analysis; and machine learning. Machine learning approaches were broken into steps of feature extraction (the detection of individual image points, edges, and “blobs”) and classifier training to detect objects using detected features. Mayer (1999) developed a list of elements of buildings such as roofs, windows, and other structural components useful for classifying damage in aerial images. In addition to detecting the objects that comprise a building, automated building extraction techniques can identify individual buildings and center them in an image for evaluation by other algorithms or human damage assessors (Shrestha, 2018). Ball et al. (2017) reviewed over 400 articles on deep learning in remote sensing. The top two open issues they identified were limitations in available training data, and challenges in human interpretability of deep learning algorithms. The set of computer vision algorithms in aerial imagery, damage assessment, and object detection reviewed for this research appears in Table 2, listing the types of data considered, the purpose of the algorithm, and types of disasters evaluated (where applicable).

Table 2 Example Computer Vision Algorithms

Article	Type, Purpose: Disaster
Attari, Ofli, Awad, Lucas, and Chawla (2017)	UAV Oblique Images, Damage Assessment: Hurricane
Cao and Choe (2018)	Satellite images, Damage Assessment: Hurricane
Duarte, Nex, Kerle, and Vosselman (2017)	UAV Oblique Images, Damage Assessment: Earthquakes
Duarte, Nex, Kerle, and Vosselman (2018)	Multiple aerial image sources, Damage Assessment: Earthquakes
Fujita et al. (2017)	Satellite images, Damage Assessment: Tsunami
Kersbergen (2018)	Multiple data sources, Damage Assessment: Hurricane
Kluckner, Mauthner, Roth, and Bischof (2009)	Aerial images, Semantic object extraction.
Moranduzzo and Melgani (2014)	UAV Images, Car Counting.
Nguyen, Ofli, Imran, and Mitra (2017)	Ground-level image damage assessment in natural disasters: Earthquakes, Hurricanes
Qi, Yang, Guan, Wu, and Gong (2017)	Satellite images, Building and land use classification.
Trekin et al. (2018)	Satellite images, building segmentation in natural disasters: Wildfire.
Vetrivel et al. (2018)	UAV Oblique Images, 3D point cloud inference in damage assessment: Earthquakes
F. Wang (2017)	Aerial images, Debris Detection and Roof Condition Assessment: Hurricanes

The ability to use existing successful models on new disasters has improved with ultra-high resolution imagery and advanced 3D point imputation; however, results still must be assessed compared to data labeled with the “correct” classification to validate performance (Vetrivel et al., 2018). Sensitivity to sparse and incorrect training information has an unknown impact on new event classifications until sufficient training labels are generated for that event (Frank, Rebbapragada, Bialas, Oommen, & Havens, 2017). Multiple authors have suggested combining human computation with automated assessments to generate labeled training data, as well as classify challenging images (Albuquerque et al., 2016; Ofli et al., 2016; Ostermann, 2015). Such architecture has been used in social media microblog classification in disaster (Imran et al., 2014) and non-disaster settings (Sadilek et al., 2016).

II.1.4 *Open Issues*

Perhaps the largest open issue is that accuracy of ratings in aerial damage assessment is not straightforward to determine. Comparisons between crowdsourced aerial damage surveys and field surveys have been found to have low agreement, with Cohen's kappa agreement scores of less than 0.4 (Westrope et al., 2014). Similarly, GIScorps (2013) found the ratings of their experts were not entirely consistent with Federal Emergency Management Agency field assessments. Damage assessments in earthquake events have particularly low agreement between field and aerial image assessments as building damage is less visible from above. Users of disaster data are accustomed to what they believe to be over-estimation of disaster damage (Westrope et al., 2014). There may not be a definitively correct classification for any single image.

There is limited knowledge of how combined human-automation systems or explanation operate in damage assessment using aerial images. A comparable system was not identified that joined human-generated labels with emerging disaster data similar to AIDR (Imran et al., 2014) for image data. Additionally, machine learning models in damage assessment have made limited efforts in interpretability where the deep learning techniques most commonly applied in this space provided only highlights of the areas the model classifies as damaged. None of the articles reported providing natural language or other forms of justification for their classifications.

The role of individual expertise and skill in crowdsourced damage assessment has conflicting findings at different levels of analysis. Kerle and Hoffman (2013) identify misperceptions about the consistency in skill levels to assess damage across crowd workers, and particularly note that they consider efforts to open damage assessment to untrained analysts as "misguided." However, the effects of these differences in expertise were not identified by

GIScorps (2013) to create meaningful differences in consensus ratings aggregated across multiple raters.

II.2 Social Cognition

II.2.1 Scope

Social cognition is an area of psychology interested in perception and information processing in the context of social interactions (Pennington, 2012). Heider (1958) put forth the concept that people are naïve psychologists that predict, understand, and control their social environment by attempting to understand why others behave the way they do. This area became the foundation of social cognition through the development of attribution theories (Reeder, 2013). Research in this area has considered and developed theories of how *observers* and *actors* explain their own behavior as compared to the behavior of those observed. In the context of this study, the simulated agent is conjectured to be an observed actor.

The social cognition approach sees behavior as a sequence of stimulus (and its interpretation), person (particularly the reasoning process within the person), and response (the behavioral outcomes of cognition), with a strong emphasis on rationality such that even erroneous inferences are a result of goal-directed thought (Fiske & Taylor, 2016). These inferences are subject to systematic failure modes, but the heuristics behind these flaws are generally efficient and accurate (Tversky & Kahneman, 1974). Attitudes about the agent and an individual's own performance are core to two of the three main principles of social cognition identified by Pennington (2012): first, people are cognitive misers and attempt to minimize cognitive effort; and second, self-esteem and positive self-evaluation are critical to being confident with one's own capacities as well as being confident in other people. The third

principle, that people have separations between spontaneous and deliberative thought, is relevant here in the form of biases and heuristics in the judgment of classifications.

Three areas within social cognition were identified with relevance to the task of interacting with the simulated agent. Intentional systems theory (Dennett, 1971), a successor of earlier naïve psychology, approaches the understanding of how humans infer mental states of observed actors, which are often technological artifacts. Self-efficacy arose from social learning theory (Bandura, 1977) and later social cognitive theory of behavior (Bandura, 1986), which has explored how humans learn from each other through observation, imitation, and modeling. Cognitive load theory developed out of attempts to understand how humans build mental schemas under human cognitive limitations (Sweller, 1988), which has been studied as a underlying mechanism that supports the development of self-efficacy (statistically modeled as a “mediator”). Explanations are likely to introduce additional cognitive load in the task, which may be beneficial to support understanding or harmful if exceeding limitations.

II.2.2 Intentional Systems Theory

II.2.2.1 Background

Some of the earliest research that informed attribution theory found that people ascribed human traits to non-human things, even simple animated shapes that seemed to interact with each other in a movie (Heider & Simmel, 1944). Intentional systems theory (Dennett, 1971) proposed that there were three mechanisms of prediction that humans use: a physical stance, where predictions of outcomes are made with one’s understanding of the physical laws of science; a design stance, where the behavior of objects is predicted by their designed purpose; and the intentional stance, where predictions of future actions are based on inferred mental states of the observed actor, such as belief and desire. The scope of the inquiry within the broad

literature here draws on learnings of how humans that interact with artificial intelligent agents on a task develop attitudes about the agent and the task.

Within XAI, attribution theory has been seen as a way to explore how belief-desire-intent agents can relate themselves to their users (Miller, Howe, & Sonenberg, 2017). The intentional stance claims that people perceive mental states even when we know that machines are incapable of them (Dennett, 1989). The perception of mental states of others has been cited as critical to observational learning with specific neural structures designated for both automatic and purposeful inference of mental states (Frith & Frith, 2012). Attributions of mental states have also been found to be stable whether they are requested as immediate impressions, or when respondents are given time to think about the nature of the observed individuals (Lobato, Wiltshire, Hudak, & Fiore, 2014).

II.2.2.2 Dimensions of Mind

Dennett (1971) noted the difficulty of differentiating the behavior of complex artifacts from systems with arguably true beliefs and desires. For instance, a chess-playing computer does not have beliefs and desires in comparison to human chess players. However, their behavior is outwardly identical, and Dennett (1971) argues that for the practical purpose of understanding behavior the difference is irrelevant. Even the designers of a chess-playing computer speak of the system as intentional in terms of an identity or person rather than in the way we speak of designed objects as toward their purposes. This thought has been extended by experiments which have found that humans perceive that robots have social traits, as well as the perception of having mental traits. This occurs even in non-anthropomorphized robots, suggesting that this arises from the interaction rather than the perception of a human (K. Gray & Wegner, 2012).

Attribution of mind has been found to be neither binary nor unidimensional across varying status and types of people and technology. H. M. Gray et al. (2007) explored how respondents perceived 12 different “characters” representing living animals or people on 18 dimensions. They found two emergent dimensions: *agency*, which is mostly made up of cognitive activities; and *experience*, which is made up of mainly biological and emotional attributes. Weisman, Dweck, and Markman (2017) expanded on this by examining 21 entities for 40 “a priori” concepts of mental capacity where entities included a broad range of subjects including a desktop stapler, a computer, a robot, various animals, and varying mental states of humans. They identified three emergent factors: *body*, physiological sensations; *heart*, affective states; and *mind*, cognitive properties of communication, planning, and making choices. In the case here, the observed behaviors that are consistent with the cognitive properties of a mind are most relevant.

II.2.2.3 Attribution of Behavior of Technological Artifacts

Several frameworks have been proposed to study the attribution of behavior for technological artifacts, as well as approaches to study the source of these perceptions. While attributions are relatively similar between humans and humanoid robots, differences have been observed as systems become increasingly non-human.

Thellman et al. (2017) evaluated how people interpret the behavior of humanoid robots compared to humans and found little difference in inferences of intelligence, which were also not modified by situational cues (Thellman et al., 2017). This study used the causal network of explanation framework proposed by Böhm and Pfister (2015). This framework provides a model of attributions which they claim classifies both behavior and explanations of behavior. They evaluate the actor’s *goals*, enduring *dispositions*, *temporary* states such as emotions, *actions*

(intentional behaviors that are goal-directed), *outcomes*, uncontrollable *events*, and *stimulus attributes* which are the “features of the person or object toward which a behavior is directed.”

Some studies have altered the appearance or portrayal of the observed entity where the behavior was otherwise equivalent. The study by de Visser et al. (2016) varied the type of agent (human, avatar, and computer) and the accuracy of its advice, evaluating trust, trust repair, and confidence in an agent-assisted cooperative synthetic number-guessing task. Similarly, Terada and Yamada (2017) varied the type of agent (a laptop computer, an intelligent toy bear, a human-like robot, or a human) as an opponent in a simple coin flipping game. In both of these experiments the portrayal of the other participant altered attributions without any change to the interaction or behavior.

Other studies have explored how differences in behavior with the same portrayal affect the perception of intelligence. Kryven, Ullman, Cowan, and Tenenbaum (2016) examined how participants attributed intelligence to videos of an animated non-human character moving around a maze. Each respondent rated many separate videos while the maze search approaches were manipulated. A single item 5-point scale of intelligence from “less intelligent” to “more intelligent” was used, and at the end of the entire sequence the participant was asked an open-ended question: “how did you make your decision?”, which was then coded as either being based on “outcome” or “strategy” per respondent. Many participants provided ratings of intelligence based not on the strategy used, but the outcome of the strategy. They speculated that evaluating outcomes was a shortcut to evaluating intelligence as a form of satisficing, since reading the mind of the animated character required more cognitive effort.

II.2.2.4 Biases and Asymmetries in Attribution

Some research has linked the cause of attributions to incorrect inferences resulting from differences in information availability between actors and observers (Malle, 2011; Robins, Spranca, & Mendelsohn, 1996). Explanations change the information available to an observer, and may interrupt systematic attribution processes based on asymmetry. Anderson (1983) found that the availability of items of information, such as having the necessary aids, using the right approach, and knowing relevant information, were cited more often in failure than success. Actors are generally able to project what observers would observe, but observers are unable to project what actors perceive (Krueger, Ham, & Linford, 1996). Other studies have not agreed with information asymmetry, such as Storms (1973), which found that actors provided similar explanations as observers when they watched themselves on videotape even though they had access to the same information as actors.

Dzindolet, Pierce, Beck, and Dawe (1999) proposed a framework of human-automation teaming based in part on cognitive biases and social processes. They used that framework to construct four experiments which found that social, cognitive, and motivational processes each impact the potential use of and reliance on automation. Van Dongen and van Maanen (2006) examined how failures were attributed in a hypothetical decision aid system. They found that that attributions of failures were related to stable permanent traits of the decision aid, rather than relating to causes outside of the decision aid; meanwhile, failures in human decisions were attributed to more temporary and external causes.

Fein (1996) found that providing suggestions of an observed actor's mental states was sufficient to decrease the role of biases in attributions of behavior and inferences about the actor. While Fein (1996) anticipated that people understand computers as objects without intentions

and therefore not subject to suspicion, intelligent agents perceived as intentional systems may encounter this phenomenon.

II.2.3 *Self-Efficacy*

II.2.3.1 *Background*

Social cognitive theory proposed that there are two conceptions that determine behavior: expectations relating to outcomes of the behavior, and beliefs about the ability to perform. This second set of beliefs was conceptualized as self-efficacy, or as defined in Bandura (1986): “judgments of their capabilities to organize and execute courses of action required to attain designated types of performances. It is not concerned with the skills one has but with one’s judgments of what one can do with whatever skills one possesses” (page 391). Bandura (1977) conceptualized self-efficacy as the expectation of performance which drives behavior to attempt an activity, predicting the amount of effort people will expend. Bandura (1986) proposed that humans acquire knowledge primarily by observing others while taking part in social interactions. Self-efficacy is a self-perception of competence and future performance as opposed to actual competence in a task. The construct has been described as the foundation of all human performance (Peterson & Arnn, 2005).

Bandura identified four ways to increase self-efficacy: successful performance, watching others succeed (also known as vicarious learning), persuasion and encouragement, and physiological/affective factors (Bandura, 1986). The first two of these are particularly relevant in interacting with an intelligent agent to perform classification, where agreement can be interpreted as successful performance, and mimicking of observed reasoning and logic for classifications similarly produces learning leading to successful performance. Gist and Mitchell (1992) examined the inter-relationships and feedback between performance and self-efficacy,

noting that interventions designed to modify self-efficacy had to address specific determinants of self-efficacy, such as perceptions of ability and task complexity.

II.2.3.2 Application to the Workplace

While originally developed in an educational context, self-efficacy has also been extensively explored in an employment setting. Studies have consistently identified positive relationships with job satisfaction and performance. The meta-analysis by Stajkovic and Luthans (1998) of 114 studies found the effect of self-efficacy on job performance to be greater than goal-setting, feedback interventions, and organizational behavior modification. General self-efficacy was found to have higher correlations with job satisfaction than other predictors (self-esteem, locus of control, and emotional stability) and comparable correlations with job performance in the meta-analysis by Judge and Bono (2001) of 135 studies. The meta-analytic path model by Brown, Lent, Telander, and Tramayne (2011) used data from 8 studies identifying that self-efficacy had similar effects as cognitive ability on work performance. The study by Koutsoumari and Antoniou (2016) identifies occupational self-efficacy as having correlations with multiple dimensions of work engagement, and they advocate for using training and development to positively influence the self-efficacy of employees.

II.2.3.3 Application to Human-Computer Interaction

An early adaptation of self-efficacy in human-computer interaction was by Compeau and Higgins (1995) which developed a scale for the use of computers based upon social cognitive theory (Bandura, 1986). Compeau and Higgins (1995) established that an individual's reactions to an information system would be impacted by their self-efficacy that they could use the system. Self-efficacy appeared in later conceptions of Technology Acceptance Model (TAM) such as the extension of TAM (Davis, Bagozzi, & Warshaw, 1989), TAM2 (Venkatesh, 2000), and UTAUT

(Dwivedi, Rana, Jeyaraj, Clement, & Williams, 2017). Within this space self-efficacy is positioned as an antecedent to use attitudes (Thompson, Compeau, & Higgins, 2006). Self-efficacy has been examined as an antecedent to automation complacency (Parasuraman & Manzey, 2010), and as a way of accounting for over-use of manual control in automation settings (Lee & See, 2004).

Self-efficacy has been examined in the context of specific systems both in case study and experimental observation, including disaster assessment. In a study by Zheng, McAlack, Wilmes, Kohler-Evans, and Williamson (2009), it was used within a computer-based learning context to evaluate the learning impact of manipulating the multimedia conditions of a tool, where more complex displays lowered cognitive load and increased self-efficacy. The design of interaction for a website to retrieve information was examined where a more complex website decreased self-efficacy (P. J.-H. Hu, Hu, & Fang, 2017). In these two cases the tasks were modified in complexity through the way information was presented. In Leaman and La (2017), concepts from self-efficacy were integrated into training plans which increased successful adoption among users of smart wheelchairs. Within crowdsourced disaster assessment, Dittus (2017) utilized self-efficacy as a framework to analyze microtask design and worker feedback, finding reinforcement led to retention of volunteers in future projects, which was interpreted as deriving from self-efficacy.

II.2.3.4 Illusory Understanding

Participants do not get graded feedback on their performance during damage assessment, leading to the potential of having high confidence in assessments which are inconsistent with the scenarios and the guideline. McKenna and Myers (1997) explored the role of accountability in ungraded tasks and found that people who were briefed to expect an assessment of their skill

level provided lower ratings of their understanding than participants that were not advised of an assessment. Kruger and Dunning (1999) performed a series of studies which examined how people rated their own performance in groups. They found that low performing participants often rated themselves as high as the highest performing participants. The illusion of explanatory understanding arises when heuristics mislead us in assessing our understanding of explanations. This is particularly deceptive for early learners, causing them to believe they understand more than they actually do through perceived surges in understanding and insight that are not, in fact, coherent (Keil, 2006). Previous studies have found that even experts can be misled by the explanations of an automated judgment into producing confident but incorrect decisions (Bussone, Stumpf, & O'Sullivan, 2015).

II.2.4 Cognitive Load Theory

II.2.4.1 Background

Cognitive load theory (Sweller, 1988) originated in cognitive psychology's studying of instruction and learning, and was later investigated together with self-efficacy to optimize learning tasks (Hesketh, 1997; Steele-Johnson, Beauregard, Hoover, & Schmidt, 2000). Cognitive load theory crossed over into the Human-Computer Interaction (HCI) literature in computer-based instruction (Nicholson, Hardin, & Nicholson, 2003) and in automated decision aids (Benbasat & Todd, 1996), and later integrated into HCI more broadly (Hollender, Hofmann, Deneke, & Schmitz, 2010). Cognitive load theory extends schema theory's perspective that knowledge is stored in mental schemata which are constructed through learning processes, where limitations in working memory interfere with mentally manipulating a task's schema (Hollender et al., 2010). Cognitive load theory has conceptualized three kinds of cognitive load: inherent (the basic mental load required for a task), germane (mental load that is not inherent to the task,

but is beneficial to learning), and extraneous (mental load that is not related to the task) (Sweller, van Merriënboer, & Paas, 1998). Key to the theory is how differences between people, tasks, and situations influence cognitive load. It is also known that cognitive load, by reducing information processing resources available to other processes, modifies attributions of observed behavior to those more consistent with the existing views of the observer (Molden et al., 2006), and prevents situational information from entering into the judgment of causal processes (Hilton, 2007).

Worked examples is an instructional method used to reduce cognitive load while increasing the rate of learning by showing the reasoning used to complete a problem-solving task (Sweller & Chandler, 1991), which in some respects is similar to the explanations of damage assessments. Example-based learning (Van Gog & Rummel, 2010) similarly has learners observe someone else performing a task to facilitate learning.

II.2.4.2 Applications in Computing Systems

Benbasat and Todd (1996) reviewed a series of experiments which manipulated the cognitive load of tasks in a decision support system to test the effect on which paths of options users followed. They found that users chose more complex and appropriate methods when cognitive load was considered in task design, and avoided the same paths otherwise. Potter and Balthazard (2004) investigated how a computing system assisted problem-solving by directing attention to causes (“cause cueing”), increasing the number of solution ideas generated. They also found that identification of causes decreased with distraction and increasing working memory requirements. Cognitive load was evaluated as a mediator for human information processing to understand how to work in split-attention and complex problem-solving environments (Oviatt, 2006). Cognitive load theory has been used to evaluate learning and performing tasks using information systems (P. J.-H. Hu et al., 2017; Zheng et al., 2009).

Worked examples have been used in a computer-based learning interface to explain chemical reactions, resulting in increased self-efficacy (Crippen & Earl, 2007). Within explanations, cognitive load was seen as a key factor in user acceptance of recommendations (Giboney, Brown, Lowry, & Nunamaker, 2015), and was also found to increase cognitive load, disrupting ease of use as the amount of explanatory information increased (Nunes & Jannach, 2017).

II.2.4.3 Expertise and Explanation

Task designs have been found to be sensitive to level of expertise with no “one size fits all” solution. In the review by Sawicka (2008), self-explanation and cognitive elaboration by learners was found to increase germane load for novices and increase extraneous load for experts. Seufert, Jänen, and Brünken (2007) found that “help” offered by an instructional system was sensitive to levels of existing knowledge, where insufficiently low levels were related to not being able to take advantage of the assistance, and levels that were too high related to “expertise reversal,” resulting in increased cognitive load without a benefit in understanding. Between these extremes, not all kinds of help were useful, and hyperlinked information and elaborations did not improve learning performance. This was further supported by Mascha and Smedley (2007) which found that computerized decision aids used by experienced people decreased the skill levels of experts on non-complex cases, even though the same aid was useful to novices approaching the same cases. As such, cognitive load can be sensitive not just to the content of the information but how and to whom it is being presented.

II.2.5 Open Issues

While research has identified multiple dimensions of mind perception across many types of entities, the task of this study imagines a non-embodied intelligent agent that is not anthropomorphized. No studies were identified that evaluated the attribution of intelligence or

the dimensionality of such attribution where the technological artifact provided explanations. Further, no studies of mind perception were identified within the context of XAI systems. Additionally, to the extent that agents are seen as intentional systems, the presence or lack of explanations may modify suspicion of the simulated output, and therefore modify the accuracy of resulting inferences about the agent. However, it is not clear whether explanations increase or decrease suspicion.

II.3 Explanation by Intelligent Agents

II.3.1 *Scope*

The goal of this portion of the review is to identify the state-of-the-art methods and key learnings from XAI to apply to the simulated agent and to identify areas to explore in the research model. More specifically, the intention of this review is to support construct validity for the explanations considered in the experiment, support content validity for the simulation of a system, and maximize face validity of the findings for AI researchers that are focused on algorithm development.

II.3.2 *Background*

While interest in XAI has recently surged, the effort to make AI interpretable and develop models able to explain their output goes at least as far back as expert systems in the 1970's (Abdul et al., 2018). Knowledge-based AI systems have employed argumentation as an extension of Toulmin's model (Moulin et al., 2002). The societal and legal requirements for explanations arises not much differently from how people are expected to explain themselves in court, as explanations act as a central tool for accountability (Doshi-Velez et al., 2017).

The target audience for explanations differs across the literature. Nunes and Jannach (2017) consider a narrow audience of data scientists. Doran et al. (2017) defined "interpretable"

as the user being able to understand the mathematical connections between inputs and outputs, which limits the potential audience and roles of explainees in many contexts. Other authors position explanations as being understandable by a broad audience with potentially no understanding of the function of the model either explicitly through their definition (Chander, Srinivasan, Chelian, Wang, & Uchino, 2018; Gilpin et al., 2018; Miller, 2019; Ribeiro, Singh, & Guestrin, 2016), or by framing understandability such that it does not require knowledge of the model's operation (Adadi & Berrada, 2018; Chander et al., 2018; Doshi-Velez & Kim, 2017; Lipton, 2016). In the case here, explanations are being used by individuals that certainly are not experts in the algorithm, and may not have proficiency in the damage assessment task.

Many current XAI approaches have not considered the utility of the explanations to humans and whether they are usable or practical in real-world situations (Abdul et al., 2018). The goals of the explainer are most often different from the explainee (Miller, 2019), creating potential mismatches in communication. Explainability supports predictability, which is key to human-agent teamwork (Ahrndt, Fähndrich, & Albayrak, 2016), however this addresses just one of the ten challenges in teamwork identified by Klein, Woods, Bradshaw, Hoffman, and Feltovich (2004). Interactive XAI, such as a dialogue of explanation, might overcome many of the remaining challenges, but this remains a technical challenge yet to be practically solved (Weld & Bansal, 2018), and relies on an elaborate process to unfold reasoning.

II.3.3 *Composition of Explanations*

Any given event has a very large number of possible explanations going back in time through a causal chain. At the greatest extreme of detail are “accounts” which are narratives developed through a process of event comprehension designed to support claims across many preceding events and contributing causes (McLaughlin, Cody, & Read, 2013). Even if

comprehensive accounts could be algorithmically generated, the cognitive burden of complete explanations is too great for most decisions (Miller, 2019). Hesslow (1988) identified a series of traits of causes which made them preferable for inclusion in an explanation. Of these, unexpected conditions, abnormal conditions, precipitating causes, variable conditions, predictive value, and deviations from the ideal are most relevant and machine-detectable in images.

Research has explored how people will expect an AI system to explain itself. When multiple people are asked to explain a shared event, they generally provide very similar and brief explanations, suggesting that there are systematic mechanisms people use to develop and select explanations (Malle, 2006). While explanations may be possible across many conceptual levels of causation, Miller (2019) claims that inferring the why-question that is provoking the need for an explanation produces the greatest relevance in explanations to the user. Gilpin et al. (2018) advocates that systems should account for trade-offs between completeness and interpretability, and err on the side of providing more detail even if it makes the explanation less understandable.

Counterfactual explanations highlight the minimum conditions which would have changed the classification (Wachter et al., 2017). This is closely related to contrasting explanations, which provide information about the features that differentiate the actual cause history from the outcome that did not happen (Van Bouwel & Weber, 2002). Counterfactual explanations may appear to be only a small semantic shift from simple causal explanations, but their framing, in contrast, has been found to make them more memorable and persuasive even when they contain the same information (Roese, 1997). When humans generate counterfactual explanations we employ the simulation heuristic (Kahneman & Tversky, 1981) to imagine how small changes in circumstances would have altered the outcome. Within the Wachter et al. (2017) formulation this is a computable proposition where the contrasting outcome is the

smallest possible change to move to another category, while Miller (2019) considers that the best counterfactual explanation addresses a contrasting categorization relevant to the user, which is generally non-explicit and not necessarily the minimum change.

Since intelligent agents do not have generalized understanding and can be easily fooled (Akhtar & Mian, 2018), the ability for an agent to detect and explain conditions where its model has not been trained or is known to fail adds factual explanatory value and can clarify why a classification is incorrect (Hoffman et al., 2018). To the extent that good explanations should be persuasive, contingencies and hedges on explanations can be perceived as lack of conviction even in scientific contexts (Hyland, 1996a). In AI argumentation a persuasive argument is considered one which defends against other competing arguments (Moulin et al., 2002), but when the purpose of the explanation is truth instead of changing someone else's position this can be counterproductive. The result is a tension between persuasive explanations and ones that are useful to detect an erroneous classification. A classification of hedges in academic writing is provided by (Hyland, 1996b), where the author claims that the purposes of hedges are "fuzzy" and that the true intent of a hedge must be inferred by the reader as they are rarely explicit.

II.3.4 *Explanation Engines in Deep Learning*

Many of the visual explanation engines were inspired or derived from image description approaches. An example adapted system was proposed by Nushi, Kamar, and Horvitz (2018). In this multi-step process a visual detector identifies objects in the image which are fed as a list into a system-generated observation engine. That engine uses a human-tuned language model to combine the terms identified; however, that engine is not able to interpret the image itself. Those candidate descriptions are then fed into a caption ranker that has access to the original image to decide which generated description is most appropriate. This approach combines training on how

people would express the relationships between objects without the machine having an explicit understanding of why the terms are combined, and what words are used to join them. Within evaluation methods for explanations, similar methods have been used to leverage humans to score automatically generated phrases for relevance to images, such as the influential scoring mechanism CIDEr proposed by Vedantam, Lawrence Zitnick, and Parikh (2015).

As an expansion on this concept, Antol et al. (2015) proposed “visual question answering” where users could ask a computer vision model questions about an image and responses would be produced tailored to the question. Due to the errors in the output of these types of systems, Ghosh, Burachas, Ray, and Ziskind (2018) and Park et al. (2018) separately explored generating explanations of the answers by combining natural language with highlighting areas of images relevant to classifications. Daniels and Metaxas (2018) proposed a context-sensitive method of classifying the content of a scene through the relationships of objects that occur together, which generated higher-accuracy descriptions of unknown images, but without producing fully natural language descriptions. Hendricks et al. (2018) developed a method for labeling images with visual explanations that detected objects within the image and explained how classifications were made, with both direct causal and counterfactual explanations generated by a neural network and selected for appropriateness by an algorithmic “phrase-critic,” shown in Figure 4.

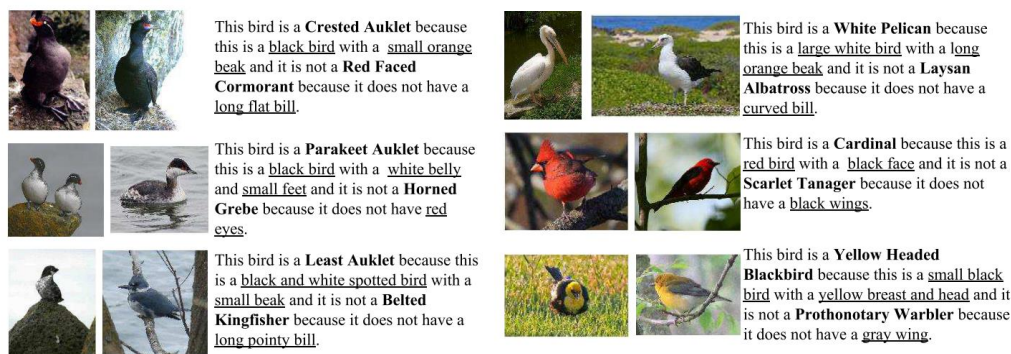


Figure 4 Example Algorithmic Counterfactual Explanations

Note. Reprinted from Figure 5 (page 12) of Hendricks et al. (2018). Copyright 2018 by Springer Nature. Reprinted with permission.

II.3.5 *Experimental Manipulation of Explanation*

Among the earliest experiments with explanations in intelligent systems were those conducted with expert and decision support systems. Suermondt and Cooper (1993) examined explanations in a decision support system in a medical context, finding that explanations improved diagnostic accuracy. Rook and Donnell (1993) compared two means of presenting information to justify a decision. That study identified that cognitive limitations in interpreting output was critical in the design of explanations. Template explanations of procedures were examined in Gregor (2001), which found relationships between their use decreasing cognitive effort and increasing usage of explanations in cooperative tasks. Rose (2005) identified that when accounting students were provided a decision-aid system to assist with decision making regarding tax-related rules, students with low cognitive loading learned the rules similarly to those without the system, but those with high load knew less about the rules than if they did not have the system. These early cases of research identified performance increases and the foundations of many of the challenges in explanation that remain today.

The recent increase in interest in explanation has resulted in reviews collecting the past findings across intelligent system technologies. The review by Nunes and Jannach (2017) was specific to experiments with explanation and analyzed 217 studies of explanation in decision support and recommender systems. They found that the performance benefits identified by early studies did not hold over time, and that the persuasiveness of explanations was a potential confound to understand the presented output. Abdul et al. (2018) analyzed the connections and isolation of areas of research and Adadi and Berrada (2018) surveyed the body of XAI research,

and both identified a lack of articles which focused on the human factor. The review of the interpretability of explanations by Doshi-Velez and Kim (2017) identified that experiments with humans have preferred simplified tasks over realistic settings to focus on algorithm and explanation performance (Doshi-Velez & Kim, 2017).

Some recent studies have focused on the subjective quality of explanations. Narayanan et al. (2018) manipulated the length and complexity of explanations and measured a subjective score from users on an artificial task, finding that long explanations were less preferred, though not as sensitive to the number of concepts in an explanation. The study by van der Waa, van Diggelen, van den Bosch, and Neerincx (2018) requested that users evaluate multiple explanations, and asked them to measure them in dimensions of preference for long or short, strategy versus policy, and sufficiency of information, with the highest preference for explanations related to policy and greater information. In the study by Lakkaraju, Bach, and Leskovec (2016), humans were employed to evaluate the interpretability of machine-generated explanation in logic form (rather than natural language) and to recreate the explanation in natural language. The responses were evaluated qualitatively by two judges, finding widely much better performance with interpretable decision sets which applied rules independently than Bayesian decision lists, which required interpreting an explanation sequentially.

Several studies evaluated compliance with system decisions based on providing an explanation. Gedikli, Jannach, and Ge (2014) examined explanations within explanation systems, identifying transparency as a key antecedent of trust and compliance. Arnold, Clark, Collier, Leech, and Sutton (2006) looked at definition, justification, rule-trace, and strategic justifications with the performance metric of acceptance of the recommendation, finding that explanations did increase compliance with differences in the types of explanations that were persuasive to novices

as compared to experts. In a classification task for microblogs, the explanation, control, and metadata available to the user were manipulated with the primary finding that system recommendations should be provided ahead of search results and tools to achieve compliance (Schaffer et al., 2015). In a robotic experiment Nikolaidis, Kwon, Forlizzi, and Srinivasa (2018) found an increase in adaptation to the robot's intentions when an explanation was provided, but with decreasing trust, and belief that the robot was not being truthful.

Among more interactive uses of explanation, three studies were identified. Soundness of recommendations and completeness of explanations in recommendations was examined by Kulesza et al. (2013) in a music preference context. They tested the impact of completeness and soundness of explanations on user's mental models, recommending that both completeness and soundness are required, but only if users believe their input is improving the intelligent agent. In Sklar and Azhar (2018) the researchers experimented with coordination of a robot that could provide an explanation in dialogue through argumentation. They found that subjective preference and objective performance criteria did not vary significantly between explanation and black box conditions. Holliday, Wilson, and Stumpf (2013) compared the effect of users being provided an explanation that they could then correct to a system that did not provide explanations. They found that participants in the condition where they were not providing explanations exhibited more control-exerting behaviors and reported such while "thinking aloud," but there were no perceptual differences in control when the same concept was tested in a questionnaire. They deduced that there were differences between perception and behavior related to perception of control.

Several experiments introduced intentional errors to evaluate conditions where users accepted erroneous output and the effects on other measures. Ribeiro et al. (2016) manipulated

the quality of the model being explained in human experiments of trust between unexplained and explained classifications, and found that explanations of erroneous performance decreased trust but increased accuracy of user perceptions about the quality of the classification model.

Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, and Wallach (2018) examined the effect of modifying the information interpretable from a linear regression model which predicted apartment prices in New York. They also introduced mistakes and measured trust. They found that the timing of offering explanations and the design of optimal interpretability was not intuitive to designers of the interface, and that different levels of local explanation did not produce statistically significant differences in perception. In the medical context a mixed-methods analysis was conducted with an experiment involving a simulated clinical decision support system. Half of the eight scenarios provided to the subjects had an incorrect recommendation. They found that that confidence statements did not sway trust but that “comprehensive why” explanations induced over-reliance, and “selective why” explanations led to better confidence in the user’s decisions (Bussone et al., 2015).

The closest experiment identified to the design of this study was conducted by Tan, Tan, and Teo (2012). They evaluated how differences in explanation impacted cognitive measures by manipulating explanations into forms of trace, justification, and strategic explanations. They measured perceived confidence in the decision and cognitive capacity (preferences on how much to think) in a consumer context with types of recommendations. No single configuration of explanations produced an optimal outcome and they instead identified trade-offs in decision quality, confidence, and speed. For instance, conditions with explanations based on a table of metrics and benchmark comparisons resulted in the fastest decision times and highest confidence, but the quality of decision was lower than that for other conditions of the

experiment. They proposed combining multiple explanation approaches into the same interaction as the solution; however, such a configuration was not experimentally evaluated.

II.3.6 *Open Issues*

There has been relatively little study done on what explanations should contain to balance persuasive power and use of logic. The most logically rigorous models consider the ability for explanations to be able to defend claims against those of competing explanations. Models based on more subjective criteria do not reject the value of compelling logic, but instead prioritize human outcomes of interpreting explanations. This is not without risk as well, as persuasive explanations may be more accepted based on subjective criteria than complete, accurate, and transparent explanations (Gilpin et al., 2018). The review by Nunes and Jannach (2017) identified open concerns over the lack of clarity in the differences between stakeholder goals, user goals, and the purposes of explanation, challenges in selecting the best content for explanations, user interface options for revealing explanations, responsive explanations that were adjusted to the needs and requirements of users, and objective evaluation protocols and metrics which do not rely on stated behavioral intentions or subjective assessment of the explanations. Additionally, while the explanations generated for images have been tested in terms of preference and trust, they have not been evaluated in terms of to what extent users process them beyond those attitudes. Experiments have found widely varying results with the most consistent “successful” outcomes based on achieving trust and compliance. Systems that introduced errors identified that compliance and trust were not always the ideal outcomes for explainable systems to achieve partnership with the human user.

III RESEARCH MODEL AND HYPOTHESIS DEVELOPMENT

III.1 Research Model

This chapter develops a research model to address the research questions by hypothesizing relationships between theoretical constructs identified from the literature that represent the concepts in the conceptual framework.

Self-efficacy in the task was selected as the ultimate dependent variable due to its predictive power for future engagement in the task, investment of mental effort, and task performance and satisfaction. Cognitive load was selected as a known mediator of self-efficacy and was expected to be impacted by adding explanations. Attribution of agent intelligence was selected as explanations should directly influence the perception of mental processes in the agent, and may be affected by cognitive load. The model focuses on the effects of adding counterfactual and hedging explanations. These selected theoretical constructs and their definitions are listed in Table 3. The developed research model appears in Figure 5.

Some prior research which used social cognitive theories has employed repeated measure designs to study within-subject effects in the development of attitudes (e.g. Compeau & Higgins, 1995) and to monitor feedback loops in cognitive processes (e.g. Leppink & van Merriënboer, 2015). However, no prior between-subjects experiments which utilized repeated measures were identified that tested the effect of pre-treatment or repeated measurements on outcomes. To avoid the effects of measurement on the experiment outcome and maximize the realism of the task for participants, the research model was developed using only post-test measurement of the constructs. This approach is similar to the sequence in the experiment by Joseph (2013), and appropriate for a between-subjects test of the effect of explanation types. While a repeated measure design would decrease the group size required for each condition (Leppink & van

Merriënboer, 2015), the complexity of the experiment in terms of number of conditions and their analysis would increase to establish the effects of measurement (Solomon, 1949).

Alternate explanations for self-efficacy outcomes were selected from the literature to understand their relative ability to explain the experiment results. Previous task experience was selected as it has potential interactions with each of the dependent variables. Trust, perceived interdependence, and perceived level of automation have been selected due to their common use across the human-computer interaction literature and will allow a comparison of their ability to explain outcomes of the experiment to the study's theoretical constructs.

In the task and survey instrument the simulated agent was referred to as “Automated Damage Assessment Machine” with the acronym “ADAM.”

Table 3 Theoretical Constructs

Construct	Definition
Causal Explanation	“A line of reasoning that explains the decision-making process of a model using human-understandable features of the input data.” (Doran et al., 2017)
Counterfactual Explanation	An explanation “crafted in such a way as to provide a minimal amount of information capable of altering a decision, and they do not require the [lay person] to understand any of the internal logic of a model in order to make use of it.” (Wachter et al., 2017)
Hedging Explanation	An explanation which provides “specific information about what the system cannot do or perform well” such as boundary conditions and failure modes (Hoffman et al., 2018).

Cognitive Load	The extent to which short-term working memory resources are utilized in the task (Sweller, 1988).
Intrinsic Cognitive Load (ICL)	The inherent and unalterable required mental effort to complete the task (Sweller et al., 1998).
Germane Cognitive Load (GCL)	Mental effort expended to construct schemas and understanding of the task (Sweller et al., 1998).
Extraneous Cognitive Load (ECL)	Mental effort which does not contribute toward the task or the learning of the task (Sweller et al., 1998).
Attribution of Agent Intelligence	“Attributing abstract causal mental states are not only the reduction in the cognitive complexity required to understand another’s behavior but also the prediction of the other’s future behavior.” (Terada & Yamada, 2017)
Self-Efficacy in Task	“Belief in one’s agentic capabilities, that one can produce given levels of attainment” (Bandura, 1997) here specific to the task of assessing structure and damage classifications together with the simulated agent.

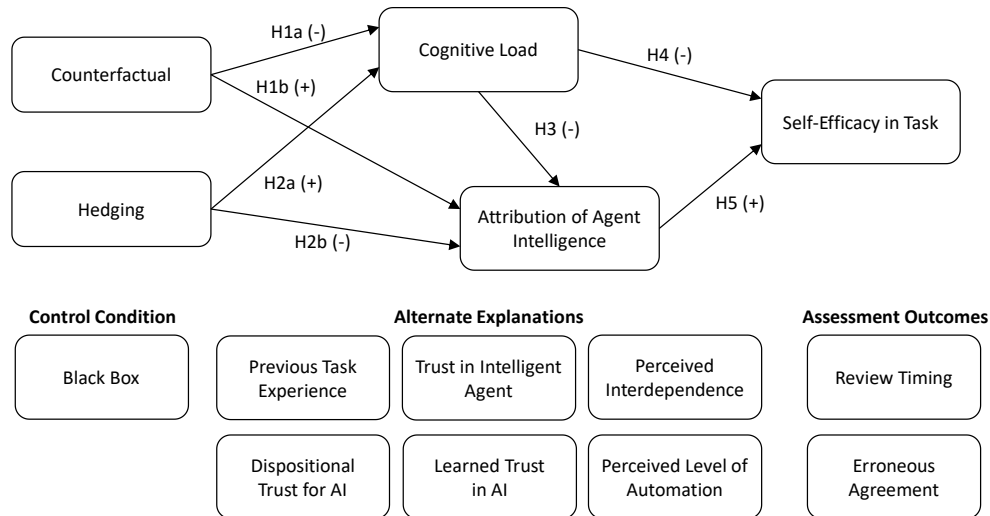


Figure 5 Research Model

III.2 Independent Variables

III.2.1 Causal Explanation

Explanations of the mathematical causes of a model's output are the most direct variety of causal explanation, but in the case of deep learning algorithms, such explanations are unlikely to be brief or useful. In comparison, explanations which contain human-understandable statements related to objects and features of the image, such as those from Hendricks et al. (2018), are useful and particularly relevant when aligned to the task guideline. In this study, the goal for the causal explanation is to provide what Doran et al. (2017) describes as "a line of reasoning that explains the decision-making process of a model using human-understandable features of the input data." As such, causal explanations were provided using human-observable attributes of the post-damage image related to the guideline. A causal explanation was included in all experiment conditions except for the "black box" control condition.

III.2.2 Counterfactual Explanation

Counterfactual explanations provide the minimal differences capable of altering a decision, and they do not require knowledge of how the agent operates (Wachter et al., 2017). They are intended to facilitate counterfactual thought by feeding the simulation heuristic (Kahneman & Tversky, 1981), and allowing the worker to form better causal connections between the model's claims and the data (Keil, 2006). The counterfactual explanations developed here are similar to the causal explanations in that they cite human-observable attributes; however, for counterfactual explanations the attributes are cited as missing, thereby making a specified alternate classification inappropriate.

III.2.3 Hedging Explanation

Explanations of why the simulated output may be incorrect are considered "hedging explanations." Unlike the causal and counterfactual explanations, an example system to model hedging explanations was not identified in the literature. The call for such explanations were used to inform the potential content. Thelisson, Padh, and Celis (2017) called for safe-guards to be designed into AI systems to expose the causes behind decisions in order to reveal potential biases and discrimination. In this context, that can be interpreted as exposing limitations in damage detection associated with socioeconomic and demographic factors. Bussone et al. (2015) called for systems to be able to explain when their results would not apply after finding that medical professionals agreed with erroneous automated diagnoses. This is comparable with the call from Hoffman et al. (2018) for explanations that help a user apply and measure trust situationally.

In formal writing it is expected to hedge claims leaving them open to being wrong, with the degree of hedging providing information on the strength of a claim (Hyland, 1996b). In the

Toulmin (1958) model of argumentation there is a provision for arguments to contain an element joined to a claim using the term “unless,” which provides a reason why an argued claim might not be true, and which is tied to something specific about the support for the claim (“a rebuttal”). The hedging explanations of this study follow this mode, referencing human-observable attributes related to a hypothetical failure mode relevant to the image. Failure modes were identified from the computer vision literature. Each is presented using the advisory phraseology “consider,” avoiding conveying confidence in a hedge.

III.3 Dependent Variables

III.3.1 *Self-Efficacy*

The seven items of the Self-Efficacy for Learning and Performance scale (Lodewyk & Winne, 2005) were adapted to the task of this study, with “will” language modified to “can,” consistent with measure development guidance from Bandura (2006). These items were rated on a 7-point Likert-Type scale from “Strongly Disagree” to “Strongly Agree” with only the ends of the scale labeled. The adapted item “I’m confident I understand how to use the damage assessment guide” was dropped after testing due to cross-loading with other concepts, reducing the scale to six items. Two additional items were developed and considered for addition: “I believe I won't overly rely on ADAM,” and “I'm confident that ADAM won't distract me;” however, these items measured reliance rather than self-efficacy. A third item, “The automated assessment prevents errors in manual damage assessments” was adapted from Singh, Molloy, and Parasuraman (1993) to evaluate a potential perceived reliance measure, but these three items did not perform well. Reliability for the six adopted items for self-efficacy was measured ($\alpha = .952$).

III.3.2 Cognitive Load

To evaluate the three theoretical types of cognitive load, the eight-item scale from Klepsch, Schmitz, and Seufert (2017) was selected and adapted to the damage assessment task. The naïve methodology was used without briefing respondents on cognitive load theory beyond the wording of the items. A ninth item was developed to balance the number of items for intrinsic cognitive load compared to the other two types, intended to be reflective with the other two: “I had to remember many things to perform the task.” Items were rated on a 7-point Likert-Type scale from “Strongly Disagree” to “Strongly Agree” with only the ends of the scale labeled. Reliability was measured for the three sub-components separately (Intrinsic $\alpha = .733$, Germane $\alpha = .676$, Extraneous $\alpha = .733$). The Paas (1992) single-item total cognitive load measure has been well-validated across many studies (Paas, Tuovinen, Tabbers, & Van Gerven, 2003) and has been shown to correlate with physiological measures (Joseph, 2013). The single item from Paas (1992) was adapted to the study and presented immediately after the completion of the ten scenarios. The item asked participants: “How much mental effort did you invest in making your assessment?” Response was rated on a 9-point scale with ends “Very, Very Low” to “Very, Very High” and intermediate scale points labeled as in the original scale. While both approaches were employed in the instrument, validity could not be assured, and the three-component model was utilized for the structural model.

III.3.3 Attribution of Agent Intelligence

Six items were adopted from Terada and Yamada (2017). The items were: thinks logically, is knowledgeable, is able to make decisions, is predictable, is insightful, has a mind. The first five items loaded on the same extracted factor (personal intelligence) and the sixth loaded on a separate factor (social intelligence). Items were rated on a 7-point Likert-Type scale

from “Strongly Disagree” to “Strongly Agree” with only the ends of the scale labeled. Reliability was measured across all of the items in this scale ($\alpha = .819$).

III.4 Alternate Explanations

III.4.1 Previous Task Experience

Previous research in attribution and self-efficacy has found a ceiling effect with high levels of pre-existing self-efficacy precluding an effect of attributions (Stajkovic & Sommer, 2000). Gist and Mitchell (1992) evaluated the development of self-efficacy across different levels of task experience with a novel task. They found that with low experience self-efficacy is informed by multiple cues in an assessment of requirements, resources, and constraints of the task. In comparison, those with high task experience form self-efficacy based on past performance and motivation rather than aspects of the task. High task experience, particularly experience with the specific damage guideline used in this study, would also greatly decrease the cognitive load. To account for this potential, participants were given three items to assess their previous experience in aerial image damage assessment: “How would you rate your experience in damage assessment of aerial images prior to this study?” with 7-point response options “None” (coded as 0) and “Very limited” (1) to “Highly Experienced” (6); “How often do you rate damage to structures in aerial images?” with 7-point response options from “Never” (coded as 0) and “Yearly, or less” (1) to “Daily” (6); and “Have you previously used a damage guideline to make ratings?” with three response options: “No” (coded as 0), “Yes – But not the same one” (3) and “Yes – The same guideline as here” (6). If “None” was answered to the first item, the remaining two items were not presented. The three components were added into an index value for the measure.

III.4.2 *Trust in the Intelligent Agent*

Trust has been the primary measure of participant attitudes in human-automation research including XAI. Since the simulated agent provides erroneous classifications in half of the scenarios, lack of trust is an appropriate response. However, even if the worker does not fully trust the simulated agent, they may have greater trust in agents that are better able to explain erroneous performance. To measure trust for comparison against previous studies a single item from Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) (“I believe I can trust the automated assessment”) was adapted here and rated on a 7-point Likert-Type scale from “Strongly Disagree” to “Strongly Agree” with only the ends of the scale labeled.

III.4.3 *Dispositional and Learned Trust for AI*

The three-level model of information systems and technology trust proposed by Marsh and Dibben (2003) defines three types of trust: “learned,” based on experience with similar systems; “situational,” where dispositions are adjusted by cues from the environment; and “dispositional” where attitudes are based on pre-existing attitudes about technology. Though participants may have previous experience in the damage assessment task, they are unlikely to have had experience making assessments with an intelligent agent. The three-level model predicts that trust in the intelligent agent will be determined by pre-existing and stable dispositions towards automation and AI. To evaluate these pre-existing attitudes, dispositional trust for automation was measured using three items adapted from Nees (2016). These items were rated on a 7-point Likert-Type scale from “Strongly Disagree” to “Strongly Agree” with only the ends of the scale labeled. For participants that have experience working with AI, the inclination to trust the intelligent agent may be better informed by learned trust (or distrust) attained from working with other AI systems. Participants rated learned trust for AI using a 5-

point scale of how well AI had met their expectations in their profession with “Far Short of My Expectations” and “Far Exceeded My Expectations” as the ends of the scale, with an option for not having used AI in their work.

III.4.4 *Perceived Interdependence*

The final rating in this task is fully dependent on the human’s judgment, and the human must actively accept the input by changing their selections after reviewing the agent’s output. As such, the task is not truly interdependent. However, explanations increase observability and predictability of the agent, which joint activity theory predicts improves performance (Johnson et al., 2012). Providing the cause of a disagreeing classification can also produce a greater understanding of the guideline, creating some level of “directability” by the agent. To assess attitudes about task interdependence, three items from Morgeson and Humphrey (2006) were adapted: “My ratings were affected by ADAM's input,” “Assessments depend on the both the human and ADAM for accuracy,” and “My ratings benefitted by working with ADAM.” Items were rated on a 7-point Likert-Type scale from “Strongly Disagree” to “Strongly Agree” with only the ends of the scale labeled.

III.4.5 *Perceived Level of Automation*

Participants that perceive a high degree of system automation are likely to rely on the agent’s assessment in erroneous cases without independently assessing the scenario themselves. Greater detail in explanation may be perceived as lower automation by creating a burden to the worker to evaluate the agent. One item was included to measure the participant’s perception of the level of automation of the microtask. Participants rated a single item “The damage assessment was” on a 7-point Likert-Type scale from “Highly Manual” to “Highly Automated” with only the ends of the scale labeled.

III.5 Scenario Measures

III.5.1 *Initial and Review Timing*

The average amount of time (in seconds) workers took to provide their ratings was measured for each step separately. The initial step time was measured between the presentation of the images and the submission of the initial rating. The review step time was measured between presentation of the simulated output (and any explanations) and the submission of any changes in the review step. The average of the ten scenarios was computed for each participant to match the unit of analysis of the other measures.

III.5.2 *Erroneous Agreement*

Some prior research has found that producing persuasive explanations causes users to accept the system's recommendation (Arnold et al., 2006; Cramer et al., 2008; Glass, McGuinness, & Wolverton, 2008). In this context there should be no presumption that the agent has the correct answer, and it is possible to be overly persuasive. Using acceptance of the system output as the performance criteria might be a measure of the absence of human cognition. While objectively correct ratings may be impossible to identify, it is possible for the agent's rating to be highly inconsistent with the image. Erroneous output, which is coherent with respect to the guidelines, but references objects that are clearly not in the image, provide an opportunity to detect overly persuasive explanations.

Five of the scenarios presented to workers were designed and selected for inclusion based on the implausibility of the objects cited in the explanation. Erroneous agreement was measured as the proportion of erroneous scenarios where the participant changed their rating to agree with the agent's damage rating in the review step. Participants that agreed with the intended erroneous

rating in their initial review (5.3% of ratings) were not counted as being in erroneous agreement, as the simulated output and explanation likely had little to no role in the worker's rating.

III.6 Manipulation Checks

Three measures were included as a check to ensure manipulation of the independent variables. The measures asked participants to rate their agreement with three statements: "The Automated Damage Assessment Machine (ADAM) explained why it made its ratings," "ADAM compared its rating to at least one other possible classification," and "ADAM pointed out features in the image that may lead to incorrect classifications." Each was rated as "Yes," "No," or "Don't Know / Don't Remember." For detail on the development and testing of the manipulation check, see Appendix F.2.7.

III.7 Demographics and Feedback

Demographics (age, gender, income, education) were requested. An optional open-text feedback item was included.

III.8 Hypothesis Development

III.8.1 *Effect of Counterfactual Explanations*

Explanations with a comparative and contrasting causal explanation will require less cognitive resources, potentially avoid referrals back to the guidelines, and provide direct contrasts to other possible classifications (Hilton, 2007; Wachter et al., 2017). These explanations will both connect to the facts of the image (which can be falsifiable), and inform the worker of the schema classifications to fill in holes and errors in their own understanding reducing the perception of information being extraneous (Bandura, 1986). Differences in self-rated cognitive load were detectable between different types of puzzles in Joseph (2013), and users that have to spend more mental effort evaluating and reviewing damage guidelines will

have higher load. The interpretation of counterfactual explanations is expected to support the participant's construction of a mental schema for the task, resulting in increased germane cognitive load in the short-run, with decreased total cognitive load as they gain experience in the task (Sweller, 1988). As these explanations are also composed in the manner in which humans generally explain themselves (Malle, 2006), counterfactual explanations are also expected to make it more likely that the agent is perceived as being predictable in terms of having a rational mind and being intelligent (Dennett, 1989).

Hypothesis 1a: Counterfactual explanations will decrease cognitive load.

Hypothesis 1b: Counterfactual explanations will increase attribution of agent intelligence.

III.8.2 Effect of Hedging Explanations

Hedging explanations provide important information on known failure modes and can highlight potential flaws and challenges in models that are not obvious to humans (Hoffman et al., 2018). At the same time, these failure modes cannot be positively detected by the model and therefore are not certain assertions, but instead notes of caution based on presumptive detection of boundary conditions which do not necessarily invalidate results. When the classifications and explanations are coherent with the image and guidelines, the presence of these explanatory warnings is nearly by definition extraneous cognitive load. They are also likely to be seen as declaring weaknesses which our biases treat as lack of competence and intelligence (Moulin et al., 2002). In conflict with this, some evaluation frameworks for explanations cite the importance of persuasiveness of explanations to human understanding (Gilpin et al., 2018; Tintarev & Masthoff, 2011). Hedging explanations are speculative in that they are not detectable with certainty, and as such are likely to increase cognitive load with “seductive details” which may be misleading and require more effort to evaluate, also consistent with extraneous cognitive load

(Klepsch et al., 2017). While framed in terms of the content of the image, hedging explanations describe conditions of the agent and are less inherently understandable to participants unfamiliar with the function of computer vision algorithms, even if they are experts in damage assessment. While literature that hasn't specifically examined hedging explanations expects a benefit, the related literature in other contexts indicates hedging explanations will increase cognitive load and be less persuasive by reducing the perception of intelligence and confidence.

Hypothesis 2a: Hedging explanations will increase cognitive load.

Hypothesis 2b: Hedging explanations will decrease attribution of agent intelligence.

III.8.3 Cognitive Load on Attribution of Agent Intelligence

Increasing cognitive load results in people selectively processing information (Sweller, 1988). By reducing information processing resources available, attributions of observed behavior conform to the pre-existing views of the observer (Molden et al., 2006). Increased cognitive load also prevents situational information from entering into the judgment of causal processes (Hilton, 2007). This has been further supported by research which manipulated cognitive load to test for the presence of implicit mind perception, which has found that perception and attention to mental state requires cognitive resources which are shed under cognitive load (Schneider, Lam, Bayliss, & Dux, 2012). The agent in this model is fallible and commits errors; however, the explanations provide a basis by which to attribute intelligence if attributions are made based on the situational causes (such as errors in interpreting a challenging image) as compared to dispositional causes. The study by Gilbert and Osborne (1989) examined recovery from incorrect inferences and found that subjects under high cognitive load made errors in attributions even if misperceptions of the events those attributions were based on were "retroactively cured." As such, while the agent's explanation may cure incorrect inferences, their attributions of the agent's behavior may

remain unchanged. Based on this previous knowledge that cognitive load disrupts the attributional processes of humans in social situations, the following hypothesis is proposed:

Hypothesis 3: Increasing cognitive load will decrease attribution of agent intelligence.

III.8.4 Effect of Cognitive Load on Self-Efficacy

Many previous studies in instruction (Hesketh, 1997; Steele-Johnson et al., 2000) and general technology-mediated task contexts (Crippen & Earl, 2007; P. J.-H. Hu et al., 2017; Zheng et al., 2009) have found relationships between cognitive load and self-efficacy. Across each of these studies, increasing cognitive load was associated with decreasing self-efficacy within their respective tasks.

Hypothesis 4: Greater cognitive load will decrease self-efficacy.

III.8.5 Effect of Attribution of Agent Intelligence on Self-Efficacy

Social cognitive theory anticipates that learning is a social process where we learn primarily by observing others rather than exclusively from our own independent experience (Bandura, 1986), and the development of self-efficacy is tied to observational learning (Bandura, 1997) which should be enhanced by explanations. Intentional systems theory states that when an observed entity appears to behave rationally, we ascribe that behavior to having a mind (Dennett, 1989). To that end, if the worker interprets the behavior of the agent to be rational, the behavior of the agent can be modeled and guide the worker in the performance of the task. To the extent that explanations of model output are analogous to example-based instructional techniques, the review by Van Gog and Rummel (2010) found that those techniques increased self-efficacy, and that effectiveness was tied to the traits of those being modeled. The review by Gist and Mitchell (1992) found that assessments of external resources are antecedents of self-efficacy, by means both automatic and intentional. A second potential mechanism combines instructional attribution

theory with early computer interaction research. Weiner (1985) theorized that stable internal attributions of success are a predictor of positive expectancy and motivation rather than external causes, and Engelbart (1962) found that people subsumed augmentation by technology into their own performance and could only differentiate augmentation through its removal. For a task where the worker is learning the particulars of an assessment guideline and the unique elements of a new disaster, an agent perceived as intelligent is much more likely to be modeled by the participant than one that is not and lead to increased expectations of future performance.

Therefore:

Hypothesis 5: Greater attribution of agent intelligence will increase self-efficacy.

IV RESEARCH METHODOLOGY

IV.1 Experiment Design

An experiment was developed to test the hypotheses of the research model by varying the explanations offered by a simulated agent in a crowdsourced damage assessment task.

Participants in the experiment were provided a briefing on the task, expectations, and rating guidelines. For each scenario participants rated the structure type and damage level, after which they were provided the simulated output and any explanations along with the opportunity to change their initial assessment. After completing a total of ten scenarios, participants provided ratings for the study measures. Only the types of explanations offered were manipulated between experiment conditions. The unit of observation and analysis for the measures of the research model was “participant” and all measurements of attitudes were made after the completion of the task.

An incomplete factorial between-subjects design for type of explanation was employed with five conditions, shown in Figure 6. One condition included only a causal explanation. The causal explanation was combined with the counterfactual and hedging explanation separately in two conditions. One condition included all three explanation types. In addition, a “black box” condition with no explanation of the simulated output was included. The probability of condition selection per participant was based on the proportion of remaining quota per condition with the goal of equal group size per condition.

		Causal Explanation		
		No	Yes	
		Counterfactual Explanation		
		No	Yes	
Hedging	No	"Black Box" Control	Causal Only	Causal + Counterfactual
	Yes	n/a	Causal + Hedging	All Three Explanations

Figure 6 Experiment Conditions

The design of the experiment attempted to ensure that differences in the formation of self-efficacy would be based solely on the explanations. Of the four pathways for developing self-efficacy through observational learning, the research focuses on “observation” (vicarious experience). By asking workers to review explanations and controlling for other pathways, the effect of the explanations on self-efficacy can be assessed. Of the other pathways, “successful performance” can only be self-assessed or inferred from rating agreement with the agent. The agent did not attempt to use “persuasion” to convince the participant about their capability to conduct a task, and any physiological or affective processes that take place outside of the experiment would be neutralized by random selection into conditions.

The study materials were submitted to the Georgia State University Institutional Review Board and approved (see Appendix A). Pre-testing was performed prior to the study to evaluate the measures, selection of scenarios, and manipulation of the independent variables by the experiment.

IV.2 Selection of Participants

Participants were recruited using the Amazon Mechanical Turk crowdsourcing platform and qualified using a separate two-question survey. The requirements for participation were having proficiency in the English language, having at least 100 approved tasks and 95% approval rating on the platform, and being an adult within the United States. The qualification survey asked participants to submit a short description of an aerial photo of an undamaged building that was not otherwise part of the study using an open-text field in order to evaluate their proficiency in English. The survey also asked them to select checkboxes if they had previous experience in crowdsourced citizen science, aerial image interpretation, and related fields. The qualification task was not compensated. Participants in any phase could be excluded from participation in later phases of data collection. Participants in the telephone interview pre-test were compensated \$10.00, and all other participants were compensated \$2.00.

The requirement to demonstrate proficiency in English was beneficial to ensure that participants would be able to understand the guideline and natural language explanations. The example used was a pre-disaster image with several easily identified features, shown in Figure 7. The qualification step also addressed the identified phenomenon of non-English proficient Amazon Mechanical Turk workers using technical means to evade the platform's United States location requirement ("farmers") who have been found to produce unintelligible or "clunky" open-text submissions (Moss & Litman, 2018).

No qualification submissions were rejected based on spelling or grammar. Errors in the interpretation such as using an incorrect shape name to describe the building or not understanding the scale of the building were not considered in the assessment of English proficiency. Examples of submissions appear in Table 4. The consistently interpretable and

efficiently constructed descriptions of the submissions indicated the participants were both proficient in English and inconsistent with the farmers identified by Moss and Litman (2018).

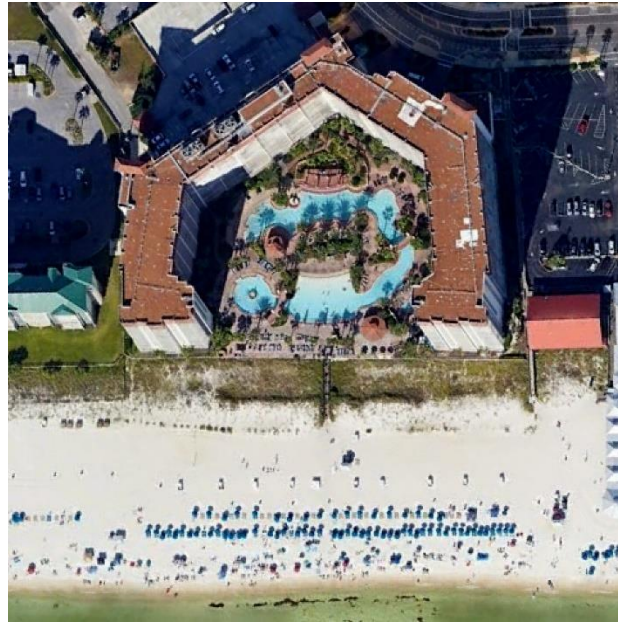


Figure 7 Image Provided for the English Proficiency Qualification

Note: Map data from Google (2018)

Table 4 English Proficiency Qualification Examples

Judgment	Submitted Description
Qualified	“This photo shows a good deal of beach erosion, which is caused by ocean waves, currents, and high winds. The energy within the water, pulls sand away from the shore, carrying it elsewhere and depositing the sediment in sandbars.”
Qualified	“This is an overhead view of a beachside mansion. You can see the beach below, which has dozens of people enjoying themselves. There may be some erosion of some of the green right outside right at the edge of the mansion grounds, likely where the high tide comes in.”

IV.3 Experiment Process Flow

The process flow for the experiment appears in Figure 8. The qualification process was employed to ensure that only participants that met the qualification requirements were recruited for the study. An attention check was utilized to validate adequate participation by participants following the briefing. Failure to select the correct answer ended the survey and excluded the participant from the study. The correct option “You may change your answer after reviewing the

automated assessment” was essential to the design of the experiment, and the remaining options included elements not mentioned or contradicted by the briefing.

The process flow of the experiment was modified for the testing rounds to support the development of the instrument. The following changes only applied during the testing: participants were asked in the initial step of each scenario after rating the image to “Please rate the difficulty of this classification” on a 7-point semantic differential with scale ends “Very Difficult” to “Very Easy.” The manipulation checks were performed immediately after the scenarios to maximize recall to support evaluating the manipulation. The failure of the attention check did not end the survey.

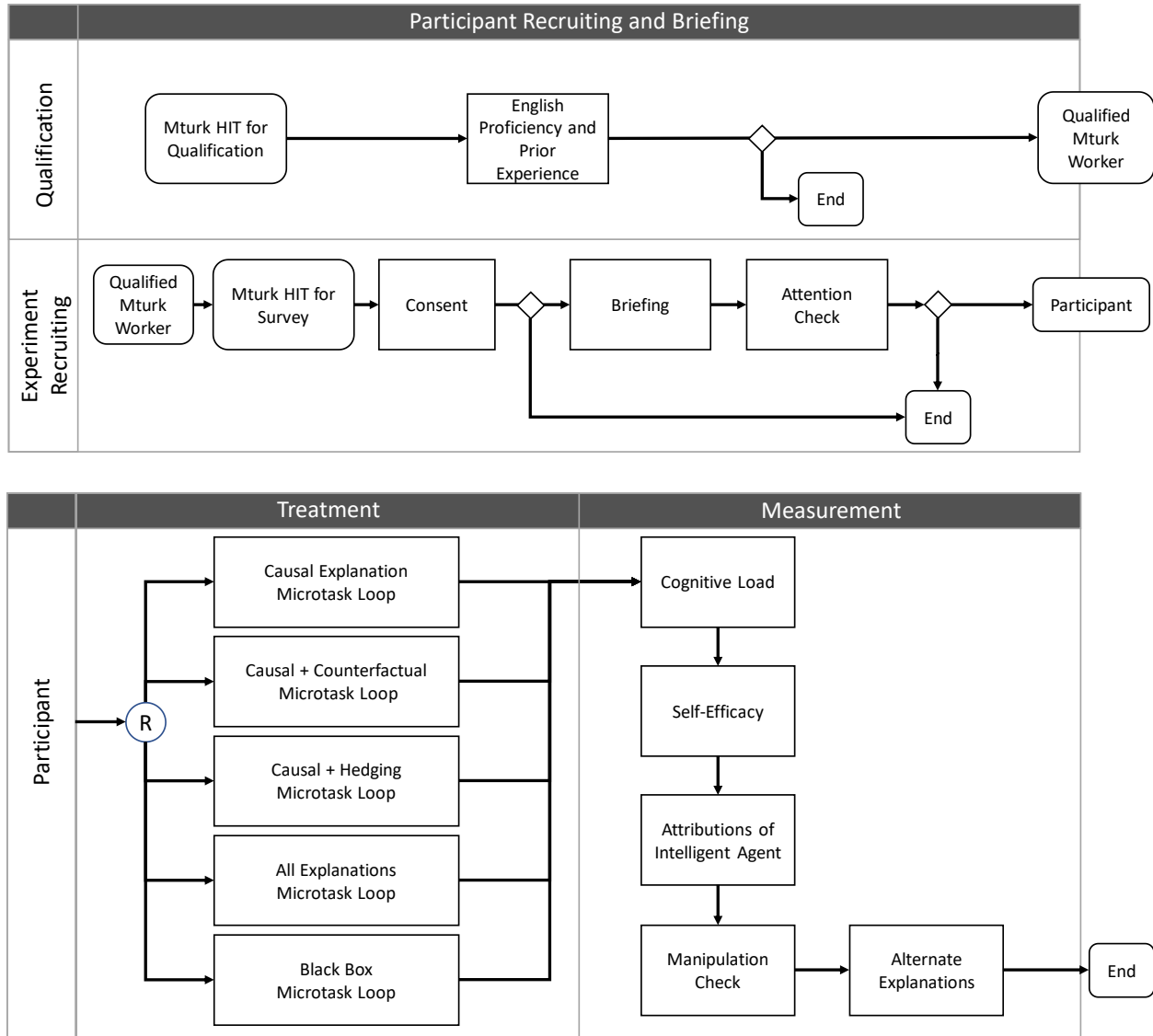


Figure 8 Experiment Process Flow

IV.4 Task Design


The findings of previous research and feedback from participants was referenced when designing the damage assessment task. The goal was to reduce any extraneous cognitive load induced by the assessment tool and environment. An example rating screen of the final instrument appears in Figure 9.

Step 1: Damage Assessment #1 of 10


Rate the structure and post-disaster damage for the centered building within the blue box.

Hover over the image to use the magnifying glass.

Pre-Disaster







Post-Disaster



[Click here to open the damage guideline.](#) or hover over the option names.

Please rate the type of structure:

No Structure  <input type="radio"/>	Light  <input type="radio"/>	Medium  <input type="radio"/>	Heavy  <input type="radio"/>
--	---	--	---

Please rate the damage level:

No Damage <input type="radio"/>	Minimal <input type="radio"/>	Significant <input type="radio"/>	Critical <input type="radio"/>
--	---	---	---

Figure 9 Scenario Interface Example

While the original intention was to recreate the “pybossa” user interface commonly used by other crowdsourcing disaster assessments, this added additional non-functional user interface elements and potentially increased the cognitive load of the task. Instead, a fully native Qualtrics-based interface was developed consistent with recommendations in the literature that minimized extraneous cognitive load, reduced situational elements of task difficulty, and supported the participant’s engagement with the rating task.

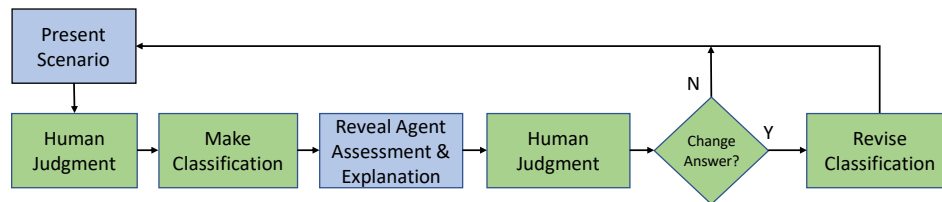
The participant rated the image before being provided the simulated output and explanation similar to a “critiquing system” per the recommendations of human-centered computing (Smith, 2018), and similar to the collaborative human-machine problem-solving approach from Tianfield and Wang (2004). Pre-centered images and a simple classification entry method were utilized per Kerle and Hoffman (2013) to minimize mapping knowledge required

for the task. An interactive magnifying glass that followed the participant's mouse cursor was implemented per GIScorps (2013) using the jquery library "magnify" (Doan, 2018), allowing users to more closely examine image details. Rating guidelines were provided via a link and by mouse "hover" over rating options per the recommendation of Albuquerque et al. (2016), which is also consistent with an integrated instructional approach (Chandler & Sweller, 1991).

Geographic information system-based damage assessment tools fall into a wide variety of designs. Many real-world crowdsourced platforms address the concerns raised in the review by Kerle and Hoffman (2013), section 3.2, where: (i) clarity of instructions is addressed by simplified classification frameworks and guidelines, (ii) the need for domain knowledge is minimized by removing the mapping aspect of the task, (iii) options for interaction are limited to the selections within the classification guideline, and (iv) the software environment is a low-barrier user interface within the user's familiar web browser containing only the task at hand. By removing the mapping element of the task and providing participants with the building to be rated centered in the image, as well as a simple rating entry tool, the microtask is focused on learning the assessment of damage to the guideline in co-production with the agent.

Integrating automation into the workflow brings the challenge of sequencing and allocating subtasks to both the human and intelligent agent. Poor outcomes are common when the human is placed in the role of "final arbiter" (Smith, McCoy, & Layton, 1997). Often these outcomes are diagnosed as confirmation bias or complacency on the part of humans, but cognitive processes such as satisficing, trust, fixation on solutions, and narrowing may better explain the state of the human (Smith, 2017). The sequencing of decision steps here is similar to a "critiquing system" which Smith (2017) claims is superior to human "final arbiter" structure. Participants may economize their cognition and bypass this requirement by entering a random

classification at the initial step to reveal the automated assessor classification and explanation. Such behavior will be detected by timing and reported as an outcome. Figure 10 shows the process flow of the microtask loop with an opportunity to change the rating after the agent assessment is revealed.



Green = Human Processes, Blue = Survey Processes

Figure 10 Microtask Process Loop

IV.5 Scenario Generation

The images for damage assessment were from a real-world disaster data set (NGS, 2018) and pre-disaster images were from sources similar to those available in real-world disaster assessment (Google, 2018). These images, while of the same structure, were potentially taken years apart at different times of day and camera angle, which is also consistent with real-world disaster assessment.

The “Wizard of Oz” approach¹ has been developed within the area of computer interaction research to experimentally evaluate near-term technologies and technologies where the artifacts are unavailable for examination, but where experiments can produce useful information about the design of future artifacts (Habibovic, Andersson, Nilsson, Lundgren, & Nilsson, 2016; Riek, 2012). Guidelines for Wizard of Oz methodologies were reviewed in the human-robot interaction area by Riek (2012), and several recommendations are explicitly

¹ This approach simulates a computer system through the actions of a human operator. The term derives from the novel “The Wonderful Wizard of Oz” by L. Frank Baum where the character “Oz” appears to other characters in multiple forms other than his own by artificial means.

accounted for in the design of the research method for this study. The capabilities and limitations of the simulated system are specified below. The simulated output and explanations were developed in an iterative process. Unlike most research involving a Wizard of Oz approach, the interaction here is not in real-time and instead is based on pre-generated scenarios which are identical between all participants. Constraints and machine-like errors in system performance are included in the simulation. “Wizard error” is attempted to be controlled by developing a specification for the behavior of the simulation and evaluation of the simulated output for consistency with that specification. That specification appears in Figure 11. The explanations were generated in an iterative process developing a total of 22 scenarios which were tested. The steps used to generate the scenarios and their components are shown in Figure 12. The full list of scenarios appears in Appendix D. See Appendix F.1. for detailed description of the development process and Appendix F.2. for testing outcomes.

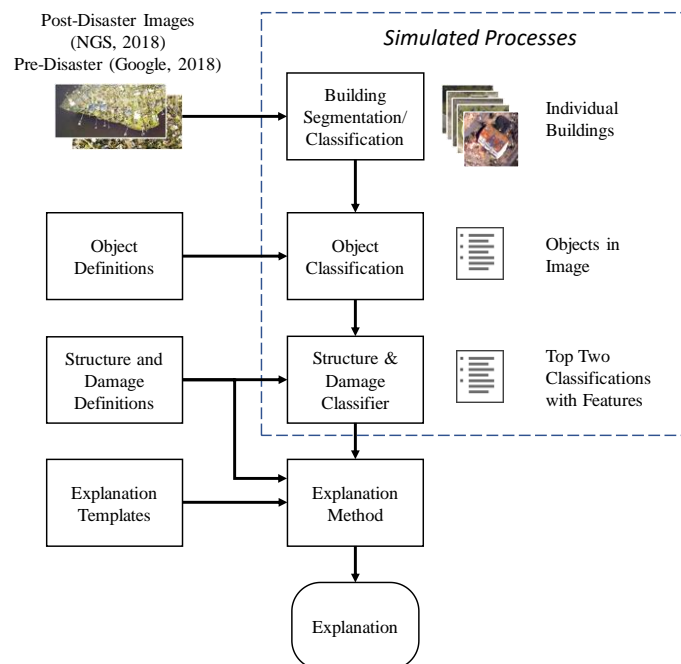


Figure 11 Simulated Output and Explanation Specification

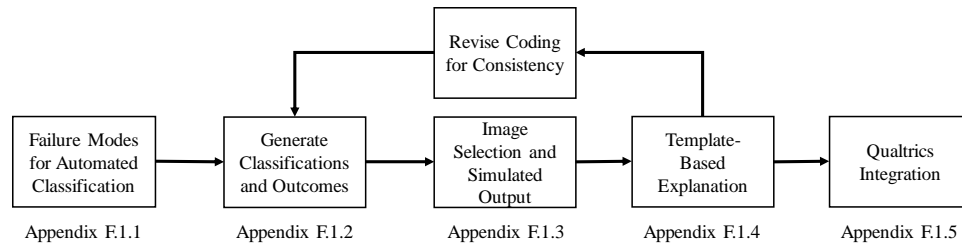


Figure 12 Process to Generate Scenarios

IV.6 Data Analysis Plan

The survey data was exported from Qualtrics in SPSS format. The data was prepared for analysis using SPSS Version 25, including restructuring and aggregation of the assessment outcome data. Partial Least Squares Structural Equation Modeling (PLS-SEM) was performed using Smart PLS 3.2.8 (Ringle, Wende, & Becker, 2015). The alternate explanations were considered as direct effects on self-efficacy.

Bagozzi and Yi (1989) proposed analyzing experiment data using a covariance structural equation modeling approach where manipulation groups were indicated with variables, which was extended by Bagozzi, Yi, and Singh (1991) to PLS-SEM. Streukens, Wetzels, Daryanto, and De Ruyter (2010) developed a method to analyze a two-factor experiment including a test of the interaction with a third dummy variable in PLS-SEM, along with mediators of the ultimate dependent variable. That approach was adopted for the data analysis method in this study, with the initial evaluation model shown in Figure 13, with single-item dummy variables indicating the counterfactual (CF) manipulation, the hedging (H) manipulation, and the interaction of the two (CF×H). To account for the control condition, a fourth dummy variable is included indicating the presence of the causal explanation (C).

PLS-SEM is a more appropriate estimation method over multivariate analysis of variance as it was developed to analyze latent constructs and mediation paths simultaneously. PLS-SEM estimates latent constructs using both reflective and formative models, and partial least squares is

well-suited to analyze the cognitive load composite-formative construct (Henseler, 2018), and is recommended for models with composite measures while covariance-based methods are inappropriate (Sarstedt, Hair, Ringle, Thiele, & Gudergan, 2016). Additionally, the review by Hair, Hollingsworth, Randolph, and Chong (2017) supports selecting PLS-SEM for research where constructs are being evaluated in a new context, and the limited assumptions of PLS-SEM make it an appropriate choice as the data was generated using Likert-type measures with unknown statistical distribution.

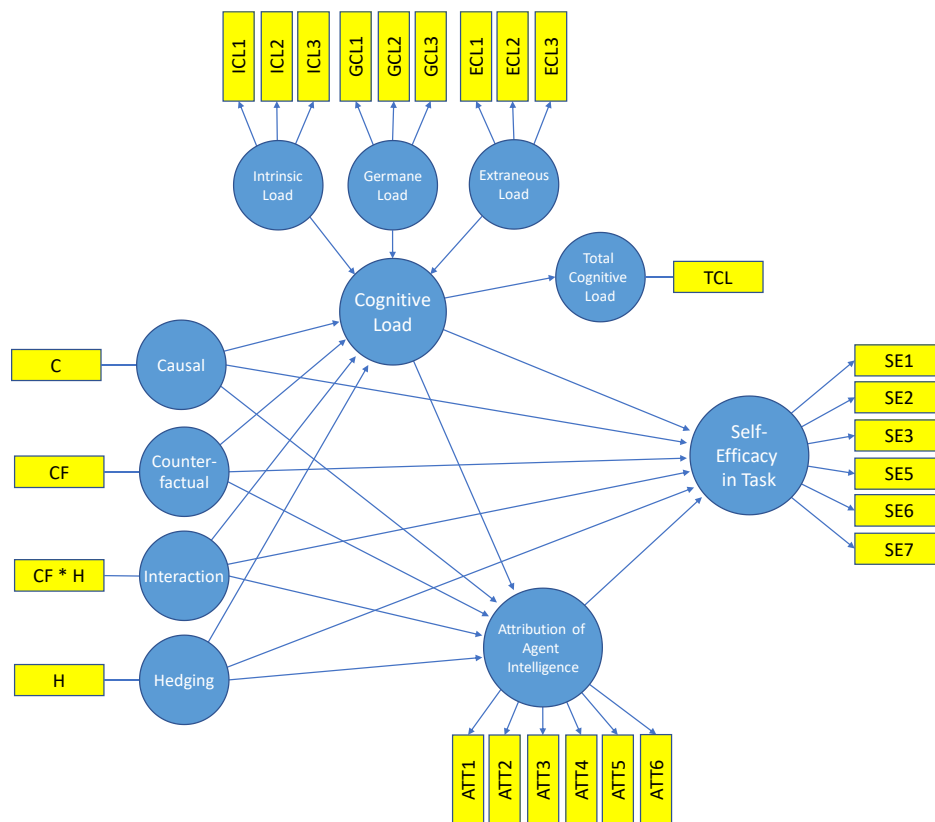


Figure 13 Second-Order Cognitive Load Structural Model

Estimates of latent constructs per participant from the PLS-SEM analysis were exported from Smart PLS and into SPSS. Bootstrap² analysis of path coefficients was conducted in SPSS and Smart PLS using bias-corrected and accelerated confidence intervals. Descriptive statistics were computed for each experiment condition, and a correlation table was constructed in SPSS. Smart PLS was used to conduct the analysis and evaluation of the structural model. Model quality criteria were adopted from Hair et al. (2016). Pairwise deletion of missing data was selected rather than mean replacement as the most appropriate treatment for missing learned trust for AI and demographic information. Mode A was used for reflective constructs and mode B was used for the formative cognitive load construct. Path weighting was selected for the analysis.

Path coefficients for relationships with the independent variables and gender are reported after transformation of the analysis output by dividing the original coefficient by the standard deviation of the indicator variable. This rescaling allows for conventional interpretation of the coefficients as changes in the mean by numbers of standard deviation. The standardized coefficients used by Smart PLS are scaled by the proportion of the indicator, which facilitates the analysis but produces path coefficients which are arbitrarily scaled and cannot be readily compared.

The intended second-order cognitive load construct was assessed first to establish the measurement model. Redundancy analysis was employed with the single-item for total cognitive load (Cheah, Sarstedt, Ringle, Ramayah, & Ting, 2018). Then the measurement model was assessed using the following metrics by the thresholds for each recommended by Hair et al. (2016): indicator loadings, multi-collinearity using variance inflation factor (VIF), discriminant

² In “bootstrap” analysis the main procedure is repeated many times with a resampled set of the same size as the original data composed of randomly drawn members with replacement. Standard errors for model parameters and quality metrics are estimated by inference of the population from the sample (Hair, Hult, Tomas, Ringle, & Sarstedt, 2016).

validity using heterotrait-monotrait ratio (HTMT), convergent validity for reflective constructs using average variance extracted (AVE), and cross-loadings of items with other constructs. Finally, the structural model was assessed for effects and predictive relevance. Manipulation checks were assessed using tests of differences using SPSS.

IV.7 Determination of Sample Size

The XAI field believes that the effect of providing explanations will be substantial and open up new areas of application for intelligent systems (Herman, 2017; Miller, 2019). Within types of explanation, no studies were identified comparing counterfactual explanations with simple causal explanations on attitudes to estimate the potential effect size. Effect sizes between explanation and “black box” conditions in the literature are generally large. However, response differences in trust and prediction error in Poursabzi-Sangdeh et al. (2018), where the number of model parameters in the explanation was varied, were not practically significant despite being statistically significant, with approximately 200 respondents per condition. The authors did not report sufficient detail to determine the effect size of types of local explanations. As the purpose of this study is to assess practical benefit, the sample size was determined on the basis of identifying or constraining the effect to at least medium effects ($f = 0.25$) (Cohen, 1988), but also sufficient power to validate the hypothesized mediation relationships.

To establish the minimum sample size required to detect medium-size direct effects between manipulations of explanation type on self-efficacy, the SPSS SamplePower 2x2 ANOVA with non-central F procedure was utilized. The required sample size per group was calculated as $n = 32$ with power of 80%. A sample size of $n = 50$ can detect an effect size of $f = 0.20$, and $n = 88$ allows $f = 0.15$. For $f = 0.10$, considered a small effect size, a group size of $n = 195$ would be required.

Power tests for mediation effects are specific to each proposed mediator, where each is expected to have unique path strength and statistical power (Thoemmes, MacKinnon, & Reiser, 2010). The inverse square root rule (Kock & Hadaya, 2018) with power threshold of 80% was utilized to evaluate sample size requirements for the structural model. This analysis finds that sample sizes appropriate for main effect size $f = 0.25$ would be sufficient to evaluate mediation paths as small as of $\beta = 0.197$; at $f = 0.20$: $\beta = 0.157$; at $f = 0.15$: $\beta = 0.119$; and at $f = 0.10$: $\beta = 0.080$.

A group size of 88 and sample size of 440 was selected for the ability to detect main effects at the medium-effect size threshold, while being able to detect statistically significant mediation for path coefficients as low as 0.119.

IV.8 Development and Testing

The study's task and damage assessment scenarios and survey instrument were developed in an iterative process. Measures of the dependent constructs, alternate explanations, and demographics were selected from the literature. The item to measure level of automation was developed by the author for this experiment. An initial pre-test using six telephone "cognitive interviews" was conducted to evaluate the task and instrument, followed by two rounds of pre-test data collection with 90 participants to evaluate the measurement model, manipulation check outcomes, and participant feedback. All composite measures were evaluated using pre-testing with scale reliability testing using Cronbach's alpha and factor loading. The development and testing results of the study materials is detailed in Appendix F.

V DATA ANALYSIS AND RESULTS

V.1 Sample Description

The study data were collected between February 4th and February 10th, 2019. The instrument used is provided in Appendix C. A summary of the results of the recruiting process are shown in Figure 14. A total of 610 potential participants were qualified, of which 90 participated in testing and 445 were recruited for the study (86% of those remaining). No submissions for qualification were rejected for lack of English proficiency. Seventy-five percent of all qualifications requested were received and granted on or before the first day of data collection. Seventy-six percent of the sample was collected in the first three days of data collection.

Of the 445 potential participants recruited into the study, 20 exited the survey at the consent screen, 7 at the briefing screen, and 27 failed the attention check. Of the remaining 418 that began the rating process, 23 dropped out during the damage assessment portion of the task. These partial completions were removed from the data as they provided no measure data. The second submission of a single participant with the same worker identification number was also removed from the data. The final sample contained 367 participants with an average of 73 per condition and between 69 and 79 per condition. While less than the intended sample size, power analysis for sensitivity of the obtained group and sample size confirmed that medium size effects could be detected with statistical power of 80% (minimum main effects: $f = 0.17$, $d = 0.34$; mediator effects: $\beta = 0.130$).

The mean age of the participants was 38.7 years with a median of 36. The minimum was 19 and the maximum was 71. Sixty percent of participants were female. A total of 55% indicated they had a college degree (either undergraduate or graduate). Participants had limited previous

experience with natural disaster damage assessment in aerial images, with 53% indicating having no prior experience; however, this rate was consistent with the experience of crowdsourced workers in previous real-world events (Dittus, 2017). Of those with experience, 76% indicated they had performed the task a few times a year or less in the last year, and 55% had not used a written guideline previously. A single participant indicated having used the same guideline as in this study. Participants had previous experience in related tasks: 12% indicated prior experience in geographic information systems, 12% with natural disaster damage assessment, 29% with aerial image interpretation, 9.3% with experience in a citizen science project, and 61% with none of the related experience categories.

Examples of participant open-text feedback appears in Table 5. Participants noted that the agent produced errors, but also was useful to identify their own errors. Few participants rejected the agent, and several indicated they felt it was essential to their performance. There was no indication that participants felt the task was artificial or that the agent was not real. Some workers provided feedback expressing that they would like the ability to indicate what the agent got wrong, similar to scrutability (Tintarev & Masthoff, 2011).

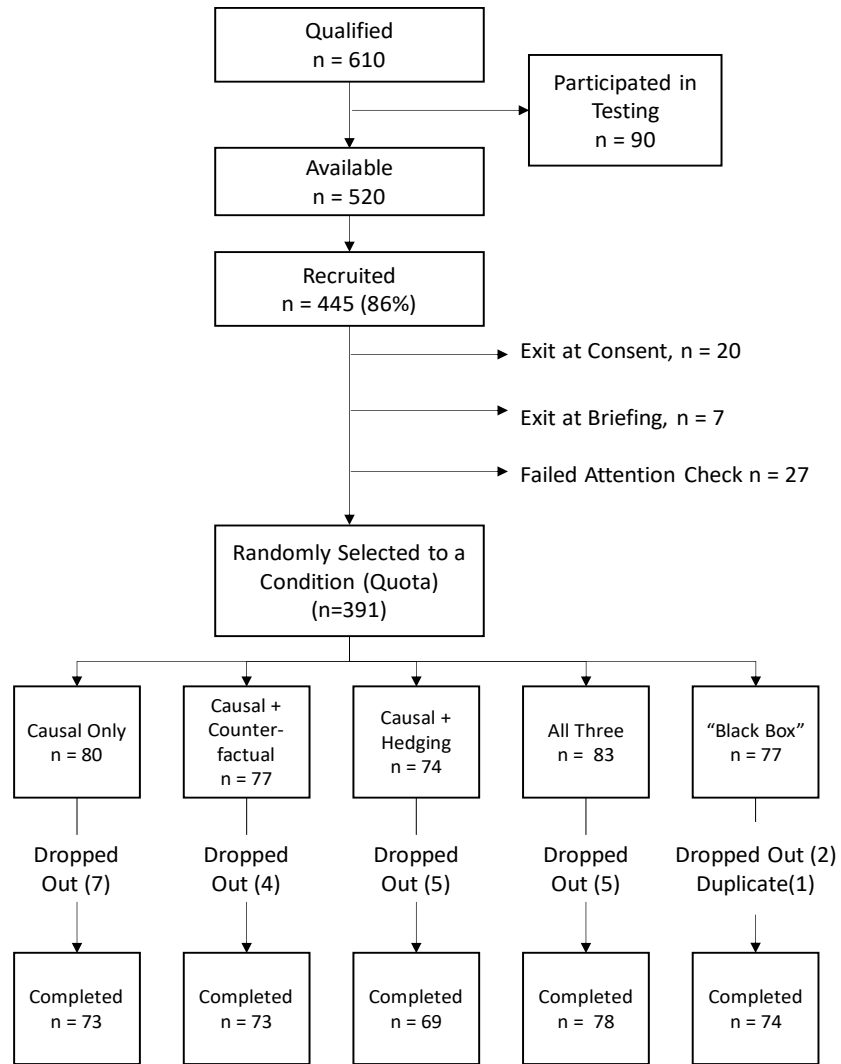


Figure 14 Recruiting Flow

Table 5 Selected Study Feedback

“I thought this was a really interesting survey and it is something I have never done before. I think that with more practice I could be really knowledgeable [*sic*] in this type of work. It was different working with ADAM because I perceived things a little differently in some cases. I felt more accurate in some cases whereas in other cases I felt ADAM was more accurate. I think that it is beneficial to work with ADAM but not heavily rely on that interpretation alone.”

“I was a bit confused by one very large building that looked like a warehouse. I would say that about 1/5 of the roof was destroyed, but I don't know enough about architecture or engineering to know if the whole roof needs to be replaced or just that section. I changed my rating to critical because of the automated rating, but I'm still not convinced.”

“Some of the time it seemed that ADAM had more information that I had access to, such as the integrity of walls inside of buildings that I only had a strictly overhead view to refer to. I couldn't see walls or possible damage so most of the time just changed to what ADAM suggested.”

“I thought that this is not only an interesting study, but it also increased my knowledge and I feel will help assist in damage assessment. It's a very meaningful study.”

“I found it very interesting. I think with the combination of AI and human intelligence, damage assessment can be a lot easier.”

“That was really hard! I am confused on the difference between minimal and significant - ADAM marked significant if just part of the roof was missing but the descriptions says entirely. Minimal says part of the roof ... so I wasn't confident in my prediction or his.”

“I enjoyed doing this hit. I think it was helpful to do practice hits like these to get a feel for how the system works, how things are labeled and scored, etc. It was good to get the feedback of ADAM so I could compare and rethink how I scored each one so I could accurately assess and change any mistakes I made but also be able to have someone check the AI's ratings for major discrepancies or mistakes too. I think with both sides on board it would give an overall better rating or picture of reality in general in the future during natural disasters and get help to those in need faster. Thanks!”

“Very interesting HIT. I enjoyed working with ADAM, and while I didn't agree with all of ADAM's assessments, it helped guide me to the correct assessment in a lot of cases. Since I'm just a beginner at this, it helped me to get better at identifying structures and assessing damage.”

“I enjoyed doing this HIT. I think that ADAM made me second guess myself but also helped me learn to do them correctly. When ADAM was wrong it was very obviously wrong to me.”

“I really enjoyed the assessment. There are evident accuracy issues with using just the ADAM [*sic*], but it is a good tool for pointing out things in the photo that may have been missed.”

V.2 Descriptive Statistics

Scores were computed for each composite measure using the sum of item ratings in SPSS based on the revised item composition of the structural measurement model. Correlations between sum of scale scores and the latent construct estimates exceeded 0.9 for all measures. The descriptive statistics shown in Table 6 are grouped by experiment condition for the study measures, alternate explanations, and demographics with tests for differences in means between experiment conditions. A correlation table was computed in SPSS using Pearson correlation with the extracted latent construct estimates from Smart PLS, with the result shown in Table 7.

Table 6 Descriptive Statistics by Condition and Differences in Means

Measure	Causal Only	Causal +	Causal +	All Three	Black Box	Test of Differences in Means	
	n = 73	Counterfactual n = 73	Hedging n = 69	Explanations n = 78	n = 74	ANOVA	p-value
	Mean (Standard Deviation)						
Total Cognitive Load	7.77 (1.11)	8.10 (0.97)	7.81 (1.10)	7.87 (1.18)	7.81 (1.11)	F(4,362) = 1.0	0.388
Intrinsic Cognitive Load	13.9 (3.95)	15.6 (3.45)	13.7 (4.07)	14.9 (3.50)	13.3 (3.83)	F(4,362) = 4.8	<0.001
Germane Cognitive Load	18.8 (2.07)	19.2 (1.93)	18.8 (2.13)	18.5 (2.52)	19.3 (1.79)	F(4,362) = 1.6	0.184
Extraneous Cognitive Load	9.4 (4.06)	9.9 (3.96)	9.2 (3.74)	10.3 (4.54)	8.0 (3.96)	F(4,362) = 3.4	0.009
Attribution of Agent Intelligence	24.6 (7.65)	24.7 (7.85)	27.0 (6.21)	24.6 (7.20)	23.4 (6.62)	F(4,362) = 2.5	0.046
Self-Efficacy in Task	33.4 (6.07)	33.0 (7.77)	34.3 (6.62)	33.0 (6.95)	35.6 (6.54)	F(4,362) = 1.9	0.106
Erroneous Agreement	0.312 (0.262)	0.290 (0.271)	0.270 (0.256)	0.287 (0.253)	0.165 (0.202)	F(4,362) = 4.0	0.004
Initial Review (Seconds)	27.8 (20.9)	30.2 (16.3)	26.4 (14.3)	31.3 (20.1)	30.4 (21.4)	F(4,362) = 0.8	0.499
Review of Automated (Seconds)	24.1 (14.1)	28.4 (14.6)	23.8 (11.5)	29.3 (19.4)	19.1 (10.0)	F(4,362) = 6.0	<0.001
Trust in Intelligent Agent	4.63 (1.40)	4.82 (1.32)	4.67 (1.52)	4.56 (1.35)	3.88 (1.42)	F(4,362) = 5.0	<0.001
Perceived Interdependence	11.8 (1.72)	11.9 (1.84)	11.8 (1.91)	11.4 (1.90)	11.0 (2.58)	F(4,362) = 2.9	0.020
Perceived Level of Automation	4.12 (1.49)	4.44 (1.47)	4.51 (1.53)	4.19 (1.49)	3.88 (1.44)	F(4,362) = 2.1	0.080
Previous Task Experience	2.45 (3.40)	2.93 (3.78)	3.23 (4.28)	2.01 (3.34)	2.58 (4.00)	F(4,362) = 1.1	0.345
Dispositional Trust	15.5 (3.32)	15.8 (3.69)	15.8 (3.57)	15.4 (3.52)	15.9 (2.94)	F(4,362) = 0.4	0.837
Learned Trust for AI	2.95 (0.82)	3.02 (0.88)	3.19 (0.77)	3.07 (0.67)	3.04 (0.78)	F(4,296) = 0.8	0.550
Gender (0 = male, 1 = female)	0.556 (0.500)	0.556 (0.500)	0.652 (0.480)	0.671 (0.473)	0.608 (0.492)	F(4,358) = 0.9	0.482
Age (years)	37.5 (9.8)	43.3 (11.4)	37.5 (9.4)	37.5 (10.3)	37.8 (7.7)	F(4,362) = 4.9	<0.001
Education	3.90 (1.04)	3.73 (1.12)	3.59 (0.93)	3.71 (1.11)	3.49 (0.95)	F(4,361) = 1.7	0.156
Income	4.55 (2.10)	4.14 (2.07)	4.25 (1.96)	4.04 (1.72)	4.36 (1.79)	F(4,360) = 0.8	0.532

Of the measures expected to be equivalent between experiment conditions, only age shows a statistically significant difference, where the causal with counterfactual experiment condition has an average age 6 years higher than the other conditions (0.47 standard deviations). The difference in median age is consistent with the difference of the mean (43 versus 35 to 36 in the other conditions), eliminating the potential role of outliers in this outcome. While this difference arising due to random chance is very rare, a causal relationship with the experimental

manipulation is not plausible given the low drop-out rates and that the Qualtrics randomization procedure has no access to the participant's age. The effect of age was included in the structural model as a direct path with self-efficacy.

There is a large skewed peak in distribution of prior experience in task measure due to 53% of respondents reporting no prior experience. The effect of this is apparent with standard deviations greater than the mean scores. This non-normality violates the assumptions of the analysis of variance test. The non-parametric Kruskal-Wallis Independent-Samples test of previous task experience by experiment condition was performed in SPSS. It also found no statistically significant differences between conditions when zero values are excluded (statistic: 3.8, $df = 4$, $p = 0.480$), or in the full data (statistic: 4.2, $df = 4$, $p = 0.385$).

Manipulation check measures were employed to validate effects of the independent variables. The "Don't Know / Don't Remember" rating for the manipulation checks were selected by 0.5%, 14%, and 6% of the participants for the causal, counterfactual, and hedging checks. For those that rated the manipulation checks, 6.5%, 19.3%, and 29.4% of participants rated them inconsistent with the experiment condition. For participants that provided response, Independent T-Tests for differences in ratings between the manipulation check between the manipulation conditions were all significant and in the appropriate direction of difference (causal $z = -27.0$, $df = 363$, $p < 0.001$; counterfactual $z = -11.58$, $df = 313$, $p < 0.001$; hedging $z = -8.4$, $df = 343$, $p < 0.001$). This indicates that participants which rated the manipulation checks with confidence were broadly able to rate their experiment condition consistent with the explanations they had been provided.

Table 7 Correlation Table

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1 Total Cognitive Load																				
2 Intrinsic Cognitive Load	.206**																			
3 Germane Cognitive Load	.201**	.088																		
4 Extraneous Cognitive Load	.031	.506**	-.277**																	
5 Attribution of Agent Intelligence	.015	.104*	.097	.063																
6 Self-Efficacy in Task	.075	-.139**	.304**	-.406**	.066															
7 Erroneous Agreement	-.007	.094	-.028	.165**	.208**	-.107*														
8 Scenarios Marked Incorrect	-.009	.010	.053	-.013	-.007	.079	-.066													
9 Initial Timing	.180**	.108*	.127*	-.034	.018	.042	-.088	-.100												
10 Review Timing	.183**	.193**	.080	.056	.066	-.086	.018	-.091	.582**											
11 Trust in the Intelligent Agent	.111*	.122*	.101	.026	.604**	.090	.332**	.036	-.096	.017										
12 Interdependence	.064	.038	.201**	-.156**	.369**	.245**	.118*	-.050	.126*	.150**	.547**									
13 Level of Automation	-.014	.051	-.046	.090	.307**	-.053	.168**	-.046	-.076	.002	.284**	.064								
14 Previous Task Experience	-.030	-.101	.017	-.084	.007	.199**	-.003	.007	-.082	-.109*	.074	.047	.066							
15 Dispositional Trust for AI	-.009	.017	.307**	-.254**	.160**	.244**	.061	.038	-.028	.003	.233**	.333**	-.036	.088						
16 Learned Trust for AI	.056	.003	.031	-.083	.152**	.067	.159**	-.102	-.005	.019	.242**	.186**	.069	.015	.201**					
17 Gender (0 = male, 1 = female)	.193**	.122*	.102	-.094	-.104*	.011	.105*	-.059	.196**	.123*	-.046	.075	-.041	-.129*	-.098	.078				
18 Age	.149**	.190**	.032	-.023	-.040	-.106*	.047	-.031	.339**	.350**	.016	.046	.012	.001	-.006	-.115*	.137**			
19 Education	.050	-.019	.045	-.030	-.082	.009	-.088	-.071	-.066	-.073	-.060	.006	-.113*	.080	.033	-.078	-.040	.051		
20 Income	.023	-.070	.019	-.077	-.099	.079	.006	-.019	-.012	-.081	-.061	-.013	-.024	-.006	.010	.010	.047	.045	.214**	
Causal Explanation	.028	.177**	-.087	.174**	.106*	-.129*	.198**	.066	-.029	.200**	.222**	.125*	.116*	.006	-.025	.010	.001	.046	.097	
Counterfactual Explanation	.083	.240**	-.018	.150**	-.024	-.106*	.079	.034	.066	.218**	.105*	.017	.049	-.038	-.051	-.007	.010	.132*	.027	
Hedging Explanation	-.021	.047	-.099	.077	.107*	-.028	.045	.022	-.012	.095	.059	-.031	.064	-.009	-.034	.080	.089	-.099	-.022	
Interaction of Explanations	.000	.105*	-.088	.109*	-.026	-.063	.046	.003	.056	.152**	.020	-.075	-.011	-.085	-.075	.013	.066	-.063	.014	

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

V.3 Assessment Outcomes

The damage assessments were evaluated to ensure the participants produced data with representative effort and repeatability. A summary table of results appears in Table 8. The participant consensus ratings in the initial step were identical between the testing and study groups for 19 of 20 ratings selections. The consensus damage rating for scenario 17 was “no damage” in the study compared to “minimal” in the testing phase. Agreement was also assessed by bootstrapped resampling with 500 resamples. Only three images changed modes for damage ratings: scenario #8 (2% of resamples), #9 (0.4%), and #17 (47%, where the consensus damage rating varies from “no damage” and “minimal”). None of the structure type rating modes varied.

Changes in ratings were assessed to evaluate differences between scenario type and agreement in the first step. Changes to agree were more likely for the faithful scenarios, and participants were unlikely to select the same damage rating as the agent in the erroneous scenarios indicating the types operated as expected. Participants changed their ratings to agree with the simulated agent in 36.6% of faithful scenarios, compared to 26.5% for erroneous scenarios, which was a statistically significant difference in proportion using a Chi-Square test for independence ($\chi^2 = 43.3$, $df = 1$, $p < 0.001$). For the erroneous scenarios, 95 of 1,835 initial participant damage ratings (5.3%) agreed with the intended erroneous agent rating. Of these, a single damage rating was changed in the following review step by the participant (to the initial step consensus rating).

The task allowed participants to indicate that the automated rating was incorrect, and this feedback was assessed for agreement with the scenario types and between-experiment conditions. Participants with no cases of erroneous agreement indicated the intended number of scenarios were incorrect, and in a limited number of cases participants changed their rating to

agree with an agent damage assessment they also had marked incorrect. Descriptive statistics for the incorrect indication in categories by the number of scenarios with which participants were in erroneous agreement appear in Table 9. Participants that did not change their ratings to agree with the erroneous agent damage assessments have a median and mode of five scenarios marked incorrect, the intended number of erroneous scenarios. Each incremental erroneous agreement outcome decreased the median number of scenarios rated incorrect by one until the fourth erroneous agreement. Binary classification analysis comparing the incorrect ratings with the scenario type (faithful/erroneous) found sensitivity of 0.53, specificity of 0.85, and precision of 0.78, indicating that participants selected the intended erroneous scenarios as incorrect most often and correctly, though with some lack of sensitivity. Usage of the checkbox was not consistent between experiment conditions: in the “black box” condition an average of 18% of participants rated the agent incorrect in faithful scenarios and 66% in erroneous scenarios, which decreased to 13% and 48% in conditions with an explanation. There were 17 damage assessments (0.5%) in which 13 participants changed their assessment to agree with the erroneous automated damage rating but also marked that they believed the agent was incorrect. Of these, seven took place in the condition with only a causal explanation, six in the condition with all three explanations, two each in the conditions where a counterfactual or hedging explanation was offered with a causal explanation, and none in the “black box” condition with no explanations.

Participants that did not indicate that any scenarios were incorrect (9%) were evaluated for their use of the “blurry” and “obscured” quality feedback items to detect any possible laziness on the part of raters. The group that rated no images to be in error marked an average of 1.39 of the 20 other image feedback options compared to 1.35 on average for people that marked

any number of images in error, indicating roughly consistent usage of the other quality feedback options between groups.

Table 8 Descriptive Results of Damage Assessments

Type	Scenario	Initial Assessment Consensus		Changed Rating During Review		Rated Agent Incorrect	Erroneous Agreement	Timing (Seconds)	
		Structure	Damage	Structure	Damage			Initial	Review
Faithful	7	Medium	Minimal	16%	43%	13%	n/a	34.6	20.1
	8	Heavy	Critical	14%	58%	7%		27.6	17.0
	10	Medium	Significant	11%	59%	10%		31.0	17.4
	18	Heavy	Critical	24%	27%	29%		20.6	24.3
	22	Heavy	Significant	35%	35%	14%		25.6	18.1
Erroneous	1	Heavy	Significant	23%	56%	59%	20%	34.8	28.5
	9	Medium	No Damage	24%	39%	46%	31%	26.7	34.2
	17	Medium	No Damage	14%	33%	44%	39%	31.1	30.9
	19	Light	No Damage	25%	38%	56%	23%	31.5	28.7
	21	Heavy	Minimal	39%	26%	59%	20%	29.1	31.0

Table 9 Erroneous Agreement and Incorrect Ratings

Erroneous Agreement			Scenarios Rated "Incorrect" by Participant			
# Scenarios	Participants	Cumulative	Mean	SD	Median	Mode
0	124	34%	4.5	2.1	5	5
1	98	60%	3.8	1.9	4	4
2	75	81%	2.7	1.6	3	3
3	47	94%	1.9	1.0	2	2
4	18	99%	0.6	0.6	0.5	0
5	5	100%	2.4	3.8	1	0
Total	367		3.4	2.1	3	3

V.4 Model Evaluation

The quality of the measurement and structural model were assessed in three steps. First, the second-order latent construct measurement model for cognitive load was assessed for validity. As a result of this check, the second-order construct was replaced with direct relationships with the three components of cognitive load, and instead the three first-order constructs were moved to direct relationships. Second, the remainder of the measurement model was evaluated to understand model quality. This process found that the items for perceived reliance did not co-vary so the concept was removed from the analysis, and one item was dropped from perceived interdependence based on low composite reliability and coherence with

the concept. Finally, the structural model was assessed identifying that cross-loading between some of the human-computer interaction concepts does not impact conclusions, and that predictive relevance for the statistically significant path relationships was confirmed.

The second-order latent construct for cognitive load was assessed to confirm its validity prior to analyzing the full measurement model. There was no significant relationship between the second-order construct and the single-item total cognitive load (path coefficient = 0.062, $p = 0.484$). The weights of the first-order constructs on the second-order were: intrinsic = 0.434, $p < 0.001$; germane = 0.247, $p = 0.146$; extraneous = 0.648, $p < 0.001$ where the items for germane cognitive load are negatively loaded, consistent with theory. While model quality metrics such as composite reliability, average variance extracted, and heterotrait-monotrait ratios were acceptable for the cognitive load components, the lack of relationship between the established measure and the second-order construct does not support its use in the measurement model.

When the repeated indicators of the second-order construct were replaced with the single-item for total cognitive load, the estimated R^2 for second-order construct was 10.5% (formative path coefficients: intrinsic 0.156, $p = 0.010$; germane = 0.225, $p < 0.001$; extraneous = 0.130, $p = 0.357$). The estimates of path coefficients were most consistent with “model 1” from Klepsch et al. (2017) which was described as the “worst fit” of the six models they evaluated in their confirmatory factor analysis of the measures used in the three-component measure adapted for this study. These results did not support the use of this measurement model.

A structural model was specified replacing the second-order construct with direct relationships for the components of cognitive load. This was consistent with “model 2” from Klepsch et al. (2017) which was deemed the best fit in that study. In this model, shown in Figure 15, the number of path relationships evaluated with attribution of agent intelligence and self-

efficacy is greatly increased. In this model each component for cognitive load had a composite reliability greater than 0.8 and each item loads with their latent constructs with statistical significance.

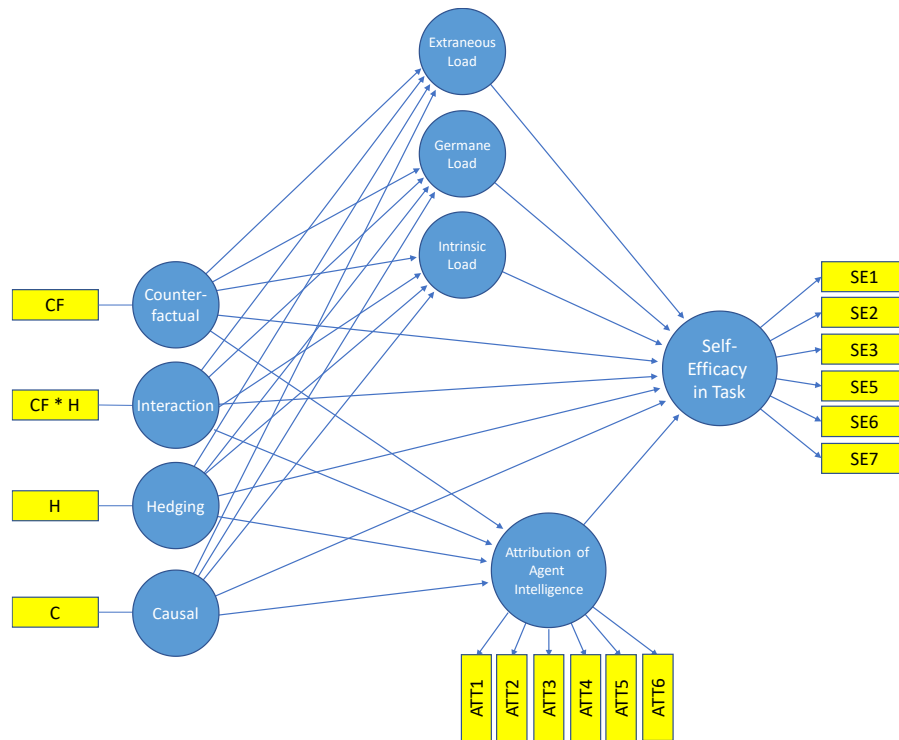


Figure 15 Revised Structural Model

Note: The three indicators of each cognitive load component are omitted for clarity.

Next, the measurement model of the revised structural model was assessed. Composite reliability was below the 0.800 threshold for perceived interdependence (0.406) and reliance (0.606). Removing the first item for perceived interdependence (loading -0.203) increases composite reliability to 0.819, though Cronbach's alpha is marginal (0.600). The first item asked for a rating of whether the agent "affected their ratings," while the third item asked if "their ratings benefited" from working with the agent. The first item was dropped from the model as "being affected" was not as coherent conceptually with interdependence compared to a benefit from working together. Removing the third item for perceived reliance (loading 0.322) increased

composite reliability to 0.680, and the remaining items were not strongly tied conceptually, with one item being about reliance and the other about distraction. The measurement of perceived reliance was deemed unreliable and the measure was dropped from the study.

Composite reliability exceeded 0.800 and average variance extracted (AVE) exceeded 0.500 with p -values less than 0.001 for all constructs. A table of each metric for model quality appears in Table 10. Outer loading for each item is listed in Table 11. The first item for intrinsic load has lower loading than is ideal for a reflective measure (0.675, $p < 0.001$), along with the second extraneous load item (0.672, $p < 0.001$). When reflective indicators load above 0.400 and the construct is otherwise acceptable measures, it is appropriate to retain the items (Hair et al., 2016).

As previous research has identified the potential for multiple dimensions of perception of mental processes, the response for the items of attribution of agent intelligence were assessed to ensure a reflective measurement model was appropriate for the construct. Principal component analysis performed in SPSS extracts a single component with eigenvalue 3.17 (52.9% of variance) where the next component has an eigenvalue of 0.849 (14.1%), with KMO sampling adequacy 0.828 and sphericity test $p < 0.001$. All items load greater than 0.600 on the single extracted component. The individual items were then evaluated. When the fourth item (“has a mind”) with loading 0.489 is dropped, the loading of the fifth item (“is predictable”) decreases to 0.622, below the desirable loading level of 0.700. Dropping both items improves model metrics slightly (composite reliability 0.869, average variance extracted 0.626, $R^2 = 5.7\%$). As the majority of the variance of the scale is explained by a single component and dropping items to increase metrics has a negligible effect on acceptable quality metrics, the items were retained as well as the reflective measurement model.

Cross-loading of items between constructs was assessed to determine if any constructs were not sufficiently distinct, with results shown in Table 12. The second intrinsic load item cross-loads above the 0.500 threshold with extraneous cognitive load. Otherwise, cross-loading above the threshold was between the items and constructs for attribution of agent intelligence, trust in the intelligent agent, and perceived interdependence. Kock and Lynn (2012) identify that cross-loadings above 0.500 indicate either incorrect association of items to construct, or to collinearity. No cross-loadings were above 0.700, and the highest outer variance inflation factor between composite constructs was 2.08 for trust in the intelligent agent and self-efficacy limiting the potential role of collinearity. To evaluate the potential that cross-loading suppressed relationships with the other alternate explanations, perceived interdependence was removed, and the model was re-evaluated. None of the other alternate explanations had a path coefficient with self-efficacy of absolute value greater than 0.10, or a 95% confidence interval that excluded zero. As such, the constructs were maintained as separate concepts without altering their composition.

Heterotrait-monotrait (HTMT) ratios were below 0.800 in the 97.5th percentile after bootstrapping analysis, with the exception of perceived interdependence with trust in the intelligent agent (0.776 original value and 0.895 at 97.5%). All other pairs were below the 0.800 threshold in the original and 97.5% bootstrapping results. Inner Variance Inflation Factors (VIF) for reflective measures were below 3.00 for all pairs except with interaction of explanation types with a maximum value of 3.55.

Finally, the explanatory and predictive quality of the model was assessed. The predictive relevance Q^2 quality metric (“blindfolding”) was computed using the cross-validated redundancy method with distance of 10. Each endogenous construct exceeded zero, meeting the benchmark. When the construct items were evaluated only the fourth item for attribution of agent intelligence

was less than zero (metric: -0.001), which also had lower than desirable loading in the measurement evaluation. When the three cognitive load constructs are removed from the model, the effect of cognitive load on self-efficacy was estimated as $f^2 = 0.167$, and the predictive effect was estimated as $q^2 = 0.115$, which are above and below the “medium” guideline of 0.150. The same test of perceived interdependence results in metrics $f^2 = 0.023$ and $q^2 = 0.016$; and, for previous task experience, $f^2 = 0.037$ and $q^2 = 0.027$, which are each of small effect.

Table 10 Latent Construct Quality Metrics

Latent Construct	Composite Reliability			Average Variance Extracted		
	Metric	T Statistic	p-value	Metric	T Statistic	p-value
Intrinsic Cognitive Load	0.823	19.6	<0.001	0.613	13.0	<0.001
Germane Cognitive Load	0.809	20.4	<0.001	0.586	12.1	<0.001
Extraneous Cognitive Load	0.847	56.7	<0.001	0.652	27.0	<0.001
Attribution of Agent Intelligence	0.861	15.9	<0.001	0.514	11.5	<0.001
Self-Efficacy in Task	0.962	210.4	<0.001	0.808	42.3	<0.001
Dispositional Trust for AI	0.800	18.2	<0.001	0.506	11.4	<0.001
Perceived Interdependence	0.820	14.9	<0.001	0.698	14.0	<0.001

Table 11 Item Outer Loading

Latent Construct	Item	Outer Loading			95% CI	
		Loading	T Statistic	p-value	Lower	Upper
Intrinsic	ICL1	0.675	7.1	<0.001	0.347	0.786
	ICL2	0.930	17.1	<0.001	0.882	0.999
	ICL3	0.720	9.0	<0.001	0.453	0.816
Germane	GCL1	0.734	8.1	<0.001	0.444	0.842
	GCL2	0.740	7.6	<0.001	0.384	0.846
	GCL3	0.821	9.3	<0.001	0.691	0.973
Extraneous	ECL1	0.869	47.2	<0.001	0.826	0.898
	ECL2	0.672	10.9	<0.001	0.515	0.765
	ECL3	0.866	38.9	<0.001	0.818	0.903
Attribution of Agent Intelligence	ATT1	0.746	8.9	<0.001	0.537	0.829
	ATT2	0.844	10.1	<0.001	0.774	0.908
	ATT3	0.818	8.9	<0.001	0.748	0.908
	ATT4	0.489	3.3	0.001	-0.347	0.661
	ATT5	0.622	6.4	<0.001	0.339	0.737
	ATT6	0.724	7.8	<0.001	0.417	0.811
Self-Efficacy in Task	SE1	0.923	81.5	<0.001	0.895	0.941
	SE2	0.932	108.4	<0.001	0.914	0.948
	SE3	0.922	83.6	<0.001	0.896	0.940
	SE5	0.912	77.4	<0.001	0.885	0.932
	SE6	0.837	29.7	<0.001	0.767	0.880
	SE7	0.862	40.7	<0.001	0.815	0.897
	Dispositional Trust for AI	DTRUST1	0.892	21.3	<0.001	0.816
DTRUST2		0.653	6.5	<0.001	0.430	0.827
DTRUST3		0.572	4.8	<0.001	0.248	0.726
DTRUST4		0.691	7.2	<0.001	0.455	0.808
Perceived Interdependence	INTERD2	0.933	19.6	<0.001	0.834	0.997
	INTERD3	0.725	6.2	<0.001	0.395	0.855

Table 12 Cross Loading of Items Between Constructs

Item	Intrinsic	Germane	Extraneous	Attribution of Agent Intelligence	Self-Efficacy in Task	Perceived Interdependence	Trust in the Intelligent Agent	Perceived Level of Automation	Dispositional Trust for AI	Learned Trust for AI	Previous Task Experience
ICL1		0.29	0.22	0.07	0.05	0.11	0.06	0.01	0.13	-0.01	-0.04
ICL2		-0.01	0.52	0.11	-0.19	0.02	0.14	0.07	0.00	0.01	-0.10
ICL3		0.10	0.34	0.03	-0.08	0.01	0.04	0.00	-0.05	-0.01	-0.08
GCL1	0.13		-0.16	0.00	0.19	0.15	0.07	-0.04	0.22	0.02	-0.01
GCL2	0.13		-0.16	-0.03	0.19	0.17	-0.01	-0.15	0.31	0.01	-0.02
GCL3	0.00		-0.28	0.17	0.29	0.15	0.13	0.03	0.21	0.03	0.04
ECL1	0.46	-0.27		0.02	-0.34	-0.16	0.04	0.06	-0.19	-0.11	-0.09
ECL2	0.30	-0.21		0.03	-0.24	-0.12	-0.04	0.05	-0.26	-0.04	-0.04
ECL3	0.44	-0.20		0.09	-0.39	-0.11	0.04	0.10	-0.19	-0.03	-0.07
ATT1	0.00	0.00	0.02		0.02	0.28	0.42	0.28	0.13	0.15	-0.01
ATT2	0.10	0.15	0.04		0.05	0.38	0.57	0.27	0.17	0.12	0.00
ATT3	0.16	0.13	0.04		0.09	0.23	0.42	0.19	0.15	0.06	-0.04
ATT4	0.07	-0.08	0.19		-0.05	0.08	0.33	0.15	-0.06	0.07	-0.05
ATT5	-0.01	-0.02	0.03		0.00	0.25	0.36	0.27	0.02	0.11	0.01
ATT6	0.05	0.03	0.09		0.06	0.28	0.50	0.21	0.11	0.14	0.10
SE1	-0.13	0.25	-0.36	0.00		0.19	0.02	-0.07	0.18	0.03	0.14
SE2	-0.14	0.33	-0.39	0.06		0.25	0.11	-0.05	0.24	0.06	0.19
SE3	-0.14	0.24	-0.40	0.07		0.25	0.12	-0.03	0.23	0.07	0.19
SE5	-0.08	0.25	-0.33	0.05		0.23	0.06	-0.08	0.23	0.04	0.19
SE6	-0.14	0.27	-0.33	0.07		0.16	0.09	-0.02	0.16	0.05	0.18
SE7	-0.12	0.29	-0.37	0.10		0.23	0.09	-0.03	0.27	0.08	0.17
INTERD2	0.02	0.20	-0.15	0.25	0.25		0.39	0.00	0.28	0.10	0.06
INTERD3	0.06	0.11	-0.11	0.44	0.13		0.62	0.16	0.29	0.23	0.00
TrustItem	0.12	0.10	0.03	0.60	0.09	0.55		0.28	0.23	0.22	0.07
LOAItem	0.05	-0.05	0.09	0.31	-0.05	0.06	0.28		-0.04	0.06	0.07
DTRUST1	0.03	0.30	-0.20	0.14	0.26	0.31	0.19	-0.04		0.11	0.08
DTRUST2	-0.08	0.15	-0.30	0.08	0.16	0.22	0.15	0.01		0.12	0.02
DTRUST3	0.04	0.20	-0.11	-0.01	0.04	0.13	0.04	-0.08		0.17	0.03
DTRUST4	0.09	0.22	-0.06	0.20	0.11	0.22	0.24	-0.04		0.21	0.14
LTrustItem	0.00	0.03	-0.08	0.16	0.07	0.19	0.25	0.07	0.20		0.02
PTE	-0.10	0.02	-0.08	0.01	0.20	0.05	0.07	0.07	0.09	0.01	

V.5 Results

The results of the path analysis appear in Figure 16, excluding the independent variables for clarity. The complete list of path results is shown in Table 13. The effects of the explanation types are shown in bar charts in Figure 17 by measure. Both the individual effect (solid bars) of the explanation type and the total effect in combination (outline only) are shown. Path coefficients are scaled by the type of measure. Continuous variables are scaled as standardized beta coefficients (β) where a variation of one standard deviation produces a shift in the mean of the dependent variable by a proportion of standard deviations. Dichotomous indicator variables are scaled as the shift in the mean of the dependent variable by proportion of standard deviations (as in Cohen's d). The comparison for the causal explanation and total explanation effects are to

the “black box” condition. The counterfactual and hedging explanations are compared to the causal explanation condition. For gender, the comparison is female compared to male participants. Statistical significance of the total effects of the explanation types was calculated using SPSS. Post hoc Tukey HSD testing was utilized to compare the conditions with explanations to the “black box” condition. The effect of explanations on self-efficacy is reported as partial eta-squared to allow direct comparison with the explanation effect on other measures.

The hypothesized relationships of the research model were assessed. The counterfactual explanation increased each component of cognitive load (intrinsic $d = 0.514$, germane $d = 0.204$, extraneous $d = 0.132$) where the increase in intrinsic was significant ($p = 0.001$), unlike the hypothesized relationship (H1a). The estimates for the effect of the hedging explanation on cognitive load (H2a) was negligible ($d = 0.032$, $d = -0.001$, $d = -0.053$) and did not reach statistical significance. The hedging explanation increased attribution of agent intelligence with statistical significance ($d = 0.380$, $p = 0.042$) which was not expected (H2b). None of the components of cognitive load had a statistically significant relationship with attribution of agent intelligence (H3) with small positive estimates for effects ($d = 0.065$, $d = 0.117$, $d = 0.056$). There is a medium-size negative effect of extraneous cognitive load on self-efficacy ($\beta = -0.339$, $p < 0.001$) and total effect of the causal explanation on self-efficacy ($d = -0.314$, $p = 0.038$). The direct effect of the causal explanation on extraneous cognitive load ($d = 0.362$, $p = 0.028$) confirms that cognitive load mediates the effect of explanations on self-efficacy in the task (H4). The role of attribution of agent intelligence in self-efficacy (H5) was limited with a very small estimated effect ($\beta = 0.017$, 95% CI $[-0.132, 0.170]$, $p = 0.826$).

The difference in the effect of cognitive load (H4) and attribution of agent intelligence (H5) on self-efficacy was assessed. The procedure from Rodríguez-Entrena, Schubert, and

Gelhard (2018) was utilized. Differences in the absolute values of the path coefficients were computed for each bootstrap sample to determine the mean and variance. The 95% confidence interval of the difference was then calculated using parametric assumptions without bias-correction and acceleration. This confirmed a greater effect of extraneous cognitive load ($\beta_{\text{difference}} = 0.276, [0.136, 0.416], p < 0.001$). The method was also utilized to compare the path coefficients of germane cognitive load and attribution of agent intelligence with self-efficacy, with less distinct differences ($\beta_{\text{difference}} = 0.136, [-0.060, 0.324], p = 0.077$).

Erroneous agreement was assessed for evidence of illusory understanding effects. The coincident decreases in self-efficacy are most consistent with participants “going along” with the simulated agent rather than being convinced by the explanation. Causal explanations increased erroneous agreement ($d = 0.580, [0.285, 0.875], p < 0.001$), but decreased self-efficacy (total effect, $d = -0.314, [-0.601, -0.016], p = 0.038$), which is inconsistent with explanations generating false understanding. The counterfactual and hedging explanations do not greatly alter erroneous agreement, though there is some indication that either other explanation type could partially offset the effect (counterfactual $d = -0.086, [-0.427, 0.258], p = 0.619$ and hedging explanations $d = -0.168, [-0.494, 0.170], p = 0.327$). Presenting all three explanations largely cancels out any potential benefit of additional explanation for erroneous agreement (interaction $d = 0.156, p = 0.519$; total effect $d = 0.482, p = 0.023$).

Finally, the alternate explanations were assessed. Direct effects on self-efficacy were identified for previous task experience ($\beta = 0.169, p < 0.001$) and perceived interdependence ($\beta = 0.166, p = 0.016$), but not for trust in the intelligent agent ($p = 0.877$), level of automation ($p = 0.577$), and dispositional ($p = 0.414$) and learned trust in AI ($p = 0.634$). The effect of the explanation types on the alternate explanations was assessed in a post hoc analysis. Only

perceived interdependence was identified as a potential mediator of explanation types on self-efficacy. Effects were detected for the causal explanation on trust in the intelligent agent ($d = 0.525$, $[0.212, 0.830]$, $p < 0.001$, $R^2 = 5.3\%$), perceived interdependence ($d = 0.400$, $[0.009, 0.737]$, $p = 0.028$, $R^2 = 3.0\%$), and level of automation ($d = 0.164$, $[-0.148, 0.485]$, $p = 0.312$, $R^2 = 2.3\%$). An analysis of comparative effect on self-efficacy between extraneous cognitive load and perceived interdependence finds less effect for perceived interdependence ($\beta_{\text{difference}} = -0.185$, $[-0.354, -0.016]$, $p = 0.016$). While the effect of the causal explanation on trust in the intelligent agent was not statistically distinct from that of extraneous cognitive load ($\beta_{\text{difference}} = 0.065$, $[-0.107, 0.238]$, $p = 0.229$), trust in the intelligent agent did not have a relationship with self-efficacy in the task ($\beta = -0.012$, $[-0.161, 0.143]$, $p = 0.877$) and therefore is not likely to have a practical effect on self-efficacy either directly or as a mediator of explanations.

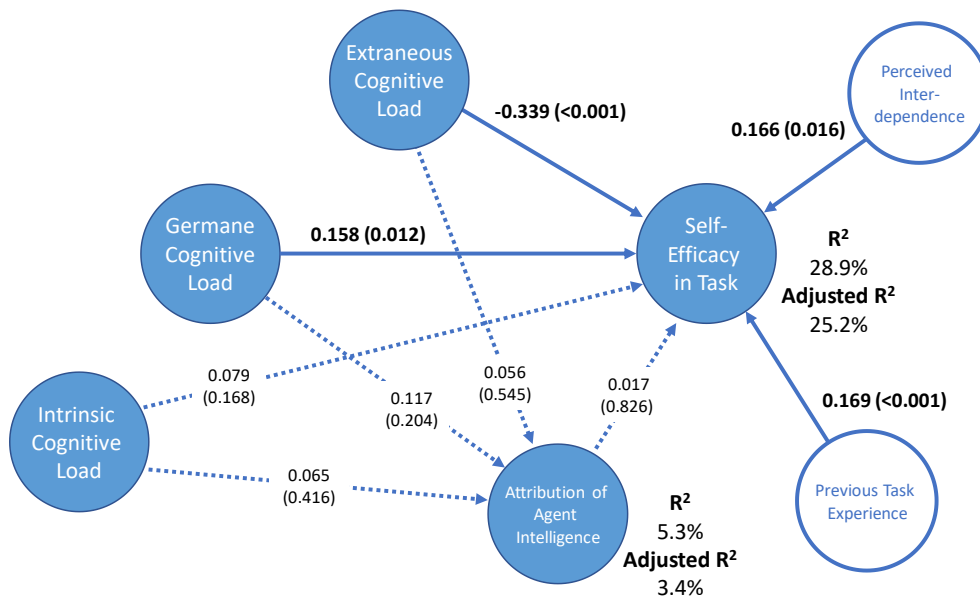


Figure 16 Structural Model Path Analysis Results

Note: Independent variable paths omitted for clarity, see Table 13.

Table 13 Model Path Coefficients

Relationship	Path Coefficient			95% CI		95% CI -1.0 -0.5 0.0 0.5 1.0
	Estimate	T Statistic	p-value	Lower	Upper	
with Attribution of Agent Intelligence						
Intrinsic Cognitive Load	0.065	0.814	0.416	-0.089	0.216	
Germane Cognitive Load	0.117	1.269	0.204	-0.090	0.265	
Extraneous Cognitive Load	0.056	0.606	0.545	-0.135	0.218	
with Self-Efficacy in Task						
Intrinsic Cognitive Load	0.079	1.380	0.168	-0.026	0.196	
Germane Cognitive Load	0.158	2.502	0.012	0.030	0.277	
Extraneous Cognitive Load	-0.339	5.938	<0.001	-0.458	-0.234	
Attribution of Agent Intelligence	0.017	0.220	0.826	-0.132	0.170	
Perceived Interdependence	0.166	2.407	0.016	0.032	0.300	
Previous Task Experience	0.169	4.192	<0.001	0.089	0.247	
Trust in the Intelligent Agent	-0.012	0.155	0.877	-0.161	0.143	
Dispositional Trust for AI	0.037	0.817	0.414	-0.058	0.115	
Learned Trust for AI	-0.024	0.476	0.634	-0.127	0.075	
Level of Automation	-0.032	0.558	0.577	-0.141	0.081	
Age	-0.129	2.535	0.011	-0.231	-0.035	
Gender	-0.048	0.453	0.651	-0.259	0.152	
Education	0.071	1.519	0.129	-0.018	0.164	
Income	-0.027	0.533	0.594	-0.126	0.072	
Causal Explanation on						
Attribution of Agent Intelligence	0.162	0.887	0.375	-0.196	0.506	
Self-Efficacy in Task	-0.226	1.678	0.094	-0.489	0.035	
Extraneous	0.362	2.205	0.028	0.025	0.667	
Germane	-0.240	1.534	0.125	-0.536	0.072	
Intrinsic	0.214	1.242	0.214	-0.123	0.552	
Erroneous Agreement	0.580	3.846	<0.001	0.282	0.872	
Counterfactual Explanation (CF) on						
Attribution of Agent Intelligence	-0.025	0.121	0.904	-0.407	0.383	
Self-Efficacy in Task	-0.038	0.232	0.816	-0.352	0.280	
Extraneous	0.132	0.820	0.413	-0.180	0.446	
Germane	0.204	1.265	0.206	-0.110	0.516	
Intrinsic	0.514	3.266	0.001	0.200	0.818	
Erroneous Agreement	-0.086	0.500	0.617	-0.427	0.254	
Hedging Explanation (H) on						
Attribution of Agent Intelligence	0.380	2.030	0.042	-0.003	0.723	
Self-Efficacy in Task	0.091	0.654	0.514	-0.185	0.357	
Extraneous	-0.053	0.320	0.749	-0.373	0.280	
Germane	-0.001	0.005	0.996	-0.336	0.344	
Intrinsic	0.032	0.169	0.866	-0.334	0.398	
Erroneous Agreement	-0.168	1.008	0.314	-0.492	0.160	
Interaction of CFxH on						
Attribution of Agent Intelligence	-0.383	1.228	0.219	-0.971	0.242	
Self-Efficacy in Task	0.035	0.157	0.875	-0.381	0.480	
Extraneous	0.115	0.492	0.623	-0.328	0.587	
Germane	-0.303	1.252	0.211	-0.774	0.172	
Intrinsic	-0.207	0.871	0.384	-0.667	0.265	
Erroneous Agreement	0.156	0.650	0.516	-0.301	0.643	

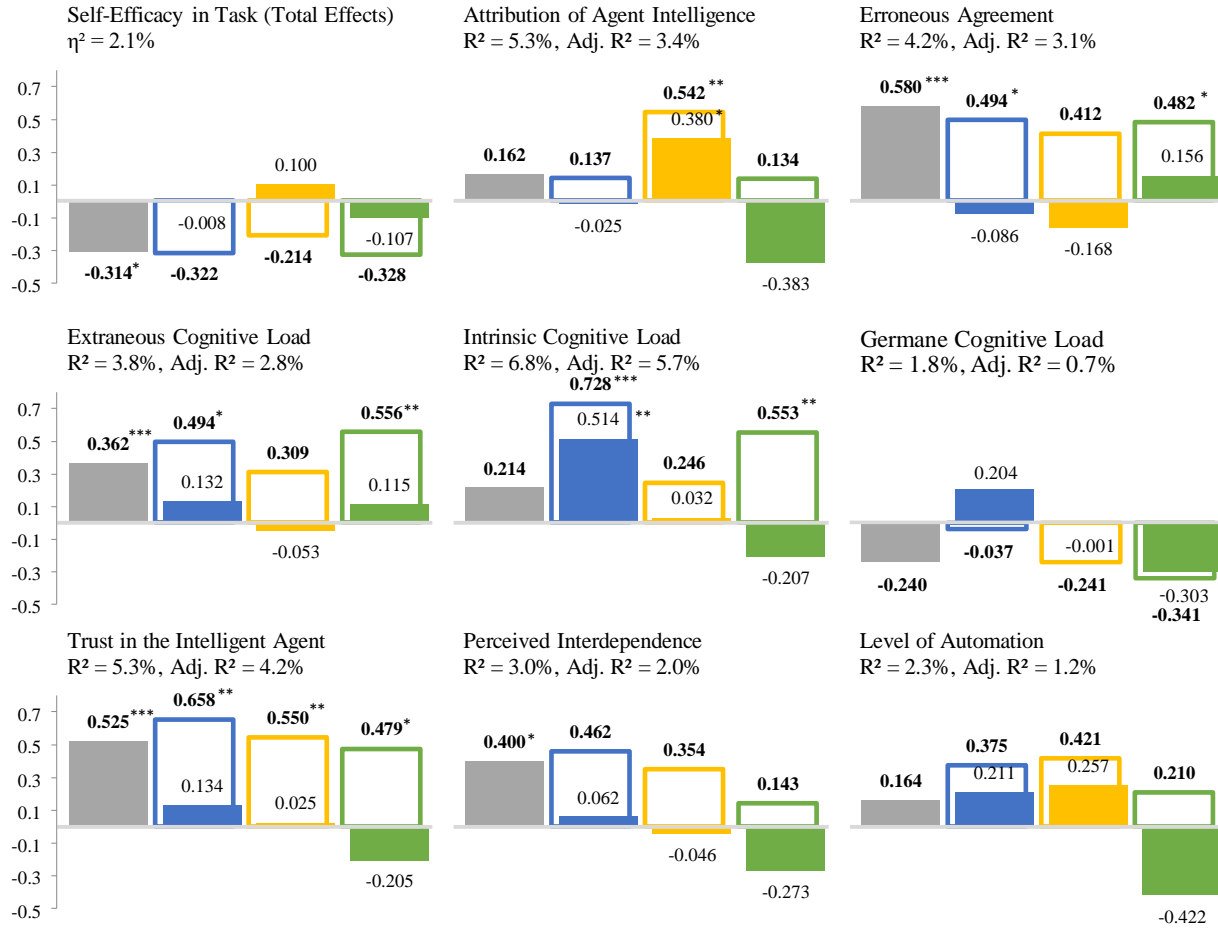


Figure 17 Effects of Explanations Relative to the “Black Box” Condition

Legend: Grey = Causal, Blue = Counterfactual, Yellow = Hedging, Green = Interaction/All Three
 Solid Fill = Individual Effect, Box/Bold = Total of Explanation Effect
 *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$
Note: Coefficients are scaled as in Cohen’s d .

V.6 Manipulation Evaluation

While the analysis of the manipulation check outcomes produced significant differences consistent with overall correct recall of the manipulated states, the size of the groups that did not recall the qualities of the explanation types in their condition was greater than anticipated. This was identified during pre-testing, and was not resolved by modifying the item wording and options (see Appendix F.2.7. for details). To understand the differences between manipulation outcomes, three post hoc analyses were conducted: first, an analysis of the manipulation

outcomes by experiment condition was evaluated to determine whether the combination of explanation types interfered with recall; then, an analysis of variance was conducted on the review step duration, attribution of agent intelligence, and self-efficacy in the task; and finally, a multi-group analysis of the structural model was employed to assess whether the manipulation outcome groups were heterogeneous with respect to any relationships. This evaluation finds that the true manipulation outcomes are consistent with attention to the explanations, with effects of explanation in the true outcome groups consistent with the overall population; however, the hedging explanation was not distinctly identified in the condition with both a causal and counterfactual explanation.

V.6.1 *Interaction with Experiment Condition*

The first step was to evaluate how manipulation outcomes differed by experiment condition. Crosstab analysis with Chi-Square test of independence was used to compare true ratings across experiment conditions. The test of the hedging manipulation identified significant differences (68.7% true ratings: $\chi^2 = 80.5$, $df = 4$, $p < 0.001$) where participants rated the check positive with the presence of other explanation types. For conditions without a hedging explanation, the true negative ratings were 91.5% for the “black box” condition, 57.4% for causal only, but 30.4% with causal and counterfactual. For conditions with a hedging explanation, true positive ratings were 76.9% for the causal with hedging and 86.1% for the all three explanation conditions where the hedging explanation was present. No differences were detected for the causal explanation (93.4% true ratings, $\chi^2 = 3.6$, $df = 4$, $p = 0.466$) and counterfactual explanation (77.5% true ratings, $\chi^2 = 3.5$, $df = 4$, $p = 0.478$).

The condition with all three explanations has the lowest false negative rate for the hedging manipulation check (13.9%) and the counterfactual manipulation check (18.6%).

However, it is not clear that the hedging explanation type was separately perceived by participants in the condition with all types from the manipulation check outcomes alone.

V.6.2 *Effect on Measures*

Next, the effect of manipulation outcome on the study measures was assessed to evaluate differences in responses between groups. Tests of differences in means for review time, attribution of agent intelligence, and self-efficacy in task were conducted using two-way analysis of variance. The counterfactual and hedging manipulation outcomes and the manipulation were entered as fixed factors in separate tests for each outcome variable and explanation type. Group sizes for the causal manipulation outcome did not support the use of this method. Effects were statistically significant for review time (outcome group, $p < 0.001$, $\eta^2 = 0.040$) and attribution of agent intelligence (interaction, $p < 0.001$, $\eta^2 = 0.069$). Effects were not statistically significant for self-efficacy in task (counterfactual outcome $p = 0.385$, interaction $p = 0.946$).

Increases in time for participants in the review step were detected for those in the true outcome groups, compared to the false groups which had little change in review time. These differences, shown in Figure 18, are apparent for each of the three explanation types. The lack of distinction between review times in the false manipulation outcome groups when an explanation was provided is consistent with lack of attention or processing of the explanation. Additionally, review times are greatest in the condition with all three explanation types for the true hedging manipulation outcome group (30.8 seconds versus 28.1 seconds for the causal with counterfactual condition), with 2.7 additional seconds taken to review the added hedging explanation. In the false outcome group, there is an 8.2 second decrease in time taken when all three explanations are presented (21.0 seconds versus 29.2 seconds). These results are consistent with incremental processing time for each added explanation in the true outcome groups

including a combination of the three types, and the false groups being inconsistent with processing the explanations.

The manipulation check outcome identifies opposing effects of the explanation types on attribution of agent intelligence, where the explanations increase the attribution in true outcomes and decrease it in false outcomes. Figure 19 shows these interactions, which also start from opposing baselines for attributions of intelligence. Participants that incorrectly recalled the presence of explanations attributed less intelligence when an explanation was present than when it was not, but participants that recalled correctly increased in ratings when an explanation was provided and had lower ratings than the incorrect group when an explanation was not present. While it is reasonable the false outcome group would attribute greater intelligence in the system, it is unclear why the presence of an explanation would decrease attributions.

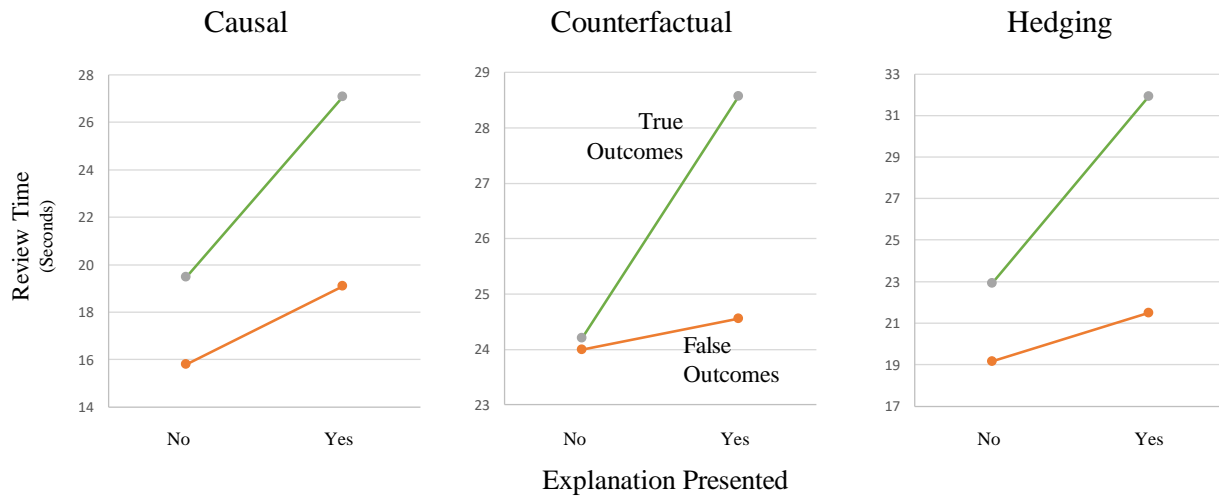


Figure 18 Review Times by Manipulation and Outcome

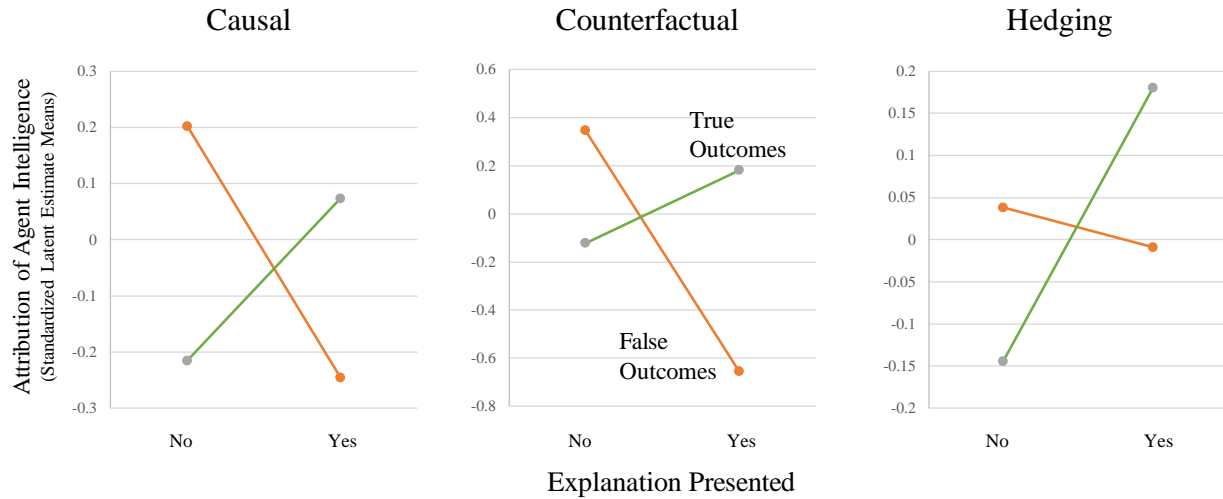


Figure 19 Attribution of Agent Intelligence by Manipulation and Outcome

V.6.3 *Effect on Relationships*

A multi-group analysis of the structural model was conducted to evaluate the potential for differences in relationships between manipulation outcome groups. Several effects of attentional and retention processes were identified, with increasing role of explanations and decreasing role of stable traits and attitudes in the true counterfactual manipulation outcome group, consistent with explanations being processed at some level by participants.

The available sample for false outcomes for the causal condition ($n = 24$) was insufficient for this analysis, and the lack of discrimination of the hedging manipulation check for the counterfactual explanation made it most attractive to evaluate the counterfactual manipulation outcome groups (false, $n = 71$; true, $n = 244$). Model parameters were estimated and compared using the multigroup analysis procedure in Smart PLS. Differences were evaluated using the parametric test. Bootstrapping with 5,000 samples was performed.

While the power of this comparison was limited and confidence intervals in the false outcome group are quite large, two statistically significant differences in path coefficients were identified. The coefficients in the true outcome groups where differences were statistically

significant are consistent with overall results and with processing of explanations. In comparison, the outcomes of the false group have greater role of pre-existing stable attitudes. The counterfactual explanation has the opposite effect on attribution of agent intelligence between groups, with path coefficient shifting from a negative relationship in the false outcome group to positive in the true outcome group ($-0.937, p = 0.002$ to $0.470, p = 0.064$; $\text{diff} = 1.407, p = 0.005$). The effect of counterfactual explanations on intrinsic cognitive load also increases in effect, becoming positive in the true group ($-0.051, p = 0.903$ to $0.740, p < 0.001$; $\text{diff} = 0.791, p = 0.049$). The differences without statistical significance have patterns consistent with lack of attention with increased path strength for stable traits and attitudes such as age, income, dispositional trust and learned trust in AI, and previous task experience. The direction of the effect changes for perceived level of automation and trust in the intelligent agent. Full results are shown in Table 14.

Given these differences, the true outcome group was assessed for differences in support for the hypotheses. While some support was identified for H1b and H5, the wide confidence intervals limit the ability to draw conclusions. The hypothesized positive relationship between counterfactual explanation and attribution of agent intelligence (H1b) has the most practically significant potential with a path coefficient estimate of 0.470 in the true manipulation outcome group. Practically significant coefficients of level 0.15 were identified at the 26th percentile, and 0.25 at the 55th percentile, and the difference with the false outcome group is statistically significant ($-1.407, p = 0.005$). However, the confidence interval includes zero (95% CI [$-0.050, 0.924$], $p = 0.064$). The hypothesized negative relationship between cognitive load and attribution of agent intelligence (H3) decreases from an estimate of 0.329 in the false group to -0.022 (difference, $p = 0.069$). This effect is less practically significant where path coefficients

of -0.15 occur at the 7.5th percentile and -0.25 at the 0.3th percentile. The hypothesized positive relationship between attribution of agent intelligence and self-efficacy in the task (H5) finds some support within the true outcome group; however, the potential effects are also limited in size. Path coefficients greater than 0.15 occur at the 77th percentile and 0.25 at 97th percentile of the true outcome group (95% CI $[-0.135, 0.266]$, $p = 0.431$).

Table 14 Multi-Group Analysis by Manipulation Outcome

Relationship	Path Coefficients		Parametric Test		95% CI		
	False	True	Difference	p-value	-1.0	0.0	1.0
with Attribution of Agent Intelligence							
Intrinsic Cognitive Load	0.043	0.035	-0.007	0.969			
Germane Cognitive Load	0.072	0.144	0.072	0.731			
Extraneous Cognitive Load	0.329	-0.022	-0.351	0.069			
with Self-Efficacy in Task							
Intrinsic Cognitive Load	0.055	0.117	0.062	0.672			
Germane Cognitive Load	0.056	0.160	0.104	0.527			
Extraneous Cognitive Load	-0.242	-0.343	-0.101	0.519			
Attribution of Agent Intelligence	-0.122	0.082	0.203	0.347			
Perceived Interdependence	0.185	0.195	0.011	0.956			
Previous Task Experience	0.199	0.143	-0.056	0.620			
Trust in the Intelligent Agent	0.088	-0.073	-0.161	0.453			
Dispositional Trust for AI	0.168	0.012	-0.156	0.326			
Learned Trust for AI	-0.072	-0.003	0.068	0.627			
Level of Automation	-0.172	0.003	0.174	0.265			
Age	-0.282	-0.104	0.178	0.225			
Gender	-0.049	-0.038	0.011	0.968			
Education	0.032	0.080	0.048	0.710			
Income	-0.102	-0.019	0.083	0.519			
Causal Explanation on							
Attribution of Agent Intelligence	-0.129	0.035	0.164	0.733			
Self-Efficacy in Task	0.125	-0.358	-0.484	0.189			
Extraneous	0.602	0.131	-0.471	0.257			
Germane	-0.159	-0.114	0.045	0.921			
Intrinsic	0.487	0.081	-0.406	0.364			
Erroneous Agreement	0.494	0.679	0.185	0.635			
Counterfactual Explanation (CF) on							
Attribution of Agent Intelligence	-0.937	0.470	1.407	0.005			
Self-Efficacy in Task	0.085	-0.128	-0.213	0.629			
Extraneous	-0.253	0.374	0.627	0.123			
Germane	0.095	0.212	0.117	0.785			
Intrinsic	-0.051	0.740	0.791	0.049			
Erroneous Agreement	0.135	-0.217	-0.352	0.426			
Hedging Explanation (H) on							
Attribution of Agent Intelligence	0.099	0.536	0.437	0.358			
Self-Efficacy in Task	-0.163	0.167	0.329	0.401			
Extraneous	-0.191	-0.119	0.072	0.859			
Germane	0.109	-0.001	-0.110	0.816			
Intrinsic	-0.138	-0.060	0.078	0.877			
Erroneous Agreement	-0.441	-0.097	0.344	0.460			
Interaction of CFxH on							
Attribution of Agent Intelligence	-0.051	-0.680	-0.629	0.390			
Self-Efficacy in Task	0.031	-0.029	-0.059	0.924			
Extraneous	0.333	0.088	-0.245	0.685			
Germane	-0.767	-0.275	0.493	0.453			
Intrinsic	-0.438	-0.111	0.327	0.599			
Erroneous Agreement	-0.093	0.242	0.335	0.599			

Counterfactual Manipulation Groups: Green = True, Orange = False

V.7 Discussion

V.7.1 *Overview of the Study*

Explaining the output of deep learning models aims to improve understanding and may improve the ability for humans to partner with intelligent agents. Natural language explanations have long been advocated. Counterfactual explanations have been touted as matching human thought processes to clarify understanding while providing a contrast to another plausible but inappropriate decision. Hedging explanations describe potentially applicable boundary conditions and known error modes which aims to make these limitations more transparent. There may be an ethical requirement to reveal such limitations, with the potential for future legal requirements to provide them. Past empirical research into explanations in AI has found mixed outcomes for explanations. However, deep learning systems may commit errors that can be readily detected when partnered with a human and explanation may be more valuable than has been identified in research on other types of AI systems. Further, there is relatively little known about how people process agent explanations regardless of type of system, with past research focused on trust, preference, and compliance. A social cognition framework was employed to focus on how explanations are processed by participants performing the task. An experiment was conducted where three types of explanations were manipulated between conditions. Participants were randomly selected into one of five conditions: A “black box” condition with no explanations (as a control), a causal explanation, two conditions where the causal explanation was augmented with a counterfactual or a hedging explanation, and a condition with all three explanation types.

Many of the effects of explanation did not operate as hypothesized and the effects were relatively subtle when present. The results did not support the hypothesis that counterfactual explanations reduced cognitive load (H1a), or that hedging explanations increased cognitive load

(H2a). The results for attribution of agent intelligence also differed from the hypothesized relationships where counterfactual explanations did not increase these attributions (H1b), while hedging explanations increased attributions of intelligence instead of decreasing them (H2b). The effects involving attribution of agent intelligence were insufficient to support the hypothesized relationships between cognitive load and self-efficacy with attribution of agent intelligence (H3 and H5). Germane cognitive load had a positive relationship with self-efficacy; however, it was not clear this was related to the explanations. Causal explanations increased extraneous cognitive load and led to decreased self-efficacy in the task, supporting the hypothesized mediation relationship between cognitive load and self-efficacy (H4). While trust and compliance increased when explanations were provided, these were not identified as having positive relationships with self-efficacy.

As in prior research, the causal explanation increased both trust and compliance (in the form of erroneous agreement). This indicates that explanations were processed similarly to prior studies. However, this study offers additional evidence of the depth to which the explanations were processed by participants to influence their judgments. The four processes of observational learning from Bandura (1986) provide specific outcomes to evaluate engagement with the task and the explanations. Attention and retention processes can be partially validated by the true manipulation outcomes where the effects are generally greater but consistent with the overall results. In comparison, those that did not correctly recall the explanation types in their condition have review times consistent with the “black box” condition without explanations, and greater relationships with stable attitudes and traits. These outcomes are consistent with participants finding functional value in explanations to trigger the storage of the nature of the explanations for later retrieval. Production processes are not as clear in the absence of a quality metric for

faithful agent assessments; however, the consensus ratings for damage assessments were highly repeatable. Only one scenario failed to reach a strong consensus, with responses split between two adjacent categories for the damage rating. This makes it plausible that workers referenced the written guideline and may have learned from the agent during the review step between scenarios. While some participants provided open-text feedback indicating they worked through explanations³, and that was similarly identified during cognitive interviews in the testing of the instrument, these participants were either induced or self-motivated to monitor and evaluate their cognitive processes during the task. The extent to which other participants engaged in these meta-cognitive processes cannot be determined. However, it is more plausible that the task design did not engage these processes rather than these other participants not having the capacity to use the causal and counterfactual explanations. The feedback also provides insight into motivation processes. Eighty-three participants (23%) provided comments in the optional field indicating the task was compelling. Additionally, 20% of participants voluntarily went beyond the 25-minute maximum stated time for the task, with some noting in the feedback that they spent extra time due to the importance of the task. Together, these results are consistent with observational learning pathways being active for at least a subset of the participants and explanations being processed to a similar extent to prior studies despite finding subtle effects.

V.7.2 Contributions to Theory

The effect of explanation on self-efficacy is greater than that of previous task experience, and was confirmed to be mediated by cognitive load rather than attribution of agent intelligence. Perceived interdependence was also identified as a potential mediator of self-efficacy. Insights from cognitive load areas may inform solutions that answer the call by Nunes and Jannach

³ Example quotes from open-text feedback are provided in Table 5.

(2017) for responsive explanations that are tailored to users. These findings enhance knowledge about the types of cognitive load imposed by the evaluated types of explanations as well as effects on attributions of intelligence. Interdependence was not originally hypothesized or manipulated extensively by the experiment and was measured with just two items. The literature on interdependence in human-robot systems offers methods of breaking apart tasks to achieve objectives that are complementary to the goals of instructional task design from cognitive load literature. While changes in the perception of interdependence did not necessarily cause increased self-efficacy, this association offers a second area to explore as tasks are broken apart to manage cognitive load.

The findings identify limited effects of the explanation types on attributions of intelligence to the agent. Much of the prior research in explanations has examined trust, and these attributions had a large correlation with trust in the intelligent agent. Explanations were expected to increase the predictability and apparent rationality of the agent by revealing hidden internal processes of the agent, especially in assessments providing erroneous output. The causal and counterfactual explanations did not result in distinct increases in attribution of agent intelligence, even compared to the “black box” condition. While it was anticipated that hedging explanations would appear to participants as weakness and lack of intelligence, the ability to describe failure modes, admit weaknesses, and be self-aware had a large positive effect on perception of mental processes. The lack of effect of attribution of agent intelligence as well as trust on self-efficacy is consistent with the known phenomenon of people generating internal attributions for causes of success (Weiner, 1985). However, the true manipulation outcome group for counterfactual explanations provides some indication that attention and more effective

integration of explanations into the task might increase the role of attribution of agent intelligence in self-efficacy.

The findings also provide evidence of the types of cognitive load induced by each of the explanation types. Counterfactual explanations were hypothesized to assist the construction of mental schema for the task by focusing participants on the relevant elements of the guideline and reducing cognitive load. The perception of intrinsic cognitive load increased, along with, to a lesser extent, extraneous cognitive load. As increased intrinsic load was not associated with decreases in self-efficacy or erroneous agreement, counterfactual explanations may have assisted participants in constructing more accurate mental schemas as expected, in order to become more aware of the inherent difficulty of the task. The lack of the hypothesized decrease in cognitive load might be explained by the short duration of the task. Hedging explanations were expected to increase extraneous cognitive load; however, there was little effect. The explanations may have been largely unprocessed, as evidenced by only slight increases in processing time despite the substantial increase in difficulty to process them compared to the other explanation types. There is also some indication that hedging explanations improved the contextual application of trust consistent with expectation in the literature, though the effect was small and did not fully offset increases in erroneous agreement. While some prior research has recommended combining types of explanations to achieve the best qualities of each, when all three explanation types were presented the result was the highest extraneous cognitive load and essentially no benefit for erroneous agreement, countering the beneficial effects of the hedging explanation.

The results also provide some insight into the nature of erroneous agreement. The increase in erroneous agreement when explanations were provided is inconsistent with an illusion of explanatory understanding and instead more consistent with workers “going along”

with the agent. Participants were at some level aware of decreased competence having rated self-efficacy in the task lower with increasing erroneous agreement, which also co-occur with increasing perceived trust, automation, and interdependence. This adverse effect was mostly driven by the causal explanation. These explanations may have decreased the participant's suspicion of the agent's output leading to greater role of cognitive biases, while counterfactual and hedging explanations provided cues to suspect the simulated output.

V.7.3 Practical Implications

This study provides insight into whether counterfactual and hedging explanations would be beneficial for partnering with an agent in contexts where both humans and agents are learning. Utilizing all of the explanation types to gain the advantages of each is tempting, but many beneficial effects of counterfactual and hedging explanation were counteracted when combined. Individually adding the counterfactual or hedging explanation to the causal explanation improved some outcomes. Counterfactual explanations increased awareness of the complexity of the task with little negative effect on cognitive load or erroneous agreement. Hedging explanations had the smallest decrease in self-efficacy as a result of explanations and the lowest erroneous agreement; however, these effects were quite small, and the large increase in attribution of intelligence by itself has little benefit in other attitudes and outcomes. While adding only a counterfactual explanation to causal achieves the greatest improvement in outcomes, the “black box” condition with no explanations has the lowest erroneous agreement and highest self-efficacy outcomes.

These results also may shed light on why projects seeking to partner humans with intelligent agents frequently experience difficulties, despite best-efforts by the stakeholders. The workers in this study rated themselves as confident in performing the task, holding trust in the

agent, and broadly provided positive open-text feedback. While self-efficacy decreased when explanations were provided, the absolute ratings were still most often greater than neutral. However, the detectable performance outcomes were less than ideal. The workers in this study did not have the opportunity to compare or select between conditions. Based on the results of other studies of preference for explanation, it is quite plausible workers would have preferred an agent that offered explanations over working with a “black box,” and thereby unknowingly advocated for a configuration with inferior outcomes. “Explainability” will also be a compelling feature in sales presentations of future intelligent systems; however, managers should evaluate the level of support for the effectiveness of explanations in the system.

In a real-world system it is likely that users will request that the review process be simplified by presenting the intelligent agent’s assessment immediately and perhaps only revealing an explanation upon request. The usage of the incorrect rating checkbox in the “black box” condition indicates that roughly half of workers might have sought an explanation of the simulated output in scenarios where the output was intendedly erroneous. Additionally, only 5% of ratings where workers agreed with the agent assessment in the initial step marked the agent incorrect across both faithful and erroneous scenarios. This suggests that agreement with the agent was sufficient to suppress critical evaluation of the simulated output. If the participants in this study performed the review in a single step, participants might be even more likely to “anchor” onto the agent’s assessment without conducting their own independent evaluation. A single-step review may be more appropriate when model output is already known or suspected to be incorrect, compared to tasks involving screening for erroneous output.

V.7.4 *Limitations*

The “Wizard of Oz” method employed in this study may not accurately represent any current or future system, and the “Wizard” in this study was the researcher. The damage rating task does not allow for certainty in the true rating of an image, compounding the challenge of replicating system performance. The simulated agent in this task had an intended error rate consistent with that of a deep learning model being deployed on a new data set. However, the error rate was higher than most users might be expected to accept from an intelligent system. While the erroneous agreement measure is useful to compare with other measures, there are substantial limitations in its interpretation. There was no quality metric to evaluate it in comparison to correct assessments, and it also does not necessarily reflect the on-going rate of error. Additionally, it is possible that the agent assisted many workers in reaching more appropriate classifications which was not detectable by this design. As such this study provides limited insight into the rating performance of the participants or explaining interaction with an intelligent agent that is most often correct.

The association between interdependence and self-efficacy was identified after the fact as an alternate explanation without a hypothesis. This measure only evaluated the perception of interdependence and used only two items which is not ideal for PLS-SEM structural models. As such, this finding requires further investigation to confirm that the manipulation of interdependence will impact self-efficacy in a task.

The workers in this study received no feedback on their performance beyond the simulated output and explanations to inform their judgment and damage assessment. While the lack of feedback is consistent with the typical crowdsourced damage assessment task, it is well-known that the feedback mechanisms in a typical work environment are a critical input to regulate behavior and establish self-efficacy.

While the outcomes of this study in terms of trust and erroneous agreement are not substantially different than in prior studies of explanations with experts, this study is unlikely to generalize to the information processing of explanations by experts who are highly familiar with a task. The information processing pathways of experts will not be loaded by learning the task. Essentially none of the participants had previously used the damage guideline provided, and only half had previously conducted damage assessments using aerial images. Participants were not calibrated in the task prior to data collection. The study also offers limited insight into long-term interactions and learned trust with an explainable intelligent system where expertise is developed with the system itself. While other studies in explanation and intelligent systems have involved interacting with a system over several hours or even multiple days, this study involved a single interaction with most participants having less than ten minutes of interaction with the task.

Finally, this study did not experimentally evaluate multiple user-interface configurations or the sequencing of user choices. The task design was unconventional in requiring the worker to conduct their assessment and provide their input before receiving the simulated output. It is possible that participants would have been inclined to analyze more deeply an explanation that they requested over one presented to them for review. The explanations were also not specifically briefed to participants and they were not highlighted by any training, which could greatly increase the strength of the manipulation. By selecting design choices considered optimal across multiple research articles that had not previously been integrated, over-specification may have resulted. It was also impractical to fully optimize the Qualtrics-based user interface, and users were required to scroll the screen more than the typical Pybossa user-interface would have required.

V.7.5 Future Research

Both cognitive load and perceived interdependence can be influenced by the design of the task, and were both found to be of similar or greater effect as previous task experience. Notably, cognitive load has been utilized in the computer-based instructional literature to develop improved tasks. The review by Hollender et al. (2010) provides a survey of methods which may be immediately transferrable to an XAI context. Interdependence also offers an existing literature within human-computer interaction to leverage. Several participants provided comments in the open-text feedback indicating that they would have liked feedback on ratings or the ability to provide specific feedback on the agent's ratings. Research in human-robot interaction has specifically considered the design of joint activity, and the "directability" requirement for interdependence from the coactive design method (Johnson et al., 2014) could be evaluated for its ability increase the effectiveness of explanations. The ability to provide feedback also has the potential to both break the task apart to decrease extraneous cognitive load in each step, while activating the production pathway of observational learning from social cognitive theory.

Researchers should consider utilizing a repeated measure design and within-subject longitudinal measurement. These designs could evaluate the effects of sequencing of multiple task configurations to determine whether configuration or competency through experience activates effective meta-cognitive monitoring and evaluation processes. This would decrease required sample size by examining effects on the same participant. Additionally, such a design offers the ability to identify the extent that "think aloud" verbal protocols induce meta-cognitive processes. Research in cognitive load and self-efficacy frequently uses a repeated measure approach to address large variance in measures between subjects (Beckmann, 2010).

This study attempted to model cognitive load as a hierarchical composite construct combining the three theoretical components of cognitive load. The resulting model estimated

path coefficients for the second-order construct that were roughly consistent with theory, but variation in the measure was not associated with study measures other than gender. The criticality of the single-item total cognitive load in the original measurement model was recognized in advance, and this item was presented as the first measure after the rating task on its own screen with the recommended wording, scale points, and labels from the literature. Further, correlations between the three-component measure cognitive load and dispositional trust might suggest some interactions between pre-existing attitudes and cognitive load. Previous research in cognitive science has directly detected a role of trust on objectively measured cognitive load (W.-L. Hu, Akash, Jain, & Reid, 2016). However, other research in explanation using mixed methods also identified that observed behaviors differed from subjective questionnaire rating results (Holliday et al., 2013). Future research that isolates the effects of changes in behavior and subjective ratings of the same concept can inform the interpretation of subjective cognitive load measurement.

VI CONCLUSIONS

A large area of research is developing deep learning systems capable of explaining their output in natural language as a means to increase transparency and improve human understanding when evaluating system output. Intelligent agents that utilize deep learning neural networks to analyze images have reached a high level of demonstrated accuracy in classification tasks, but their performance is known to decrease substantially when utilized on a new data set. Explanations are expected to improve the ability for humans to evaluate model performance, and where the model is more capable may improve human understanding. However, much of the empirical research on the effectiveness of explanations has evaluated trust in the system and compliance with system output, which may not be appropriate when humans are intended to partner with a system to produce judgments.

In this study participants reviewed images taken before and after a natural disaster and classified the type of structure and degree of damage. After rating each image, a simulated agent's rating was provided for their review along with the opportunity for the participant to modify their rating. A written guideline was provided, which the explanations referenced. Three types of explanations of those ratings were manipulated between participants: "causal" which offer reasons the classification selected was appropriate, "counterfactual" which offer why another classification was not appropriate, and "hedging" which offer hypothesized failure modes and boundary conditions related to an image. The research and measurement model focused on the effects of the explanations on the participant. Self-efficacy was chosen as the key outcome due to its ability to predict future engagement and investment of mental effort. Manipulation checks and testing of the instrument indicated that participants largely retained and recalled the qualities of the types of explanations they received.

Causal explanations increased trust and compliance with the simulated agent's ratings, consistent with prior research. However, they also increased extraneous cognitive load, increased agreement with erroneous agent ratings, and decreased self-efficacy in the task. When a counterfactual explanation was added to the causal explanation the perception of intrinsic cognitive load increased without an effect on self-efficacy, suggesting that they assisted the participant in assessing the complexity of the task. However, they did not substantially offset the increase in erroneous agreement of causal explanations. When a hedging explanation was provided with the causal explanation the perception of agent intelligence increased but had little benefit on other outcomes. Providing all three types of explanation resulted in adverse outcomes on most study measures. Participants that changed ratings to agree with an erroneous agent assessment rated lower self-efficacy indicating they were likely "going along" with the provided rating rather than experiencing an illusion of explanatory understanding.

Two areas of existing research were identified with the potential improve the effectiveness of explanations. The limitations of the human cognitive architecture can be managed by breaking tasks into smaller elements to optimize demands on working memory, utilizing existing knowledge within human-computer interaction and cognitive load theory (Hollender et al., 2010). Perceived interdependence was also found to have a positive relationship with self-efficacy, and methods to increase interdependence also start by breaking tasks apart with the goal of increasing the human and the agent's ability to observe, predict, and direct each other (Johnson et al., 2014). The effects of cognitive load and perceived interdependence were greater than that of previous experience in the task but are under the control of the designer of the task. In comparison, the most commonly evaluated construct, trust in the intelligent agent, was not found to have a significant relationship with self-efficacy.

APPENDICES

Appendix A: IRB Approval

INSTITUTIONAL REVIEW BOARD

Mail: P.O. Box 3999
Atlanta, Georgia 30302-3999
Phone: 404/413-3500
Fax: 404/413-3504

In Person: Dahlberg Hall
30 Courtland St, Suite 217



January 08, 2019

Principal Investigator: Pamela Ellen

Key Personnel: Donthu, Naveen; Dougherty, Sean E; Ellen, Pamela

Study Department: Georgia State University

Study Title: Crowd-sourced Damage Assessment in Natural Disasters

Submission Type: Exempt Protocol Category 2

IRB Number: H19315

Reference Number: 352782

Approval Date: 12/31/2018

Expiration Date: 12/30/2021

The above referenced study has been determined by the Institutional Review Board (IRB) to be exempt from federal regulations as defined in 45 CFR 46 and has been evaluated for the following:

1. determination that it falls within one of more of the six exempt categories allowed by the institution; and
2. determination that the research meets the organization's ethical standards

If there is a change to your study, you should notify the IRB through an Amendment Application before the change is implemented. The IRB will determine whether your research protocol continues to qualify for exemption or if a new submission of an expedited or full board application is required.

Exempt protocols must be renewed at the end of three years if the study is ongoing. When the study is complete, a Study Closure Form must be submitted to the IRB.

Any unanticipated/adverse events or problems resulting from this investigation must be reported immediately to the University Institutional Review Board. For more information, please visit our website at www.gsu.edu/irb.

Sincerely,

Shelia L. White, IRB Member

Federal Wide Assurance Number: 00000129

Appendix B: Test Instrument

See following page.

Consent

Georgia State University Informed Consent

Title: Crowd-sourced Damage Assessment in Natural Disasters
Principal Investigator: Dr. Pam S. Ellen
Student Principal Investigator: Sean E. Dougherty

Procedures

You are being asked to take part in a research study. If you decide to take part, you will be asked to rate the damage in ten images. We will also ask your opinions related the task. You have been invited because of your background in either crowd-sourced citizen science or aerial image interpretation. This survey should take less than 25 minutes of your time in one sitting. As this study will be completed online, your IP address will be recorded. However, this data will be destroyed when data collection is complete. The results of the study will be summarized and reported in group form. You will not be identified personally. Your name and other facts that might identify you will not appear in this study.

Benefits

Your participation may provide data that improves the speed and quality of disaster relief efforts.

Risks

There is no physical risk in undertaking the survey beyond a normal adult day. The disaster images are no more disturbing than would appear in a newspaper or on network television news.

Compensation

The compensation will be \$2.00 for your participation in the study. Survey completions with obvious signs of lack of participation by completion time or failed attention checks will be rejected.

Voluntary Participation and Withdrawal

You do not have to be in this study. You may stop participating at any time by closing this window.

Contact Information

Sean Dougherty at (813) 344-5408 or sdougherty5@student.gsu.edu.

Consent

If you are willing to volunteer for this research, please start the survey by clicking next below.

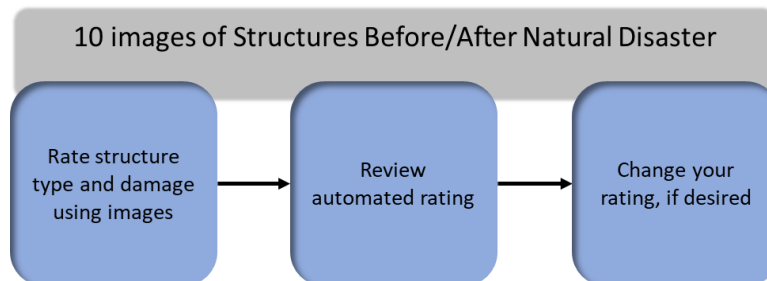
Briefing

Disaster Damage Assessment

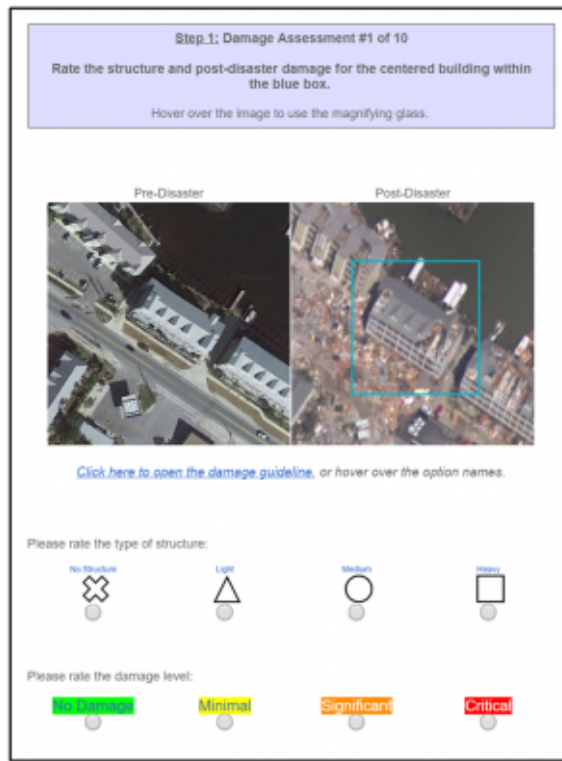
Crowdsourcing has been used for many years to rapidly assess damage after natural disasters. The results have been used to inform disaster relief officials of the extent of damage and areas to focus relief efforts. Because of the volume of data and urgent time sensitivity, automated damage assessments are being tested. Combining crowdsourcing with automated damage assessments is expected to increase quality and speed, but both the automated system and crowd workers must learn the unique aspects of each disaster in a partnership.

Your task is to review 10 images of buildings and classify the type of building and the extent of damage using the guideline below. Rate only the building at the center of the image as there may be more than one building. If you place your mouse cursor over the image you can magnify sections for more detail. After you submit your rating, you will receive the automated damage assessment for your review. Please evaluate this assessment and, if necessary, review the images again. You can change your ratings, if desired. If you do not believe the image is of a building, please mark "no structure".

Overview of Process






Screenshot Example of a Damage Assessment



In previous disasters it has been found that having a clear damage rating guideline is very important for data quality. The Harvard Humanitarian Institute's guidelines are shown below. You can recall this guide at any time while reviewing images.

Harvard Humanitarian Initiative's Aerial Imagry Interpretation Guide

Structure Type	Symbol	Description
Light		Structures that are built predominantly from light material or locally sourced materials. These structures may be mobile or possess no real hard roof, in some cases, roofs are made of metal or light material; they are often small in size. As such, these structures are likely to be the most vulnerable structures in any impacted region. Examples of these types of structures can include huts, tukuls or mobile trailers.
Medium		Structures that are built from semi-hard materials or mixed products. These structures have solid frames built using wood, steel or cement. These type of structures are fixed and possess hardened walls and roofs which can be made out of wood or cement. Unlike light structures, these types of structures are able to withstand moderate level of wind, with no to little damage, while maintaining their structural integrity. These types of structures can be individual or multi-family houses, small stores, places of worship and similar structures.
Heavy		Structures that are built from hard materials such as reinforced cement and steel. Infrastructure of this type is the least structurally vulnerable in any observed region. These structures are designed to withstand high level winds without receiving heavy damage or endangering the structural integrity of the structure. In many areas, these may include multiple story buildings, strip malls, hospital buildings, or public utilities.

Damage Classification	Color	Description

No Visible Damage		The roof is virtually undamaged and the walls, in effect, remain standing. The structure appears to have complete structural integrity and does not appear to need repair.
Minimal Visible Damage		The roof remains largely intact, but presents partial damage to the roof's surface, with minimal exposure beneath. In oblique aerial and satellite imagery, minimal damage may be able to be observed within the structure and to the exterior walls. The structure appears to have general structural integrity but needs minor repairs.
Significant Visible Damage		The roof is entirely damaged or missing. The walls of the structure remain upright. However, the interior wall partitions can be partially damaged. Debris inside the structure can also potentially be visible. The structure does not appear to have complete structural integrity and is in need of significant repair.
Critical Visible Damage		The roof is completely destroyed or missing, and the walls have been destroyed or collapsed. The support structures are completely leveled, and interior objects have also suffered visibly heavy damage or destruction. The structure does not appear to have any structural integrity and requires comprehensive reconstruction or demolition of the entire structure

Briefing Attention Check

Based on the briefing, select the statement which is true.

- You are to ignore the type of structure when making your assessment.
- You may change your rating after reviewing the automated assessment.
- You will be asked to estimate a value for the property.
- You will upload pictures you've taken of damaged buildings.

Sharing Notice

Please note:

This is a research study of the damage assessment task. Please do not share the contents of this study with other potential participants.

Rating Block

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

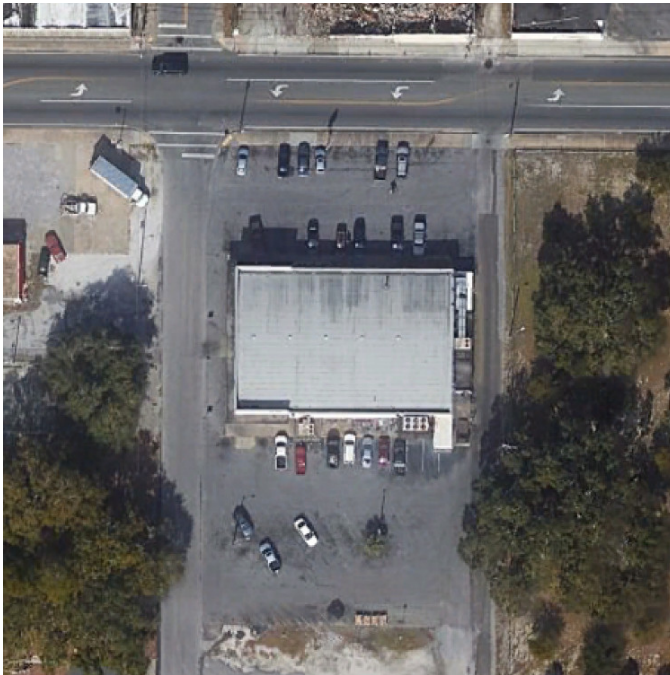
Click Count: 0 clicks

Step 1: Damage Assessment #1 of 10

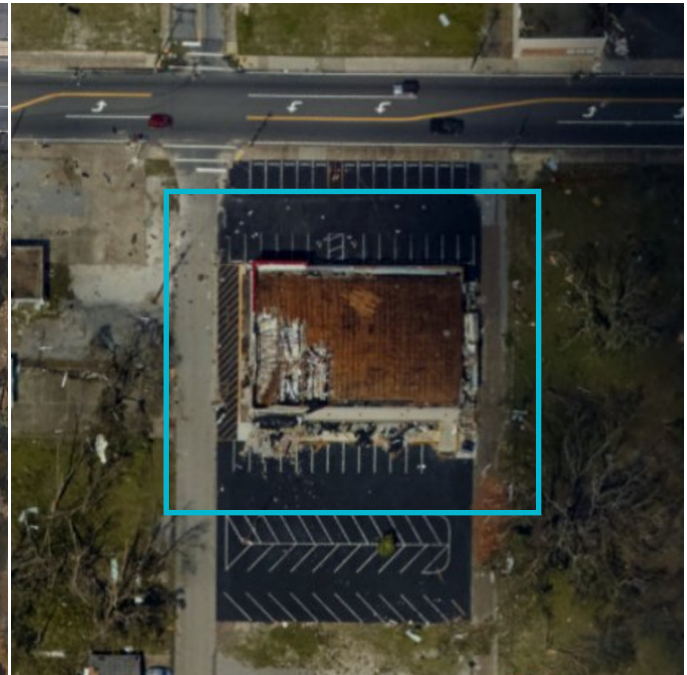
Rate the structure and post-disaster damage for the centered building within the blue box.

Hover over the image to use the magnifying glass.

Pre-Disaster



Post-Disaster



[Click here to open the damage guideline](#), or hover over the option names.

Please rate the type of structure:

No Structure



Light



Medium



Heavy



Please rate the damage level:

No Damage



Minimal



Significant



Critical



Please rate the difficulty of this classification:

Very Difficult



Very Easy



These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds

Last Click: 0 seconds

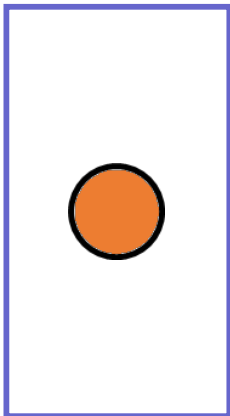
Page Submit: 0 seconds

Click Count: 0 clicks

Step 2: Review Automated Assessment #1

Hover over the image to use the magnifying glass.

Your Assessment



Automated Assessment

This structure was automatically classified Medium Critical.

Because it is a Medium Commercial Building with collapsed roof and internal visible debris and collapsed wall with adjacent debris.

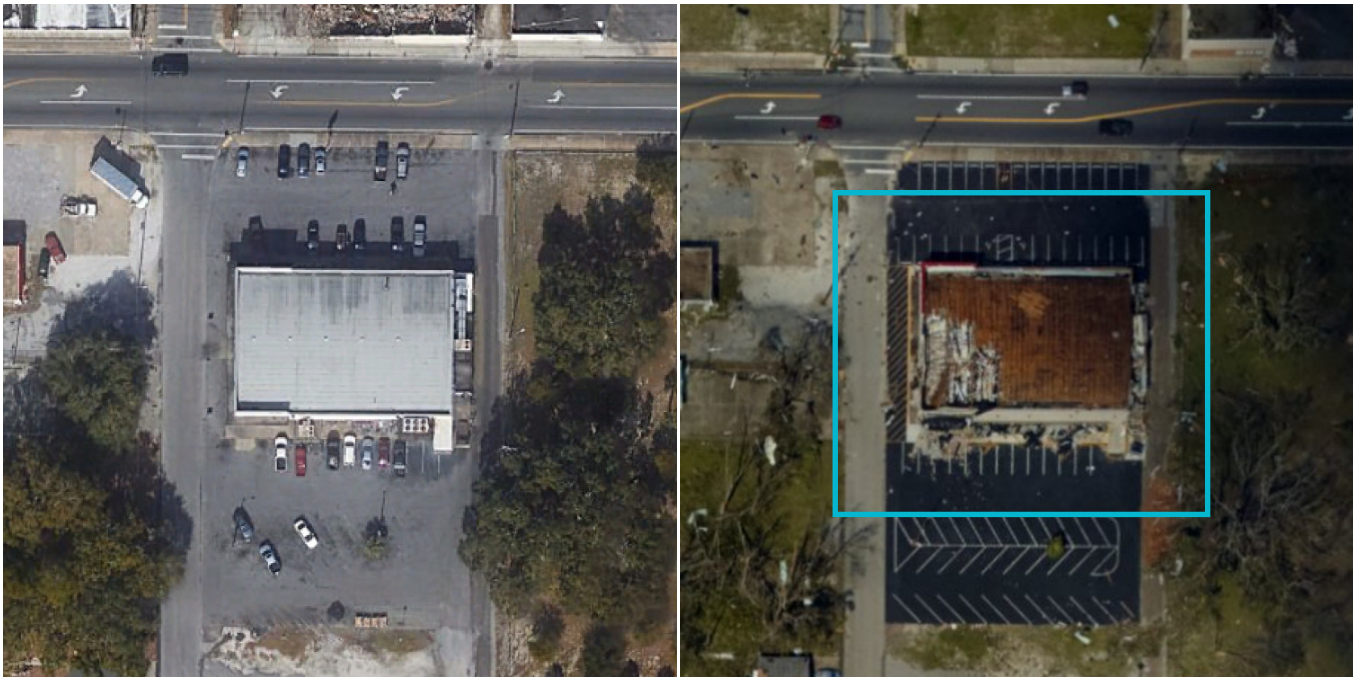
It is not Heavy Significant because of wooden building materials and it does not have intact walls.

Consider: Bodies of water, parking lots, and shadows may be misclassified as damage.

The images are below if you would like to review them again:

Pre-Disaster

Post-Disaster



[Click here to open the damage guideline \(new tab\).](#)

If you'd like to change your answer, please select a new structure type: **(OPTIONAL)**

» No Structure



» Light



» Medium



» Heavy



If you'd like to change your answer, please select a new damage rating: **(OPTIONAL)**

» No Damage



» Minimal



» Significant



» Critical



Please indicate if any of the following apply in your opinion:

- The structure was hidden or obscured
- The image was blurry
- The automated assessment was incorrect

Manipulation Checks

Please rate your opinion of the following statements.

The automated assessor explained why it made its ratings.

Strongly disagree Somewhat disagree Neither agree nor disagree Somewhat agree Strongly agree

The automated assessor contrasted its rating to another possible classification.

» Strongly disagree » Somewhat disagree » Neither agree nor disagree » Somewhat agree » Strongly agree

The automated assessor provided features of the image to consider which could cause incorrect classifications.

» Strongly disagree » Somewhat disagree » Neither agree nor disagree » Somewhat agree » Strongly agree

Attitudes

How much mental effort did you invest in making your assessments?

Very, Very Low Very Low Low Rather Low Neither Low nor High Rather High High Very High Very, Very High

Please rate your opinion on the following statements about the damage assessment task:

	Strongly Disagree						Strongly Agree
Many things needed to be kept in mind simultaneously when rating damage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The damage assessment task was very complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I made an effort to understand the overall task and not just on the details.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I wanted to make sure I understood everything I was provided while completing the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Strongly Disagree						Strongly Agree
I was provided information which supported my ability to assess damage ratings.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was exhausting to find the important information to assess damage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The design of this assessment tool was very inconvenient for making ratings.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was difficult to recognize and link the crucial information while assessing damage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I had to remember many things to perform the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Based on your experience, please rate your expectations if you did more damage assessments using the tool:

	Strongly Disagree						Strongly Agree
Knowing my skills and abilities, I think I can do well assessing damage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I expect I can do well on future damage assessments with this tool	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe I can produce high quality assessments with this tool	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm confident I understand how to use the damage assessment guide	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm certain I have the skills necessary for damage assessment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm confident I understand the most difficult categories	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm confident I know how to use the damage assessment tool	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe I won't overly rely on the automated assessment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm confident the automated assessment won't distract me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please rate the automated assessor:

	Strongly Disagree						Strongly Agree
Thinks logically	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is knowledgeable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Strongly Disagree						Strongly Agree
Is able to make decisions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has a mind	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is predictable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is insightful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Alternate Explanations

Please estimate the number of aerial damage images you've assessed prior to this study.

No Experience
 Less Than 10 Images
 Between 10 and 50 Images
 Between 50 and 100 Images
 Between 100 and 500 Images
 Over 500 Images

How often do you rate damage to structures in aerial images?

Never
 Yearly, or less
 A few times a year
 Monthly
 A few times a month
 Weekly
 Daily

Have you previously used a written damage guideline to make ratings?

No
 Yes - But a different guideline.
 Yes - The same guideline as here.

Please rate the automated assessor on the following statements:

	Strongly Disagree						Strongly Agree
I believe I can trust the automated assessment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My assessment was affected by the automated assessor's input	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assessments depend on both worker and automated assessor for accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My ratings benefitted by working with the automated assessor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The damage assessment was:

Highly Manual



Highly Automated

Please rate your opinion of the following statements:

	Strongly Disagree						Strongly Agree
It is difficult for me to contain my feelings when I see people in distress.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am often reminded by daily events how dependent we are on one another.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel sympathetic to the plight of disaster victims.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please rate your opinion of the following statements:

	Strongly Disagree						Strongly Agree
I like to use technology to make tasks easier for me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have bad experiences when I try to use new technology instead of doing things "the old-fashioned way"	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
There are tasks in my life that have been made easier by computers doing the work for me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have a positive view of the potential for robots and artificial intelligence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Demographics

The following questions are for descriptive purposes only.

In what year were you born?

Are you... ?

- Male
 Female

What is the highest level of formal education you have completed?

- Less than high school degree
 High school graduate (high school diploma or equivalent including GED)
 Some post-high school
 Bachelor's degree
 Some graduate school
 Graduate degree

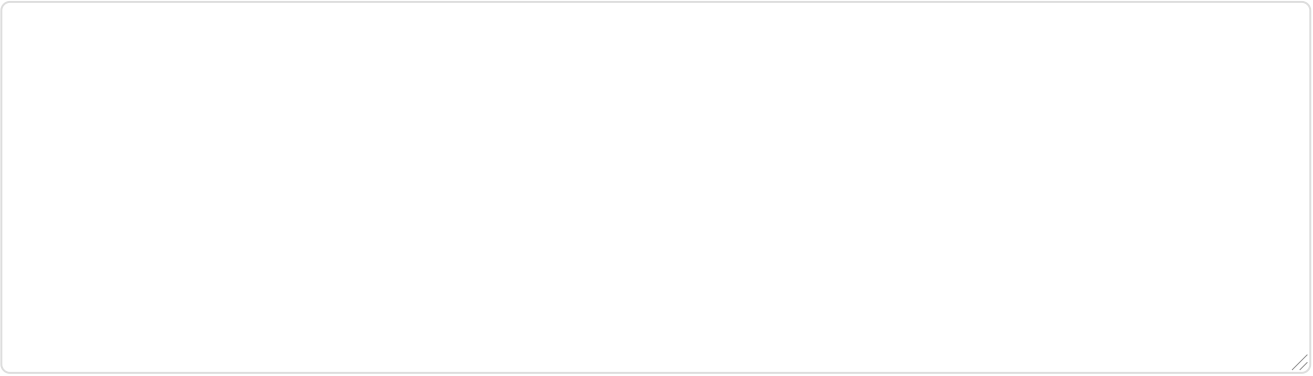
Which of the following best describes your total family income?

- | | | |
|--|--|--|
| <input type="radio"/> Less than \$15,000 | <input type="radio"/> \$45,000 to \$59,999 | <input type="radio"/> \$100,000 to \$149,999 |
| <input type="radio"/> \$15,000 to \$29,999 | <input type="radio"/> \$60,000 to \$79,999 | <input type="radio"/> Over \$150,000 |
| <input type="radio"/> \$30,000 to \$44,999 | <input type="radio"/> \$80,000 to \$99,999 | <input type="radio"/> Prefer not to Say |

Feedback

How did you find out about this HIT? if you found out on a web forum, please paste a link to the page.

Finally, we value any feedback concerning this survey. Please let us know your thoughts and or suggestions.



Appendix C: Study Instrument

See following page.

Consent

Georgia State University Informed Consent

Title: Crowd-sourced Damage Assessment in Natural Disasters
Principal Investigator: Dr. Pam S. Ellen
Student Principal Investigator: Sean E. Dougherty

Procedures

You are being asked to take part in a research study. If you decide to take part, you will be asked to rate the damage in ten images. We will also ask your opinions related the task. You have been invited because of your interest in crowd-sourced citizen science. This survey should take less than 25 minutes of your time in one sitting. As this study will be completed online, your IP address will be recorded. However, this data will be destroyed when data collection is complete. The results of the study will be summarized and reported in group form. You will not be identified personally. Your name and other facts that might identify you will not appear in this study.

Benefits

Your participation may provide data that improves the speed and quality of disaster relief efforts.

Risks

There is no physical risk in undertaking the survey beyond a normal adult day. The disaster images are no more disturbing than would appear in a newspaper or on network television news.

Compensation

The compensation will be \$2.00 for your participation in the study. Survey completions with obvious signs of lack of participation by repeated random damage ratings or failed attention checks will be rejected.

Voluntary Participation and Withdrawal

You do not have to be in this study. You may stop participating at any time by closing this window.

Contact Information

Sean Dougherty at (813) 344-5408 or sdougherty5@student.gsu.edu.

Consent

If you are willing to volunteer for this research, please start the survey by clicking next below.

Briefing

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

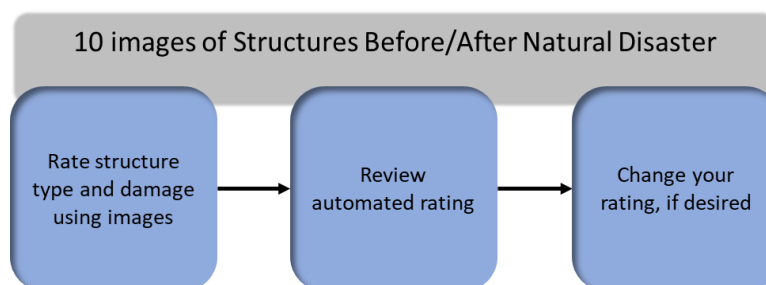
Click Count: 0 clicks

Disaster Damage Assessment

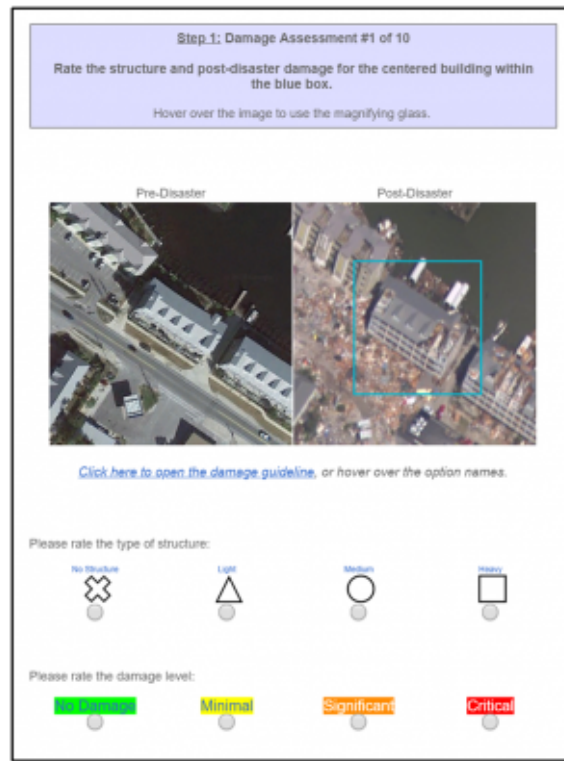
Crowdsourcing has been used for many years to rapidly assess damage after natural disasters. The results have been used to inform disaster relief officials of the extent of damage and areas to focus relief efforts. Because of the volume of data and urgent time sensitivity, automated damage assessments are being tested. Combining crowdsourcing with automated damage assessments is expected to increase quality and speed, but both the automated system and crowd workers must learn the unique aspects of each disaster in a partnership.

Your task is to review 10 images of buildings and classify the type of building and the extent of damage using the guideline below. Rate only the building at the center of the image as there may be more than one building. If you place your mouse cursor over the image you can magnify sections for more detail. After you submit your rating, you will receive a rating from the Automated Damage Assessment Machine (ADAM) for your review. Please evaluate this assessment and, if necessary, review the images again. You can change your ratings, if desired. If you do not believe the image is of a building, please mark "no structure".

Overview of Process






Screenshot Example of a Damage Assessment



In previous disasters it has been found that having a clear damage rating guideline is very important for data quality. The Harvard Humanitarian Institute's guidelines are shown below. You can recall this guide at any time while reviewing images.

Harvard Humanitarian Initiative's Aerial Imagery Interpretation Guide

Structure Type	Symbol	Description
Light		Structures that are built predominantly from light material or locally sourced materials. These structures may be mobile or possess no real hard roof, in some cases, roofs are made of metal or light material; they are often small in size. As such, these structures are likely to be the most vulnerable structures in any impacted region. Examples of these types of structures can include huts, tukuls or mobile trailers.
Medium		Structures that are built from semi-hard materials or mixed products. These structures have solid frames built using wood, steel or cement. These type of structures are fixed and possess hardened walls and roofs which can be made out of wood or cement. Unlike light structures, these types of structures are able to withstand moderate level of wind, with no to little damage, while maintaining their structural integrity. These types of structures can be individual or multi-family houses, small stores, places of worship and similar structures.
Heavy		Structures that are built from hard materials such as reinforced cement and steel. Infrastructure of this type is the least structurally vulnerable in any observed region. These structures are designed to withstand high level winds without receiving heavy damage or endangering the structural integrity of the structure. In many areas, these may include multiple story buildings, strip malls, hospital buildings, or public utilities.

Damage Classification	Color	Description

No Visible Damage		The roof is virtually undamaged and the walls, in effect, remain standing. The structure appears to have complete structural integrity and does not appear to need repair.
Minimal Visible Damage		The roof remains largely intact, but presents partial damage to the roof's surface, with minimal exposure beneath. In oblique aerial and satellite imagery, minimal damage may be able to be observed within the structure and to the exterior walls. The structure appears to have general structural integrity but needs minor repairs.
Significant Visible Damage		The roof is entirely damaged or missing. The walls of the structure remain upright. However, the interior wall partitions can be partially damaged. Debris inside the structure can also potentially be visible. The structure does not appear to have complete structural integrity and is in need of significant repair.
Critical Visible Damage		The roof is completely destroyed or missing, and the walls have been destroyed or collapsed. The support structures are completely leveled, and interior objects have also suffered visibly heavy damage or destruction. The structure does not appear to have any structural integrity and requires comprehensive reconstruction or demolition of the entire structure

Briefing Attention Check

Based on the briefing, select the statement which is true.

- You may change your rating after reviewing the automated assessment.
- You will be asked to estimate a value for the property.
- You are to ignore the type of structure when making your assessment.
- You will upload pictures you've taken of damaged buildings.

Sharing Notice

Please note:

This is a research study of the damage assessment task. Please do not share the contents of this study with other potential participants.

Rating Block

These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds

Last Click: 0 seconds

Page Submit: 0 seconds

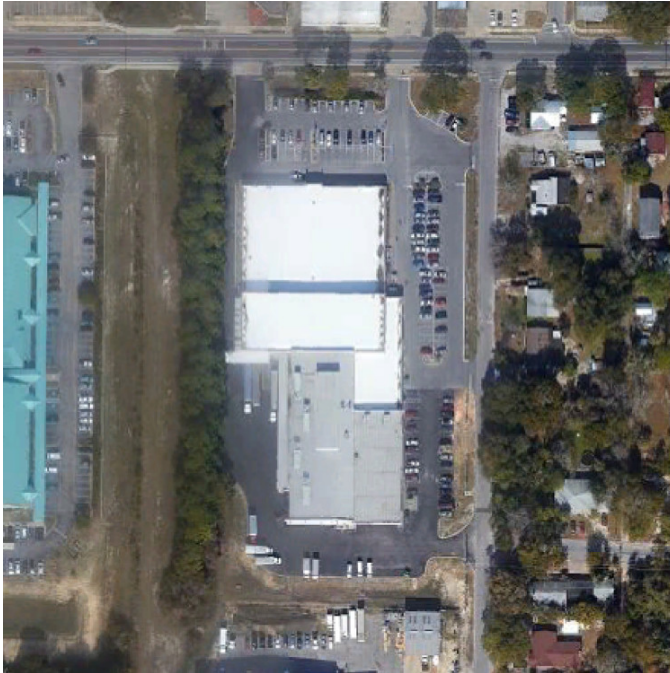
Click Count: 0 clicks

Step 1: Damage Assessment #1 of 10

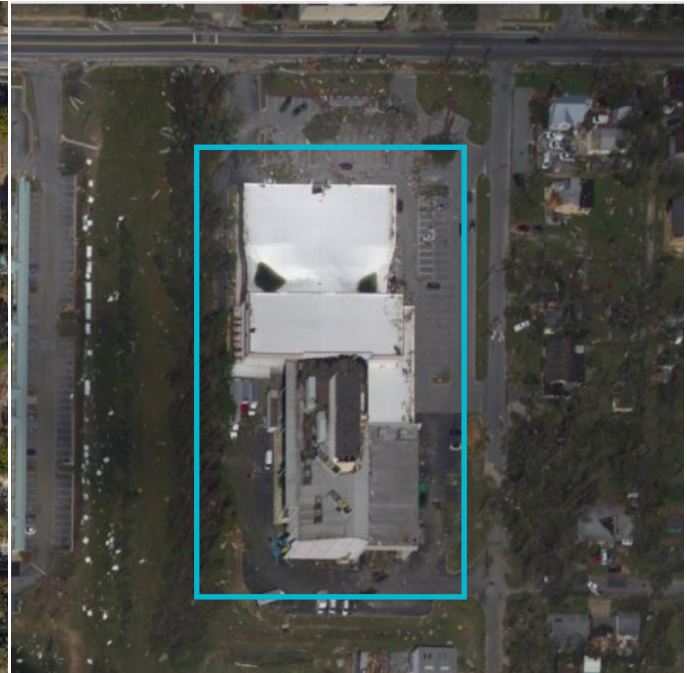
Rate the structure and post-disaster damage for the centered building within the blue box.

Hover over the image to use the magnifying glass.

Pre-Disaster



Post-Disaster



[Click here to open the damage guideline](#), or hover over the option names.

Please rate the type of structure:

No Structure



Light



Medium



Heavy



Please rate the damage level:

No Damage



Minimal



Significant



Critical



Please rate the difficulty of this classification:

Very Difficult



Very Easy



These page timer metrics will not be displayed to the recipient.

First Click: 0 seconds

Last Click: 0 seconds

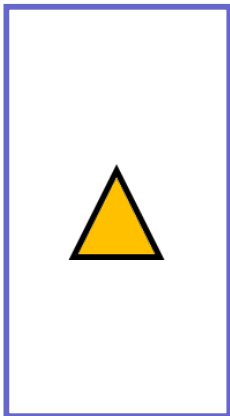
Page Submit: 0 seconds

Click Count: 0 clicks

Step 2: Review the automated rating #1

Hover over the image to use the magnifying glass.


Your
Assessment



ADAM's Rating
Automated Damage Assessment Machine

This structure was automatically classified Heavy No Damage.

Because it is a Large Commercial Building with multiple stories and intact roof.



It is not Heavy Minimal because it does not have light roof damage and does not have adjacent debris.

Consider: Large numbers of different features may decrease the accuracy of assessments.

The images are below if you would like to review them again:

Pre-Disaster

Post-Disaster



[Click here to open the damage guideline \(new tab\).](#)

If you'd like to change your answer, please select a new structure type: **(OPTIONAL)**

» No Structure



» Light



» Medium



» Heavy



If you'd like to change your answer, please select a new damage rating: **(OPTIONAL)**

» No Damage



» Minimal



» Significant



» Critical



Please indicate if any of the following apply in your opinion:

- The structure was hidden or obscured
- The image was blurry
- The automated rating by ADAM was incorrect

Advisory

In the next section we want to ask for your honest opinion.

Your answers have no effect on your eligibility for future HITS.

Attitudes

How much mental effort did you invest in making your assessments?

Very, Very Low Very Low Low Rather Low Neither Low nor High Rather High High Very High Very, Very High

Please rate your opinion on the following statements about the damage assessment task:

	Strongly Disagree							Strongly Agree
The damage assessment task was very complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was provided information which supported my ability to assess damage ratings.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I wanted to make sure I understood everything I was provided while completing the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I made an effort to understand the overall task and not just on the details.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Many things needed to be kept in mind simultaneously when rating damage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was exhausting to find the important information to assess damage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The design of this assessment tool was very inconvenient for making ratings.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was difficult to recognize and link the crucial information while assessing damage.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I had to remember many things to perform the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Based on your experience, please rate your expectations if you did more damage assessments using the tool:

	Strongly Disagree							Strongly Agree
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Strongly Disagree						Strongly Agree
I'm confident I understand the most difficult categories	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Knowing my skills and abilities, I think I can do well assessing damage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm certain I have the skills necessary for damage assessment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I'm confident I know how to use the damage assessment tool	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I expect I can do well on future damage assessments with this tool	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe I can produce high quality assessments with this tool	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please rate the Automated Damage Assessment Machine (ADAM):

	Strongly Disagree						Strongly Agree
Thinks logically	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is predictable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is able to make decisions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is insightful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is knowledgeable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has a mind	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Manipulation Checks

Please rate your opinion of the following statements.

The Automated Damage Assessment Machine (ADAM) explained why it made its ratings.

- Yes
- No
- Don't Know / Don't Remember

ADAM compared its rating to at least one other possible classification.

- » Yes
- » No

» Don't Know / Don't Remember

ADAM pointed out features in the image that may lead to incorrect classifications.

» Yes

» No

» Don't Know / Don't Remember

Alternate Explanations

Please estimate the number of aerial damage images you've assessed prior to this study.

No Experience
 Less Than 10 Images
 Between 10 and 50 Images
 Between 50 and 100 Images
 Between 100 and 500 Images
 Between 500 and 5,000 Images
 Over 5,000 Images

How often have you rated damage to structures in aerial images in the last year?

Never
 Yearly, or less
 A few times a year
 Monthly
 A few times a month
 Weekly
 Daily

Have you previously used a written damage guideline to make ratings?

No
 Yes - But a different guideline.
 Yes - The same guideline as here.

Please rate the Automated Damage Assessment Machine (ADAM) on the following statements:

	Strongly Disagree						Strongly Agree
I'm confident that ADAM won't distract me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe I won't overly rely on ADAM	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My ratings benefitted by working with ADAM	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ADAM prevents errors in manual damage assessments	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Strongly Disagree						Strongly Agree
My ratings were affected by ADAM's input	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I believe I can trust ADAM	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assessments depend on the both the human and ADAM for accuracy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The damage assessment task was:

Highly Manual Highly Automated

Please rate your opinion of the following statements:

	Strongly Disagree						Strongly Agree
I like to use technology to make tasks easier for me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have bad experiences when I try to use new technology instead of doing things "the old-fashioned way"	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
There are tasks in my life that have been made easier by computers doing the work for me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have a positive view of the potential for robots and artificial intelligence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

In your own work or profession, how has artificial intelligence (AI) performed for you?

Far exceeded my expectations Exceeded my expectations Equaled my expectations Short of my expectations Far short of my expectations Have not used AI in my work

Demographics

The following questions are for descriptive purposes only.

In what year were you born?

Are you... ?

- Male
- Female

What is the highest level of formal education you have completed?

- Less than high school degree
- High school graduate (high school diploma or equivalent including GED)
- Some post-high school
- Bachelor's degree
- Some graduate school
- Graduate degree

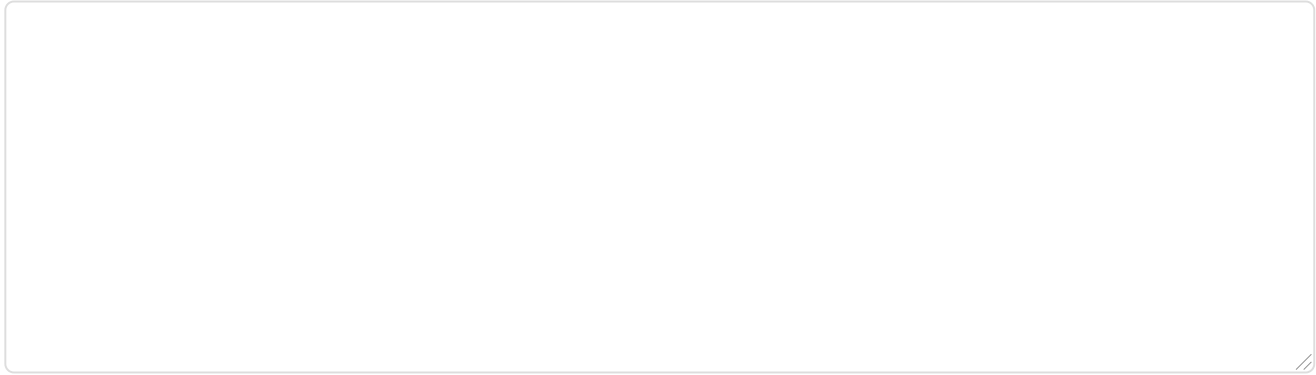
Which of the following best describes your total family income?

- | | | |
|--|--|--|
| <input type="radio"/> Less than \$15,000 | <input type="radio"/> \$45,000 to \$59,999 | <input type="radio"/> \$100,000 to \$149,999 |
| <input type="radio"/> \$15,000 to \$29,999 | <input type="radio"/> \$60,000 to \$79,999 | <input type="radio"/> Over \$150,000 |
| <input type="radio"/> \$30,000 to \$44,999 | <input type="radio"/> \$80,000 to \$99,999 | <input type="radio"/> Prefer not to Say |

Feedback

How did you find out about this HIT? if you found out on a web forum, please paste a link to the page.

Finally, we value any feedback concerning this survey. Please let us know your thoughts and or suggestions.

A large, empty rectangular box with a thin grey border, intended for the respondent to provide feedback or suggestions. The box is currently blank.

D. Scenarios and Explanations by Condition

See following page.

Scenarios: Images and Explanations

Contents

Template for Scenarios (Definitions)	2
Scenario 1 – Factory	3
Scenario 2 – Grocery Store	4
Scenario 3 – Government	5
Scenario 4 – Clear Trailer	6
Scenario 5 – Obstructed Trailer	7
Scenario 6 – Trailer Roof Error	8
Scenario 7 – Intact Trailer	9
Scenario 8 – Shipyard.....	10
Scenario 9 – Coastal Condo.....	11
Scenario 10 – Tarp Colored Roof	12
Scenario 11 – Large Condo.....	13
Scenario 12 – Day Care	14
Scenario 13 – Boat Dealer.....	15
Scenario 14 – House with Pool	16
Scenario 15 – Artificial Island.....	17
Scenario 16 – Funeral Home	18
Scenario 17 – Recreation Center	19
Scenario 18 – Hangar	20
Scenario 19 – Former Forest.....	21
Scenario 20 – Retention Ponds	22
Scenario 21 – Place of Worship	23
Scenario 22 – Supermarket.....	24

Template for Scenarios (Definitions)

Scenario [X]

Image [XXX] – [Pre-Disaster Image Source] / [Post-Disaster Image Source]



Type: [Faithful (*consistent with guideline*) / Erroneous (*object not in the image*)]

Model Output: *Structure and Damage Classification provided to user*

Next Best: *Classifications used as the counter-factual contrast*

Failure Mode: *Plausible boundary condition or outlier in the image*

Objects: List of detected objects and features (**bold items are simulated erroneously detected objects**)

All Conditions (*Model Output*): This structure was automatically classified [C-BEST-SLOT].

C (*Causal*): Because it is a [O-STRUCTURE-SLOT] with [O-FACTUAL-SLOT1] [and ...].

CF (*Counterfactual*): It is not a [C-NEXT-SLOT] classification because it [does not] have [O-COUNTERF-SLOT1] [and ...].

H (*Hedge*): Consider: [H-ERROR-MODE-SLOT].

Scenario 1 – Factory

Image 008 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Heavy No Damage
Next Best:	Heavy Minimal
Failure Mode:	Distractions
Objects:	Large Commercial Building. Changed roof structure (intact roof). No debris. Multi-story.

This structure was automatically classified Heavy No Damage.

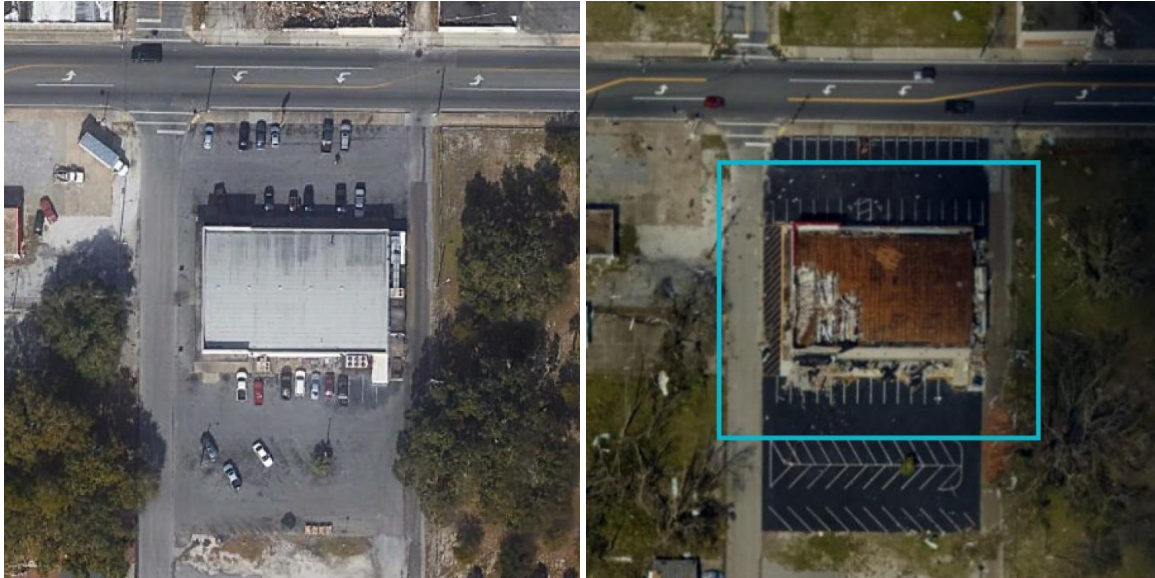
Because it is a Large Commercial Building with multiple stories and intact roof.

It is not Heavy Minimal because it does not have light roof damage and does not have adjacent debris.

Consider: Large numbers of different features may decrease the accuracy of assessments.

Scenario 2 – Grocery Store

Image 007 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Medium Critical
Next Best:	Heavy Significant
Failure Mode:	Illumination
Objects:	Medium Commercial Building. Roof damage. Wooden Roof. Debris adjacent to structure (classified as wall collapse).

This structure was automatically classified Medium Critical.

Because it is a Medium Commercial Building with collapsed roof and internal visible debris and collapsed wall with adjacent debris.

It is not Heavy Significant because of wooden building materials and it does not have intact walls.

Consider: Bodies of water, parking lots, and shadows may be misclassified as damage.

Scenario 3 – Government

Image 009 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Heavy Minimal
Next Best:	Heavy Significant
Failure Mode:	Feature Count
Objects:	Large Government Building. Multiple story. Metal roof. Roof damage. No debris. Intact walls. No visible debris.

This structure was automatically classified Heavy Minimal.

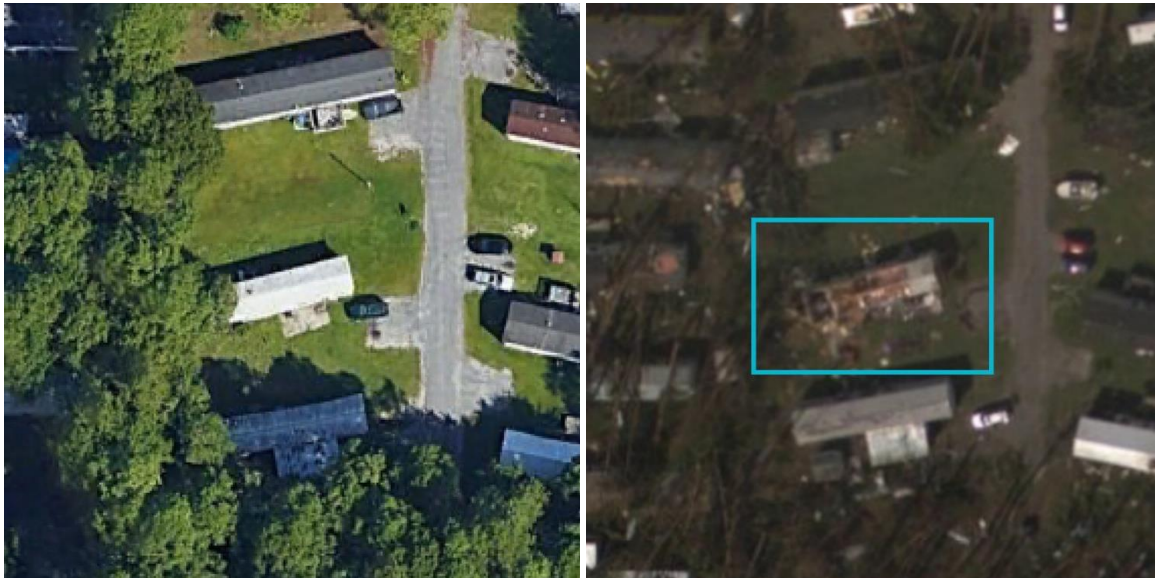
Because it is a Large Government Building with multiple stories and metal roof and light roof damage.

It is not Heavy Significant because it does not have missing roof.

Consider: Large numbers of the same object may decrease the accuracy of assessments.

Scenario 4 – Clear Trailer

Image 010 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Light Critical
Next Best:	Light Significant
Failure Mode:	Illumination
Objects:	Residential Trailer. Roof damaged. Interior debris. Adjacent debris.

This structure was automatically classified Light Critical.

Because it is a Residential Trailer with collapsed roof and internal visible debris.

It is not a Light Significant classification because it does not have intact walls.

Consider: Shadows, reflections, and color changes may decrease the accuracy of assessments.

Scenario 5 – Obstructed Trailer

Image 012 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Light Minimal
Next Best:	Light Significant
Failure Mode:	Obscuration
Objects:	Residential Trailer. Roof damaged. Intact walls. Wooden roof materials. Trailer shape.

This structure was automatically classified Light Minimal.

Because it is a Residential Trailer with a trailer shape and damaged roof.

It is not a Light Significant classification because it does not have a missing roof.

Consider: Overlapping objects may decrease the accuracy of assessments.

Scenario 6 – Trailer Roof Error

Image 013 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Light Significant
Next Best:	Light Minimal
Failure Mode:	Obscuration
Objects:	Damaged roof (classified as internal visible debris). Adjacent debris. Intact walls. Missing roof. Trailer shape.

This structure was automatically classified Light Significant.

Because it is a Residential Trailer with collapsed roof and internal visible debris.

It is not a Light Minimal classification because it does not have light roof damage.

Consider: Overlapping objects may decrease the accuracy of assessments.

Scenario 7 – Intact Trailer

Image 002 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Light No Damage
Next Best:	Light Minimal
Failure Mode:	Obscuration
Objects:	Residential Trailer. Intact roof. Intact walls. Trailer shape.

This structure was automatically classified Light No Damage.

Because it is a Residential Trailer with intact walls and no roof damage.

It is not a Light Minimal classification because it does not have light roof damage.

Consider: Overlapping objects may decrease the accuracy of assessments.

Scenario 8 – Shipyard

Image 021 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Heavy Critical
Next Best:	Heavy Significant
Failure Mode:	Spectral Distinction
Objects:	Large Industrial Building. Metal roof. collapsed wall with adjacent debris. collapsed roof and internal visible debris.

This structure was automatically classified Heavy Critical.

Because it is a Large Industrial Building with a metal roof and collapsed roof and internal visible debris.

It is not a Heavy Significant classification because it does not have intact walls.

Consider: Bodies of water, parking lots, and shadows may be misclassified as damage.

Scenario 9 – Coastal Condo

Image 005 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Medium Critical
Next Best:	Heavy Critical
Failure Mode:	Feature Density
Objects:	Multi-Family Residential Building. Intact roof. adjacent debris (classified as collapsed wall). Multiple stories. (seen as single story)

This structure was automatically classified Medium Critical.

Because it is a Multi-Family Residential Building with collapsed wall with adjacent debris.

It is not a Heavy Critical because it does not have multiple floors.

Consider: Tightly grouped features may decrease the accuracy of assessments.

Scenario 10 – Tarp Colored Roof

Image 003 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Medium Significant
Next Best:	Heavy Critical
Failure Mode:	Feature Density
Objects:	Multi-Family Residential Building. Missing roof. Wooden roof materials. adjacent debris. Multiple stories.

This structure was automatically classified Medium Significant.

Because it is a Multi-Family Residential Building with missing roof.

It is not a Heavy Critical because it has wooden roof materials and intact walls.

Consider: Tightly grouped features may decrease the accuracy of assessments.

Scenario 11 – Large Condo

Image 006 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Heavy Significant
Next Best:	Heavy Critical
Failure Mode:	Feature Distinction
Objects:	Multi-Family Residential Building. Multi-story. Adjacent debris. Missing roof sections. Intact walls.

This structure was automatically classified Heavy Significant.

Because it is a Multi-Family Residential Building with multiple stories and missing roof sections.

It is not a Heavy Critical classification because it does not have collapsed wall with adjacent debris.

Consider: Complex textures may decrease the accuracy of assessments.

Scenario 12 – Day Care

Image 015 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Medium Significant
Next Best:	Heavy Minimal
Failure Mode:	Feature Distinction
Objects:	Attached Building Commercial Complex. Adjacent debris. Light roof damage. Wood Roof Materials. (collapsed wall with adjacent debris)

This structure was automatically classified Medium Significant.

Because it is a Medium Commercial Complex with collapsed wall with adjacent debris.

It is not a Heavy Minimal because it has a wooden roof and does not have intact walls.

Consider: Complex textures may decrease the accuracy of assessments.

Scenario 13 – Boat Dealer

Image 016 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Medium Significant
Next Best:	Heavy Significant
Failure Mode:	Object Count
Objects:	Medium Commercial Building. Light roof damage. Adjacent debris (boats on building side classified as collapsed wall).

This structure was automatically classified Medium Significant.

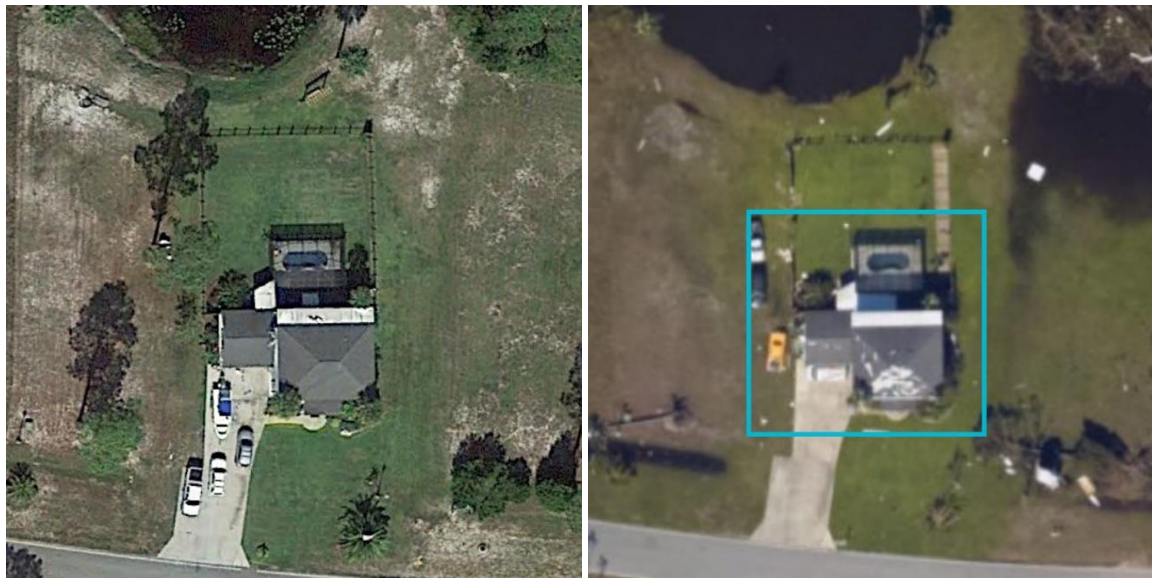
Because it is a Medium Commercial Building with collapsed wall with adjacent debris.

It is not a Heavy Significant because of it does not have multiple stories.

Consider: Large numbers of the same object may decrease the accuracy of assessments.

Scenario 14 – House with Pool

Image 001 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Medium Minimal
Next Best:	Medium Significant
Failure Mode:	Roof Complexity
Objects:	Single-Family Home. Light roof damage. No debris.

This structure was automatically classified Medium Minimal.

Because it is a Single-Family Home with light roof damage.

It is not a Medium Significant because of no missing roof sections.

Consider: Windows, chimneys, and roof textures may be misclassified as damage.

Scenario 15 – Artificial Island

Image 004 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Medium Minimal
Next Best:	Medium Significant
Failure Mode:	Illumination
Objects:	Single-Family Home. Light roof damage. No debris.

This structure was automatically classified Medium Minimal.

Because it is a Single-Family Home with light roof damage.

It is not a Medium Significant because of no missing roof sections.

Consider: Shadows, reflections, and color change may decrease the accuracy of assessments.

Scenario 16 – Funeral Home

Image 014 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Medium Significant
Next Best:	Heavy Significant
Failure Mode:	Spectral Distinction
Objects:	Medium Commercial Building. Missing roof segment. Adjacent debris. Wooden roof.

This structure was automatically classified Medium Significant.

Because it is a Medium Commercial Building with missing roof segment and adjacent debris.

It is not a Heavy Significant because of wood roof materials.

Consider: Bodies of water, parking lots, and shadows may not be identified properly.

Scenario 17 – Recreation Center

Image 017 – Google 2018 / NGS Michael 2018



Type: Erroneous

Model Output: Medium Critical

Next Best: Medium Minimal

Failure Mode: Feature Density

Objects: Attached Building Commercial Complex. Collapsed wall with adjacent debris (**classified as wall collapse**). Missing roof section. Wooden roof construction.

This structure was automatically classified Medium Critical.

Because it is a Medium Commercial Complex with collapsed wall with adjacent debris.

It is not a Medium Minimal because of missing roof section.

Consider: Tightly grouped features may decrease the accuracy of assessments.

Scenario 18 – Hangar

Image 018 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Heavy Significant
Next Best:	Medium Critical
Failure Mode:	Unusual Scale
Objects:	Large Industrial Building. Intact walls. Collapsed roof and internal visible debris. Metal roof materials.

This structure was automatically classified Heavy Significant.

Because it is a large industrial building with collapsed roof and internal visible debris.

It is not a Medium Critical because it does not have collapsed walls and adjacent debris.

Consider: Very large and very small structures may decrease the accuracy of assessments.

Scenario 19 – Former Forest

Image 019 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Medium Significant
Next Best:	Heavy Critical
Failure Mode:	No Pre-Disaster Image
Objects:	Medium Commercial Building. Intact walls. collapsed roof and internal visible debris. Wooden construction. (entire description)

This structure was automatically classified Medium Significant.

Because it is a Medium Commercial Building with collapsed roof and internal visible debris.

It is not a Heavy Critical because of wood roof construction and intact walls.

Consider: Lack of pre-disaster image may decrease the accuracy of assessments.

Scenario 20 – Retention Ponds

Image 020 – Google 2018 / NGS Michael 2018



Type:	Erroneous
Model Output:	Medium No Damage
Next Best:	Medium Minimal
Failure Mode:	Spectral Distinction
Objects:	Attached Building Commercial Complex. pre-existing poor condition roof. Intact walls. (entire description)

This structure was automatically classified Medium No Damage.

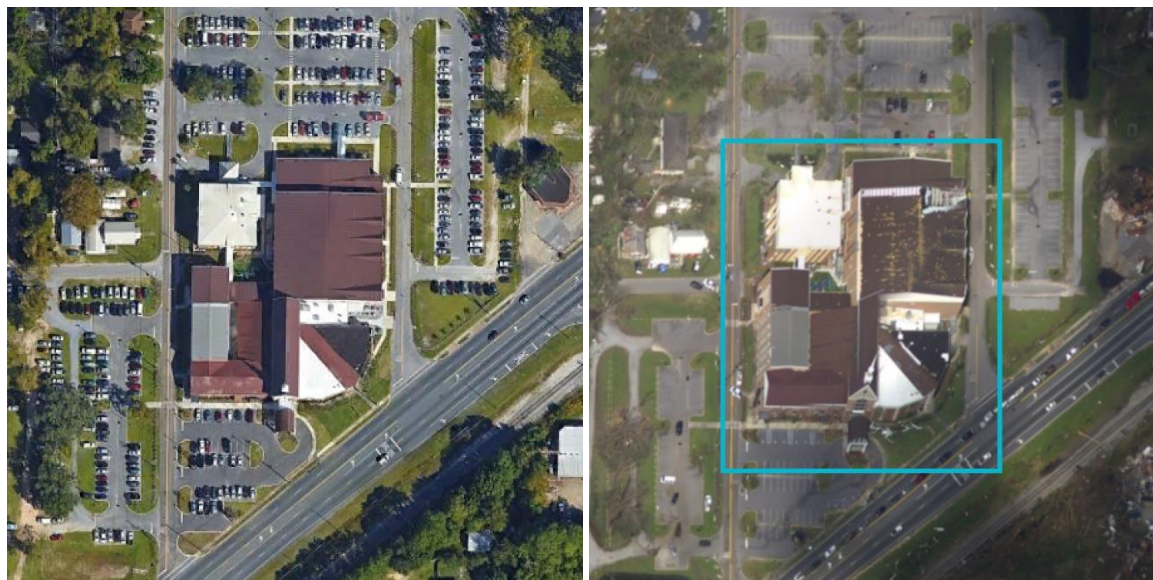
Because it is a Medium Commercial Complex with pre-existing poor condition roof.

It is not a Medium Minimal because of no visible debris and intact roof.

Consider: Bodies of water, parking lots, and shadows may not be identified properly.

Scenario 21 – Place of Worship

Image 011 – Google 2018 / NGS Michael 2018



Type: Erroneous

Model Output: Medium No Damage

Next Best: Medium Significant

Failure Mode: Obscuration

Objects: Place of Worship Complex. Missing roof shingles. Intact walls. **(Missing roof section)**. Multiple stories. No debris.

This structure was automatically classified Medium No Damage.

Because it is a Place of Worship Complex with intact walls and missing roof shingles.

It is not a Medium Significant because of no missing roof sections.

Consider: Shadows, reflections, and color changes may decrease the accuracy of assessments.

Scenario 22 – Supermarket

Image 022 – Google 2018 / NGS Michael 2018



Type:	Faithful
Model Output:	Medium Significant
Next Best:	Medium Critical
Failure Mode:	Object Count
Objects:	Medium Commercial Building. Missing roof segment. Adjacent debris. Intact walls.

This structure was automatically classified Medium Significant.

Because it is a Medium Commercial Building with missing roof segment and adjacent debris.

It is not a Medium Critical because of intact walls.

Consider: Large numbers of the same object may decrease the accuracy of assessments.

Appendix E: Pre-Test Interview Guide

See following page.

INTERVIEW GUIDE: Crowd-sourced Damage Assessment in Natural Disasters

December 21, 2018

Student Principal Investigator: Sean Dougherty (sdougherty5@student.gsu.edu)

Principal Investigator: Pam Ellen (pellen@gsu.edu)

1. Recruiting

Directions: Ensure the participant can see the consent on their screen. Read the participant the consent form and ask them to verbalize that they agree and then to click the next button on the consent page of the survey.

2. Interview

Briefing

Directions: Allow the participant to read the briefing on their own without describing the survey or the experiment.

Interview Questions:

1. Did you find any part of the briefing unclear?
2. Did you find the structure and damage guideline to be clear?
3. Can you describe the task as you understand it?

Initial Rating Step

Directions: Ask the following question when the image appears for them to rate in the initial step. Allow them to complete 7 of the 10 ratings on their own. Ask them to verbalize as much of their decision process as possible for all images.

Interview Questions:

1. Did the images appear clearly and a useful size?
2. Did the magnifying glass functionality work for you?

Review Rating Step

Directions: Ask the following questions for the first three image ratings.

Interview Questions:

1. How confident were you in selecting your damage rating for this image?
2. How do you feel about the automated rating for this image?
3. Is the information provided useful to evaluate the rating?

Survey Measures

Directions: Allow the participant to complete the survey measures and ask any questions as they complete the items on their own.

Interview Questions:

1. Did the questions make sense?
2. Did the options for rating your answers make sense?
3. How do you feel about the number of questions we asked you to rate?

Close Out

Directions: Instead of the participant completing the optional feedback questions on the survey, ask them the following questions.

Interview Questions:

1. Do you have any suggestions to improve the damage assessment task?
2. Do you think any of the images should not have been used in the study?
3. Do you have any feedback about how the automated assessment was presented?
4. Do you have any other feedback we did not cover?

Ending Script:

Please remember to complete the survey and enter the completion code into the HIT. Thank you for participating in this study and helping us to test and improve this survey.

Appendix F: Instrument Development and Testing

See following page.

F. INSTRUMENT DEVELOPMENT AND TESTING

F.1 Task Development

The scenarios for the task were generated in a multi-step process to create the elements needed for the specification of the simulated agent. The steps used to generate the scenarios and their components are shown in Figure 1 and discussed in detail below. The full list of scenarios appears in Appendix C.

The literature was reviewed to identify the required components to support an explanation engine at the current state-of-the-art capability. The simulated output and explanations are synthetic with the goal of plausibility and agent behavior authentic to what the participant would experience if the agent were real, but not intended to replicate a single specific system. Building damage and object classes were generated by author evaluation of the images using the damage assessment guideline and land use classifications in a multiple-cycle process. Failure modes were collected from the literature considering the images selected for this experiment. Natural language explanations were generated by transforming example explanations from the XAI literature into templates. The model simulated here most closely resembles definitions of objects of buildings as in Mayer (1999), roof damage detection as in F. Wang (2017), general damage detection by Vetrivel et al. (2018), and semantic object recognition and classification as in Hendricks et al. (2016) to classify portions of an image and generate explanations. The specification for the simulated agent appears in Figure 2.

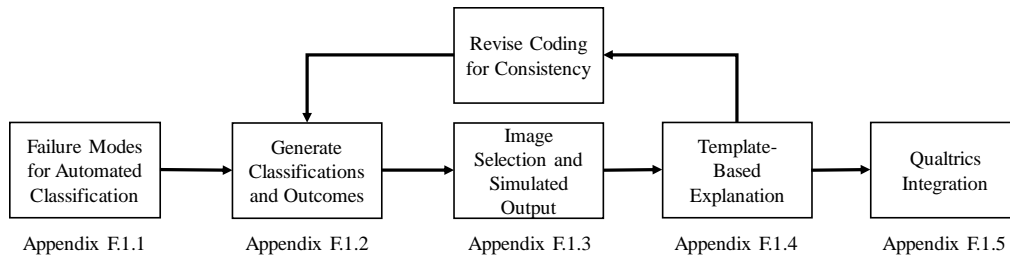


Figure 1 Process to Generate Scenarios

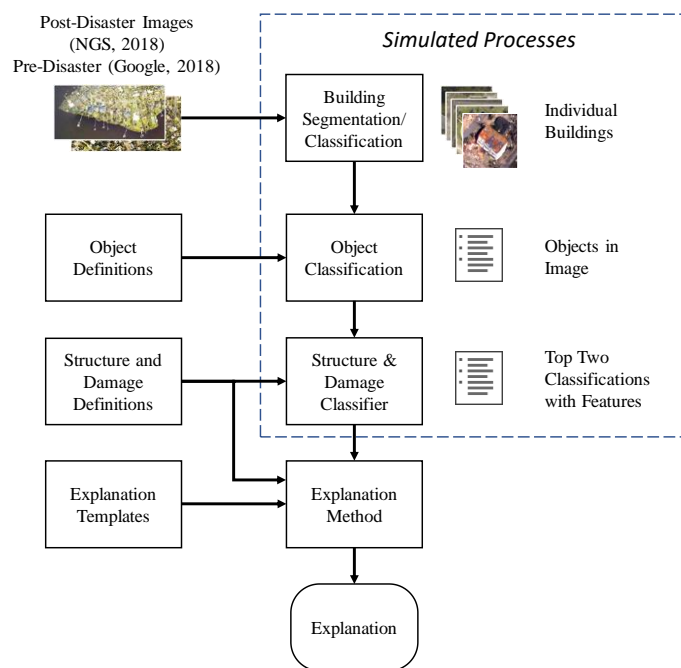


Figure 2 Simulated Output and Explanation Specification

F.1.1 Failure Modes for Automated Classification

A set of real-world failure modes and boundary conditions for automated classification was identified by a literature search of “convolutional neural network” and “aerial” and “damage” since 2014 with the goal of covering the components of the specification rather than an exhaustive review of the literature. Articles not related to the assessment of structures were

excluded. When an article was included, a “snowball” approach was used for references relevant to building damage assessment. A two-cycle coding method was used to collect the disclosed failure modes with second-cycle analytic codes generated based on the causes within the image leading to the failure. The finished list is shown in Table 1. Two of the failure modes in the literature (“unexplained error,” “imbalanced classes”) were not used to generate scenarios as they are not readily detectable by an intelligent agent or by the participant.

Table 1 Failure Modes Identified in Aerial Damage Assessment Literature

ID	Failure Mode	References	Hedging Slot
1	Atmosphere	[D,N]	Clouds, fog, and smoke may decrease the accuracy of assessments.
2	Distraction	[C,N]	Large numbers of different objects may decrease the accuracy of assessments.
3	Object Count	[C,D,I]	Large numbers of the same object may decrease the accuracy of assessments.
4	Feature Density	[A,E]	Tightly grouped features may decrease the accuracy of assessments.
5	Feature Distinction	[G,H,J]	Complex textures may decrease the accuracy of assessments.
6	Illumination	[B,E]	Shadows, reflections, and color changes may decrease the accuracy of assessments.
7	Obscuration	[E,N]	Overlapping objects may decrease the accuracy of assessments.
8	Pre-Existing Conditions	[A]	Poorly maintained structures may be misclassified as disaster damage.
9	Roof Complexity	[A,K,M]	Windows, chimneys, and roof textures may be misclassified as damage.
10	Spectral Distinction	[F,N]	Bodies of water, parking lots, and shadows may not be identified properly.
11	Unusual Scale	[H,L]	Very large and very small structures may decrease the accuracy of assessments.
12	No Pre-Disaster Image	[I]	Lack of pre-disaster image may decrease the accuracy of assessments.
13	Unexplained Error	[B]	n/a - not used in this research
14	Imbalanced Classes	[G]	n/a - not used in this research

[A] Vetrivel, A., Gerke, M., Kerle, N., Nex, F., & Vosselman, G. (2018); [B] Kersbergen, D. (2018); [C] Attari, N., Ofli, F., Awad, M., Lucas, J., & Chawla, S. (2017); [D] Mather, P. M., & Koch, M. (2011); [E] Moranduzzo, T., & Melgani, F. (2014); [F] Kluckner, S., Mauthner, T., Roth, P. M., & Bischof, H. (2009); [G] Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017); [H] Qi, K., Yang, C., Guan, Q., Wu, H., & Gong, J. (2017); [I] Fujita, A., Sakurada, K., Imaizumi, T., Ito, R., Hikosaka, S., & Nakamura, R. (2017); [J] Duarte, D., Nex, F., Kerle, N., & Vosselman, G. (2017); [K] Duarte, D., Nex, F., Kerle, N., & Vosselman, G. (2018); [L] Wang, F. (2017); [M] Cao, Q. D., & Choe, Y. (2018);

F.1.2 Generate Classifications and Outcomes

Targeted classifications for structure images for twenty scenarios were selected in advance of identifying images to ensure a diversity of structure type and damage classification categories. Half of the scenarios were selected at random for erroneous performance of object detection prior to image selection. This failure rate is not implausible for a model being transferred to a new disaster or geographic region without labeled training data (Vetrivel et al., 2015), especially where classifications are sought for images with classification disagreements. Two scenarios were designated to have no structure in the image as a means to detect low

participatory effort; however, this phenomenon is also consistent with the user experience reported by GIScorps (2013) where some images crowd workers are exposed to are not appropriate for damage assessment.

F.1.3 *Image Selection and Simulated Output*

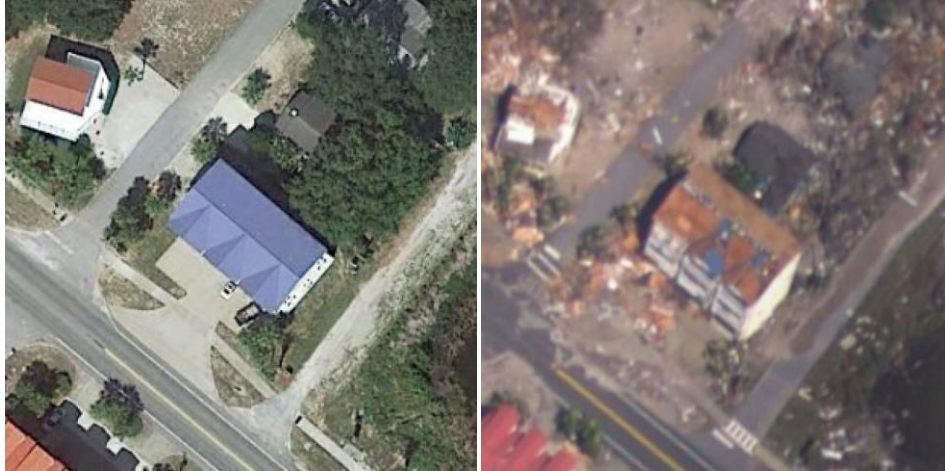
Images were selected for the target structure type and damage classification by an initial review by the author. Aerial imagery for post-disaster images was sourced from the National Geodetic Survey for Hurricane Michael, which was collected aurally from an altitude of 5,500 feet (NGS, 2018). Pre-disaster images were sourced from Google Earth with images originally from Google Satellite as disclosed by the user interface (Google, 2018) in “2D” overhead orientation at an eye altitude of less than 500 meters. Images did not include overlays of street names or provide geographic coordinates identifying the location depicted. Images were selected to reflect a mix of levels of difficulty and plausibility of the identified failure modes in the literature.

The selected images were coded for structure and objects in a two-step process. For structure type, land use classifications for structures were identified based primarily on pre-disaster conditions; however, some structure type classifications benefited from exposed construction materials visible in post-disaster images. Damage visible in images was coded referencing the definitions contained in the damage guideline (Achkar et al., 2016) and the damage detection literature referenced in Chapter 2, Table 2. For erroneous scenarios, plausible erroneous detection of objects was used to generate an incorrect image object description and hedging explanation from Table 1. In the first step codes were expanded based on objects contained in the images and preliminary explanations were drafted. In the second step, the list of generated object codes was reviewed with codes combined and modified for parsimony, class

discrimination, consistency with the guidelines, and considerations of algorithmically generating explanations using templates. The final list of codes generated are shown in Table 2.

Simulated output was revised for consistency between coded damage and structural features and the generated output classifications. The “next best” classification for counterfactual explanation was selected to be that classification which would result from testing for what addition or subtraction of an object would modify the classification. A decision tree was computed using R version 3.3.2 “rpart” library to confirm that the simulated output was consistent with the detected objects (decision tree visualized in Figure 4). Building structure classification was similarly compared to the land use classification specification. The list of scenarios generated for the pilot is listed in Table 3.

Example images are shown in Figure 3 for pre-disaster and post-disaster conditions from Scenario 10. The simulated output for the scenario is a medium structure (multi-family home with multiple stories and wooden roof) and significant visible damage (the walls remain intact, but the roof material is nearly entirely missing, and there is adjacent debris). This example demonstrates the value of the pre-disaster imagery. The blue areas on the roof of the post-disaster image could be interpreted as having tarps covering roof damage or revealing underlying roofing material, where the pre-disaster image makes clear that the blue portions of the roof are likely the only remaining small areas of undamaged roof. The challenge of automatically comparing images taken of different angles and color balance is also apparent.



Map Data Pre-Disaster: (Google, 2018); Post-Disaster: (NGS, 2018)

Figure 3 Example Pre-Disaster and Post-Disaster Images

Table 2 Table of Structures, Objects, and Classifications**Objects for Structure Classification**

O-STRUCTURE-SLOT	Classification
Residential Trailer	Light
Medium Commercial Building	Medium
Medium Commercial Complex	Medium
Multi-Family Residential Building	Medium
Single-Family Home	Medium
Place of Worship Complex	Medium
Large Commercial Building	Heavy
Large Government Building	Heavy
Large Industrial Building	Heavy

Objects Supporting Structural Classification

O-SUPPORTING-SLOT	Classifications
metal roof materials	Heavy
wooden building materials	Light,Medium
multi-story	Medium,Heavy
trailer shape	Light

Objects Supporting Damage Classification

O-DAMAGE-SLOT	Classifications
intact walls	No Damage,Minimal,Significant
collapsed wall with adjacent debris	Critical
intact roof	No Damage,Minimal
changed roof structure	No Damage
missing roof shingles	No Damage
light roof damage	Minimal
missing roof sections	Significant
missing roof	Significant
collapsed roof and internal visible debris	Significant
adjacent debris	Significant,Critical
no visible debris	No Damage
pre-existing poor condition roof	No Damage

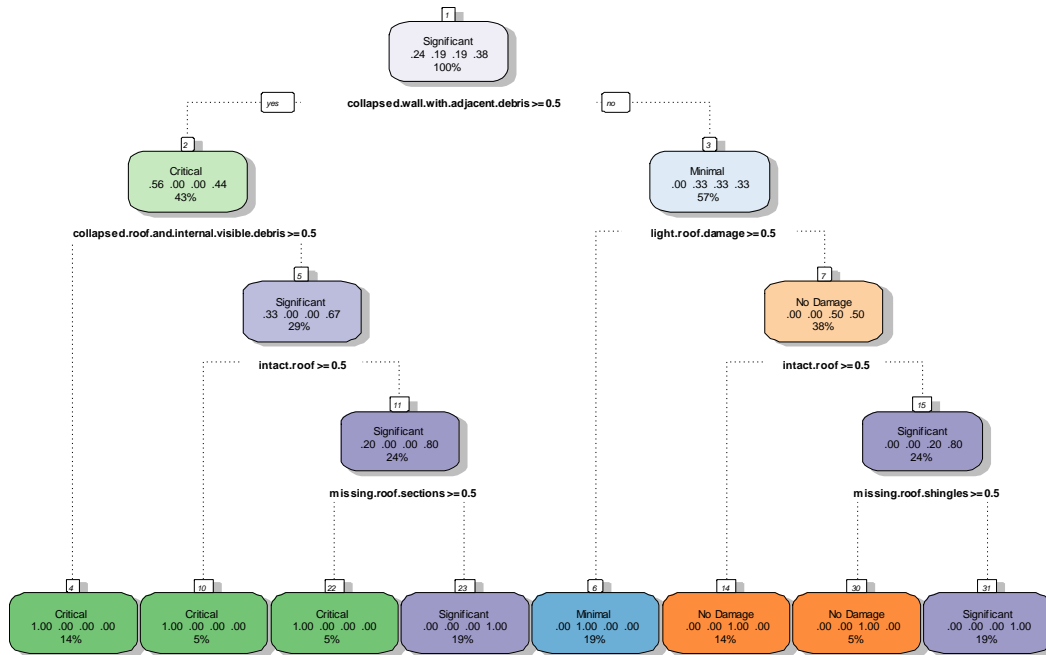


Figure 4 Validation of Simulated Output Classifications

Table 3 Pilot Scenario Listing

Scenarios		Simulated Model Output					
ID	Land Use	Intended Model Output	Error Mode	Model Output		Next Best	
				Structure	Damage	Structure	Damage
1	Large Commercial	Erroneous	9	Heavy	No Damage	Heavy	Minimal
2	Medium Commerical	Faithful	6	Medium	Critical	Heavy	Significant
3	Large Government	Faithful	3	Heavy	Minimal	Heavy	Significant
4	Residential Trailer	Faithful	6	Light	Critical	Light	Significant
5	Residential Trailer	Faithful	7	Light	Minimal	Light	Significant
6	Residential Trailer	Erroneous	7	Light	Significant	Light	Minimal
7	Residential Trailer	Faithful	7	Light	No Damage	Light	Minimal
8	Large Commercial	Faithful	10	Heavy	Critical	Heavy	Significant
9	Multi-Family Residential	Faithful	4	Medium	Critical	Medium	No Damage
10	Multi-Family Residential	Faithful	4	Medium	Significant	Medium	Critical
11	Multi-Family Residential	Faithful	5	Heavy	Significant	Heavy	Critical
12	Medium Commercial Complex	Erroneous	5	Medium	Significant	Heavy	Minimal
13	Medium Commerical	Erroneous	3	Medium	Significant	Heavy	Significant
14	Single-Family Home	Faithful	2	Medium	Minimal	Medium	Significant
15	Single-Family Home	Faithful	6	Medium	Minimal	Medium	Significant
16	Medium Commercial	Faithful	10	Medium	Significant	Heavy	Significant
17	Medium Commercial Complex	Erroneous	4	Medium	Critical	Medium	Minimal
18	Large Industrial	Faithful	11	Heavy	Significant	Medium	Critical
19	[no building]	Erroneous	12	Medium	Significant	Heavy	Critical
20	[no building]	Erroneous	10	Medium	No Damage	Medium	Minimal
21	Place of Worship	Faithful	7	Heavy	No Damage	Heavy	Significant
22	Medium Commercial	Faithful	3	Medium	Significant	Medium	Critical

F.1.4 *Template-Based Explanation*

The neural network structures used by state-of-the-art explanation engines such as those by R. Hu, Rohrbach, Andreas, Darrell, and Saenko (2017), Mao, Xu, Yang, Wang, and Yuille (2014), Donahue et al. (2015), and Hendricks et al. (2018) use varieties of recurrent neural networks which predict the word sequences of explanations based on training the network on exemplar explanations. Natural language generation researchers looking to increase available training data beyond human-generated exemplars have developed a method of replacing keywords from existing exemplars into a template algorithmically to transfer grammar between domains (Wen et al., 2016). That method is used here to produce templates for explanations without the use of a neural network.

In the visual explanation approach by Hendricks et al. (2016) *visual explanations* are generated from *image descriptions* and *class definitions*.¹ To be consistent with other literature, the human-recognizable elements of images which are termed “features” in Hendricks et al. (2016) are termed “objects” here while individual measurable properties of images which may not be measurable or recognizable by humans are termed “features” (e.g. histograms of oriented gradients). Class definitions are comprised of lists of objects which are indicative of the *class*. Explanation templates for causal and counterfactual explanations were created in a manner similar to Wen et al. (2016) with *slots* where object labels could be substituted into delexified statements. Fixed strings were used for hedging explanations. *Classification slots*, *object support slots*, and *object slots* were identified based on the *classes* from the guideline and *objects* in the images. For each post-disaster image, objects were identified with classifications to feed the slots

¹ Italics are used in this section to note terminology adopted from the literature to define the behavior of the simulated agent. Terms in quotes differ from the cited literature for reasons cited.

and slot-value pairs for explanations. Class definitions were created for building damage based on extractable objects in an iterative process through images selected for the scenarios. The list of structures, classes, and objects are listed in Table 2.

The causal explanation template was framed in the form “Because” then listing objects present as in the image, which was discriminative to the next best classification using a supporting structure classification slot in the “with” form. The counterfactual explanation was in the form “It is not [next best class] because” with objects listed with “does not have” where an object is distinctive of and contained within the next best class definition and “has” when an object is not contained within the next best class definition. The slots listed for the objects in Table 2 were substituted into the templates joined with “and” when more than one object was inserted into a slot.

An approach to algorithmically generate hedges for image explanations was not identified in the literature. Hedges can be made in a continuum which includes disclaimers, warnings, cautions, alerts, precautions, advisories, considerations, notices, and messages. Hyland (1996a) analyzed mitigating statements to claims in academic writing, and three categories identified were relevant to explanations as claims here: epistemic adjectives (*possible, consistent with*), speculative judgment (*suggest, indicate*), and modal verbs (*may, could*) used in an epistemic sense. Where feasible, the hedge was placed in the form of “Consider: [failure mode] may decrease the accuracy of assessments.” Each hedge used the modal verb “may” with the mitigating “consider,” allowing the hedge to be either true or false and to convey minimal information on the probability of the hedging statement being true. The final hedging explanations are listed in Table 1.

F.1.5 Qualtrics Integration

Simulated output and explanations for the five conditions of each of the 22 scenarios were transformed into static images programmatically. A python script was developed for this purpose. These static images were referenced by the Qualtrics survey based on the scenario and experiment condition. Images were hosted on Amazon S3. The Qualtrics “Loop and Merge” feature was utilized to randomly select and sequence scenarios for presentation.

F.1.6 Participant Briefing




The following briefing was provided to workers:

Crowdsourcing has been used for many years to rapidly assess damage after natural disasters. The results have been used to inform disaster relief officials of the extent of damage and areas to focus relief efforts. Because of the volume of data and urgent time sensitivity, automated damage assessments are being tested. Combining crowdsourcing with automated damage assessments is expected to increase quality and speed, but both the automated system and crowd workers must learn the unique aspects of each disaster in a partnership.

Your task is to review 10 images of buildings and classify the type of building and the extent of damage using the guideline below. Rate only the building at the center of the image as there may be more than one building. If you place your mouse cursor over the image you can magnify sections for more detail. After you submit your rating, you will receive a rating from the Automated Damage Assessment Machine (ADAM) for your review. Please evaluate this

assessment and, if necessary, review the images again. You can change your ratings, if desired. If you do not believe the image is of a building, please mark "no structure".

An adapted presentation of the text of the damage assessment guide by Achkar et al. (2016) was included below the briefing, shown in Figure 5. The guide could be recalled for review while assessing images in a separate user-interface window. At the end of the survey participants were debriefed to inform them that their contribution supported future efforts in disaster damage assessment, but ensured they understood it was a research project.

Structure Type	Symbol	Description
Light		Structures that are built predominantly from light material or locally sourced materials. These structures may be mobile or possess no real hard roof, in some cases, roofs are made of metal or light material; they are often small in size. As such, these structures are likely to be the most vulnerable structures in any impacted region. Examples of these types of structures can include huts, tukuls or mobile trailers.
Medium		Structures that are built from semi-hard materials or mixed products. These structures have solid frames built using wood, steel or cement. These type of structures are fixed and possess hardened walls and roofs which can be made out of wood or cement. Unlike light structures, these types of structures are able to withstand moderate level of wind, with no to little damage, while maintaining their structural integrity. These types of structures can be individual or multi-family houses, small stores, places of worship and similar structures.
Heavy		Structures that are built from hard materials such as reinforced cement and steel. Infrastructure of this type is the least structurally vulnerable in any observed region. These structures are designed to withstand high level winds without receiving heavy damage or endangering the structural integrity of the structure. In many areas, these may include multiple story buildings, strip malls, hospital buildings, or public utilities.





Damage Classification	Color	Description
No Visible Damage		The roof is virtually undamaged and the walls, in effect, remain standing. The structure appears to have complete structural integrity and does not appear to need repair.
Minimal Visible Damage		The roof remains largely intact, but presents partial damage to the roof's surface, with minimal exposure beneath. In oblique aerial and satellite imagery, minimal damage may be able to be observed within the structure and to the exterior walls. The structure appears to have general structural integrity but needs minor repairs.
Significant Visible Damage		The roof is entirely damaged or missing. The walls of the structure remain upright. However, the interior wall partitions can be partially damaged. Debris inside the structure can also potentially be visible. The structure does not appear to have complete structural integrity and is in need of significant repair.
Critical Visible Damage		The roof is completely destroyed or missing, and the walls have been destroyed or collapsed. The support structures are completely leveled, and interior objects have also suffered visibly heavy damage or destruction. The structure does not appear to have any structural integrity and requires comprehensive reconstruction or demolition of the entire structure.

Figure 5 Classification Guideline

Note. Categories and descriptions adopted from Achkar et al. (2016).

To ensure the validity of the task and instrument a series of tests were conducted with multiple improvement cycles. All participants were recruited from Mechanical Turk, as in the study. In the first round a telephone interview was conducted while the participant concurrently completed the survey. “Individual debriefing” (Ruel, 2015) was employed to evaluate the briefing and study measures, and “cognitive interviewing” (Ruel, 2015) was employed for the damage assessment task. The ten scenarios selected for this initial test had the same 50% mix of faithful and erroneous automated classification performance and variation in structure types. The erroneous automated classification scenario with square retention ponds and retaining walls (Scenario 20) was selected to evaluate how participants reacted to a “no structure” true rating. For erroneous automated classifications two “heavy,” one “light,” and two “medium” structure scenarios were selected. For faithful classification scenarios, one “heavy,” four “medium,” and three “light” scenarios were selected. The order of scenarios was fixed for all participants in the pre-test in the same randomized order to simplify the interview process. The later pre-tests recruited participants, as in the study, completing the instrument independently. The second round of testing assessed the recruiting and qualification procedures, the assessment outcomes for scenarios, the effectiveness of the manipulation of the independent variables, and the measurement of the reliability of scales. Ten scenarios were selected randomly per participant from the full set. As such the number of erroneous scenarios varied between participants, which allowed evaluating the effect of the number of erroneous scenarios during testing. Later rounds primarily refined manipulation checks, and the final test was the initial set of respondents for the final instrument as used in the study.

F.2 Instrument Testing

F.2.1 *Interview Pre-Test*

The initial pre-test using a telephone interview was performed between January 11th and 13th, 2019 to evaluate the simulation task and assess how participants interpreted the measures. A total of six participants were recruited from Mechanical Turk. The recruiting advertisement was modified to add the telephone interview requirement, but was otherwise identical to the later phases. Each experiment condition was utilized in the pre-test, with the control condition appearing twice. The briefing, rating task, and attitude survey questions were separately discussed for impressions and feedback. The interview guide appears in Appendix E. Two participants had no prior damage assessment experience, two had some prior experience, and two had evaluated many hundreds of images using multiple platforms. The interviews were between 16 and 47 minutes long. Feedback on the overall task about the length and number of images and survey questions was positive with no participants reporting fatigue. None of the participants had previously rated images with the assistance of an automated assessment.

Participants with prior experience reported the tool as “similar to” or “better than” other tools they had used. The images were expressed by participants to be better quality with clearer and less obscured images. More experienced raters had the longest interviews and verbalized their decision-making process in the greatest detail. None of the participants was familiar with the exact rating guideline used in this study, but those with prior written guideline experience felt the guideline in this study was clearer and more useful.

F.2.2 *Pre-Test Sample Description*

The second round of pre-testing was performed from January 15th to 18th, 2019. Participants were recruited as in the study to test overall procedures and measurement reliability.

The instrument utilized appears in Appendix B. Data from this second round was used to evaluate rating of the scenarios and selection for the study. No submissions for qualification were eliminated by the English proficiency requirement. A total of 58 participants were recruited, of which 5 failed the attention check (91.4% pass rate). The median completion time of the survey was 16 minutes. Each condition had between 9 and 14 responses. Scenarios were assessed by between 17 and 32 participants with an average of 24.

The composition of the sample was 42% female, 58% male with an average age of 37 years (median 36). A total of 15% of participants had a high school diploma or equivalent, 34% had some post-high school education, 43% Bachelor's degree, and 7.5% with some graduate schooling or a graduate degree. Seven participants reported no prior experience in damage assessment of aerial images. The 46 respondents with previous experience self-rated a mean 6.98 on an 18-point composite scale, standard deviation 3.04. The highest category of experience was reported by 26% of respondents (more than 500 images), indicating the potential to meaningfully increase discrimination among the most highly experienced.

F.2.3 *Scenario Assessment Outcomes*

The consensus of participant assessments was compared to individual assessments in the initial step (prior to seeing the automated assessment). The goal was to check for participants that answered randomly in the initial step, evaluate any role of expertise in rating outcomes, and evaluate any relationship between time for submission and consensus agreement. Because there are no assumed "correct" ratings for the scenarios, consensus among participants was used to evaluate repeatability of outcomes rather than their accuracy. The number of participants by number of scenarios they agreed with the consensus appears in Table 4.

Table 4 Pilot Agreement with Consensus

Pilot Participant Agreement with Consensus
(# of pilot participants)

# of Scenarios	Damage	Structure
2		1
3	1	2
4	4	3
5	9	3
6	19	6
7	11	10
8	8	14
9	1	12
10		2

Most participants (48 out of the 53) produced damage assessments in agreement with the consensus in at least 50% of the scenarios they assessed. Within this group there was no statistically significant relationship between completion time and agreement with the consensus (Pearson correlation, $p = 0.61$). Within the group of five low-consensus participants, the respondent with the least agreement and ratings consistent with random chance also did not have high agreement with the automated assessment in the review step. While that participant completed the survey in five minutes and 23 seconds, other participants that completed the survey between five and six minutes achieved 80% agreement. While it is likely the single low agreement participant input random ratings, the time to complete the survey was not predictive of rating agreement with the consensus. These results support the validity of the damage assessment task and respondent pool achieve repeatable structure and damage assessments.

F.2.4 Effect of Number of Erroneous Scenarios

The number of erroneous scenarios participants received varied as the ten scenarios were randomly selected from the full set for pilot participants. The median number of erroneous scenarios was 4, with minimum of 2 and maximum of 7. The number of participants for each

count in increasing order was 4, 17, 14, 13, 4, and 1. The conclusions that can be reached are limited due to differences in experimental manipulation and small group sizes; however, any collapse in trust as a result of error rate would be observable. The study variables were assessed for correlation with the number of erroneous scenarios received as well as general linear models using the count as a covariate, and none of the differences in means or correlations were statistically significant. Means with the greatest differences between conditions were the single-item measures of the trust in the intelligent agent and cognitive load. These measures show a potential greater trust and lower cognitive load for the four participants with only 2 erroneous scenarios (6.25 on a 7-point scale and 7.5 on the 9-point scale, respectively), but from 3 to 6 there were only small differences in means (between 4.5 and 4.8, and 7.75 and 8.24). The single participant that received 7 erroneous scenarios rated trust in the intelligent agent “5” and total cognitive load “6.” Without evidence to suggest adjusting the error rate, 50% was retained.

F.2.5 *Scenario Selection for the Study*

The consensus “initial step” participant assessment prior to being presented the automated assessment for each of the evaluated scenarios appears in Table 5. Scenarios are grouped into those selected for the study to represent faithful scenarios, erroneous scenarios, and those not selected. Ratings highlighted in green indicate agreement of the participant consensus with the simulated automated assessment. The table also shows sample size per scenario and the automated assessment presented in the “review step.” The percentages indicated are the proportions of participants that agreed with the consensus for structure type and damage classification, and the proportion that disagreed with the automated damage assessment. In the initial step the consensus was shared by an average of 72% of participants on structure type and 62% for damage classification. One-way analysis of variance was used to evaluate differences in

outcomes between scenarios. The test of perceived difficulty was significant $F(21,529)=1.783$, $p = 0.018$; however, a Tukey HSD test did not differentiate means between scenarios ($p = 0.058$). Differences in the duration of the initial assessment were not significant $F(21,529)=1.425$, $p = 0.100$.

Table 5 Pilot Scenario Ratings and Outcomes

Type	Scenario #	n	Initial Step Assessment				Automated Disagree Damage	Time (seconds)	Perceived Difficulty	Automated Assessment	
			Consensus Rating		Participant Agreement					Structure	Damage
Faithful	7	19	Light	Minimal	74%	58%	79%	57	3.1	Light	No Damage
	8	25	Heavy	Critical	48%	52%	48%	39	3.2	Heavy	Critical
	10	25	Medium	Significant	80%	52%	48%	41	3.7	Medium	Significant
	18	28	Heavy	Critical	61%	79%	79%	26	2.5	Heavy	Significant
	22	21	Heavy	Significant	57%	62%	38%	42	3.1	Medium	Significant
Erroneous	1	17	Heavy	Significant	76%	53%	100%	30	3.3	Heavy	No Damage
	9	32	Medium	No Damage	84%	56%	100%	36	3.4	Medium	Critical
	17	29	Medium	Minimal	90%	41%	100%	36	3.2	Medium	Critical
	19	23	No Structure	No Damage	91%	74%	96%	36	3.3	Medium	Significant
	21	21	Heavy	Minimal	62%	76%	100%	29	3.5	Medium	No Damage
Not Selected	2	29	Heavy	Significant	62%	59%	69%	40	3.2	Medium	Critical
	3	24	Heavy	Minimal	83%	46%	54%	28	2.5	Heavy	Minimal
	4	22	Light	Critical	82%	50%	50%	29	3.3	Light	Critical
	5	22	Medium	Minimal	59%	45%	55%	33	3.4	Light	Minimal
	6	24	Light	Critical	71%	83%	17%	21	2.8	Light	Significant
	11	27	Heavy	Significant	70%	59%	41%	33	3.2	Heavy	Significant
	12	25	Medium	Significant	60%	72%	28%	27	3.2	Medium	Significant
	13	20	Medium	Minimal	50%	60%	60%	27	3.7	Medium	Significant
	14	22	Medium	Minimal	73%	86%	14%	24	2.2	Medium	Minimal
	15	26	Medium	Minimal	88%	69%	31%	32	2.8	Medium	Minimal
16	29	Medium	Significant	62%	48%	52%	29	3.4	Medium	Significant	
20	20	No Structure	No Damage	95%	90%	10%	23	2.4	Medium	No Damage	

Green highlight indicates agreement of participant consensus and the automated assessment.

Erroneous scenarios were selected on the basis of: (a) inconsistency of the automated output with the guideline, and (b) lack of agreement with the automated assessment in the initial review. None of the participants selected the same damage classification as the automated assessment in the initial step for four of the erroneous scenarios selected, and one scenario had one participant (4%) select the erroneous rating. While the correct answer is not necessarily determinable, this will ensure the erroneous answers were unlikely to be agreeable to the vast majority of participants.

Faithful scenarios were selected on the basis of being (a) most clearly consistent with the damage assessment guideline compared to the other faithful scenarios, and (b) exhibiting a level of disagreement with participants in the initial review such that the automated assessment acts as more than a confirmation of initial assessments. The selected faithful scenarios had average consensus agreement of approximately 64% for structure and 60% damage classifications; however, that consensus was not in agreement with the yet-to-be-revealed automated assessment for three scenarios. Disagreement with the automated damage classification was 59% on average, ranging between 38% to 79% by scenario.

F.2.6 Evaluation of the Measurement Model

Scale reliability was assessed with Cronbach alpha calculated using SPSS version 25 using the pre-test data collection of participants after the interviews were completed. Each of the dependent and alternative explanation composite measures was found to have Cronbach alpha metrics of 0.708 or greater. Exploratory factor analysis using principal components analysis with varimax rotation was employed to analyze the multi-item dependent measures.

For the self-efficacy items, SE8 and SE9 extracted into a second component with an eigenvalue of 1.27, item loadings greater than 0.850, and first component item loadings of 0.133 and 0.192. These two items were most closely related to the automated assessor's impact on the ability to perform the task. Reliance has been found as a separate dimension from both confidence and trust in automation complacency (Lee & See, 2004; Singh et al., 1993). Item SE4 had the lowest loading on the first component (0.700) and greatest cross-loading (0.524) and as least distinct on either dimension, was dropped from the study. By removing items SE4, SE8, and SE9 alpha increased from 0.905 to 0.930.

The deleted self-efficacy items SE8 and SE9 were relabeled RE1 and RE2 to evaluate the explanatory potential of “automation reliance.” Their composite score (by simple addition) had correlations greater than an absolute value of 0.200 with germane cognitive load ($r = 0.307, p = 0.025$), extraneous cognitive load ($r = -0.250, p = 0.071$), previous task experience ($r = -0.265, p = 0.056$), and dispositional trust ($r = 0.240, p = 0.084$). A third item RE3, “I think the automated assessor prevents manual assessment errors” was added to meet recommendations of having at least three-items for PLS-SEM analysis of a composite construct (Hair et al., 2016), and the scale was retained as an alternate explanation.

Cronbach’s alpha for the three components of cognitive load were: intrinsic 0.786, germane 0.829, and extraneous 0.708. The nine items of cognitive load extracted three components with 73% of the variance, with the items loading 0.654 or greater on their respective dimensions. ICL2 was the highest cross-loading item loaded 0.690 on the intrinsic load component and 0.493 on germane load component, followed by ECL1 loading 0.654 on extraneous load component and 0.295 on intrinsic component load. The remaining cross loadings were below 0.200. All items were retained since they loaded most strongly on their original dimensions as developed by Klepsch et al. (2017). Total cognitive load did not have a statistically significant correlation with the sum of all cognitive load scores ($r = 0.172, p = 0.217$), however it is known that the three components are not best modeled as additive with equal weight (Klepsch et al., 2017). The single item for total cognitive load had a statistically significant correlation with the germane load composite score ($r = 0.504, p < 0.001$). Across all pilot participants only the top three points of the 9-point scale were utilized.

The six items of the attribution of agent intelligence scale extracted a single component with eigenvalue of 3.402 representing 56.7% of the variance of the items. The lowest loaded item

was ATT5 “is predictable” with loading of 0.575. Cronbach’s alpha was 0.839 for the whole scale, however deleting ATT5 would increase to 0.847. With limited improvement and consistent membership from the source study (Terada & Yamada, 2017), the item was retained.

The viability of the alternate explanations was assessed. With the exception of previous task experience, each had a correlation with dependent variable of an absolute value of $r = 0.2$ or greater. As a result, all of the alternate explanations were retained.

F.2.7 Manipulation Check Refinement

The test of manipulation of the independent variables for explanation types produced significant differences in means using Independent T-Tests in the correct association with the presence of the manipulation with the original items: causal explanation difference=3.20 ($t = 9.81, df = 51, p < 0.001$), counterfactual explanation difference=1.38 ($t = 3.71, df = 51, p = 0.001$), and hedging explanation difference=1.16 ($t = 3.20, df = 51, p = 0.002$).

The original measures were opinion statements about the agent rated on a 7-point Likert-type scale. Binary classification tests² were utilized to analyze participant response to the manipulation items making the responses Binary with a threshold of “Agree” and above being considered a “Yes.” A score of 1.00 on a test metric indicates perfect matching of participant response and the manipulation state and 0.00 indicates answering inverse to the intended state. The initial item wording found sensitivity for the causal explanation was 0.91 indicating a high ability for participants to detect the explanation, and specificity was 1.00 indicating that no participants incorrectly identified the presence of the explanation. Further, precision was 1.00 showing that no participants falsely indicated the condition while also indicating the presence of

² Sensitivity = Count of true positives divided by the sum of count of true positive and false negative.
 Specificity = Count of true negatives divided by the sum of count of true negatives and false positives.
 Precision = Count of true positives divided by the sum of count of true positives and false positives.

the condition correctly. For counterfactual and hedging conditions sensitivities were 0.78 and 0.83, specificities were 0.67 and 0.53, and precision was 0.64 and 0.58. These results indicate that the majority of respondents recalled the presence of the explanations consistently with the intended manipulation, with some lack of sensitivity, specificity, and precision for counterfactual and hedging explanations. While these results were consistent with successful manipulation, revisions to the items and rating options were tested to understand the false positive ratings for the presence of counterfactual and hedging explanations when they were not present.

In the final revision of the items wording was simplified further and more directly connected to the manipulation, and responses were made binary “Yes” or “No” with a third option “Don’t Know/Don’t Remember” which was treated as a missing response. Reviewing the feedback from testing, some participants were viewing the task as a qualification for future damage assessment work. To address any potential of participants rating the tool and the agent with future qualification in mind, a message was added at the start of the measures which asked the participants to rate their opinion honestly and that their answers would have no effect on their eligibility for future tasks. The first 56 valid respondents for the study in the final form were used to evaluate the final revisions. Sensitivity, specificity, and precision for the explanation types were: causal 89%, 90%, 98%; counterfactual 76%, 60%, 57%; hedging 83%, 69%, 67%. Five respondents (9%) reported not knowing/remembering whether another classification was compared, and no respondents of the other explanation types used that option. These revisions reflected small improvements for hedging explanations, and small decreases in classification matching for causal and counterfactual explanations. With a similar response pattern across multiple revisions and false responses occurring in the lack of presence of the explanation type, data collection was completed without further modifications.

F.2.8 *Improvements to the Task*

Over the course of testing the following improvements were made to the task: a notice was placed after the briefing to ensure that participants understood the task was part of a study, and a reminder was given not to share the contents with other potential participants. Each of the participants interpreted the simulation as a genuine damage assessment tool and two participants discussed how they expected participants to compare interpretations of guidelines and images on public forums in an effort to improve their understanding and the quality of their results. The notice was added to limit the extent that the manipulation might be exposed by comparison across participants. Some participants evaluated buildings in the image other than the intended building. This was primarily occurring in the second step following the automated assessment output to rationalize erroneous assessments. To increase clarity of the task, a box with a blue boundary was added around the subject building in the post-damage image to highlight the subject of the classification but not to highlight specific elements of the explanation. Checkboxes were added in the second step on each image to provide feedback for images that were (a) obscured or (b) blurry, and to (c) indicate that the automated assessment was incorrect. The ADAM (Automated Damage Assessment Machine) acronym was added to simplify references to the simulated agent. An advisory statement that the answers would not impact eligibility for future tasks was included.

Only half of the participants passed the original attention check despite otherwise being attentive in the telephone interview. The distractors were made less challenging and the “none of the above” option was replaced with a fourth distractor. No revisions were suggested or made to

the wording of the briefing or to the study measures and participants expressed meanings behind the items of the measures that was consistent with their constructs.

REFERENCES

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). *Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda*. Paper presented at the 2018 CHI Conference on Human Factors in Computing Systems, Montreal.
- Accenture. (2017). Reworking the Revolution. Retrieved April 23, 2018 from https://www.accenture.com/t20180125T024403Z_w_us-en/acnmedia/PDF-69/Accenture-Reworking-the-Revolution-Jan-2018.pdf
- Achkar, Z. A., Baker, I. L., & Raymond, N. A. (2016). Imagery Interpretation Guide: Assessing Wind Disaster Damage to Structures. *Harvard Humanitarian Initiative*. Retrieved October 23, 2018 from <https://hhi.harvard.edu/publications/satellite-imagery-interpretation-guide-assessing-wind-disaster-damage-structures>
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160. doi:10.1109/ACCESS.2018.2870052
- Ahrndt, S., Fähndrich, J., & Albayrak, S. (2016). *Human-agent teamwork: what is predictability, why is it important?* Paper presented at the 31st Annual ACM Symposium on Applied Computing, Pisa.
- Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6, 14410-14430. doi:10.1109/ACCESS.2018.2807385
- Albuquerque, J., Herfort, B., & Eckle, M. (2016). The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping. *Remote Sensing*, 8(10), 859.
- Anderson, C. A. (1983). The causal structure of situations: The generation of plausible causal attributions as a function of type of event situation. *Journal of Experimental Social Psychology*, 19(2), 185-203.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). *Vqa: Visual question answering*. Paper presented at the 2015 IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile.
- Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2006). The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS Quarterly*, 30(1), 79-97.
- Attari, N., Ofli, F., Awad, M., Lucas, J., & Chawla, S. (2017). *Nazr-cnn: Fine-grained classification of uav imagery for damage assessment*. Paper presented at the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA).
- Bagozzi, R. P., & Yi, Y. (1989). On the use of structural equation models in experimental designs. *Journal of Marketing Research*, 26(3), 271-284.
- Bagozzi, R. P., Yi, Y., & Singh, S. P. (1991). On the use of structural equation models in experimental designs: Two extensions. *International Journal of Research in Marketing*, 8(2), 125-140.
- Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, 11(4), 042609.
- Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2), 191.

- Bandura, A. (1986). *Social foundation of thought and action: A social-cognitive view*. Englewood Cliffs: Prentice-Hall, Inc.
- Bandura, A. (1997). *Self-efficacy : the exercise of control*: New York : W.H. Freeman and Company.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents*, 5(1), 307-337.
- Beckmann, J. F. (2010). Taming a beast of burden – On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, 20(3), 250-264. doi:<https://doi.org/10.1016/j.learninstruc.2009.02.024>
- Benbasat, I., & Todd, P. (1996). The effects of decision support and task contingencies on model formulation: A cognitive perspective. *Decision Support Systems*, 17(4), 241-252.
- Biran, O., & Cotton, C. (2017). *Explanation and justification in machine learning: A survey*. Paper presented at the IJCAI 2017 Workshop on Explainable Artificial Intelligence, Melbourne.
- Böhm, G., & Pfister, H.-R. (2015). How people explain their own and others' behavior: a theory of lay causal explanations. *Frontiers in Psychology*, 6(139). doi:10.3389/fpsyg.2015.00139
- Bradshaw, J. M., Feltovich, P. J., Johnson, M., Breedy, M., Bunch, L., Eskridge, T., . . . van Diggelen, J. (2009). *From tools to teammates: Joint activity in human-agent-robot teams*. Paper presented at the International conference on human centered design, San Diego.
- Brown, S. D., Lent, R. W., Telander, K., & Tramayne, S. (2011). Social cognitive career theory, conscientiousness, and work performance: A meta-analytic path analysis. *Journal of Vocational Behavior*, 79(1), 81-90.
- Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). *The role of explanations on trust and reliance in clinical decision support systems*. Paper presented at the 2015 International Conference on Healthcare Informatics, Dallas.
- Cao, Q. D., & Choe, Y. (2018). Deep Learning Based Damage Detection on Post-Hurricane Satellite Imagery. *arXiv preprint arXiv:1807.01688*.
- Chamales, G. (2013). Towards trustworthy social media and crowdsourcing. Wilson Center Commons Lab. Retrieved October 21, 2018 from https://www.wilsoncenter.org/sites/default/files/TowardsTrustworthySocialMedia_FINA_L.pdf
- Chander, A., Srinivasan, R., Chelian, S., Wang, J., & Uchino, K. (2018). *Working with Beliefs: AI Transparency in the Enterprise*. Paper presented at the 23rd International Conference on Intelligent User Interfaces, Tokyo.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4), 293-332.
- Cheah, J.-H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM: On using single-item versus multi-item measures in redundancy analyses. *International Journal of Contemporary Hospitality Management*, 30(11), 3192-3210. doi:10.1108/IJCHM-10-2017-0649
- Cheng, G., & Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117, 11-28.
- Cohen, J. (1988). Statistical power analysis for the behavioural sciences. In Hillsdale: Erlbaum.
- Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2), 189-211.

- Corbane, C., Saito, K., Dell’Oro, L., Bjorgo, E., Gill, S. P., Emmanuel Piard, B., . . . Spence, R. J. (2011). A comprehensive analysis of building damage in the 12 January 2010 MW7 Haiti earthquake using high-resolution satellite and aerial imagery. *Photogrammetric Engineering & Remote Sensing*, 77(10), 997-1009.
- Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., . . . Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5), 455.
- Crippen, K. J., & Earl, B. L. (2007). The impact of web-based worked examples and self-explanation on performance, problem solving, and self-efficacy. *Computers & Education*, 49(3), 809-821.
- Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1), 7.
- Daniels, Z. A., & Metaxas, D. (2018). *ScenarioNet: An Interpretable Data-Driven Model for Scene Understanding*. Paper presented at the 2nd Workshop on Explainable Artificial Intelligence, Stockholm.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management science*, 35(8), 982-1003.
- de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331-349. doi:10.1037/xap0000092
- Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68(4), 87-106.
- Dennett, D. C. (1989). *The intentional stance*: MIT press.
- Dittus, M. S. (2017). *Analysing Volunteer Engagement in Humanitarian Crowdmapping*. (Doctoral Dissertation), UCL (University College London), Retrieved from <http://discovery.ucl.ac.uk/id/eprint/10024735>
- Doan, T. H. (2018). jquery.magnify (Version 2.3.2). github. Retrieved from <https://github.com/thdoan/magnify/blob/master/dist/js/jquery.magnify.js>
- Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description*. Paper presented at the IEEE conference on computer vision and pattern recognition, Boston.
- Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.
- Došilović, F. K., Brčić, M., & Hlupić, N. (2018). *Explainable artificial intelligence: A survey*. Paper presented at the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia.

- Duarte, D., Nex, F., Kerle, N., & Vosselman, G. (2017). Towards a More Efficient Detection of Earthquake Induced FAÇADE Damages Using Oblique Uav Imagery. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, 42(2), W6.
- Duarte, D., Nex, F., Kerle, N., & Vosselman, G. (2018). Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4(2).
- Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2017). Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a Revised Theoretical Model. *Information Systems Frontiers*. doi:10.1007/s10796-017-9774-y
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697-718. doi:[https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and disuse of automated aids. *Human Factors and Ergonomics Society Annual Meeting*, 43(3), 339-339.
- Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. (Vol. Summary Report AFOSR-3223 under Contract AF 49 (638)-1024. SRI Project 3578 for Air Force Office of Scientific Research). Menlo Park, CA.: Stanford Research Institute.
- Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology*, 70(6), 1164-1184. doi:10.1037/0022-3514.70.6.1164
- Fiske, S. T., & Taylor, S. E. (2016). *Social cognition: From brains to culture*: Sage.
- Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation*, 80, 38.
- Frank, J., Rebbapragada, U., Bialas, J., Oommen, T., & Havens, T. (2017). Effect of Label Noise on the Machine-Learned Classification of Earthquake Damage. *Remote Sensing*, 9(8), 803.
- Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual review of psychology*, 63, 287-313.
- Fujita, A., Sakurada, K., Imaizumi, T., Ito, R., Hikosaka, S., & Nakamura, R. (2017). *Damage detection from aerial images via convolutional neural networks*. Paper presented at the IAPR International Conference on Machine Vision Applications, Nagoya, Japan.
- Gartner. (2018). *CEO Survey: CIOs Should Guide Business Leaders Toward Deep-Discipline Digital Business*. Retrieved from <https://www.gartner.com/doc/3870869/-ceo-survey-cios-guide>
- Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4), 367-382.
- Ghosh, S., Burachas, G., Ray, A., & Ziskind, A. (2018). *Generating Natural Language Explanations for Visual Question Answering Using Scene Graphs and Visual Attention*. Paper presented at the 2nd Workshop on Explainable Artificial Intelligence, Stockholm.
- Giboney, J. S., Brown, S. A., Lowry, P. B., & Nunamaker, J. F. (2015). User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems*, 72, 1-10. doi:<https://doi.org/10.1016/j.dss.2015.02.005>

- Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, 57(6), 940.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069*.
- GIScorps. (2013). Aerial Damage Assessment Following Hurricane Sandy. Retrieved October 29, 2018 from http://www.giscorps.org/sandy_109/
- Gist, M. E., & Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *Academy of Management Review*, 17(2), 183-211.
- Glass, A., McGuinness, D. L., & Wolverton, M. (2008). *Toward establishing trust in adaptive agents*. Paper presented at the 13th international conference on Intelligent user interfaces, Los Angeles.
- Google (Cartographer). (2018). Google Earth Satellite Images (Map Data: Google). Retrieved from <http://www.earth.google.com>
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, 315(5812), 619-619. doi:10.1126/science.1134475
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125-130.
- Green, B., & Chen, Y. (2019). *Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments*. Paper presented at the Proceedings of the Conference on Fairness, Accountability, and Transparency.
- Gregor, S. (2001). Explanations from knowledge-based systems and cooperative problem solving: an empirical study. *International Journal of Human-Computer Studies*, 54(1), 81-105. doi:<https://doi.org/10.1006/ijhc.2000.0432>
- Habibovic, A., Andersson, J., Nilsson, M., Lundgren, V. M., & Nilsson, J. (2016, 19-22 June 2016). *Evaluating interactions with non-existing automated vehicles: three Wizard of Oz approaches*. Paper presented at the 2016 IEEE Intelligent Vehicles Symposium (IV), Gothenburg.
- Hagar, C. (2010). Introduction to the Special Section. *Bulletin of the American Society for Information Science and Technology*, 36(5), 10-12. doi:doi:10.1002/bult.2010.1720360504
- Hair, J. F., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management & Data Systems*, 117(3), 442-458.
- Hair, J. F., Hult, G., Tomas, M., Ringle, C. M., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*: Sage Publications.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York: John Wiley & Sons.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2), 243-259.
- Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). *Generating visual explanations*. Paper presented at the European Conference on Computer Vision, Amsterdam.
- Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). *Grounding visual explanations*. Paper presented at the 2018 European Conference on Computer Vision, Munich.

- Henseler, J. (2018). Partial least squares path modeling: Quo vadis? *Quality & Quantity*, 52(1), 1-8.
- Herman, B. (2017). The Promise and Peril of Human Evaluation for Model Interpretability. *arXiv preprint arXiv:1711.07414*.
- Hesketh, B. (1997). Dilemmas in training for transfer and retention. *Applied Psychology*, 46(4), 317-339.
- Hesslow, G. (1988). The problem of causal selection. In D. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11-32): Harvester Press.
- Hilton, D. (2007). Causal explanation. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 232-253): Guilford Publications.
- Hoffman, R. R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining Explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intelligent Systems*, 33(3), 87-95. doi:10.1109/MIS.2018.033001421
- Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human-computer interaction. *Computers in Human Behavior*, 26(6), 1278-1288.
- Holliday, D., Wilson, S., & Stumpf, S. (2013). *The effect of explanations on perceived control and behaviors in intelligent systems*. Paper presented at the CHI '13 Extended Abstracts on Human Factors in Computing Systems, Paris, France.
- Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., & Zatloukal, K. (2017). Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. *arXiv preprint arXiv:1712.06657*.
- Hu, P. J.-H., Hu, H.-f., & Fang, X. (2017). Examining the Mediating Roles of Cognitive Load and Performance Outcomes in User Satisfaction with a Website: A Field Quasi-Experiment. *MIS Quarterly*, 41(3), 975-A911.
- Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K. (2017). *Modeling relationships in referential expressions with compositional modular networks*. Paper presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu.
- Hu, W.-L., Akash, K., Jain, N., & Reid, T. (2016). Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine*, 49(32), 48-53.
- Hyland, K. (1996a). Talking to the academy: Forms of hedging in science research articles. *Written communication*, 13(2), 251-281.
- Hyland, K. (1996b). Writing without conviction? Hedging in science research articles. *Applied linguistics*, 17(4), 433-454.
- Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). *AIDR: Artificial intelligence for disaster response*. Paper presented at the 23rd International Conference on World Wide Web, Seoul.
- Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B., & Sierhuis, M. (2012). Autonomy and interdependence in human-agent-robot teams. *IEEE Intelligent Systems*, 27(2), 43-51. doi:10.1109/MIS.2012.1
- Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B., & Sierhuis, M. (2014). Coactive design: designing support for interdependence in joint activity. *J. Hum.-Robot Interact.*, 3(1), 43-69. doi:10.5898/JHRI.3.1.Johnson
- Joseph, S. (2013). *Measuring cognitive load: A comparison of self-report and physiological methods*. (Doctoral Dissertation), Arizona State University Tempe, AZ, Retrieved from

- https://repository.asu.edu/attachments/110550/content/SchinkJoseph_asu_0010E_12971.pdf
- Joshi, A. R., Tarte, I., Suresh, S., & Koolagudi, S. G. (2017). *Damage identification and assessment using image processing on post-disaster satellite imagery*. Paper presented at the 2017 IEEE Global Humanitarian Technology Conference (GHTC), San Jose.
- Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, 86(1), 80.
- Kahneman, D., & Tversky, A. (1981). *The simulation heuristic*. Retrieved from <https://apps.dtic.mil/docs/citations/ADA099504>
- Kamar, E., & Manikonda, L. (2017). *Complementing the Execution of AI Systems with Human Computation*. Paper presented at the 31st AAAI Conference on Artificial Intelligence, San Francisco.
- Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, 57, 227-254.
- Kerle, N., & Hoffman, R. R. (2013). Collaborative damage mapping for emergency response: the role of Cognitive Systems Engineering. *Natural Hazards & Earth System Sciences*, 13(1).
- Kersbergen, D. (2018). *Automated Building Damage Classification using Remotely Sensed Data: Case study: Hurricane Damage on St. Maarten*. (Masters Thesis), University of Delft, Retrieved from <https://repository.tudelft.nl/islandora/object/uuid%3A43a82eb1-cc37-48fc-ba39-f6a1375ba3c7>
- Kiatpanont, R., Tanlamai, U., & Chongstitvatana, P. (2016). Extraction of actionable information from crowdsourced disaster data. *Journal of emergency management*, 14(6), 377-390.
- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6), 91-95.
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, 8, 1997.
- Kluckner, S., Mauthner, T., Roth, P. M., & Bischof, H. (2009). *Semantic classification in aerial imagery by integrating appearance and height information*. Paper presented at the 9th Asian Conference on Computer Vision, Xi'an, China.
- Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal*, 28(1), 227-261.
- Kock, N., & Lynn, G. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the association for information systems*, 13(7).
- Kolbe, A. R., Hutson, R. A., Shannon, H., Trzcinski, E., Miles, B., Levitz, N., . . . Muggah, R. (2010). Mortality, crime and access to basic needs before and after the Haiti earthquake: a random survey of Port-au-Prince households. *Medicine, conflict and survival*, 26(4), 281-297.
- Koutsoumari, M., & Antoniou, A.-S. (2016). Self-Efficacy as a Central Psychological Capacity within the Construct of Positive Organizational Behavior: Its Impact on Work. *New Directions in Organizational Psychology and Behavioral Medicine*, 147.
- Krueger, J., Ham, J. J., & Linford, K. M. (1996). Perceptions of behavioral consistency: Are people aware of the actor-observer effect? *Psychological science*, 7(5), 259-264.

- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121.
- Kryven, M., Ullman, T., Cowan, W., & Tenenbaum, J. (2016). *Outcome or Strategy? A Bayesian Model of Intelligence Attribution*. Paper presented at the CogSci 2016, Austin, TX.
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). *Too much, too little, or just right? Ways explanations impact end users' mental models*. Paper presented at the 2013 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), San Jose.
- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). *Interpretable decision sets: A joint framework for description and prediction*. Paper presented at the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco.
- Lallemant, D., Soden, R., Rubinyi, S., Loos, S., Barns, K., & Bhattacharjee, G. (2017). Post-disaster damage assessments as catalysts for recovery: A look at assessments conducted in the wake of the 2015 Gorkha, Nepal, earthquake. *Earthquake Spectra*, 33(S1), S435-S451.
- Leaman, J., & La, H. M. (2017). A Comprehensive Review of Smart Wheelchairs: Past, Present, and Future. *IEEE Transactions on Human-Machine Systems*, 47(4), 486-499.
doi:10.1109/THMS.2017.2706727
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.
- Leppink, J., & van Merriënboer, J. J. (2015). The beast of aggregating cognitive load measures in technology-based learning. *Educational Technology & Society*, 18(4), 230-245.
- Licklider, J. C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics*(1), 4-11.
- Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.
- Lobato, E. J., Wiltshire, T. J., Hudak, S., & Fiore, S. M. (2014). No time, no problem: Mental state attributions made quickly or after reflection do not differ. *Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1341-1345.
- Lodewyk, K. R., & Winne, P. H. (2005). Relations among the structure of learning tasks, achievement, and changes in self-efficacy in secondary students. *Journal of educational psychology*, 97(1), 3.
- Malle, B. F. (2006). *How the mind explains behavior: Folk explanations, meaning, and social interaction*: Mit Press.
- Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In *Advances in experimental social psychology* (Vol. 44, pp. 297-352): Elsevier.
- Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.
- Marsh, S., & Dibben, M. R. (2003). The role of trust in information science and technology. *Annual Review of Information Science and Technology*, 37(1), 465-498.
- Mascha, M. F., & Smedley, G. (2007). Can computerized decision aids do “damage”? A case for tailoring feedback and task complexity based on task experience. *International Journal of Accounting Information Systems*, 8(2), 73-91.
doi:<https://doi.org/10.1016/j.accinf.2007.03.001>

- Mayer, H. (1999). Automatic object extraction from aerial imagery—a survey focusing on buildings. *Computer vision and image understanding*, 74(2), 138-149.
- McKenna, F. P., & Myers, L. B. (1997). Illusory self-assessments—Can they be reduced? *British Journal of Psychology*, 88(1), 39-51.
- McLaughlin, M. L., Cody, M. J., & Read, S. J. (2013). *Explaining one's self to others: Reason-giving in a social context*: Routledge.
- Michelucci, P. (2016). Human computation and convergence. *Handbook of science and technology convergence*, 455-474.
- Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 267, 1-38.
- Miller, T., Howe, P., & Sonenberg, L. (2017). *Explainable AI: Beware of inmates running the asylum*. Paper presented at the IJCAI-17 Workshop on Explainable Artificial Intelligence, Stockholm.
- Molden, D. C., Plaks, J. E., & Dweck, C. S. (2006). “Meaningful” social inferences: Effects of implicit theories on inferential processes. *Journal of Experimental Social Psychology*, 42(6), 738-752.
- Moranduzzo, T., & Melgani, F. (2014). Automatic car counting method for unmanned aerial vehicle images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(3), 1635-1647.
- Morgeson, F. P., & Humphrey, S. E. (2006). The Work Design Questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology*, 91(6), 1321.
- Moss, A. J., & Litman, L. (2018). After the bot scare: Understanding what’s been happening with data collection on MTurk and how to stop it. Retrieved February 4, 2019 from <https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it>
- Moulin, B., Irandoust, H., Bélanger, M., & Desbordes, G. (2002). Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review*, 17(3), 169-222.
- Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1802.00682*.
- Nees, M. A. (2016). Acceptance of Self-driving Cars: An Examination of Idealized versus Realistic Portrayals with a Self-driving Car Acceptance Scale. *Human Factors and Ergonomics Society Annual Meeting*, 60(1), 1449-1453.
- NGS. (2018). October 2018: Hurricane Michael. Retrieved October 22, 2018 from <https://storms.ngs.noaa.gov/>
- Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). *Damage assessment from social media imagery data during disasters*. Paper presented at the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Sydney.
- Nicholson, J., Hardin, A., & Nicholson, D. (2003). *Test Performance and the Medium: Unearthing Differences That Make a Difference*. Paper presented at the 2003 Americas Conference on Information Systems, Tampa.
- Nikolaidis, S., Kwon, M., Forlizzi, J., & Srinivasa, S. (2018). Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(3), 22.

- Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3-5), 393-444.
- Nushi, B., Kamar, E., & Horvitz, E. (2018). *Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure*. Paper presented at the Sixth AAAI Conference on Human Computation and Crowdsourcing, Zürich.
- Ofli, F., Meier, P., Imran, M., Castillo, C., Tuia, D., Rey, N., . . . Parkan, M. (2016). Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data*, 4(1), 47-59.
- Ostermann, F. (2015). *Hybrid geo-information processing: Crowdsourced supervision of geo-spatial machine learning tasks*. Paper presented at the 18th AGILE International Conference on Geographic Information Science, Lisbon, Portugal.
- Oviatt, S. (2006). *Human-centered design meets cognitive load theory: designing interfaces that help people think*. Paper presented at the 14th ACM international conference on Multimedia, Santa Barbara.
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of educational psychology*, 84(4), 429.
- Paas, F. G., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1), 63-71.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3), 381-410.
- Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*. Paper presented at the 31st IEEE Conference on Computer Vision and Pattern Recognition, Long Beach.
- Pennington, D. C. (2012). *Social cognition*: Routledge.
- Peterson, T. O., & Arnn, R. B. (2005). Self-Efficacy: The Foundation of Human Performance. *Performance Improvement Quarterly*, 18(2), 5-18. doi:doi:10.1111/j.1937-8327.2005.tb00330.x
- Poblet, M., García-Cuesta, E., & Casanovas, P. (2014). Crowdsourcing tools for disaster management: A review of platforms and methods. In *AI Approaches to the Complexity of Legal Systems* (pp. 261-274): Springer.
- Potter, R. E., & Balthazard, P. (2004). The role of individual memory and attention processes during electronic brainstorming. *MIS Quarterly*, 28(4), 621-643.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.
- Qi, K., Yang, C., Guan, Q., Wu, H., & Gong, J. (2017). A Multiscale Deeply Described Correlations-Based Model for Land-Use Scene Classification. *Remote Sensing*, 9(9), 917.
- Reeder, G. D. (2013). Attribution as a gateway to social cognition. In *The Oxford handbook of social cognition* (pp. 95-117): Oxford University Press, Oxford.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why should i trust you?: Explaining the predictions of any classifier*. Paper presented at the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco.
- Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1), 119-136.

- Ringle, C. M., Wende, S., & Becker, J. M. (2015). SmartPLS 3.2.7. Boenningstedt: SmartPLS GmbH. Retrieved from <http://www.smartpls.com>
- Robins, R. W., Spranca, M. D., & Mendelsohn, G. A. (1996). The actor-observer effect revisited: Effects of individual differences and repeated social interactions on actor and observer attributions. *Journal of Personality and Social Psychology*, *71*(2), 375.
- Rodríguez-Entrena, M., Schuberth, F., & Gelhard, C. (2018). Assessing statistical differences between parameters estimates in Partial Least Squares path modeling. *Quality & Quantity*, *52*(1), 57-69.
- Roese, N. J. (1997). Counterfactual thinking. *Psychological bulletin*, *121*(1), 133.
- Rook, F. W., & Donnell, M. L. (1993). Human cognition and the expert system interface: Mental models and inference explanations. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*(6), 1649-1661.
- Rose, J. M. (2005). Decision aids and experiential learning. *Behavioral Research in Accounting*, *17*(1), 175-189.
- Ross, C., & Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News* <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments>.
- Ruel, E. (2015). Pretesting and Pilot Testing. In E. Ruel, W. E. Wagner III, & B. J. Gillespie (Eds.), *The practice of survey research* (pp. 101-119): Sage.
- Sadilek, A., Kautz, H. A., DiPrete, L., Labus, B., Portman, E., Teitel, J., & Silenzio, V. (2016). Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. *AI magazine*, *38*, 3982-3990.
- Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of business research*, *69*(10), 3998-4010.
- Sawicka, A. (2008). Dynamics of cognitive load theory: A model-based approach. *Computers in Human Behavior*, *24*(3), 1041-1066. doi:<https://doi.org/10.1016/j.chb.2007.03.007>
- Schaffer, J., Giridhar, P., Jones, D., Höllerer, T., Abdelzaher, T., & O'Donovan, J. (2015). *Getting the message?: A study of explanation interfaces for microblog data analysis*. Paper presented at the 20th International Conference on Intelligent User Interfaces, Atlanta.
- Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological science*, *23*(8), 842-847.
- SEC. (2013). *Administrative Proceeding, File No. 3-15570, Release 70694*. Retrieved from <https://www.sec.gov/litigation/admin/2013/34-70694.pdf>
- See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., . . . Laakso, M. (2016). Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS international journal of geo-information*, *5*(5), 55.
- Seufert, T., Jänen, I., & Brünken, R. (2007). The impact of intrinsic cognitive load on the effectiveness of graphical help for coherence formation. *Computers in Human Behavior*, *23*(3), 1055-1071. doi:<https://doi.org/10.1016/j.chb.2006.10.002>
- Shrestha, S. (2018). Improved fully convolutional network with conditional random field for building extraction. *Remote Sensing*, *10*(7), 1135.

- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the complacency-potential rating scale. *The international journal of aviation psychology*, 3(2), 111-122.
- Sklar, E. I., & Azhar, M. Q. (2018). *Explanation through Argumentation*. Paper presented at the 6th International Conference on Human-Agent Interaction, Southampton, United Kingdom.
- Smith, P. J. (2017). Making brittle technologies useful. In *Cognitive Systems Engineering* (pp. 181-208): CRC Press.
- Smith, P. J. (2018). Conceptual frameworks to guide design. *Journal of Cognitive Engineering and Decision Making*, 12(1), 50-52.
- Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(3), 360-371.
- Solomon, R. L. (1949). An extension of control group design. *Psychological bulletin*, 46(2), 137.
- Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological bulletin*, 124(2), 240.
- Stajkovic, A. D., & Sommer, S. M. (2000). Self-efficacy and causal attributions: Direct and Reciprocal Links. *Journal of Applied Social Psychology*, 30(4), 707-737.
- Steele-Johnson, D., Beauregard, R. S., Hoover, P. B., & Schmidt, A. M. (2000). Goal orientation and task demand effects on motivation, affect, and performance. *Journal of Applied Psychology*, 85(5), 724.
- Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology*, 27(2), 165.
- Streukens, S., Wetzels, M., Daryanto, A., & De Ruyter, K. (2010). Analyzing factorial data using PLS: application in an online complaining context. In *Handbook of partial least squares* (pp. 567-587): Springer.
- Suermondt, H. J., & Cooper, G. F. (1993). An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research*, 26(3), 242-254.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285.
- Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and instruction*, 8(4), 351-362.
- Sweller, J., van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, 10(3), 251-296.
- Tan, W.-K., Tan, C.-H., & Teo, H.-H. (2012). Consumer-based decision aid that explains which to buy: Decision confirmation or overconfidence bias? *Decision Support Systems*, 53(1), 127-141. doi:<https://doi.org/10.1016/j.dss.2011.12.010>
- Terada, K., & Yamada, S. (2017). Mind-Reading and Behavior-Reading against Agents with and without Anthropomorphic Features in a Competitive Situation. *Frontiers in Psychology*, 8(1071). doi:10.3389/fpsyg.2017.01071
- Thelisson, E., Padh, K., & Celis, L. E. (2017). *Regulatory mechanisms and algorithms towards trust in AI/ML*. Paper presented at the 2017 Workshop on Explainable Artificial Intelligence, Melbourne.
- Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-Psychological Interpretation of Human vs. Humanoid Robot Behavior: Exploring the Intentional Stance toward Robots. *Frontiers in Psychology*, 8(1962). doi:10.3389/fpsyg.2017.01962

- Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling, 17*(3), 510-534.
- Thompson, R., Compeau, D. R., & Higgins, C. A. (2006). Intentions to use information technologies: An integrative model. *Journal of Organizational and End User Computing (JOEUC), 18*(3), 25-46.
- Tianfield, H., & Wang, R. (2004). Critic Systems—Towards Human—Computer Collaborative Problem Solving. *Artificial Intelligence Review, 22*(4), 271-295.
- Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In *Recommender systems handbook* (pp. 479-510): Springer.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge university press.
- Trekin, A., Novikov, G., Potapov, G., Ignatiev, V., & Burnaev, E. (2018). *Satellite Imagery Analysis for Operational Damage Assessment in Emergency Situations*. Paper presented at the International Conference on Business Information Systems, Berlin.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.
- Van Bouwel, J., & Weber, E. (2002). Remote Causes, Bad Explanations? *Journal for the Theory of Social Behaviour, 32*(4), 437-449. doi:doi:10.1111/1468-5914.00197
- Van de Ven, A. H. (2007). *Engaged Scholarship : A guide for organizational and social research*. Oxford, New York: Oxford University Press.
- van der Waa, J., van Diggelen, J., van den Bosch, K., & Neerinx, M. (2018). *Contrastive explanations for reinforcement learning in terms of expected consequences*. Paper presented at the 2nd Workshop on Explainable Artificial Intelligence, Stockholm.
- Van Dongen, K., & van Maanen, P.-P. (2006). Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. *Human Factors and Ergonomics Society Annual Meeting, 50*(3), 225-229.
- Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational psychology review, 22*(2), 155-174.
- Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). *Cider: Consensus-based image description evaluation*. Paper presented at the 2015 IEEE conference on computer vision and pattern recognition, Boston.
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research, 11*(4), 342-365.
- Vetrivel, A., Gerke, M., Kerle, N., Nex, F., & Vosselman, G. (2018). Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 140*, 45-59. doi:<https://doi.org/10.1016/j.isprsjprs.2017.03.001>
- Vetrivel, A., Gerke, M., Kerle, N., & Vosselman, G. (2015). Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. *ISPRS Journal of Photogrammetry and Remote Sensing, 105*, 61-78. doi:<https://doi.org/10.1016/j.isprsjprs.2015.03.016>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2).

- Wang, F. (2017). *Understanding High Resolution Aerial Imagery Using Computer Vision Techniques*. (Doctoral Dissertation), Rochester Institute of Technology, Retrieved from <https://scholarworks.rit.edu/theses/9553/>
- Wang, R. Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences*, *111*, 139-147. doi:<https://doi.org/10.1016/j.cageo.2017.11.008>
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, *92*(4), 548.
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374-11379.
- Weld, D., & Bansal, G. (2018). The Challenge of Crafting Intelligible Intelligence. *arXiv preprint arXiv:1803.04263*.
- Wen, T.-H., Gasic, M., Mrksic, N., Rojas-Barahona, L. M., Su, P.-H., Vandyke, D., & Young, S. (2016). Multi-domain neural network language generation for spoken dialogue systems. *arXiv preprint arXiv:1603.01232*.
- Westrope, C., Banick, R., & Levine, M. (2014). Groundtruthing OpenStreetMap Building Damage Assessment. *Procedia Engineering*, *78*, 29-39. doi:<https://doi.org/10.1016/j.proeng.2014.07.035>
- Wiener, N. (1950). *The human use of human beings: Cybernetics and society*: Perseus Books Group.
- Woods, D. D. (1985). Cognitive technologies: The design of joint human-machine cognitive systems. *AI magazine*, *6*, 86-92.
- Yu, M., Yang, C., & Li, Y. (2018). Big Data in Natural Disaster Management: A Review. *Geosciences*, *8*(5), 165.
- Zheng, R., McAlack, M., Wilmes, B., Kohler-Evans, P., & Williamson, J. (2009). Effects of multimedia on cognitive load, self-efficacy, and multiple rule-based problem solving. *British Journal of Educational Technology*, *40*(5), 790-803.
- Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy*, *2*(2), 7-33.

VITA

Sean E. Dougherty was born in Tampa, Florida. He holds a Bachelor of Science in Information Systems from the University of South Florida, a Master of Business Administration from Stetson University, and an Executive Doctorate in Business from Georgia State University. For over twenty years his focus in both professional practice and academic research has been integrating intelligent technology into real-world operations. His research background began at the University of South Florida with computer vision technologies, where his work was published in the journal article “Edge Detector Evaluation Using Empirical ROC Curves” in 1999. For eighteen years he was a professional in the heavy industrial field services industry with director responsibilities over functions including information systems, supply chain, and commercial operations. His most recently published research is the conference paper “Will Automated Trucks Trigger the Blame Game and Socially Amplify Risks?” which conducted an experiment to understand how future autonomous truck incidents might be interpreted and re-communicated on social media and word-of-mouth networks. He was awarded top solution in the Santa Fe Institute’s Complexity Explorer Challenge 2018 with the paper “Envious Agents and the Tragedy of the Digital Commons” which used agent-based modeling and intelligent agents to explore strategies for cooperation in a resource-scarce setting. Sean is a Senior Member of the Institute for Electrical and Electronics Engineers and is a certified Lean Six Sigma Black Belt. He is a member of the Florida Chapter Board for Gift of Adoption, and a volunteer financial literacy instructor for Project Prosper, teaching refugees and recent immigrants to the United States.