NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

INFLUENCING TRUST IN HUMAN AND ARTIFICIAL INTELLIGENCE TEAMING THROUGH HEURISTICS

by

Joel E. Thompson

June 2021

Thesis Advisor: Mollie R. McGuire
Second Reader: Michael Senft

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2021	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE INFLUENCING TRUST IN HUMAN AND ARTIFICIAL INTELLIGENCE TEAMING THROUGH HEURISTICS	**5. FUNDING NUMBERS**
6. AUTHOR(S) Joel E. Thompson	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

This thesis analyzes potential methods intended to influence trust within military units and their use of artificial intelligence (AI) systems. AI systems are being developed to enhance the human decision-making process and when employed properly can greatly increase the rate at which actions are taken, a key requirement for generating combat power. Human and AI teams rely on the user's trust for the AI system, and that trust is influenced by rational, affective, and normative trust factors. This thesis examines those trust factors and determines that only rational trust factors are directly connected to the trustworthiness of the AI and that the user's trust can be influenced independently of the AI's trustworthiness through affective and normative trust factors. Influencing the user's trust of the AI through substitution of affective and normative trust factors in place of rational trust factors produces unjustified trust because this trust is not dependent on the trustworthiness of the AI.

14. SUBJECT TERMS heuristic, bias, decision-making, artificial intelligence, AI, human and artificial intelligence team, trust, trustworthy, influence, deception, unjustified trust, unjustified mistrust	**15. NUMBER OF PAGES** 87
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

INFLUENCING TRUST IN HUMAN AND ARTIFICIAL INTELLIGENCE TEAMING THROUGH HEURISTICS

Joel E. Thompson
Captain, United States Marine Corps
BA, Texas A&M University, 2015

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION WARFARE SYSTEMS ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
June 2021

Approved by: Mollie R. McGuire
Advisor

Michael Senft
Second Reader

Alex Bordetsky
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This thesis analyzes potential methods intended to influence trust within military units and their use of artificial intelligence (AI) systems. AI systems are being developed to enhance the human decision-making process and when employed properly can greatly increase the rate at which actions are taken, a key requirement for generating combat power. Human and AI teams rely on the user's trust for the AI system, and that trust is influenced by rational, affective, and normative trust factors. This thesis examines those trust factors and determines that only rational trust factors are directly connected to the trustworthiness of the AI and that the user's trust can be influenced independently of the AI's trustworthiness through affective and normative trust factors. Influencing the user's trust of the AI through substitution of affective and normative trust factors in place of rational trust factors produces unjustified trust because this trust is not dependent on the trustworthiness of the AI.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DNN	Deep Neural Network
NN	Neural Network
RL	Reinforcement Learning
SL	Supervised Learning
UL	Unsupervised Learning

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

> The rapidly improving ability of computer systems to solve problems and to perform tasks that would otherwise require human intelligence is transforming many aspects of human life and every field of science. It will be incorporated into virtually all future technology. The entire innovation base supporting our economy and security will leverage AI. How this "field of fields" is used—for good and for ill—will reorganize the world. (National Security Commission on Artificial Intelligence, 2021, p. 20)

As great power rivals and the two leaders in AI development, the United States of America (USA) and the People's Republic of China (PRC) are seeking to compete *on* and *using* AI. The competition on AI reflects a desire to be the global leader of AI research, development, and employment. This competition is most pronounced in the information and economic arenas, where the USA currently holds the pole position for AI development (National Security Commission on Artificial Intelligence, 2019, p. 20). This reality is not overlooked by the PRC, which is committed to a strategy of heavily resourced and highly emphasized endeavors to close that competitive gap in the coming decade (State Council, 2017, pp. 5–7). The competition using AI will also occur in the information and economic arenas but will also include the military arena. Both country's view the incorporation of AI into warfare as totally transformational and fundamentally necessary to its future (National Security Commission on Artificial Intelligence, 2021, p. 79) (Office of the Secretary of Defense, 2020, p. 161). This perspective is because of the incredible value that AI offers to the competitor that can most effectively employ it first.

AI is attractive to the military because of its high potential to increase the rate at which vast quantities of data can be processed at minute levels of detail into actionable information. This increased rate of action directly translates to a higher operational speed which is desirable because speed is one of the foundational aspects of combat power (U.S. Marine Corps, 2018, p. 2.19). Being faster than an adversary means having the initiative against them, a position where they must react and they cannot choose the time nor place of confrontation (U.S. Marine Corps, 2018, pp. 2.11-2.12). Considering that speed produces combat power, it should come as no surprise that AI, as a source of speed, is being heavily prioritized by militaries. For the foreseeable future, it is most likely that both

countries will seek to employ personnel and AI together in human-AI teams (HAT). The HAT benefits from the accuracy and rapidity of processing possessed by the AI paired with the creative, normative, and perceptive abilities of the human; properly teamed, each performs tasks best suited to it that results in better overall performance (National Security Commission on Artificial Intelligence, 2021, p. 80). Thus, the military which can deploy the best HAT will likely gain a significant combat advantage over the other.

The potential combat power advantages gained by an adversary through the use of HAT requires development of countermeasures to mitigate that potential advantage. This potential adversary advantage can be overcome by developing faster friendly capabilities, degrading the adversary's capabilities, or by choosing a strategy that makes the adversary's advantage irrelevant (e.g., utilizing diplomacy, information, or economic power in such a way that the adversary will not choose to utilize its military power). This research will focus on investigating exploitation of vulnerabilities to disrupt the adversary's HAT.

Critical to the functioning of the HAT is the need for the AI to be trustworthy, which is the basis for the user's trust of the AI. Without trust, the user will likely either monitor the outputs of the AI so closely that the speed advantage of the HAT is lost or will stop using the AI altogether (J. Lee & See, 2004, p. 50). Thus, friendly efforts to decrease the mission performance of the adversary HAT should focus on undermining trust.

This research will define a method for decreasing an adversary's HAT performance by determining how to use heuristics to influence the user's trust level in the AI. Heuristics are the focus of Chapter II, which introduces the cognitive processes of perception and intuitive decision making. It will be shown that an intuitive decision relies upon both the available data in the environment and the subconscious substitution of a related judgement in order to make that decision. Making decisions based upon a substituted judgement, the decisionmaker is exposed to the potential for errors. Exploring AI systems is the focus of Chapter III, which describes its fundamental subcomponents and the potential for an AI to be competent or compromised. These two states represent the high potential of AIs described in the opening paragraphs and the alternate potential for AIs to be faulty in design or vulnerable to deliberate attack. Chapter IV combines these topics and outlines that user's trust can vary independently of the AI's competency, and that the variance is subject to

influence using substituted trust factors. Chapter IV describes methods to influence a user's trust within a HAT using heuristics will be defined. The conclusion provides study limitations and areas for future research identified while conducting this research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. SCHEMA AND INTUITIVE DECISION MAKING AS COGNITIVE PROCESSES

A. INTRODUCTION

This chapter provides a foundation for the cognitive processes and concepts that will be applied in Chapter IV. A discussion of these concepts helps to scope the research, focusing specifically on intuitive judgements as described by heuristics. These are intertwined with other cognitive processes and concepts that are discussed to provide the foundation for understanding their role in our judgments about AI, which will be covered in Chapter IV. The concepts discussed are not meant to be an exhaustive exploration in this space but rather are constrained to the scope of this thesis.

B. INFORMATION AND SCHEMA

Using the single term 'information' lacks the specificity necessary to properly develop an understanding of decision making. This is recognized in Chaim Zins' "Conceptual Approaches for Defining Data, Information, and Knowledge":

> The field of Information Science (IS) is constantly changing. Therefore, information scientists are required to regularly review—and if necessary—redefine its fundamental building blocks. (Zins, 2007, p. 479)

His article then collects definitions for the terms data, information, and knowledge from a panel of 57 scholars, representative of most fields within Information Science. The expansion from the singular term of 'information' to 'data', 'information', and 'knowledge' permits a more nuanced approach that recognizes a progression from data to knowledge, as seen in these definitions provided by Donald Kraft:

> **Data** are atomic facts, basic elements of "truth," without interpretation or greater context. It is related to things we sense. **Information** is a set of facts with processing capability added, such as context, relationships to other facts about the same or related objects, implying an increased usefulness. Information provides meaning to data. **Knowledge** is information with more context and understanding, perhaps with the addition of rules to extend definitions and allow inference. (Zins, 2007, p. 484)

For the purposes of this research, these definitions will be used. It is immediately obvious from this expanded definition that data comprises everything that is able to be sensed in the environment around the individual. Even with this nearly infinite source of data, humans are able to consume only limited quantities of data for transformation into information and knowledge. This uptake of data into the mind in meaningful ways is enabled by schemas.

As it relates to cognition, a schema is "a memory structure capable of representing extremely complex constructs employing this information [knowledge] to influence encoding and retrieval of episodic memory, and guide elaborate, context-specific patterns of behavior" (Ghosh & Gilboa, 2014, p. 113). Described another way, a schema is a memory framework that guides perception for anticipated data based upon knowledge and previous experiences that are associated with the present situation. Schemas are broadly characterized by Ghosh and Gilboa as necessarily having an associative network structure, a basis on many episodes, a lack of unit detail, and being adaptable across those many episodes (Ghosh & Gilboa, 2014, p. 105). In addition to these necessary features, schemas consider chronological relationships and hierarchical organizations, assign the same schema-units to multiple schemas, and contain embedded-response options (Ghosh & Gilboa, 2014, p. 105). As a broad constructive example, consider ordering and eating at restaurant C. A 'restaurant schema' enables an individual to automatically apply memories from previous trips to restaurants A and B (many episodes) when visiting restaurant C for the first time because restaurants A, B, and C are categorically similar, even if they are not identical (associative network structure). The drink machine schema-unit developed in previous visits to restaurant A will guide recognition of the drink machine in restaurant C, even if the location in the restaurant or variety of drinks offered is different (lack of unit detail). The payment schema-unit changes across schema as the individual encounters different means of payment (adaptability). Payment is a schema-unit present in the 'restaurant schema' that is also present in the 'retail schema' (same schema-units to multiple schemas). If the general process at restaurant A is order, eat, pay, leave table and the process at restaurant B is order, eat, leave table, pay the difference will likely prompt the individual at restaurant C to inquire about when it is appropriate to pay (chronological

relationship and embedded-response options). The ‘restaurant schema’ may itself be a sub-schema available under a larger schema related to local vehicle travel (hierarchical organizations). Summarily, a schema is the cognitive learning aspect of gaining experience in a general situation.

The volume of data available at any one time, that could potentially turn into knowledge is too vast for humans to attend to all of it. Schemas help to narrow down the data that is attended to and potentially acquired as information, all the way to what is turned into knowledge through selecting, abstracting, interpreting, and integrating the incoming data (Alba & Hasher, 1983, p. 225). Because of this, schemas are intertwined with perception by guiding the senses through the available sensory inputs, or data based upon anticipation of specific data associated with that specific schema, which can be seen in Ulrich Neisser’s Perceptual Cycle, shown in Figure 1 (Neisser, 1976, p. 21).

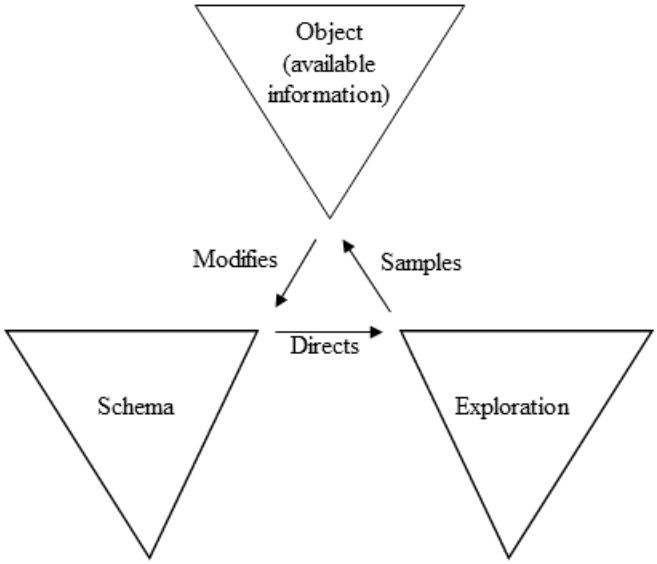

Figure 1. Perceptual Cycle. Source: Neisser (1976, p. 21).

The development and change of a schema is shown by the ‘modifies’ arrow, an interaction that is also represented in Figure 2 (Ghosh & Gilboa, 2014, p. 108).

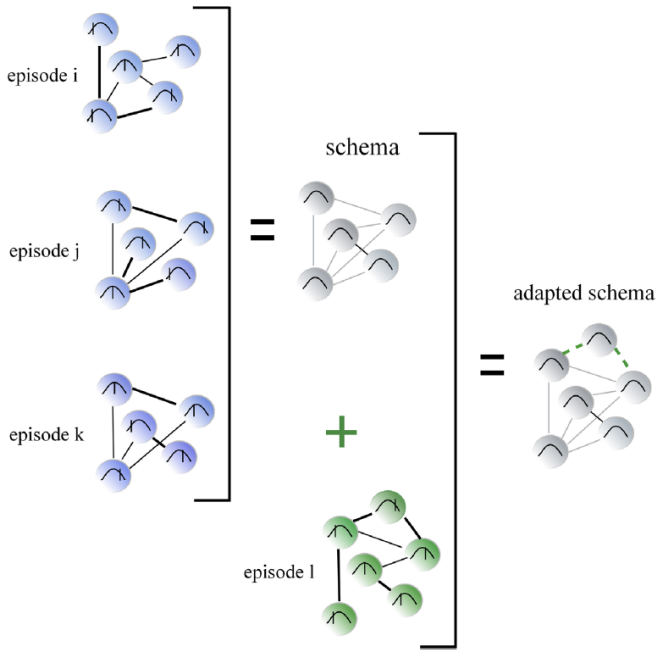

Figure 2. Schema Development and Adaptation. Source: Ghosh and Gilboa (2014, p. 108).

Because of the contextual nature of schemas, a single schema is activated for any one episode. If sensory data that is associated with a schema arrives at an individual, that schema will activate to guide the continuing collection of other pieces of data which are aligned with that schema. Plant and Staunton described this as bottom up activation that drives a top down schema response (Plant & Stanton, 2012, p. 302). The development and activation of a schema is not dependent on a single type of incoming data. In an experiment conducted by Johnson et al., participants who recalled sentences the best had received a contextually appropriate word before the sentence, compared to those who received a contextually inappropriate word or those who only heard "ready" (Johnson et al., 1974, pp. 358–359). For those who received the contextually inappropriate word, the lack of associating schema reduced the ability to recall the sentence. In another study, King et al. demonstrated that schema can be created and activated through motor sequence learning (King et al., 2019, p. 963). Differing groups of study participants were taught the same sequence of button pressing during their first session; during a second session, the different groups were either evaluated on performing the same sequence or a variety of other

sequences (King et al., 2019, pp. 965–967). The new sequences varied the order and transition between buttons, and those who were assigned the most incompatible session two sequences performed worse compared to those who were assigned new and compatible sequences (King et al., 2019, p. 974). This difference can be attributed to the schema developed during session one, because its associations aided those with familiar sequences but hindered those who received completely unfamiliar sequences. In both of these studies, it is apparent that when schemas are activated by data, the individual performs better than those for whom there is no activation or inaccurate activation.

A specific type of schema activation is known as priming. Priming refers to the identification of perceptual objects in an associative manner that is disconnected from consciousness but is not necessarily unconscious (Tulving & Schacter, 1990, p. 302). This means that priming automatically happens as a part of the perceptual cycle and is not interrupted if the individual becomes aware of the stimulus. Priming extends through many psychological disciplines including social psychology where social priming of an individual can take place by presenting cultural identities as the perceptual object. (Molden, 2014, pp. 4–5). What is critical to recognize is that an individual that is primed will more likely act in accordance with the prime than one who is not primed (Weingarten et al., 2016, p. 490). In a study conducted with Asian Americans (n=116), the participants were separated into three groups (American self-priming, Asian self-priming, and control) and asked to complete ten sentences before writing in detail about two important memory events from their life (Wang, 2008, p. 745). The priming manipulation was achieved during the sentence completion, where the self-priming groups responded to five sentence prompts of "As an American (Asian), I am…" and five sentence prompts of "In general, Americans (Asians) are…"; the control group completed sentence prompts related to things in nature (Wang, 2008, p. 745). The effect of priming based on cultural identity appeared in the recall of memories, when those who were self-primed with an American identity were more likely to recall their own perspectives, views, and actions in the memory (I did…, I saw… I thought…) whereas those with an Asian identity were more likely to recall the interactions and behaviors of those involved in the memory (A told B…, C and D walked to…, E helped me…) (Wang, 2008, p. 747).

Priming is important to this thesis because of its compliance effect, where the stimulus guides the following perceptions. Chapter IV will utilize this idea when discussing the trust factors associated with AI.

C. HEURISTICS AND BIASES

The study of heuristics and their attendant biases grew out of an inability for the classical model of rational choice to explain consistently irrational decisions. In the rational model,

> the "rational actor" (i.e., the typical person) chooses what options to pursue by assessing the probability of each possible outcome, discerning the utility to be derived from each, and combining these two assessments. The option pursued is the one that offers the optimal combination of probability and utility. (Gilovich & Griffin, 2002, p. 1)

To arrive at a truly rational choice, the decision maker would need a well filtered collection of data to produce information for their knowledge to then apply the formal rules of probability so that an accurate assessment can be made of the likelihood of outcomes. Filtering data and information is necessary because there is data and information that will produce knowledge that is irrelevant to a rational decision. For the model of rational choice to hold true, the decision maker would need to recognize this and discard that which does not apply.

Common experience and research have shown that humans are not best represented by the rational actor. Herbert Simon is credited with the concept of bounded rationality, the idea that humans and other organisms are 'satisficing' instead of optimizing like in classical rationality (Simon, 1956, p. 136). Satisficing is the meeting of the first minimum threshold of acceptability in the available options, which permits the time limited decider to move onto the next task instead of considering every single option as would be necessary to rationally optimize a decision (Simon, 1956, p. 136). An indirect development of bounded rationality is Tversky and Kahneman's research on intuitive decisions made in uncertain conditions.

Tversky and Kahneman's 1974 "Judgement under Uncertainty: Heuristics and Biases" provides evidence against the 'rational actor' (Tversky & Kahneman, 1974, p.

1124). In one of the originating experiments, Kahneman and Tversky presented study participants with a group of 100 people that were either engineers or lawyers; half of the participants were told that there were 70 engineers and 30 lawyers and the other half were told that there were 30 engineers and 70 lawyers (Kahneman & Tversky, 1973, p. 241). When participants were asked what the probability was of an individual belonging to a specific profession, the responses mirrored the given rate of occurrence for that profession unless there was an uninformative description of the individual (Kahneman & Tversky, 1973, p. 242). An example of a description is as follows:

> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well like by his colleagues. (Kahneman & Tversky, 1973, p. 242)

This description contains no statistically relevant information but still altered the participants response from its correct value of 0.3 or 0.7 to a statistically incorrect value of 0.5 (Kahneman & Tversky, 1973, p. 242). This consistent and predictable error, or bias, was termed "insensitivity to prior probability of outcomes" or neglect of base rates because it biased the value of the description of the individual over the given rates (Tversky & Kahneman, 1974, p. 1124). Clearly, some form of irrationality is impacting the outcome of a rational evaluation and that indicates a fallibility of the rational actor.

These results have been reinforced by numerous studies and the process has been standardized to that of substituting a heuristic attribute for the target attribute within the decision (Kahneman, 2003, p. 707). Although there are others, the most widely recognized heuristics are representativeness, availability, and affect (Gilovich & Griffin, 2002, p. 17). Representativeness is a substitution based on the similarity of a choice option to the available categories (Tversky & Kahneman, 2002, p. 22). Availability is a substitution based on the ease of arrival of mental occurrences (Tversky & Kahneman, 1974, p. 1127). Affect is a substitution base on the inherent goodness or badness of the choice option (Slovic et al., 2002, p. 397). It should be noted that each of these descriptions is the tip of separate academic-and-research icebergs, which are massive and full of nuanced similarities and differences. The importance of heuristics to this thesis comes in the substitution of a separate judgement that carries the potential for a flawed outcome.

In any decision, using the heuristic does not guarantee a flawed outcome, but it does expose the individual to the risk of violating a logical relationship by substituting a contextually related but illogical judgement. Cognitive biases describe the flawed outcomes which can result from the use of heuristics when making decisions. In the case of the previously described engineer and lawyer categorization study, base rate neglect is the identified bias because the study participants subconsciously ignored the given rate of occurrence and instead made evaluations based upon the apparent representativeness of the described individual for the two categories (Tversky & Kahneman, 1974, p. 1125). Another cognitive bias is the omission bias, which results from an affective judgment of the potential consequences of an action as worse than the potential consequences of inaction, even if they consequences are objectively identical (Brown et al., 2010, p. 4182). From the perspective of the decision maker, neither use of these heuristics would seem to be unreasonable, but the results are contradictory to a logical consideration of the information and choices. Thus, careful consideration must be given as to *how* a decision is being made to avoid falling into a cognitive bias.

Heuristic substitution in intuitive decision making is both aligned with and reinforced by dual process theory. Dual process theory is a refinement of bounded rationality that separates intuitive thinking and decision-making from analytical thinking and decision-making and seeks to define their relationship to the other (Sloman, 2002, p. 379). (Sloman, 2002, p. 379). The distinction is a recognition that the individual can think and decide either through a deliberately analytical and logical process or an intuitive process that minimizes cognitive effort. In this research, the two processes will be labelled System 1 and System 2 which respectively represent the separate intuition and analytical reasoning processes (Stanovich & West, 2000, p. 658). Figure 3 provides descriptions of the two processes and perception against the types of content that they predominantly handle (Kahneman, 2003, p. 698). It should be noted that in the assignment of numbers to the two systems, intuition comes first (System 1) and reasoning comes second (System 2). This is reflective of the automatic and pervasive nature of System 1 and indicates that System 2 outputs are produced more slowly than those of System 1 (Epstein et al., 1992, p. 334). This position after System 1 allows System 2 to function as a check on the

reasonableness of the System 1 output and override it when necessary (Stanovich & West, 2000, p. 662). There are situations where System 2 is engaged immediately after perception receives and recognizes data from the environment around that requires analytical thinking (e.g., a calculus test, assessing a mortgage contract, etc.), but in general the effortless nature of System 1 will engage first (Kahneman, 2003, p. 698).

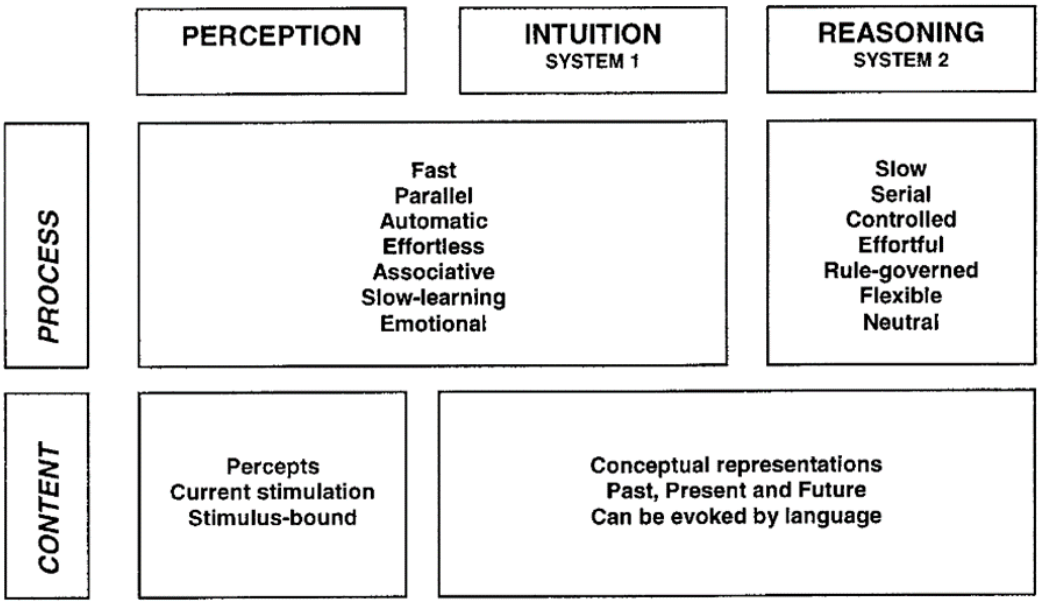

Figure 3. Processes and Content of Perception, Intuition, and Reasoning. Source: Kahneman (2003, p. 698).

These dual processes are illustrated in Figure 4, with emphasis being placed on the unknown heuristic which System 1 used. The decision maker automatically utilizes the System 1 heuristic substitution process to produce a decision. In the illustration, the cloud indicates that the System 1 output is not based upon the true target attribute (e.g., probability) but instead by the heuristic attribute (e.g., representativeness). System 2 provides a rapid assessment of the System 1 output, checking for reasonableness which does not inherently mean that it is checking why System 1 produced that output. If the output is deemed reasonable, the decision is used; if not, the slow and effortful System 2 is used to develop a decision.

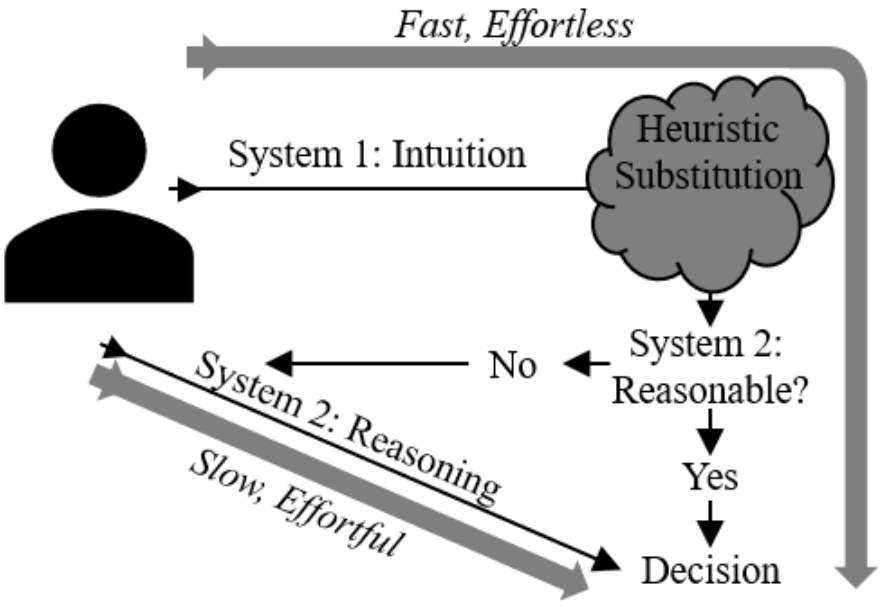

Figure 4. Heuristic Substitution as a Component of Dual Process Theory.

A discussion of heuristics and biases must also recognize their limitations as a means of identifying when to be concerned about their use. Voluminous reading of heuristics and biases literature can cause the reader to believe that individuals have no ability to intuitively arrive at correct responses; a brief moment of self-reflection will show numerous occasions where intuitive decisions produce correct outcomes. A longer moment of self-reflection will show that the belief about the total fallibility of individuals and intuitive thinking is itself the confirmation bias because the reader is steeped in literature primarily describing decision errors and not successes. In an article co-authored by the leading proponent for heuristics and biases, Daniel Kahneman, and the leading proponent for natural decision making (a theory which treats intuitive decision making as a skill that can be honed), Gary Klein, both authors agree that intuitive decisions can produce excellent outcomes and predictable errors and that these are reflective of the individuals making the decisions and the environments around them (Kahneman & Klein, 2009, p. 515). This congruence could be surprising given that the term 'bias' is described as a controversial term in the same article (Kahneman & Klein, 2009, p. 515). The underpinning of their agreement is that certain decisions are placed within highly valid settings, that is there are consistently predictable outcomes from consistent decisions that are consistently cued

(Kahneman & Klein, 2009, p. 520). Chess can be considered a highly valid situation; all pieces are visible, their movement capabilities are known, and these factors do not differ between games of chess. The unpredictability in chess, the opponent, does not invalidate the situation; over time, the s*killed* player develops a familiarity with the situation that incorporates opponent actions as part of the situational cues. At the other end of the continuum are highly invalid environments, where the complexity of action-reaction further inhibits situational cues from priming an appropriate decision which prevents truly skilled intuition from developing (Hogarth, 2001, p. 90). Long-term political forecasting falls within an invalid environment due to the scale and complexity of the system (Kahneman & Klein, 2009, p. 520). Of particular relevance to this thesis, the tactical level of war is specifically listed as a highly valid environment despite its inherent uncertainty because consistently advantageous actions exist that produce further opportunities to gain an advantage (Kahneman & Klein, 2009, p. 524). Tactical actions such as flanking, suppressing, or isolating generally produce the same outcomes, allowing the tactical leader to learn and apply experience across future fights. Thus, military decision makers can develop intuitive decision making as a skill, but should also recognize that there is a limit to using intuition. One such instance is when the validity of the environment decreases, which should be expected when fighting against a new adversary with new systems in a new battlespace (e.g., cyberspace or in orbit). As the novelty of the situation declines through continued exploration and operations, its validity is revealed and intuition can again be counted as a useful tool.

D. CONCLUSION

Human cognition is an immensely complex collection of processes, of which this chapter has provided an introductory understanding of some related to judgments and decision making. The perception of a situation and subsequent judgments and decisions made in that space are based on the understanding of the data available to the decision maker. Schemas help make sense of incoming data by using pre-existing knowledge and experience to attach meaning and turn it into information. Schemas can be primed, and therefore create opportunity to influence someone's perception of data. The perception of that data informs the individual's use of either System 1 intuition or System 2 analysis. The

effortlessness of intuitive decision making comes from automatic heuristic substitution, which introduces a potential for cognitive biases to produce a flawed outcome that is not caught by the System 2 check for reasonableness. Understanding these processes permits extension onto how humans assess and develop trust in AIs, and how that trust can be deliberately influenced to decrease the effectiveness of a HAT.

III. ARTIFICIAL INTELLIGENCE

A. INTRODUCTION

This chapter will introduce artificial intelligence, describing both its competence and ability to be compromised. This binary classification provides the foundation for Chapter IV to determine whether trust within a HAT is justified or unjustified. AI represents the forefront of computing technologies and is appropriately complex given that position. Thus, the following sections in Chapter III should be read for conceptual understanding of AI as a system and its subcomponents.

B. NARROW AI

1. Narrow versus General

AI is a holistic term that comprises several major components and can be conceptually organized in a number of ways. Fundamentally, AI is a combination of datasets, algorithms, software, and hardware (Denning, 2019a, p. 18). It should be noted that the term 'data' and 'dataset' are not interchangeable; 'data' retains the definition assigned in Chapter II while 'dataset' is more aligned with the Chapter II definition of information because it is a collection of data that contains meaning. The interplay of these components varies based upon the AI, but they are the fundamental building blocks present in all AI. In this research, AI will be used as a system level term as well as an adjective for components, namely algorithms, that bear additional significance within the AI.

Within the scope of this thesis, the term AI will solely represent a Narrow AI and not a General AI. A Narrow AI is a computer system (comprising the components of the Canonical Architecture) that performs tasks which augment human intelligence, such as perceiving, learning, classifying, abstracting, reasoning, and acting (Martinez et al., 2019, p. 9). General AI, also called Aspirational AI, is the science fiction representation of AI. No longer limited to a single task, it can seamlessly exist and navigate in physical and informational environments. These definitions set a limit to the extent of the capabilities of the AI: by augmenting human intelligence, a narrow AI does not replicate or replace the total capabilities of the brain. The human brain does serve as an inspiration for certain types of algorithms, a design that has

enabled those AIs to exceed the performance of the human mind in specific tasks. Implied by the name, Narrow AIs are developed for a particular task or set of tasks and often become fragile and useless when applied outside that task (Apte, 2019, p. 15). One analogy for Narrow AI is a razor blade; it is excellent in a limited number of situations such as shaving or performing surgical cutting, but would be completely useless if used to cut down a tree. A similar analogy for General AI is to consider it as a replacement lumberjack; it will be able to decide and then use the proper tool to cut down the tree. There are currently no known General AIs, therefore they will not be considered in this thesis. It cannot be ruled out that General AIs will be developed in the future as Narrow AIs have consistently met and surpassed developmental hurdles (Rowe, 2019, p. 9).

2. Basic Structure

Within Narrow AI, there are a number of fundamental components that can be assumed to be part of any AI. One concept of these core components is the Lincoln Laboratory's AI Canonical Architecture, seen in Figure 5 (Martinez et al., 2019, p. 27). The extension elements are Human-Machine Teaming, which in this thesis is referred to as Human-AI Teaming (HAT), Robust AI, and Users. These were added to the AI structure because of the necessity of the AI to integrate with humans. HAT is a design and operation paradigm that assists in identifying tasks are most suitable for humans or most suitable for the AI. Robust AI is the Lincoln Laboratory term for a trustworthy AI that demonstrates its trustworthiness by being explainable, meaningfully measurable, verifiably tested, resilient to attack, and deployed within organizational policies that empower the use of AI within HAT (Martinez et al., 2019, pp. 62–65). Chapter IV will highlight that AIs can be trusted even in the absence of the Robust AI traits because of trust factors that are not based on these traits, making it all the more critical for designers to produce systems which meet this expectation. Lincoln Laboratory's architecture will be used to frame the remainder of this chapter, with specific emphasis on the data sets and algorithms.

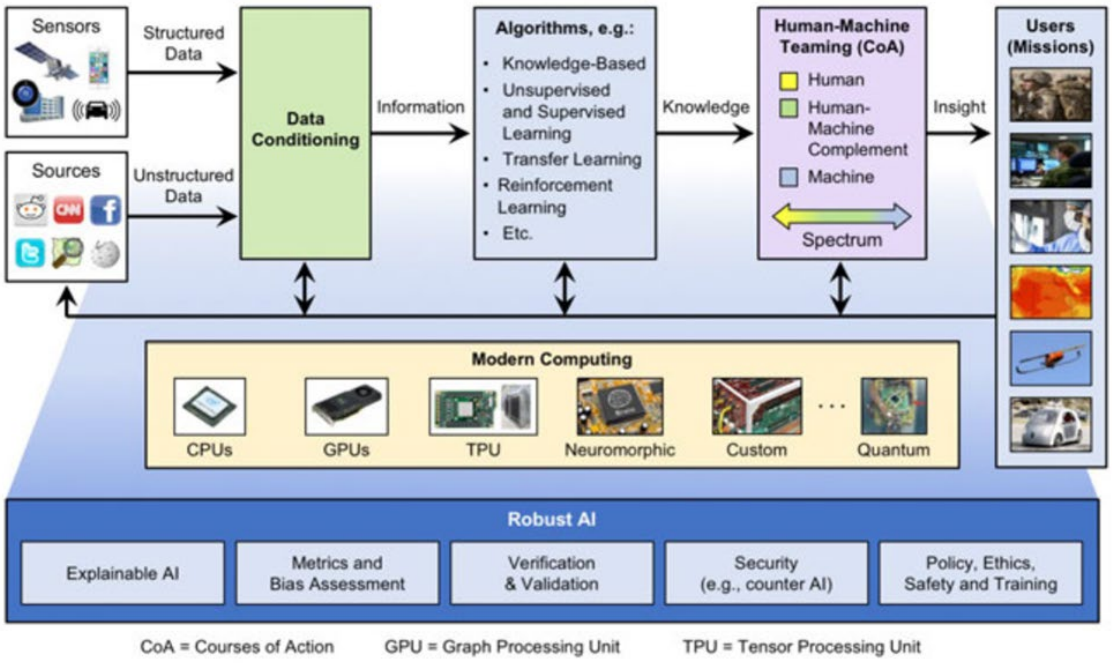

Figure 5. AI Canonical Architecture. Source: Martinez et al. (2019, p. 27).

The three major phases for an AI's lifetime are the design, training, and operational phases. The training phase is where the machine learning takes place and how AI differs from traditional computer programs. Training depends upon the type of algorithm being used, but generally means that the designed algorithm processes relationships amongst a variety of specific data inputs to learn which relationships achieve the designer's identified desired output. This training is often conducted through multiple iterations of data input and data output, which allow the designers to adjust parameters within the design to better guide the learning. Once the model has achieved the desired output, it can be transitioned into its operational mode where it will process its data inputs in accordance with the structure that it acquired during training.

C. DATA, DATA SCIENCE, AND DATABASES

1. Data

The first major component that will be discussed is data. Datasets are the collection of inputs that are evaluated by the AI algorithm that then produces the output. Datasets and their relation to AI algorithms are often described in food terms like 'fed' or 'raw' and

these parallels are instructive to the necessity of datasets to AI. The two major types of data that fill databases are unstructured and structured data. Unstructured data are collections of single objects without context that are often the outputs of sensors (Martinez et al., 2019, p. 43). Structured data are more closely related to the Chapter II definition for information because they are typically packets that contain various pieces of datum, often referred to as metadata (Martinez et al., 2019, p. 43). Structured data predominantly comes from information systems like social media. The expected environment and desired goal of the AI will determine which type of data the AI will need.

2. Data Science and Databases

Data Science, as an academic discipline and profession, is best suited for those determinations because it concerns itself with data analysis, the building of models that correctly apply datasets to accurately predict events, and the validation of those models (Schuchard, 2019, p. 3). Data scientists are needed because of the challenges associated with data conditioning and database construction. Data conditioning, or data munging, intends to modify the 'raw' data into a form that retains its necessary value while being manageable in the dataset and can also be manipulated by the algorithm. An important modification that is required for supervised-learning algorithms is the assignment of labels, which in most cases requires manual assignment at the outset of a database development (Martinez et al., 2019, p. 49). Other data modifications are necessary so that the data can be organized, stored, and accurately recalled within the datasets according to the software language used to code the dataset; examples of those languages include the Structured Query Language (SQL), Not SQL (NoSQL), New SQL (NewSQL), and a variety of combinations meant to store both structured and unstructured data from multiple sources within the same dataset (Martinez et al., 2019, p. 39). The era of Big Data has demanded some of these developments. Three major factors, the volume of data being received, the rate at which that volume arrives at the dataset (i.e., periodic deliveries or continuous flow), and the variety of data being received and assigned to the dataset (Martinez et al., 2019, p. 38). Figure 6 provides a visualization for the scale of Big Data by describing a portion of the data produced in the average minute in 2020 (DOMO, 2021, end of article). Expanding the food parallel, datasets generally follow the 'garbage in, garbage out' maxim: poorly

conditioned data will likely produce useless outputs. Such is the criticality of data conditioning that weeks and potentially months of an AI development project can be consumed by the data munging process, potentially 80% of the development project's lifespan (Martinez et al., 2019, p. 38).

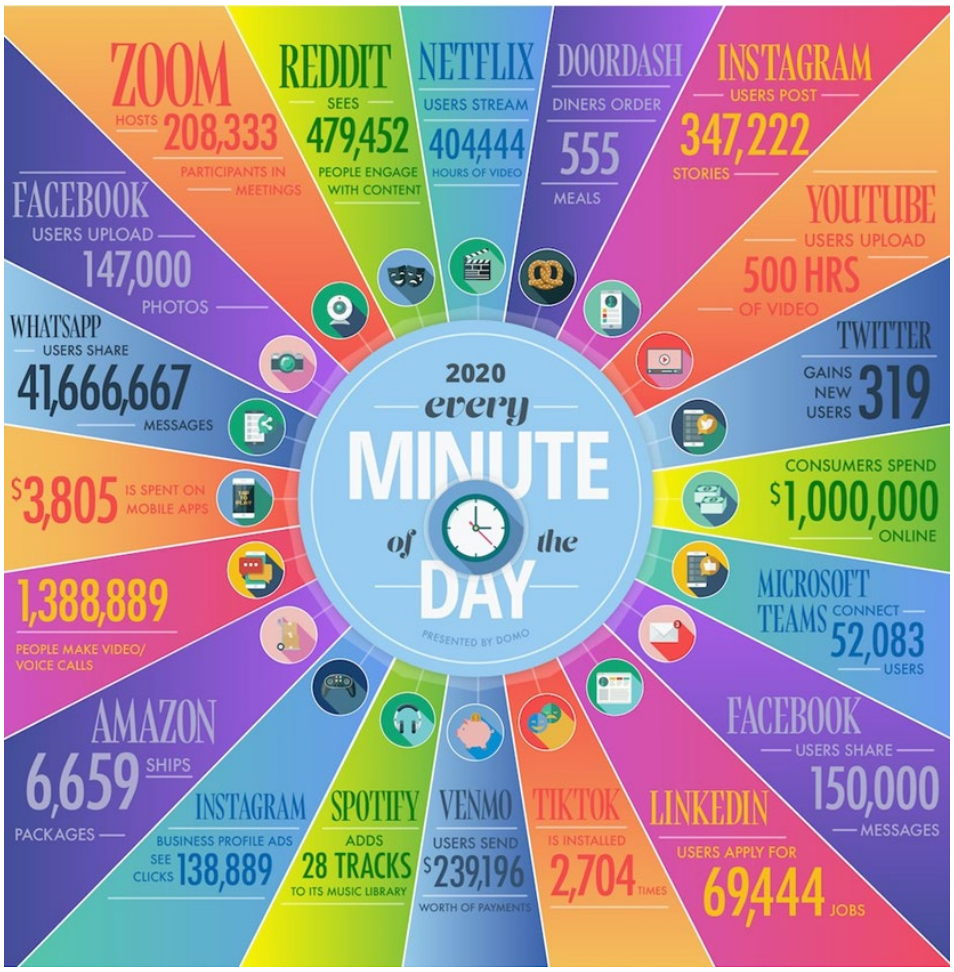

Figure 6. The Average Data per Minute in 2020. Source: DOMO (2021).

D. ALGORITHMS

Algorithms are the mathematical models which are used to evaluate the data and datasets that are fed into the AI. Attempting to differentiate amongst the types of algorithms on the basis of 'intelligence' is inexact because of the multitude of ways to define intelligence and its intrinsic linkage with human cognition (Denning, 2019b, p. 6). A better means of

describing the hierarchy of AI algorithms is by their learning power which is the ability to acquire new capacity for action (Denning, 2019b, pp. 2–3). Using this method, the AI Group at Naval Postgraduate School produced the following six hierarchical levels of learning with their respective algorithm types and dataset requirements, shown in Table 1.

Table 1. Hierarchical Levels of Learning and their Respective Algorithm Types and Dataset Requirements. Adapted from Denning (2019b, p. 2).

NPS Levels of Learning			Algorithm Type		Dataset Requirement
0	Basic Automation	*operates as*	a response to defined inputs	*based upon*	programmed outputs
1	Rule-based System	*learns by using*	designer determined hierarchical decision processes		designer organized network of knowledge
2	Supervised Learning		neural network based on designer's framework		labeled dataset
3	Unsupervised Learning				unlabeled dataset
4	Human-Machine Teaming		any combination of levels 1-3		the level 1-3 appropriate dataset(s)
5	Aspirational AI		unknown		unknown

In general, the table shows broad similarity with the AI Canonical Architecture. Level 0 is basic automation and is represented by systems like traditional computers which receive inputs to perform a task that will produce a consistent output (Denning, 2019b, p. 13). Level 1 is rule-based or expert systems, which produce outputs based upon a manually structured logical deduction which considers the input against the static dataset (Denning, 2019b, p. 14). Level 2 is supervised learning, which utilizes neural networks and a labeled training dataset to learn a desired output and then evaluate inputs based upon the learned features of the training dataset (Denning, 2019b, p. 15). In relation to this hierarchy, Level 2 is where the term 'Machine Learning' begins because it is the first level where the system develops its own process for achieving its task. Level 3 is unsupervised learning, which utilizes neural networks and an un-labeled dataset to generate a desired output based upon associations inferred by the AI algorithm (Denning, 2019b, p. 16). The AI Canonical Architecture identifies a specific usage of unsupervised learning called reinforcement learning, which will be discussed separately in this thesis. Level 4 is Human-Machine Teaming (HAT in this thesis) and contains in its use of Levels 0–3 the ability to learn how

to augment and aid its human companion or user (Denning, 2019b, p. 17). This is the only level that would not be regarded by the Canonical Architecture of AI as an AI algorithm because the Canonical Architecture separates HAT from the algorithms because of its task assignment requirement. This thesis will discuss HAT in alignment with the Canonical Architecture. Level 5 is Aspirational or General AI which was discussed in Section B of this chapter and has yet to be achieved. The following subsections will describe Levels 1–3 with an additional subsection for both neural networks and reinforcement learning. Level 4 will topically be covered in the opening sections of Chapter IV.

1. Level 1: Rule-Based System

Ruled-Based or Expert Systems are the most basic form of AI, despite their immense complexity in design. The major separation between a rule-based system and a Level 2 and higher system based on neural networks is that the rule-based system has a fixed knowledge base and can only learn when taught specific new knowledge relationships by its designers (Monaco, 2019, p. 13). This means that the knowledge base must be created by a knowledgeable expert who must prescribe not only the data objects but their relationships to each other in a logical manner that the algorithm can then use to reason for its outputs (Martinez et al., 2019, p. 46). This requirement is both an advantage and a major hindrance to the use of rule-based systems. The time requirement for the development and updating of the knowledge base is extensive, as seen with the preeminent rule-based system developer Cycorp, which has spent 35 years producing and managing its knowledgebase; a simplified representation of the design of their knowledge base and how it moves from high level concepts to arbitrary facts is shown in Figure 7 (Cycorp, 2021, third heading). There are also issues with the ability of an expert to identify an association among large quantities of data with minute differences in one dimension and massive differences in another. Viewed by advocates of neural networks, these sort of complex problems are best solved using learning instead of programming, if there is sufficient computation power and data available (Krizhevsky et al., 2017, p. 84) The use of a manually constructed knowledge base does produce an AI that is computationally cheap, because the system must only follow the logical relations written into the knowledge base instead of developing its own processes that require numerous iterations of complex,

parallel operations (Monaco, 2019, p. 18). Because these rules are written by the human designer, the relationships can be used by the rule-based system to explain why it made a specific decision (Monaco, 2019, p. 18). This is a highly desirable trait that presents a major hurdle for most neural network-based AI algorithms.

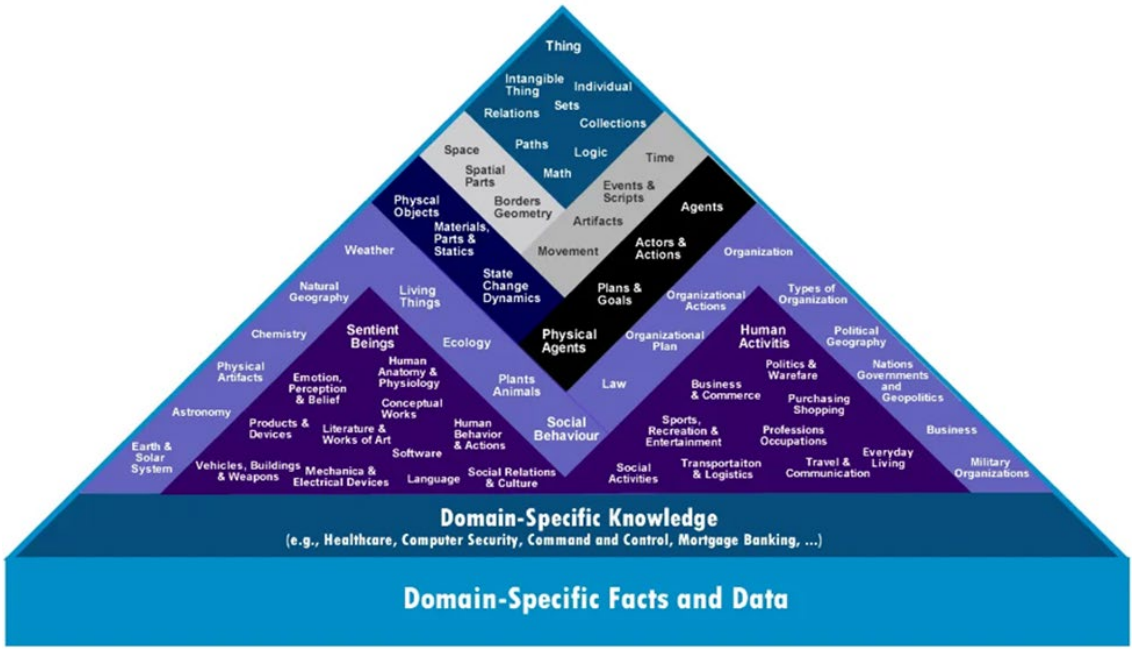

Figure 7. Cycorp's Knowledgebase. Source: Monaco (2019, p. 17).

2. Level 2 and Beyond: Neural Networks

Neural Networks (NN) provide the mathematical foundation for computers to learn as a capability instead of simply receiving knowledge through teaching. A brief introduction is needed as NNs are fundamental to the subsections on level 2 supervised learning, level 3 unsupervised, and level 3 reinforcement learning. NNs are filled with mathematical neurons, which are modeled on the biological brain's neurons and their connections to each other. A visualization of the mathematical neuron and its inspiration, the biological neuron, is shown in Figure 8.

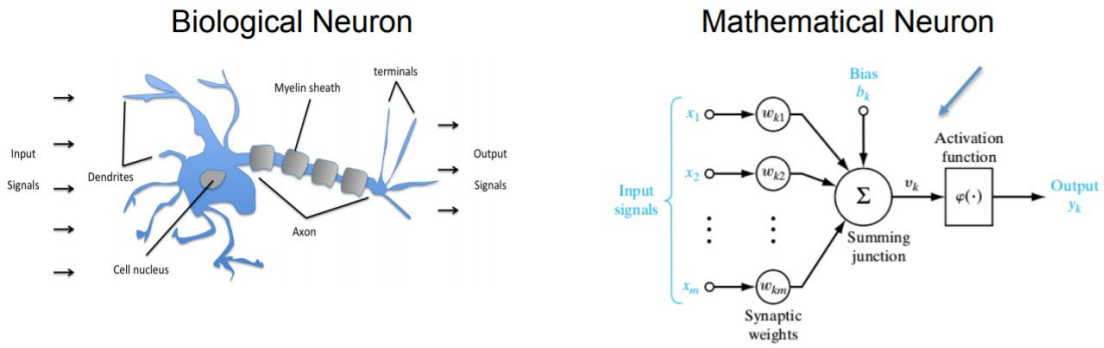

Figure 8. Comparison of a Biological Neuron and a Mathematical Neuron. Source: Orescanin (2019, p. 10).

The mathematical neuron is a summation function within a network layer that receives multiple input 'signals' or values; this reception is modeled on the biological neuron's dendrites receipt of electrical impulses from other neurons or nerves. Those values are assigned a scaling weight and added with a scaled bias, the sum of which is evaluated at an activation function that produces an output value for the neuron. The outcome of that activation function is either the final algorithm output or the input into the next layer (Orescanin, 2019, p. 10). This function is shown in the equation below (Martinez et al., 2019, p. 51).

$$y_{l+1} = f\left(w_l y_l + b_l\right)$$

$$y_l = \text{ layer } l \text{ output}, \; w_l = \text{ weight between } l \text{ and } l+1,$$

$$b_l = \text{ corresponding layer bias}$$

Two advanced implementations of NN are the deep neural network (DNN) and the convolutional neural network (CNN). DNNs expand the NN by inserting a number of hidden layers between the input and output layers that each perform evaluations that then pass through the following layers and eventually arrive at the output layer (DeepAI, 2019, second heading). Because of the increased number of layers available, DNNs can incorporate 10^8 parameters, permitting non-deterministic and extremely under-constrained optimizations (Kölsch, 2019, p. 7). It is generally understood that more hidden layers will produce better outputs (Martinez et al., 2019, p. 52). Figure 9 shows a NN and a DNN (DeepAI, 2019, second heading).

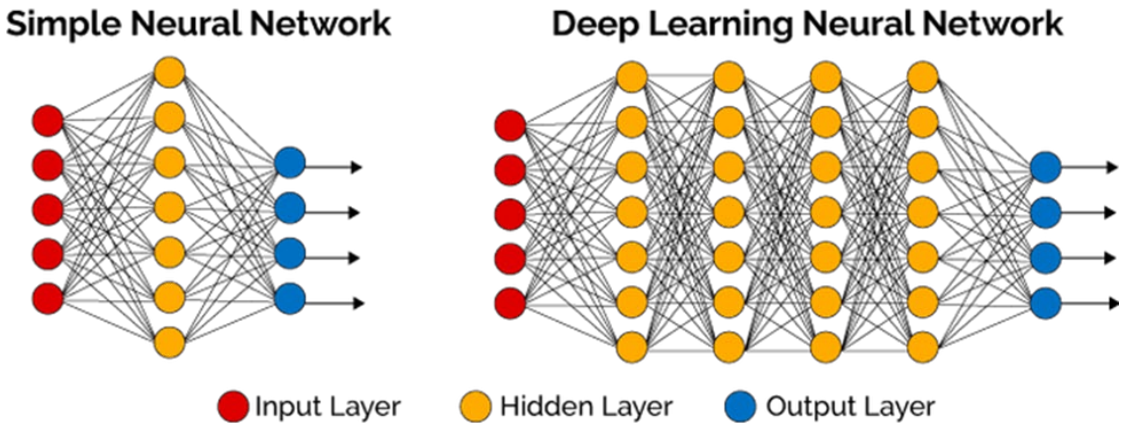

Figure 9. Comparison of a Simple Neural Network and a Deep Neural Network. Source: DeepAI (2019).

The CNN is a DNN that uses the mathematical convolution operation as a hidden layer where it manipulates the input data dimensions and then uses a pooled collection of those convolved outputs as the input for additional convolutional layers and finally into traditional, fully-connected neural network hidden layers (Li et al., 2020, architecture overview heading). Figure 10 shows the architecture of a CNN used for image classification; the salient feature of this figure is the change of dimension between each layer due to the convolution operations performed at the hidden layers (Krizhevsky et al., 2017, p. 87). This specific architecture was one of the first successful CNNs and significantly outperformed DNNs in image classification tasks (Krizhevsky et al., 2017, p. 88).

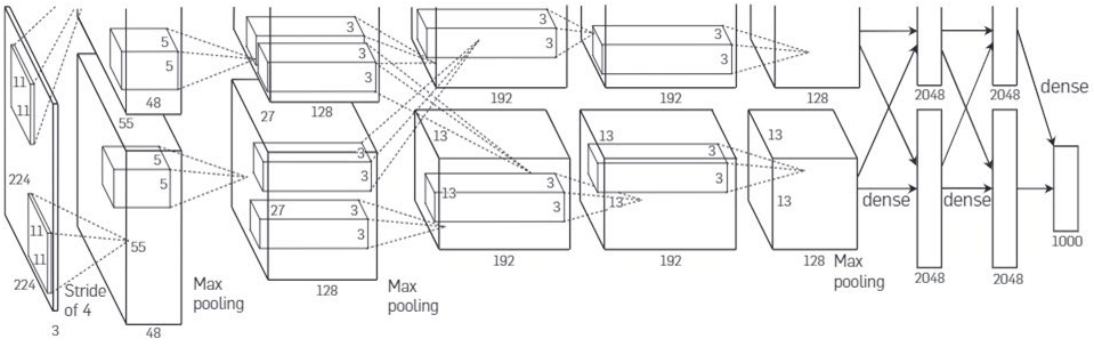

Figure 10. Change of Dimensions within a Convolutional Neural Network. Source: Krizhevsky et al. (2017, p. 87).

3. Level 2: Supervised Learning

Supervised Learning (SL) is a NN usage that utilizes labeled training datasets and defined labels to produce a learned model capable of classifying an object (Orescanin, 2019, p. 6). The focus is on the dataset and its labeling before being fed into the AI algorithm for the algorithm to learn the associative links between the object and its assigned labels (Orescanin, 2019, p. 6). The labeling of a dataset for supervised learning is one of the major time consuming factors associated with data conditioning, because of the necessity for high quality labeling (Martinez et al., 2019, p. 49). After the algorithm is trained with the labeled data set, it is able to use k-nearest neighbor operations to classify the object based upon the available labels (Peterson, 2009, p. 1883). As seen in the Neural Network section in the CNN example, supervised learning is exceptional in object recognition, if trained on a high-quality dataset. This high performance is also seen in the use of supervised DNNs trained to identify fractures, which improved physician identification of fractures in x-rays from 80.8% to 91.5% (Lindsey et al., 2018, p. 11591).

4. Level 3: Unsupervised Learning

Unsupervised learning (UL) utilizes NNs with training datasets which do not have explicit data labels provided (Martinez et al., 2019, p. 49). This does not mean that a dataset does not need to be conditioned, only that labels have not been assigned to the data objects. Considering that dataset preparation is a time consuming and labor intensive task, the removal of the requirement for labeled data is a significant advantage for unsupervised learning. One of the primary uses for unsupervised learning is the clustering of dataset objects based upon similarities within a dimension, which is useful for data conditioning as a starting place for the creation of a labeled data set (Martinez et al., 2019, p. 49). Dimensionality reduction can take place using an unsupervised NN to determine how a dataset distribution could be manipulated into a lower dimension while retaining its data values (Martinez et al., 2019, pp. 50–51). This idea, called Principal Component Analysis, can be seen in Figure 11, where the data originally distributed across axes X1 and X2 could instead be represented across axis Y1, with minor variation along axis Y2; since the

distribution along Y1 is much greater than the distribution along Y2, the values along Y2 can be assumed to be trivial (Martinez et al., 2019, p. 50).

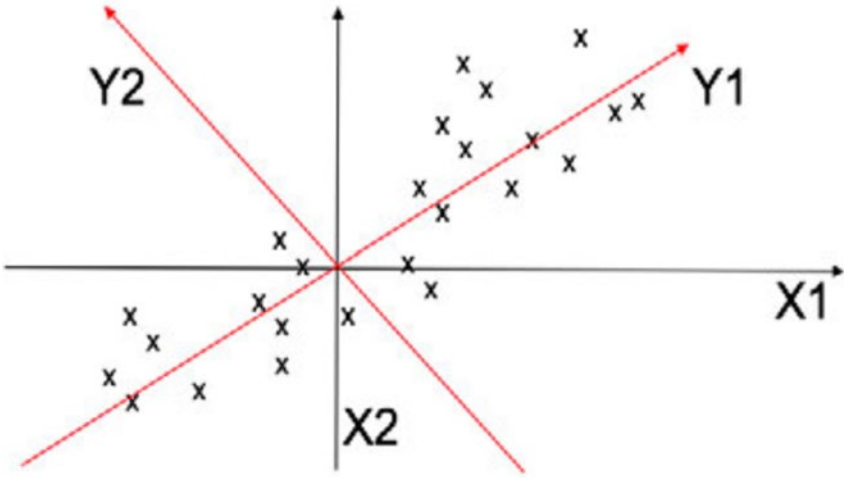

Figure 11. Reduction of Dimensions through Principal Component Analysis. Source: Martinez et al. (2019, p. 50).

Separate from data conditioning tasks, two applications of unsupervised learning applications are neural autoencoders and Generative Adversarial Networks (GANs). A neural autoencoder is an algorithm that is designed to reproduce the input dataset identically, without being able to 'copy and paste' the input as the output; (C. Darken, 2019, p. 8). Once the algorithm is trained and accurately reproducing the training dataset, it can then be given data sets for inference; if the new dataset is not identical to the training set, the output of the algorithm will not be identical to the training set and the system will flag that new dataset as anomalous (C. Darken, 2019, p. 7). Neural autoencoders are useful for predicting system failures because of their ability to indicate when that system is no longer functioning nominally but has not yet failed. By recognizing anomalous behavior through an imperfect AI output, catastrophic failure can be avoided (C. Darken, 2019, p. 5). GANs take this a step further and utilize two NNs, one as a generator and the other as a discriminator which assesses the output of the generator, a function that is generally performed by the user when utilizing a neural autoencoder. Conceptually, these two unsupervised NNs are in a competition; the generator attempts to produce a training set

object which is then evaluated by the discriminator to determine the probability that its input is from the training dataset or from the generator (Martinez et al., 2019, p. 52). Both NNs will improve as training continues until the generator produces an output that causes the discriminator to assign a probability of 0.5 to the object (Goodfellow et al., 2014, p. 1).

Reinforcement learning (RL) is another form of unsupervised learning that divides the overall task for the AI into individual states with specific actions for that state and then uses a designer defined output to guide the algorithm through those states towards the defined output. The state representation is the information status of the task at a given moment in time; for a game like chess, each turn would be its own state and includes the available actions. A score is assigned to the AI's performance after every action, indicating how well the AI is performing. The NN is designed with the objective to complete the overall task and to maximize the score value at task completion; it is not enough to complete the task, it must complete it well (C. Darken, 2019, pp. 12–15). This design performs well in a highly valid (well defined actions and rules) and highly variable (many sequences and combinations of actions) environment, where each iteration of training permits the AI algorithm to apply the learned score values of the previous iteration to its current attempt (C. Darken, 2019, p. 11). A significant challenge of properly developing a RL algorithm is how to best parameterize the NN and what score values should be assigned to a state representation (C. Darken, 2019, p. 19). Part of the challenge comes from delayed feedback because the algorithm must complete the overall task to receive a final score and then attempt to improve its score by changing actions during the various state representations (Martinez et al., 2019, p. 54). The delay in feedback and the total length of time to train increases greatly as the variability and the number of possible states increases (Martinez et al., 2019, p. 54). An example of a successful RL algorithm is AlphaZero, which used a DNN-RL design to learn how to play chess, Shogi, and Go simply by random play with no domain knowledge except for the rules (Silver et al., 2018, p. 1140). Most notable in its performance is that this single algorithm was able to beat the reigning AI champions for each game, all of which had been specifically crafted for that particular game; the chess champion was beaten with 9 hours of training, the Shogi champion with 12 hours of training, and the Go champion with 13 days of training (Silver et al., 2018, pp.

1142, 144). The significant increase in training time needed reflects the increased variability associated with each game.

E. SOFTWARE AND HARDWARE

For this thesis, the value of advancements in software and hardware in relation to AI must be acknowledged for completeness but will not be investigated. Because AI is fundamentally a computational task, advancements in computing have benefited as the technologies related to both software and hardware have developed. Because all computers rely upon physical circuitry at their most basic layer, the design of a processing chip can greatly improve performance in the completion of its task provided the chip design is paired to the types of operations that will be performed. The advent of graphics processing units (GPU) meant primarily for improving the performance of computers in rapidly rendering digital images have also allowed AI developers to take advantage of hardware designed for parallel computations of vectors and matrices (Orescanin, 2019, p. 15). Application Specific Integrated Circuits (ASIC) such as Google's Tensor Processing Unit (TPU) have also been developed with a focus on rapid parallel vector calculations, except that these are specifically meant for complementary software developed by the same companies to best enable performance in AI tasks (Martinez et al., 2019, p. 57). These types of hardware and software developments have permitted higher performance and will likely continue to enable the trend of more and more capable AIs.

F. COMPROMISED AI

AI is not infallible and must be carefully designed, implemented, and secured to ensure that it is consistently able to perform its assigned tasks. Lincoln Laboratory considers an AI that meets these standards to be a Robust AI, resilient in the presence of security threats and properly designed to ensure that it integrates well with the humans that use it (Martinez et al., 2019, p. 62). This section will address how an AI can be compromised, either intentionally or unintentionally.

1. Poor Design and Implementation

AIs are highly complex systems that require thorough and careful design and proper usage in order to function as intended. Functioning *as intended* is a critical qualification for this research, because it is possible for the AI to complete the task without procedural errors but to produce a useless output because the AI was not adequately designed to produce the correct output. Poorly managed databases and training sets can produce a variety of errors, including selection and labeling. Selection bias results from having the wrong data or not knowing that the wrong data is present in the database, which means the training set is non-representative of the test data or that the trends within the data are misrepresented (Schuchard, 2019, p. 10). Labeling bias can result from human error, due to input mistakes or a lack of data understanding, as well as from machine error, due to data misrepresentation in the analysis algorithm (Schuchard, 2019, p. 11). AI can also be compromised by faults in the measurements of the outputs meant to determine if the assigned task has been sufficiently accomplished (Martinez et al., 2019, p. 64). These measurements, known as metrics, must be defined by the designers to indicate the level of system or component performance. It is not enough to make a valid measurement; it must be the correct-valid measurement. Within supervised learning, combinations of true and false positives and negatives must be adequately described to ensure that recommendations made by the system are always true (Martinez et al., 2019, p. 64). Within unsupervised learning, the correct ratios for the desired output must be what are actually computed by the AI algorithm, be it the intra-cluster distances versus the inter-cluster distance or the distance between cluster centroids (Martinez et al., 2019, p. 64). While the development of component level metrics is a task with an engineering solution, system level metrics, those which measure the holistic performance of the AI, are more ambiguous. Both flaws in the datasets and the metrics can prevent an AI from performing as intended, even without the presence of an adversarial threat.

2. Security

Machine and network security are longstanding practices that extend to the development and use of AI. Because AI is a component of a larger information system,

good AI security must begin with good network security. This is especially true for AIs designed for use in military HAT structures because they will likely be targeted by advanced persistent threats, which are typically nation-state funded attackers with the resources, skills, and time needed to plan and execute highly successful cyber operations (National Security Commission on Artificial Intelligence, 2021, p. 12). Each core tenet of the information security triad of confidentiality, integrity, and availability can be degraded within both the larger network and the AI. Specific to the AI, confidentiality violations occur when the attacker is able to steal information and derive knowledge about the dataset, the algorithm, or potentially both (Tabassi et al., 2019). Integrity violations occur when the attacker is able to manipulate the algorithm, producing misclassifications for SL, meaningless data representations in UL, and unintelligent or degraded performance in RL (Tabassi et al., 2019, p. 10). Availability violations occur when the attacker is able to slow the testing speed or prevent access to the AI (Tabassi et al., 2019, p. 10). Threats are the means by which attackers cause these violations, which the designer and operator want to prevent. Because AIs are themselves components of larger information systems and networks, many of the attacks on those systems can also be used as attacks on the AI, as seen in Table 2 (MITRE, 2020a, middle of post). The white cells represent attack methods found in MITRE's ATT&CK Enterprise Matrix, a 204 entry matrix that catalogues attack types against non-AI based networks and computers (MITRE, 2021, interactive website).

Table 2. Adversarial Machine Learning Threat Matrix. Adapted from MITRE (2020a).

Reconnaissance	Initial Access	Execution	Persistence	Model Evasion	Exfiltration	Impact
Acquire OSINT information (various techniques)	Pre-trained ML model with backdoor	Execute unsafe ML models (various techniques)	Execute unsafe ML models (various techniques)	Evasion attack (various techniques)	Exfiltrate training data (various techniques)	Defacement
ML model discovery (various techniques)	Valid account	Execution via API	Account manipulation	Model poisoning	Model stealing	Denial of service
Gathering datasets	Phishing	Traditional software attacks	Implant container image	Data poisoning (various techniques)	Insecure storage (various techniques)	Stolen intellectual property
Exploit physical environment	External remote services					Data encrypted for impact defacement
Model replication (various techniques)	Exploit public facing application					Stop system shutdown/reboot
Model stealing	Trusted relationship					

Cells with a gray fill indicate attack methods specific to AI.

Cells with a white fill indicate attack methods that are derived from attack methods against non-AI networks and computers.

3. Artificial Intelligence Specific Threats

Due to their unique designs and functions, there are also attack threats that are specific to AI. The National Institute of Standards and Technology developed a taxonomy for classifying AI attacks, defenses, and consequences, seen in Figure 12 (Tabassi et al., 2019, p. 4). Of specific interest to this subsection is the left half of the diagram, which contains the targets, techniques, and knowledge subcomponents which comprise an attack on AI. Not all combinations are valid, but any attack must have a level of knowledge of its target to inform which technique can be applied. In a collection of case reports on attacks on AI, one of the trends to emerge is the consistency of multiple attack types being used to achieve a successful attack on the AI (MITRE, 2020b, case studies page). The following paragraphs will briefly describe attacks against supervised and reinforcement learning algorithms. While these three attacks are not exhaustive, they represent the ability for attackers to corrupt the proper functioning of the AI.

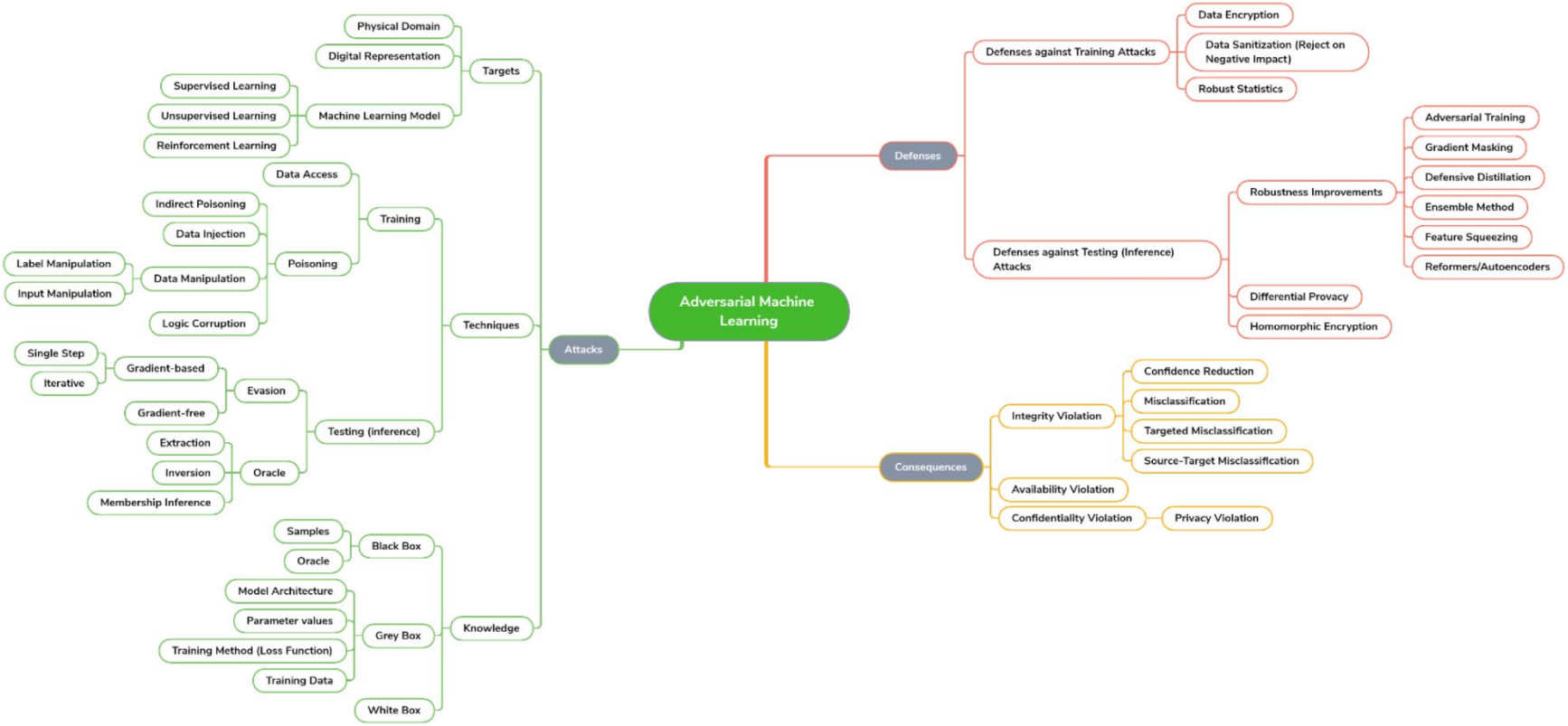

Figure 12. Adversarial Machine Learning Attacks, Defenses, and Consequences. Source: Tabassi et al. (2019, p. 4).

One such attack against a supervised learning algorithm highlights the combination of traditional cyber attacks and those specifically focused against AIs. In this attack, the attackers were able to determine that the designers outsourced the training of their model to a networked third-party because of the computational requirements of the dataset and algorithm (Gu et al., 2019, p. 47230). Using that knowledge, the attackers were able to maliciously train an algorithm "BadNet" that provided excellent performance on the training and validation sets yet behaved erratically on deliberate data triggers (Gu et al., 2019, p. 47230). This poisoned model was then delivered to the original designers, instead of a trained and validated version of their AI. This represents an abuse of external remote services and an exploitation of a public facing application, because the public knowledge of the training and validation datasets permitted the attackers sufficient knowledge to then attack the targeted model. These techniques are not unique to attacks targeted at AI, but their use permits the poisoning of the algorithm.

In another attack against a supervised learning algorithm, attackers conducted a data poisoning attack against an AI performing image identification. These attackers were able to insert specifically generated image perturbations based on other specific labels within the dataset to consistently and confidently misclassify tested objects (Graves, 2020, pp. xvii–xviii). While the tested images appear identical to the human eye, they are classified or misclassified differently as illustrated in Figure 13 (Graves, 2020, p. 51). This disconnect produces a significant problem if the AI is part of a HAT, because this misclassification is immediately obvious to the user. Chapter IV will highlight that task performance is a significant trust factor for the development of the user's trust in AI and an error such as this could significantly damage the trustworthiness of the AI.

Figure 13. Comparison of the Classification Results of the Original Image and a Color Aware Targeted Attack Image. Source: Graves (2020, p. 51).

Reinforcement learning algorithms are also susceptible to attack, albeit of a different type. In one such attack, the attacker was able to develop an enchanting attack against a RL algorithm, which lures the AI towards a designated target state by using a generative model to predict the possible future states and a planning algorithm to determine the actions necessary to draw the AI towards that designated target state (Tabassi et al., 2019, p. 15). These attackers were able to successfully divert the trained AI towards the designated target state in 70% of attempts across three Atari games the AI was trained to successfully complete (Lin et al., 2019, p. 6).

G. CONCLUSION

Narrow AIs represent a major technological step forward for computers. While AI designers must apply significant effort to properly developing datasets and algorithms, the inherent learning capability creates a system that can maximize the performance of those architectures beyond the direct capability of a user. This incredible potential should not obfuscate a fundamental truth; Narrow AI is a highly specialized machine that can be poorly designed or maliciously compromised. Military planners should recognize the vulnerabilities associated with AIs and consider them as part of an adversary's total attack surface. The juxtaposition of competent or compromised AIs will be examined in Chapter IV to form a basis for how military planners can attack a human-AI team.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. INFLUENCING TRUST IN ARTIFICIAL INTELLIGENCE

A. INTRODUCTION

Human-AI trust is a critical requirement to achieve the maximum team performance and is developed through a variety of factors. It is the objective of this chapter to describe that these factors can also be targeted to decrease the overall HAT effectiveness. The focus will be on applying the Chapter III assertion that AIs can be categorized as either competent or compromised based on the user's perception of the AI's trustworthiness. Matched correctly, the changes in trust will degrade the HAT's accomplishment of its mission. As a component of an adversary's military effort, this degradation will decrease the combat power of the adversary relative to the attacker.

B. HUMAN AND ARTIFICIAL INTELLIGENCE TEAMING

1. Complementary Teaming

The fundamental reason for human and AI teaming is the potential for an increase in overall performance through the complementary pairing of human and AI strength and weaknesses. Many of the strengths that AI have in comparison to humans are a result of their computational and logical foundation, which enables high performance in calculations, comparisons, and logic application (R. Darken, 2019, p. 11). It is the ability of an AI to be able to perform these tasks on massive datasets across many iterations that make them the stronger member of the team in this regard (R. Darken, 2019, p. 11). An exemplar of this can be seen in the application of DNN algorithms to assist doctors in identifying fractures in x-rays; performance of this task improved from a detection rate of 80.8% to 91.5% with the assistance of the algorithm (Lindsey et al., 2018, p. 11591). Insurance fraud detection is another area where the ability of AI algorithms to detect minute trends in large quantities of data over time has provided insurance investigators a tool capable of 75% accuracy in its detection, leading to a better application of investigator time and effort (Shift, 2021, p. 1).

Conversely, there are a variety of tasks where the *intelligence* of humans stands out in comparison to the *learning* of AIs. One example is an AI's misclassification of street

signs due to the presence of small squares of tape on the face of the sign (Eykholt et al., 2018, p. 1). The placement of the tape on the sign followed deliberate patterns produced by another algorithm meant to trick the detection algorithm into misclassification a stop sign as a speed limit sign, at a success rate ranging from 67% to 100% (Eykholt et al., 2018, pp. 6–9). As seen on the right in Figure 14, the placement of the tape does not impede a human's proper classification of the object (Eykholt et al., 2018, p. 2). This sort of error stems from the narrow nature of AI; it is only capable of high task performance on datasets that were representative of its training datasets. AIs becomes fragile the further inputs become from the training dataset and lacks the contextual intelligence and learning that a human is able to employ when encountering a novel situation (Apte, 2019, p. 15).

Figure 14. Tape Placement that Caused AI Mis-classification. Source: Eykholt et al. (2018, p. 2).

Lack of context also extends beyond a physical recognition of objects within their surroundings and into the appropriateness of an action based upon societal or interpersonal norms. In 2016, Microsoft developed a chatbot named MS Tay that utilized AI to generate tweets and replies from other Twitter accounts (Schwartz, 2019, para. 1). The AI was designed to continue learning after its public debut based upon interactions with other twitter users, a feature which became a vulnerability once a loosely coordinated group of trolls discovered how to exploit a developer command that caused MS Tay to repeat whatever was tweeted at it (Schwartz, 2019, para. 8). This group then flooded MS Tay with offensive and violent tweets which the AI retweeted under its own handle and incorporated into its lexicon for use in constructing replies to other Twitter users (Davis, 2016, p. 23). Sixteen hours and 95,000 tweets later, Microsoft suspended the MS Tay account (Schwartz,

2019, para. 6). The Microsoft apology for MS Tay acknowledged that the AI's tweets did not reflect the values or design principles of the company or its design team (P. Lee, 2016, para. 1). As a part of the design process, the developers had identified topics which MS Tay would bypass, but these topics were not generalized throughout the AI as norms and could thus be superseded by the tweets learned from the malicious attackers (Davis, 2016, p. 23). This incident underscores the current inability for AIs to adequately grasp what constitute good and bad influences which are necessary for the formation of an understanding of societal or interpersonal norms. Thus, serious consideration must be given to determining the capabilities of the user and the AI and then aligning those capabilities to specific tasks within the HAT's mission.

2. Task Suitability and Mapping

To maximize the performance of the HAT, tasks must be properly evaluated and assigned to the team-member best suited to perform them. The division of tasks reflects a gradient of the seriousness of the consequence of the action and the confidence in the system's performance, which is represented in Figure 15 (Martinez et al., 2019, p. 68). Within this gradient, an organization's willingness to accept risk is inherently a component of the consequences of the action. During the design of the team, the organization must identify what tasks it holds to be sufficiently important or beyond the ability of the AI and then align those with the human; all remaining tasks should be given to the AI (R. Darken, 2019, pp. 7–9). Human Computation is a field of research which seeks to identify tasks which require human intelligence, be it contextualization or emotional, because they cannot yet be accomplished by a computer (Quinn & Bederson, 2011, p. 1404). Whatever the method of determination is, a clear understanding of task mapping is necessary to ensure that the HAT is able to make complementary use of the strengths of each.

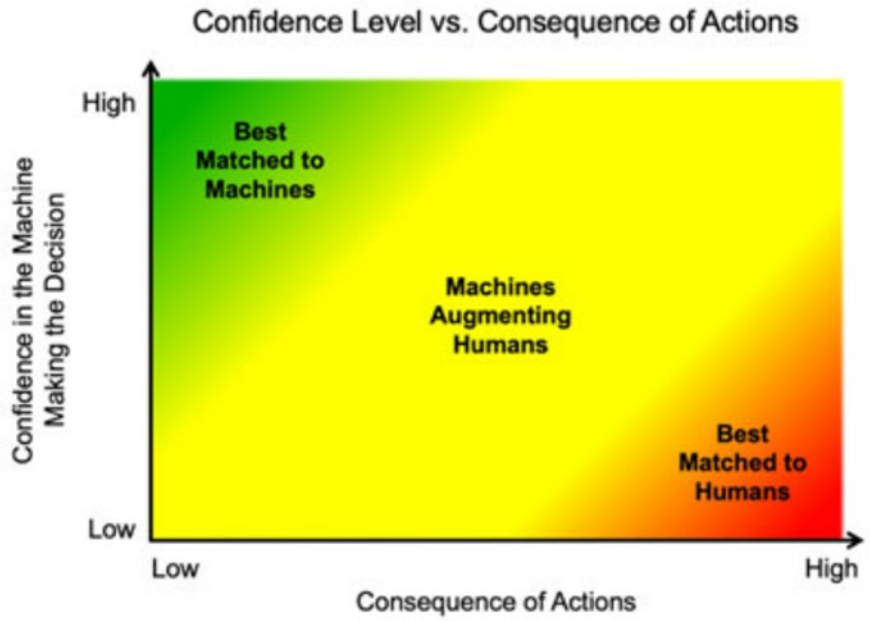

Figure 15. Severity of Consequences and Confidence in System's Performance. Source: Martinez et al. (2019, p. 68).

3. Metrics

A component of properly assigning tasks to an AI is determining how well it performs on those tasks. These metrics are typically considered to be at the component (e.g., database, algorithm, et cetera) or system level (Martinez et al., 2019, p. 64). At a component level, metrics are relatively simple to develop. Within learning algorithms, supervised learning can be measured on accuracy, precision, or recall performance whereas unsupervised learning can be evaluated on its intra-cluster distance versus inter-cluster distance, distance between cluster centroids, or mutual information present in the data objects (Martinez et al., 2019, p. 64). At a system level, metrics become more difficult to identify and the practice becomes a function of project management's triple constraint on scope represented by the equation below (Heagney, 2011, p. 9). In system level metrics for AI, the scope represents the tasks mapped to it and must be supported by an accurate understanding of those levels of component performance. These metrics are project dependent, based on the actual interactions of performance, cost, and time, but the key concept with this representation is that a change in one will cause a change in at least one

other. Thus, a choice must be made about the scope of the AI tasks based upon its performance of those tasks and the cost and time requirements to achieve the desired levels of performance.

$$S = f\left(P, C, T\right)$$

$S =$ scope, $P =$ performance, $T =$ Time or schedule

C. TRUST IN ARTIFICIAL INTELLIGENCE

1. Human Trust

Trust is a complex part of human-human interaction which can be carefully extended onto HAT. Mayer, Davis, and Schoorman define trust as

> the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (Mayer et al., 1995, p. 712)

Key to this definition is exposure to risk, the idea of willingly accepting a level of vulnerability. (Mayer et al., 1995, p. 712). If there is no potential for a loss due to the failure of the trustee, then there is no need for trust. While this definition was proposed for organizational trust, the tenets apply to the human within HAT as well. As a part of a HAT with maximized task mapping, the human expects the AI to produce certain categorical outputs and will be unable to complete the objectives assigned to the HAT without them. Within a HAT, the human trustor does not have the ability to usefully monitor the processes that the AI algorithm is performing. This inability to usefully monitor an AI algorithm in progress is seen in Figure 16, which is the output of the first convolution layer in a CNN designed for image identification (Krizhevsky et al., 2017, p. 89). The output of this layer is critical to the performance of the CNN, but is indecipherable to a human beyond identifying that it is the output of a convolutional function. Thus, there is risk present for the human within the HAT because of an inherent inability to usefully monitor the AI.

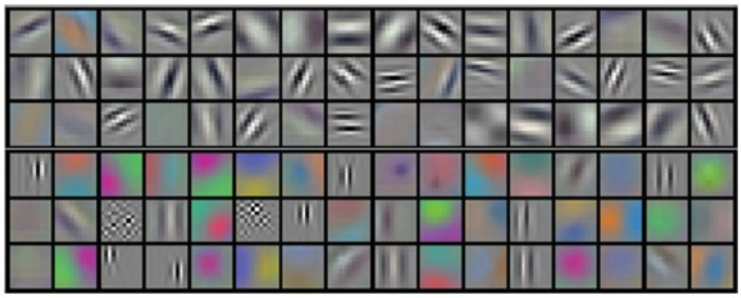

Figure 16. Visual Output of the First Layer in a Convolutional Neural Network. Source: Krizhevsky et al. (2017, p. 89).

2. Trustworthiness of Artificial Intelligence

Trustworthiness forms the basis for trusting another party. Within human-human interactions, Ryan (2020) asserts that trustworthiness is a belief about the other party that is based upon a combination of rational, affective, and normative trust factors ascribed to the other party (Ryan, 2020, pp. 2752–2753). Trustworthiness from rational traits characterizes trust as a decision based upon a logical determination of the ability of the trustee to perform the necessary task: a quarterback rationally trusts the offensive line to block because that quarterback knows that the linemen have practiced those skills and have the physical ability to block the defensive players (Ryan, 2020, p. 2752). Trustworthiness from affective traits characterizes trust as a feeling based upon the trustor's belief that the trustee is motivated upon goodwill towards the trustor: a quarterback affectively trusts the offensive line because the quarterback believes that the linemen are motivated to want to keep the quarterback to unhurt and unhurried (Ryan, 2020, p. 2752). Trustworthiness from normative traits characterizes trust as an expectation by the trustor that the trustee will do what they should do because that behavior is expected of them and to not do so would be considered a betrayal: a quarterback normatively trusts the offensive line to block against the defense because the linemen are on the quarterback's team and to not block for the quarterback would be a betrayal of the quarterback, the team as a whole, and the role of the offensive lineman (Ryan, 2020, p. 2753). These forms of trust can be seen in Table 3. Viewed through this construct, trustworthiness is a relational foundation consisting of a complex mix of trust factors.

Table 3. Relationship of the Trustor and Trustee Depending upon the Type of Trust.

The trustor will	Rationally	*trust the trustee because*	the trustor is capable of performing the action that the trustee is expecting the trustor to perform.
	Affectively		the trustor believes that the trustee is motivated by goodwill towards the trustor.
	Normatively		there is an expectation of the trustee that if not upheld results in a feeling of betrayal.

Key to this research is Ryan's argument that AIs are inherently untrustworthy because trustworthiness within HAT can only fulfill the requirements for only a single trust factor, rational trust (Ryan, 2020, p. 2750). True affective and normative trustworthiness require the trustee to hold a relational motivation for why they will do what the trustor believes that they will do (Ryan, 2020, p. 2753). As stated in Chapter III, narrow AI is a highly specialized computer system which is capable of exceptional performance on specific, computational or logical tasks. Nevertheless, an AI fundamentally remains an inanimate machine that possesses no individuality or personal identity. "Decisions made by the AI do not matter to it," because it is only doing what it has been programmed to do (Ryan, 2020, p. 2762). It is a matter of rational trust to know that the AI possesses the capability to perform the correct tasks. The lack of intentionality in the existence of the AI prevent it from possessing traits of benevolence, loyalty, or value congruence which are central features in human-human trust (J. Lee & See, 2004, p. 66). **Thus, the correct basis for the trustworthiness of an AI is on its ability to reliably and adequately perform only its properly assigned tasks in such a manner that the user is able to accurately monitor its performance.** Despite this, it is likely that the affective and normative forms of trust will still be present in the user (J. Lee & See, 2004, p. 76). The relationship of these different forms of trust to justified and unjustified trust will be assessed in the following section.

D. TRUST FACTORS AS HEURISTIC OR TARGET ATTRIBUTES

1. Extension of Kahneman and Frederick's General Definition for Heuristic to Trust Factors

The notion of a correct basis for trustworthiness, or warranted trustworthiness, implies that there is also an incorrect basis, or unwarranted trustworthiness. Jacovi et al. assert that justified trust is the matching calibration of the trust level of the user to the warranted trustworthiness of the system (Jacovi et al., 2021, p. 627). Unjustified trust is the presence of trust in a system that does not have the capability of completing its assigned task (Jacovi et al., 2021, p. 627). The trustworthiness of a system is evaluated on an individual task level and is predicated by the AI's ability to fulfill the requirements of that task, which can be seen by the user through the outputs of the verification and validation of the AI during training or through direct personal experience working alongside the AI (Jacovi et al., 2021, p. 628). Conversely, users can have mistrust of a system and this can also be justified or unjustified (Jacovi et al., 2021, p. 626). Mistrust is an equally important consideration because it often leads to misuse or disuse (J. Lee & See, 2004, p. 55). This concept of warranted trust is also seen in the National Security Commission on Artificial Intelligence's (NSCAI) Final Report, which dedicates a full chapter to the idea of developing justified confidence in the development and use of AI (National Security Commission on Artificial Intelligence, 2021, p. 131). The NSCAI derives this term from an internationally recognized definition for assurance, which is the "grounds for justified confidence that a claim has been or will be achieved" (International Organization for Standardization et al., 2019, p. 2). Ultimately, a responsible user should ascertain that their trust or mistrust of a system is warranted.

The use of non-rational trust factors in HAT can be described by the general definition of heuristics presented in Chapter II. The key to the use of heuristics in intuitive decision making is the subconscious substitution of a heuristic attribute in place of the legitimate target attribute (Kahneman, 2003, p. 707). This concept can be extended to human trust in AI; when assessing the trustworthiness of the AI, the user can subconsciously substitute affective and normative trust in place of rational trust when the user lacks knowledge of the AI's performance. The user's lack of knowledge can be caused

by a lack of transparency regarding the AI's processes, a lack of experience with using that AI, or a combination of both. This substitution process is seen in Figure 17. The specific traits and characteristics that influence rational, affective, or normative trust are called trust factors. Rational trust factors are directly linked to the development of warranted trustworthiness that is the basis for justified trust. This is contrasted with affective and normative trust factors, which will produce unjustified trust unless the AI is sufficiently transparent for the user to obtain performance-related knowledge and monitor the AI's performance. This does not invalidate the benefit of affective or normative traits being present within the AI, because their presence can amplify the trustworthiness of an AI in concert with the rational trust factors. It is when rational trust factors are not visible to the user that the presence of affective or normative trust factors can exert an influence on the user that is disconnected from the task performance of the AI. Thus, affective and normative trust factors should be considered heuristic substitutes for the rational trust factors, which are the target attributes. The remainder of the section will introduce AI trust factors that are identified in two systemic literature reviews, which will then be categorized as either a target or heuristic trust factor based upon Ryan's three types of trust in the following section.

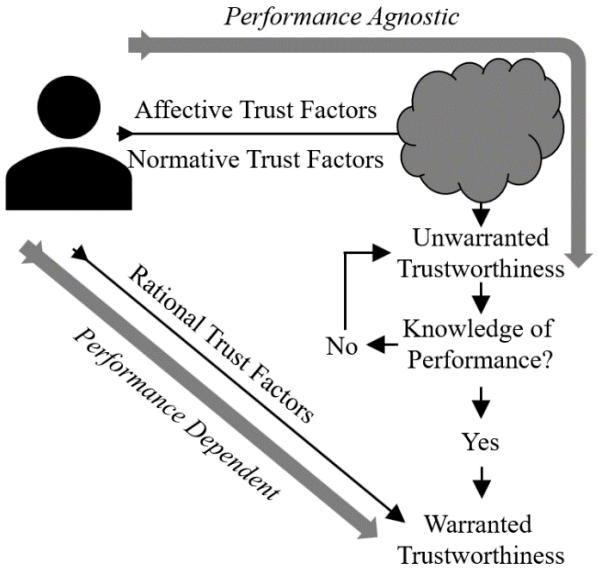

Figure 17. Heuristic Substitution of Affective and Normative Trust Factors for Rational Trust Factors.

Hoff and Bashir (2015) provide a foundational approach to trust in automation that identifies three areas in which variabilities of trust factors occur. The three broad categories of trust factors identified are: (a) dispositional trust factors, which are characteristics of the user's personhood that forms tendencies about automation; (b) situational trust factors, which are characteristics of the environment around the user and automation; and (c) learned trust factors, which are characteristics of the automation that the user has previously experienced. (Hoff & Bashir, 2015, p. 413). These categories and their interrelation with trust are illustrated in Figure 18. Dispositional trust factors include culture, age, gender, and user personality traits (Hoff & Bashir, 2015, p. 413). Situational trust factors are further split into external and internal variabilities; external variabilities include the type of system, the system complexity, the task difficulty, the workload, perceived risks, perceived benefits, the organizational setting, and the task framing, while the internal variabilities include user self-confidence, subject-matter expertise, mood, and attentional capacity (Hoff & Bashir, 2015, p. 415). Learned trust factors begin with the initial trust level formed by preexisting knowledge which includes attitudes and expectations for the system, the reputation of the system, previous experience with the system or similar technology, and the user's understanding of the system (Hoff & Bashir, 2015, p. 421). Learned trust factors are then supplemented by dynamic trust factors that are learned during use including system performance based upon the reliability, validity, predictability, dependability, usefulness, and error factors as well as specific user-interface design features including the appearance, ease-of-use, communication style, the transparency or feedback, and the level of user control (Hoff & Bashir, 2015, p. 421).

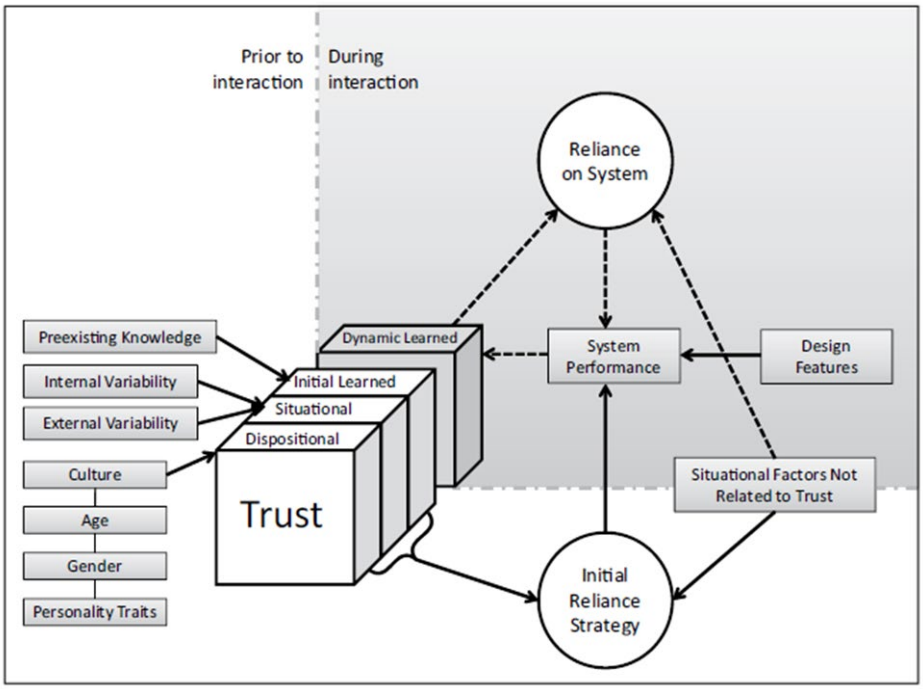

Figure 18. The Interrelation of Dispositional, Situational, and Dynamic Trust for a System. Source: Hoff & Bashir (2015, p. 427).

A systemic literature review of human trust in AI by Glikson and Woolley provides a collection of common trust factors used to increase trust in AI. The list is divided into two categories, cognitive and emotional trust factors, utilizing definitions from Lewis and Weigert (1985) for these categories. Cognitive trust is the decision to trust based upon information available to the trustor about the trustee that is deemed to be "good reasons" for evidence of the trustee's trustworthiness (Lewis & Weigert, 1985, p. 970). It should be noted from Chapter II that decisions made by humans often utilize heuristic substitution, meaning that a reason may not be rational, even though it may be considered to be reasonable. Glikson and Woolley use the following cognitive trust factors based on this definition: tangibility, the physical presence and ability to be touched; transparency, the ability for the user to understand the AI's decisions and actions; reliability, the basis for the user's ability to predict the AI's decisions and actions; task characteristics, the indicators that the AI is performing the correct task; and immediacy behaviors, the proactive actions like active listening and responsiveness that make the AI more of a presence for the user (Glikson & Woolley, 2020, pp. 13–14). Lewis and Weigert define

emotional based trust as the affective bond among the involved parties that is developed to the level that a failure of the trust results in a feeling of betrayal in the trustee (Lewis & Weigert, 1985, p. 971). Based upon that definition for emotional trust, Glikson and Woolley use the following emotional trust factors: tangibility, anthropomorphism, and immediacy behaviors (Glikson & Woolley, 2020, p. 12). Tangibility and immediacy behaviors are also considered to be emotional trust factors because of the affective responses that they can produce in users (Glikson & Woolley, 2020, p. 39). Glikson & Woolley broadly categorize anthropomorphism as design features and user behaviors that project human-like traits onto the appearance and behaviors of AIs (Glikson & Woolley, 2020, pp. 40–42).

Considering that Ryan asserts that two of the three forms of trust are inappropriate for human-AI trust, the trust factors identified by Hoff and Bashir and by Glikson and Woolley should be categorized as contributing to rational, affective, or normative trust. This delineation will allow for specific trust factors to be emphasized so that the user's trust level for the AI is unjustified, a desirable state when attempting to degrade the performance of an adversary's HAT.

2. Determination of Target or Heuristic Trust Factors

In accordance with the basis for trustworthiness in AI identified in the previous section, the target trust factors must be derived from or contribute to one of the following: *AI transparency*, *AI task assignment*, or *AI performance*. These three factors are the rational trust factors because the user must be able to gain data and information (transparency) what the AI is doing (task assignment) and how well it is doing it (performance). The trust factors which are not directly linked to the above should be considered as affective or normative trust factors. The separation between rational and the combination of affective and normative is the primary effort: it is of secondary importance within this research to differentiate between affective and normative trust factors because they can contribute to both justified and unjustified trust.

AI transparency is the gateway for the user to gain data and information leading to knowledge about the AI's tasking and performance. It is the clear communication to the

user of the AI's overall and current task assignments or performance metrics; without this, any trust from the user is unjustified because the user has no information to base their trust upon. This transparency can take the form of clearly communicated performance metrics and task assignments given to the user by the AI developers or the form of feedback from the AI that permits the user to monitor its processes and outputs. Both Glikson & Woolley and Hoff & Bashir directly identify transparency as a trust factor. Hoff & Bashir make a distinction about when the system is transparent, considering the understanding of the system to be a component of the initial-learned trust (i.e., knowledge about metrics and task assignments) and the feedback or transparency of the system to be a component of dynamic-learned trust (i.e., the ability to monitor and understand the processes and outputs) (Hoff & Bashir, 2015, p. 420). They also identify the system complexity as a situational variable; the more complex the AI, the more difficult it is for the user to be able to interpret what the AI is doing, which decreases the usefulness of the transparency (Martinez et al., 2019, p. 52). As seen in Chapter III, neural networks are highly complex constructs that can produce exceptional performance yet do so in a manner than is nearly opaque to the user and even the designer. A potential remedy for this is explainable AI, which is a developmental effort that seeks to design AIs that can maintain those high levels of performance and simultaneously communicate to the user how and why the AI is performing its tasks and producing its outputs (Defense Advanced Research Projects Agency, 2016, p. 7).

AI task assignment is the output of the proper task mapping conducted during the design of the HAT. It is a rational trust factor because of the narrow nature of AI. AIs are designed and trained for a specific task or range of tasks with little potential to transfer that training onto a task for which it was not specifically trained (Martinez et al., 2019, p. 53). Glikson & Woolley call this task characteristics, which are the actions and decisions which the system is assigned to perform (Glikson & Woolley, 2020, p. 14). Hoff & Bashir identify several trust factors across situational and learned trust that contribute to the user's understanding of the AI's task assignments. Within situational trust, the general type of the system indicates to the user what its designed purpose is, which can be aided by the user's level of subject-matter expertise on the tasks and the overall workload assigned to the HAT

(Hoff & Bashir, 2015, p. 416). Within areas of initial-learned trust, the user's understanding of the AI and previous experiences with similar systems affects trust because the user has either been told or experienced the AI's task suitability (Hoff & Bashir, 2015, pp. 420–421). These three trust factors all contribute to the user's level of trust in the AI before usage because they describe what the AI is trained to do and what the AI is assigned to do. If training and assignment align, then the user has a basis for justified trust; if they do not, then the user has a basis for justified mistrust. Dynamic-learned trust factors are also incorporated into the system task assignment as the user views the validity, the "degree to which an automated system performs the intended task," and experiences and attempts to mitigate system errors (Hoff & Bashir, 2015, p. 424). This performance feedback specifically regards whether or not the pre-usage comparison of trained and assigned tasks are actually aligned as anticipated.

System performance is the collective measurements of outputs as defined within the designer's metrics. These are rational trust factors because the AI must be able to accomplish its tasks in proper synchronicity with the user for the HAT to achieve its team mission. Within Glikson & Woolley, performance is most characterized by the idea of reliability (Glikson & Woolley, 2020, p. 14). Hoff & Bashir underscore the criticality of performance based trust factors by declaring that "trust depends on results" (Hoff & Bashir, 2015, p. 424). Those results include the reliability, predictability, dependability, and usefulness to the user (Hoff & Bashir, 2015, pp. 424–426).

Separate of these rational trust factors are affective and normative trust factors that can produce trust in the AI. Anthropomorphism is the most broad and crucial because its use of human-like features can prime the user to perceive the AI less like a machine and more like a person which increases trust development (Schaefer et al., 2016, p. 383). As a machine, the AI is incapable of truly filling the requirements to uphold affective or normative trust because it can neither possess good will towards the user nor can it choose whether or not it will perform its tasks (Ryan, 2020, pp. 2752–2753). Anthropomorphizing an AI is also important to consider because the dissonance between the presence of human-like features on an otherwise non-human machine can cause a user to feel uncomfortable or unsettled when in proximity to the AI (Glikson & Woolley, 2020, p. 38). This affective effect can

potentially cause the user to unjustifiably mistrust an AI solely because of its uncanny and eerie presence. Closely associated to anthropomorphizing designs and behaviors are immediacy behaviors, which are meant to engender a feeling of social closeness to the AI through the use of complementary interactions with the user and proactivity that anticipates the needs of the user (Glikson & Woolley, 2020, p. 14). Hoff & Bashir refer to this as dynamic-learned trust from the system's appearance and communication style (Hoff & Bashir, 2015, pp. 422–423). Collectively, this behavior has affective and normative effects on the user, which are avenues of trust development (Hoff & Bashir, 2015, pp. 422–423). Mimicry of social interaction and attentiveness adds to the n anthropomorphism, further increasing the humanness of the AI. Beyond trust factors directly associated with the AI, Hoff and Bashir also note that the external situational factor of the organization that the HAT is resident within will affect the trust for the AI (Hoff & Bashir, 2015, p. 417). This is especially critical, because some organizations may require the AIs usage, regardless of the user's trust (Hoff & Bashir, 2015, p. 419). In the context of targeting an adversary's HAT, it must be targeted at the personnel who are permitted to choose to use or not use.

These categorizations are neither meant to be exhaustive nor inflexible. They are indicative of the variety of trust factors present within the HAT construct and their presence will vary depending upon the specifics of the HAT and the broader situation around it (Hoff & Bashir, 2015, pp. 418–419). What will not vary is the need for rational trust factors to form the basis for warranted trustworthiness. It is important to recall from Chapter II that humans do not fit the rational actor model and therefore it is unrealistic to expect users to be able to only base their trust upon the rational trust factors, especially when specific design decisions are made to incorporate affective or normative trust factors like anthropomorphism into the AI. Thus, it is critical to understand the user's perception of the AI because that is how the user will assess its trustworthiness to determine if the AI should be trusted. The following section will conceptually explore how to leverage the user's perception of the AI's trustworthiness in order to decrease the overall effectiveness of the HAT.

E. USER PERCEPTION AND MILITARY DECEPTION AGAINST HUMAN AND ARTIFICIAL INTELLIGENCE TEAMS

Military Deception is a component of Operations in the Information Environment that is meant to "deliberately mislead adversary decision makers" (Office of the Chairman of the Joint Chiefs of Staff, 2020, p. 141). It can either be ambiguity increasing, meant to create doubt about an event or capability, or ambiguity decreasing, meant to create confidence about an event or capability (Office of the Chairman of the Joint Chiefs of Staff, 2017, p. I–9). The operation is meant to generate a difference between the actual situation and the intended target's perception of the situation. Employed effectively, the ruse can generate surprise against adversary, a key means of gaining combat power relative to the adversary (U.S. Marine Corps, 2018, p. 2.22). This separation between the point of view of the adversary and the actual truth of the situation provides an application for the idea of a competent or compromised AI and the justified or unjustified trust in AI based upon the user's perception of the AI as trustworthy or untrustworthy. Figure 19 illustrates the intersection of these relationships. Because deception is meant to mislead an adversary, the desired cases for the adversary must be either unjustified trust or mistrust. This focus on misleading an adversary's trust is in direct conflict with the adversary's AI designers, who must strive to ensure that all adversary users only possess justified trust or mistrust.

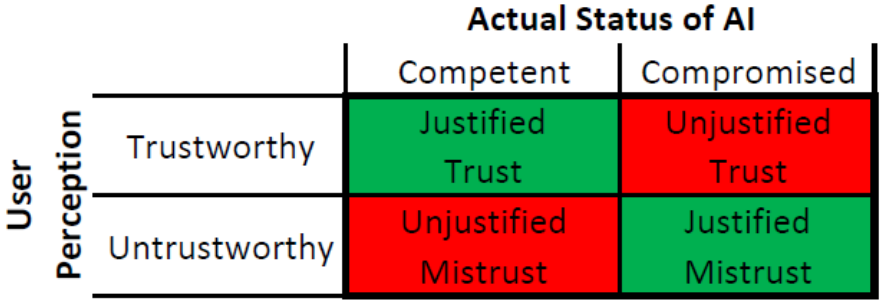

Figure 19. Mapping of the User's Perception of the Status of the AI onto the AI's Actual Status.

1. Unjustified Trust

The goal of a deception operation focused on increasing unjustified trust in the compromised AI is to convince the adversary that their current military capability is sufficient and secure. Effectively executed, the adversary will continue to make plans that utilize the full capability of the HAT because the adversary trusts that the AI will be able to accomplish its mission. The adversary will then be surprised when the AI is unable to accomplish its mission, and the clever deception planner will attempt to synchronize this surprise with other friendly actions so that the adversary must now simultaneously contend with both an unexpected failure and the friendly action.

The two primary considerations for unjustified trust are the specific details of the compromised status of the AI and the unintended potential consequences of a successful deception. As discussed in Chapter III, AIs can be compromised due to errors in design, poor performance, as well as due to deliberate attacks and each of these forms of compromise could produce a different failure condition. Deception planners must understand how the AI will fail to determine which trust factors should be targeted by a deception operation to generate unjustified trust. Understanding the failure conditions associated with the compromise will also permit the deception planner to selectively utilize rational trust factors to increase the user's trust. If the attackers in the "BadNet" compromise example discussed in Chapter III were confidently able to control the release of specific trigger inputs, they could have used messaging to emphasize the legitimate performance that the AI showed during verification and validation, heightening the user's trust in that AI. In this instance, control of the release of the triggering inputs is a critical requirement for the attacker, because accidental release to the compromised AI would begin to produce the flawed outputs which would indicate to the user that the AI is compromised. Because deception relies on the user's unjustified trust of a compromised system, the deception planner must not introduce information that is incongruent with the deception narrative Doing so would increase the likelihood that the deception is discovered by the adversary (Whaley, 1982, p. 189). The unintended potential consequences of a successful deception are particularly important for unjustified trust because they may cause the adversary to commit an action that neither they nor friendly forces desire. An example

of this would be the case of an AI being used to identify objects in satellite imagery to be attacked, similar to the deliberate image perturbation attack seen in Chapter III (Graves, 2020, p. 51). Because the deception effort would have the adversary continue to trust their AI even though friendly forces have rendered it compromised, the adversary would likely continue to employ it as designed until the compromise is discovered. This could have disastrous effects if the flawed outputs of the AI result in the adversary conducting a weapons strike that produced civilian casualties or other collateral damage; such real-world consequences could trigger a demand for retaliation that ultimately escalates the conflict above the desired level. It must also be considered that the adversary will always attempt to ensure they are using a competent and uncompromised system. Thus, the adversary may be able to detect the issues related to the competency and correct it. In this event, the friendly efforts that were based on an adversary's compromised AI are now being carried out on an uncompromised system. Instead of degrading the adversary's HAT, friendly efforts are strengthening it which may give that adversary an advantage.

2. Unjustified Mistrust

The goal of a deception operation focused on increasing unjustified mistrust in the compromised AI is to convince the adversary that their current military capability is insufficient or insecure. Effectively executed, the adversary will continue to expend time and resources attempting to determine the specific nature of the compromise. The adversary may also delay or avoid operational use until the AI is perceived as trustworthy. This may also prompt the adversary to pursue alternative methods to accomplish the mission that would be assigned to the HAT; if the alternative means reassigning the tasks to humans instead of the AI, the outputs will likely be slower or less accurate because of the performance difference between a human and an AI.

The two primary considerations for unjustified mistrust are the means of countering justified trust through adversarial use and the longevity of the deception. Hoff & Bashir's assertion that results are highly influential to trust levels means that friendly efforts to deceive the adversary must be able to counter the consistent reinforcement of justified trust produced by regular exposure to the AI's competency (Hoff & Bashir, 2015, p. 424).

Consider again the example of the AI being used to identify objects in satellite imagery, except in this instance friendly forces are unable to compromise the AI. To create doubt in the adversary's mind regarding the AI's competency and trustworthiness, deception planners could utilize physical decoys with coordinated messaging to amplify the alleged misclassification. The decoy would be designed to match the signature of a legitimate military system or unit but would be able to be rapidly hidden following its imaging by the adversary's satellite. When the adversary AI categorizes the decoys as the matching system or unit, messaging efforts could then describe the actual location of the matching system or unit (e.g., an aircraft decoy is imaged on a runway in country X, the decoy is hidden, and the messaging describes and confirms the actual aircraft is being refueled in country Y which is sufficiently far to prevent the aircraft from flying between countries X and Y). Performed multiple times with synchronized messaging, the adversary would likely begin to doubt the AI's performance and potentially begin an investigative effort to determine how the AI is compromised. Another serious consideration is the length of time associated with the deception effect. This is especially critical if the deception is considered to be a necessary condition to enable an operation against the adversary; it is likely that the adversary's behavior will react to the operation, so the deception planner must consider how to continue reinforcing the unjustified mistrust of the AI throughout the entire timeline. The deception operations surrounding the amphibious assault at Normandy during World War II is a good example of this; the deception plan mislead the Nazis to believe that the main effort of the amphibious assault would land at Pas de Calais and that any other amphibious assault was a feint (Breuer, 1993, p. 117). The deception needed to last long enough to prevent the Nazis from committing their armored reserves against the Normandy beachhead while the Allies lacked the combat power ashore to defeat that force. Had the deception succeeded in surprising the Nazis about the amphibious assault location been discovered to be a deception in the week following D-Day, it is likely that the Nazis would have been able to destroy the Allied forces ashore (Breuer, 1993, pp. 220–221). Similarly, the length of the deceptive effect in unjustified mistrust must be possible for as long as friendly forces predominantly depend upon the mistrust of the AI to maintain combat power. If the deception planners believe that the deception effort will be discovered

by the adversary before the friendly forces are able to gain an advantage, mitigating measures should be implemented to ensure that those friendly forces are not left exposed.

F. CONCLUSION

Complementary Human-AI Teams will likely be the most advantageous employment of AI because they intentionally assign tasks to the team-member best suited to accomplish them. To do this, AIs will need to be able to demonstrate their warranted trustworthiness to the user so that the user will actually use and rely upon the performance benefits provided by the AI. The user's trust is a critical link that can be influenced for both the good or the bad of the HAT's performance. Military deception planners should closely consider how to ensure that the adversary user always holds unjustified trust or mistrust for the AI so that their HAT's performance is degraded. Doing so will generate a competitive advantage relative to the adversary that can translate into a combat power advantage if necessary.

V. CONCLUSION

A. RESEARCH LIMITATIONS

The most significant limitation for this research is the lack of experimentation to test the assertion that trust in AI can be influenced independently of the AI's trustworthiness through heuristic substitutions. Hypothesis testing is a fundamental step in the scientific process and is critical in evaluating the arguments presented in this thesis.

Another critical consideration in research is to ensure that the study population is representative of the targeted population. Within this research, very few of the cited studies specifically utilized military personnel as the study participants. The concept of heuristic substitution was specifically chosen because of its nearly universal usage in intuitive decision making, so it is unlikely to discover that heuristic substitution does not apply at all to the military population. It does however allow for the possibility of another unknown and unconsidered factor to also make a significant contribution to the user's trust in AI.

Closely related to the representativeness of the study populations for the targeted military population are the limitations associated with cultural differences between the study and target population. Culture is the amalgamation of the societal inputs on individuals and groups and will have a profound impact on outlooks. Hoff and Bashir recognize this by identifying the dispositional trust present in the user well before the first use of the system. (Hoff & Bashir, 2015, p. 413). Because this research is focused on influencing a foreign adversary's military personnel, the unique culture of that foreign military must be considered and adequately represented in the study population. A specific culture was not chosen for this research in order to develop a broader method for influencing trust without solely assuming a specific cultures tendencies. Thus, in order for this research to be operationalized, the target audience's culture must be identified and captured within any study participants.

B. AREAS FOR FUTURE RESEARCH

To test the hypothesis of this research, there is a need for a study framework that assesses the influencing of trust development involving military personnel within HAT.

This framework will assist in reducing some of the limitations present within this research, namely confirmation or disconfirmation of the hypothesis and the lack of studies using military personnel as study participants. Because the ultimate goal of this research is to reduce the effectiveness of an adversary's HAT, the testing of the methods described should utilize a consistent approach that attempts to approximate the likely human target as closely as possible. The study should specify a single rational trust factor and single affective or normative trust factor to treat as deliberately variable. The study should also assess the level of trust over the full length of the interaction to determine how direct experience affects the user's trust.

Chapter III introduced the consideration of AI as either being competent or compromised for the purposes of matching its status with the desired trust level of the user. The examples cited in Chapter III, Section F, all utilized either full knowledge of the algorithm or conducted their attack against an otherwise undefended system. It is naïve to assume that an adversary will leave an AI system unguarded at any point and will likely attempt to protect the datasets, algorithms, and other system design considerations. Thus, it will be necessary to develop a framework for intelligence collection that aids planners in guiding intelligence collection efforts towards critical information about the adversary's HAT.

Another area for future research is how to best protect HATs employed by the USA against a similar form of deception described within this research. This presents two areas for investigation: protecting the human and protecting the AI. Ultimately it is the user's perception that matters in regard to trust, but the AI should receive significant focus to determine what is the best combination of the trust factors to ensure that the trustworthiness of the AI is always accessible to the user.

C. CONCLUSION

The impact of AI on the military as an instrument of national power should not be understated (National Security Commission on Artificial Intelligence, 2021, p. 9). Complementary teaming of humans and AIs will deliberately harness the advantages of each and likely offset the inherent disadvantages each brings to the team. This strength will generate combat power through its speed and will threaten a force that has either no HAT

or a less optimized one. In either case, it should be the objective of friendly forces to degrade the performance of the HAT to gain a combat power advantage relative to the adversary. This research has defined a method on how to do so based upon two preliminary observations.

The first is that intuitive decision making uses a subconscious heuristic substitution in place of the actual decision that can often result in systemic errors (Kahneman, 2003, p. 707). It is the prevalence of heuristic substitutions that pass the System 2 reasonableness check on the output of System 1 that makes this such a critical component of the research. This is the basis for the extension to substitution of affective and normative trust factors in place of or in equal consideration to rational trust factors.

The second observation is the contrast between a competent AI and a compromised AI. Chapter III provided a cursory knowledge level of the functioning and performance of AIs to show that a well-designed AI is a marvel of modern computing technology. It should also be apparent from Chapter III that an AI can be compromised as easily as it is designed. The example of MS Tay is an illuminating example of a vulnerability within an otherwise exquisite system being identified and exploited culminating in an undesired and unpredicted state of affairs. The juxtaposition of competence and compromise is the basis for determining whether and AI is trustworthy or not.

The combination of these two observations provides the key assertion of this research: **the overall effectiveness of the HAT can be degraded by influencing the user's trust for the AI through heuristic trust factors in order to produce unjustified trust or mistrust.** This separation of the user's perception of the AI from its actual status drives a wedge between the two, disrupting the integrated nature of the HAT. Separating the user's trust from the AI's trustworthiness results in the degradation of the HAT's critical capability to generate combat power through speed of action, resulting in the gain of a combat power advantage for friendly forces.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Alba, J., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203–231.

Apte, U. (2019, October 30). *Management and AI* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Breuer, W. (1993). *Hoodwinking Hitler*. Praeger Publishers.

Brown, K. F., Kroll, J. S., Hudson, M. J., Ramsay, M., Green, J., Vincent, C. A., Fraser, G., & Sevdalis, N. (2010). Omission bias and vaccine rejection by parents of healthy children: implications for the influenza A/H1N1 vaccination programme. *Vaccine*, *28*(25), 4181–4185. http://dx.doi.org/10.1016/j.vaccine.2010.04.012

Cycorp. (2021). *The Cyc Platform* [Commercial]. Cycorp. https://www.cyc.com/platform

Darken, C. (2019, October 16). *Unsupervised Learning* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Darken, R. (2019, October 21). *Human-Machine Teaming AI* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Davis, E. (2016). AI amusements: the tragic tale of Tay the chatbot. *AI Matters*, *2*(4), 20–24. https://doi.org/10.1145/3008665.3008674

DeepAI. (2019, May 17). *Hidden Layer*. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning

Defense Advanced Research Projects Agency. (2016). *Broad agency announcement 16–53: explainable artificial intelligence*. DARPA. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Denning, P. (2019a, September 30). *Harnessing Artificial Intelligence* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Denning, P. (2019b, October 2). *A Hierarchy of AI Machines* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

DOMO. (2021). *Domo Resource—Data Never Sleeps 8.0*. Learn Center. https://www.domo.com/learn/data-never-sleeps-8

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *62*(2), 328. https://search.proquest.com/docview/1295949068?pq-origsite=primo

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning models. *ArXiv:1707.08945 [Cs]*, 11. http://arxiv.org/abs/1707.08945

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Gilovich, T., & Griffin, D. (2002). Introduction—Heuristics and Biases: Then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgement* (First Edition, pp. 1–19). Cambridge University Press.

Glikson, E., & Woolley, A. (2020). Human trust in artificial intelligence: review of empirical research. *Academy of Management Annals*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Neural Information Processing Systems*, *27*, 9. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Graves. (2020). *Image Perturbation Generation: Exploring new ways for adversaries to interrupt neural network image classifiers* [Master's Thesis]. Naval Postgraduate School.

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). Badnets: evaluating backdooring attacks on deep neural networks. *IEEE Access*, *7*, 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

Heagney, J. (2011). An overview of project management. In *Fundamentals of Project Management* (Fourth Edition). American Management Association.

Hoff, K., & Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hogarth, R. M. (2001). *Educating intuition*. University of Chicago Press. https://books.google.com/books?hl=en&lr=&id=fsffJGkpwVIC&oi=fnd&pg=PR7&dq=Educating+intuition&ots=_fOXQImbZi&sig=GCUmgEzBQrbt_XmMYiAPMFA0Qbo#v=snippet&q=wicked%20environment&f=false

International Organization for Standardization, International Electrotechnical Commission, & Institute of Electrical and Electronic Engineers Standards Association. (2019). *International Standard ISO/IEC/IEEE 15026-1 systems and software engineering—systems and software assurance—Part 1: concepts and vocabulary*. ISO/IEC/IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8657410

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923

Johnson, M. K., Doll, T. J., Bransford, J. D., & Lapinski, R. H. (1974). Context effects in sentence memory. *Journal of Experimental Psychology*, *103*(2), 358–360. https://search.proquest.com/docview/1290475028?pq-origsite=primo

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. http://dx.doi.org/10.1037/a0016755

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://search.proquest.com/docview/1290900057?pq-origsite=primo

King, B. R., Dolfen, N., Gann, M. A., Renard, Z., Swinnen, S. P., & Albouy, G. (2019). Schema and motor-memory consolidation. *Psychological Science*, *30*(7), 963–978. https://doi.org/10.1177/0956797619847164

Kölsch, M. (2019, November 4). *Computer vision and AI* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lee, J., & See, K. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *Vol. 46*(No. 1), 50–80. https://journals.sagepub.com/doi/pdf/10.1518/hfes.46.1.50_30392

Lee, P. (2016, March 25). Learning from Tay's introduction. *The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, *63*(4), 967–985. https://doi.org/10.2307/2578601

Li, F.-F., Krishna, R., & Xu, D. (2020). *CS231n Convolutional Neural Networks for Visual Recognition* [Course website]. CS231n Github. https://cs231n.github.io/convolutional-networks/

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., & Sun, M. (2019). Tactics of adversarial attack on deep reinforcement learning agents. *ArXiv:1703.06748 [Cs, Stat]*. http://arxiv.org/abs/1703.06748

Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., & Potter, H. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, *115*(45), 11591–11596. https://doi.org/10.1073/pnas.1806905115

Martinez, D., Malyska, N., Streilein, B., Caceres, R., Campbell, W., Dagli, C., Gadepally, V., Greenfield, K., Hall, R., King, A., Lippmann, R., Miller, B., Reynolds, D., Richardson, F., Sahin, C., Tran, A., Trepagnier, P., & Zipkin, J. (2019). *Artificial intelligence: Short history, present developments, and future outlook* (p. 135) [Final Report]. Massachusetts Institute of Technology. https://www.ll.mit.edu/r-d/publications/artificial-intelligence-short-history-present-developments-and-future-outlook-0

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

MITRE. (2020a, October 23). *Adversarial ML Threat Matrix*. GitHub. https://github.com/mitre/advmlthreatmatrix

MITRE. (2020b, December 15). *Adversarial ML Threat Matrix: Case Studies Page*. GitHub. https://github.com/mitre/advmlthreatmatrix

MITRE. (2021). *MITRE ATT&CK*. ATT&CK Matrix for Enterprise. https://attack.mitre.org/

Molden, D. C. (2014). Understanding priming effects in social psychology: What is "social priming" and how does it occur? *Social Cognition*, *32*(Supplement), 1–11. http://dx.doi.org/101521soco201432supp1

Monaco, V. (2019, October 9). *Rule-Based AI* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

National Security Commission on Artificial Intelligence. (2021). *Final report* [Congressional Report]. National Security Commission on Artificial Intelligence. https://www.nscai.gov/reports

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. W. H. Freeman and Company.

Office of the Chairman of the Joint Chiefs of Staff. (2017). *Military Deception*. The Joint Staff.

Office of the Chairman of the Joint Chiefs of Staff. (2020). *DOD Dictionary of Military and Associated Terms*. The Joint Staff.

Orescanin, M. (2019, October 14). *Supervised Learning* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1. https://doi.org/10.4249/scholarpedia.1883

Plant, K., & Stanton, N. (2012). Why did the pilots shut down the wrong engine? Explaining errors in context using Schema Theory and the Perceptual Cycle Model. *Safety Science*, *50*(2), 300–315. https://doi.org/10.1016/j.ssci.2011.09.005

Quinn, A. J., & Bederson, B. B. (2011). Human computation: A Survey and taxonomy of a growing field. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1403–1412. https://doi.org/10.1145/1978942.1979148

Rowe, N. (2019, October 23). *Aspirational AI or 'Extreme" AI Ideas* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schuchard, R. (2019, October 28). *Data Science and AI* [Lecture]. CS4000: Harnessing Artificial Intelligence, Naval Postgraduate School, Monterey CA. https://nps.edu/web/ai-group/harnessing-ai-course

Schwartz, O. (2019, November 25). In 2016, Microsoft's racist chatbot revealed the dangers of online conversation. *IEEE Spectrum: Technology, Engineering, and Science News*, *History of Natural Language Processing*(Part 5), 1. https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Shift. (2021). *Shift claims fraud detection: Identify more fraud with greater accuracy and efficiency*. Shift. http://2cdn6p3duuk84f0po2489ytp-wpengine.netdna-ssl.com/wp-content/uploads/2021/03/Shift_2Pager_ClaimsFraudDetection_EN_RGB-1.pdf

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140. https://doi.org/10.1126/science.aar6404

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. http://dx.doi.org/10.1037/h0042769

Sloman, S. (2002). Two systems of reasoning. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 379–396). Cambridge University Press.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgement* (First Edition, pp. 397–420). Cambridge University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665; discussion 665–726. https://www.proquest.com/docview/212294496/B13CCB2CA8754072PQ/1

Tabassi, E., Burns, K., Hadjimichael, M., Molina-Markham, A., & Sexton, J. (2019). *A taxonomy and terminology of adversarial machine learning* (Draft NISTIR No. 8269; p. 36). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8269-draft

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science (Washington)*, *247*(4940), 301–301. https://search.proquest.com/docview/79589230?pq-origsite=primo

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1130.

Tversky, A., & Kahneman, D. (2002). Extensional versus intuitive reasoning. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 19–48). Cambridge University Press.

U.S. Marine Corps. (2018). *Warfighting*. Headquarters United States Marine Corps.

Wang, Q. (2008). Being American, being Asian: the bicultural self and autobiographical memory in Asian Americans. *Cognition*, *107*(2), 743–751. https://doi.org/10.1016/j.cognition.2007.08.005

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: a meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, *142*(5), 472–497. http://dx.doi.org/10.1037/bul0000030

Whaley, B. (1982). Toward a general theory of deception. *The Journal of Strategic Studies*, *5*(1), 178–192. https://doi.org/10.1080/01402398208437106

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, *58*(4), 479–493. https://doi.org/10.1002/asi

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California