2019-09

TRUST AND UNDERSTANDABILITY IN AUTONOMOUS AND UNMANNED SURFACE VEHICLES

Adesanya, Kehinde A.; Shivashankar, Santhosh K.

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

TRUST AND UNDERSTANDABILITY IN AUTONOMOUS AND UNMANNED SURFACE VEHICLES

by

Kehinde A. Adesanya and Santhosh K. Shivashankar

September 2019

Thesis Advisor: Shelley P. Gallup
Co-Advisor: Douglas J. MacKinnon

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2019	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE TRUST AND UNDERSTANDABILITY IN AUTONOMOUS AND UNMANNED SURFACE VEHICLES	**5. FUNDING NUMBERS** W9A77
6. AUTHOR(S) Kehinde A. Adesanya and Santhosh K. Shivashankar	
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)

Within the human-machine relationship, distrust can arise. The Department of Defense utilizes automation, autonomous systems, and artificial intelligence to reduce cognitive workload and improve mission capabilities; however, adoption rates of autonomous unmanned surface vehicles (USVs) remain low. This thesis asks how human distrust of machines and machine learning relates to adoption rates. First, we identify trust components by building upon a model created by Gari Palmer, Anne Selwyn, and Dan Zwillinger in 2016. Then, we identify components that apply to the military environment that could affect the adoption rate such as smoothing time, policies and regulations, competition, robustness, understandability, subjective norm, human interaction, policy effect, risk to force, time sensitivity, war, time between wars, and catastrophic failure. Through S-curve and smoothing modeling, we find that trust components can be quantified in the human machine relationship as positive or negative trust, and that a relationship exists between understandability and adoption. While autonomous system components generally undergo rigorous testing to verify suitability and operability, human-machine trust is not usually incorporated into design and testing phases. When trust is built into the design and acquisition process, adoption of autonomous USVs is more likely to increase. Researchers can apply our trust model to future autonomous systems to mitigate distrust and human-machine teaming.

14. SUBJECT TERMS trust, understandability, autonomous, autonomy, transparency, adoption, unmanned, quantitative, USV	**15. NUMBER OF PAGES** 75
	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.

TRUST AND UNDERSTANDABILITY IN AUTONOMOUS AND UNMANNED SURFACE VEHICLES

Kehinde A. Adesanya
Lieutenant, United States Navy
BS, National University, 2007

Santhosh K. Shivashankar
Lieutenant Commander, United States Navy
BSEE, The University of Texas at Austin, 2003

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN NETWORK OPERATIONS AND TECHNOLOGY

from the

NAVAL POSTGRADUATE SCHOOL
September 2019

Approved by: Shelley P. Gallup
Advisor

Douglas J. MacKinnon
Co-Advisor

Dan C. Boger
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Within the human-machine relationship, distrust can arise. The Department of Defense utilizes automation, autonomous systems, and artificial intelligence to reduce cognitive workload and improve mission capabilities; however, adoption rates of autonomous unmanned surface vehicles (USVs) remain low. This thesis asks how human distrust of machines and machine learning relates to adoption rates. First, we identify trust components by building upon a model created by Gari Palmer, Anne Selwyn, and Dan Zwillinger in 2016. Then, we identify components that apply to the military environment that could affect the adoption rate such as smoothing time, policies and regulations, competition, robustness, understandability, subjective norm, human interaction, policy effect, risk to force, time sensitivity, war, time between wars, and catastrophic failure. Through S-curve and smoothing modeling, we find that trust components can be quantified in the human machine relationship as positive or negative trust, and that a relationship exists between understandability and adoption. While autonomous system components generally undergo rigorous testing to verify suitability and operability, human-machine trust is not usually incorporated into design and testing phases. When trust is built into the design and acquisition process, adoption of autonomous USVs is more likely to increase. Researchers can apply our trust model to future autonomous systems to mitigate distrust and human-machine teaming.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AI	artificial intelligence
AIS	automatic identification system
CIWS	close-in weapon system
CJCSM	*Chairman of the Joint Chiefs of Staff manual*
DoD	Department of Defense
DSB	Defense Science Board
FAR	Federal Acquisition Regulation
LUSV	large unmanned surface vehicle
ML	machine learning
SGR	Sentry Guard Robot
UAV	unmanned aerial vehicle
USV	unmanned surface vehicle
VCR	videocassette recorder

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

We would like to thank our advisors Dr. Shelley Gallup and Dr. Douglas MacKinnon for their diligent work and guidance.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The very idea of humans being replaced by machines causes fear in many of us, yet most of us also know that machines can not only make our lives easier but also save lives. Within the Department of Defense (DoD), systems such as the Phalanx close-in-weapon system (CIWS), M-48 advanced capability, and AEGIS weapon system provide evidence that the use of unmanned systems is replacing humans at the tactical level of warfare. However, trusting and relying solely on autonomous systems' decisions without fully understanding their decision-making process has sometimes proven fatal. Such was the case with the USS *Vincennes* (CG-49), a Ticonderoga-class guided missile cruiser that accidentally shot down an Iranian civilian airliner in 1988 although the crew had significant doubts about the air-defense system's classification of a radar contact. When it comes to machines that can act independently let alone autonomous weapons systems, it is no wonder we have some distrust. However, just as any new technology seems strange at first, we may find ourselves better able to trust advanced autonomous systems if we find a way to determine and to measure which factors increase trust in the human-machine relationship. Today, most autonomous systems require humans in the loop, making human-machine teaming essential to successful combat operations. The use of autonomous unmanned surface vehicles (USV) faces similar trust issues. Evidence that these machines can perform better than humans, especially with complex and risky tasks, is crucial for building trust in human-machine team operations.

A. PROBLEM STATEMENT

Although DoD has rigorously tested autonomous USVs (Federal Acquisition Regulation, 2019), widespread adoption remains slow (Klein, 2018). While the Federal Acquisition Regulation (FAR) provides a tailorable testing environment to achieve the end goal of adequately testing autonomous systems, managers with expertise must design and tailor the testing for various new technology, which is challenging as technology shifts. DoD incorporates automation and Artificial Intelligence (AI) into its platforms to reduce the cognitive workload of junior officers and other personnel and to improve mission capabilities. Naval operations, for example, require an almost omniscient awareness of surroundings so that personnel must track friendly ships, possible hazards, atmospheric conditions, aircraft positions, and civilian shipping traffic, among others.

The use of AI reduces that workload. AI has, in fact, become indispensable in recognizing and differentiating variables to detect potentially lethal vessels. However, commanding officers are often reluctant to trust artificial systems with decision-making and reporting. Even when a system delivers the correct answer, officers and other personnel are unsure of how the system derived its solution or what tactics the AI system employed and why.

B. PURPOSE STATEMENT

Many academic papers have qualitatively addressed the topic of trust of machines. The purpose of this thesis is to collect quantitative data on operational risk, system transparency, system understandability, and trust, and then to analyze and determine whether a relationship between trust and the slow adoption of unmanned surface vehicles exists within DoD. By quantifying the main components of trust—trust factors—into a model that can be used in the design and acquisitions phase of automated, autonomous, and artificial intelligence (AI) systems, we hope to extend the possible uses for DoD. To build our model, we research, describe, and then build upon the components of the Trust V model (Palmer et al., 2016) because it covers the attributes most relevant to our discussion of trust in autonomous systems. We quantify some of Palmer's specific attributes of automated systems, which are "benevolence, direct-ability, false-alarm rate, perceived competence, reliability, robustness, understandability, utility, validity" and some of Palmer's specific characteristics of autonomous systems, which are "adaptive/learning, adversarial, dynamic, human interaction, self-directed, self-governed, uncertain, unstructured" (Palmer et al., 2016, p. 62). Our effort also seeks to support quantitative comparisons of future systems in terms of trust to aid in make versus buy decisions.

C. RESEARCH QUESTIONS

- Primary: How can quantifying trust in DoD's design and acquisition phase of autonomous unmanned surface vehicles help DoD's adoption rate?
- Secondary: How can we quantify trust in terms of acquisition, design, and understandability of a given technology?

D. SCOPE

We utilize a quantitative model to create an overall equation to determine how trust correlates with the adoption of autonomous USVs. Future research might include testing of metrics, for example, understandability, human interaction, and robustness, with regard to an actual autonomous system to verify applicability and efficacy of our findings.

E. THESIS OVERVIEW

In Chapter I, we raised the possibility that lack of trust may be related to DoD's slow adoption of autonomous USVs. We also identified attributes of automated and autonomous systems that a trust model could quantify to potentially increase DoD's adoption of autonomous USVs. In Chapter II, we define trust as it relates to human-machine relationships and delineate the difference between automated, autonomous, and Artificial Intelligence (AI) systems. We also review previous research on trust, discuss surrounding controversies, describe how our research differs, and discuss why our quantitative approach may be more appropriate to explore the trust problem. In Chapter III, we present our methodological framework, including why we use the attributes and characteristics from the Trust V model and why we add our own attributes. Chapter III also explains what makes our framework valid as well as the possible expected outcomes. In Chapter IV, we summarize collected data, present our final model, analyze key findings with respect to research questions, and discuss how our results may fill noted knowledge gaps. In Chapter V, we summarize the thesis's contributions to new knowledge and practice. We conclude with recommendations for future research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

Chapter I outlined both the usefulness of autonomous unmanned surface vehicle (USV) and the challenges humans face in trusting technology to make decisions. To help determine how trust correlates with DoD's rate of adoption of USVs, Chapter II reviews relevant literature in seven sections. Following an introduction, the chapter defines trust and then differentiates between automated systems, autonomous systems, and Artificial Intelligence (AI). The chapter then discusses the need for technology and the limitations of testing, the surrounding controversy, which trust factors could impact adoption, and, finally, how a quantitative measure of trust may fill a gap in the existing literature around DoD adoption of autonomous USVs.

A. INTRODUCTION

Humans question whether to trust autonomous unmanned surface vehicles as they question any new technology before it is familiar, before they know if it is reliable, and before they determine its overall usefulness. Autonomous USVs engender even deeper trust concerns than other new technologies (Hartig & Vanhoose, 2019). For systems without offense capabilities, such as a surveillance USV, distrust can stem from the question of responsibility if the USV causes a collision. For weapons systems, distrust can stem from concerns that a machine could take a life without understanding situational ethics. In both scenarios, concerns come from the possibility that a machine could act in ways that humans would not. Possibly, this lack of trust has led to DoD's slow adoption of autonomous unmanned systems.

One of the major challenges is machine learning (ML), which allows a system to learn from its environment and independently adapt the way it operates based on that learning without human intervention. This can possibly lead to distrust because humans are not usually privy to a ML system's decision-making process. Humans have less trust in autonomous USVs with ML capability because their actions can change even when given the same set of inputs so that understanding the system becomes more challenging and, therefore, confidence that it will achieve the desired end decreases. Modern navigation software illustrates this challenge. The software can use ML to predict delays and route around them, but, because the human operator may not understand why the system recommends a different route at any given time, the person may not be

able to determine if it is malfunctioning or, in other words, whether to trust the alternate route provided. While the risk is minimal in civilian navigation compared to military operations, we still see consequences. For example, beyond the questionable directions most of us have received at one time or another while driving, driverless cars have caused accidents like the Tesla autopilot malfunction that caused a crash and killed an Apple engineer (Brown, 2019). Similar malfunctions happen in the air like the Ethiopian Airline Boeing 737 Max 8 crash where the readings from the two onboard sensors did not agree. The differing readings activated the automated safety system, which repeatedly pushed the plane's nose down despite the pilot and the first officer's desperate attempts at manual overrides (BBC News, 2019).

While it makes sense to use unmanned systems to decrease loss of life, rushing an unmanned system into use can also negatively affect trust. In October 2011, a Federal Advisory Committee called the Defense Science Board (DSB), established to provide independent advice to the Secretary of Defense, completed its information gathering on The Role of Autonomy in DoD Systems. The DSB Task Force concluded that unmanned systems are making significant contributions to DoD, but adoption of unmanned systems—especially when it comes to their autonomous capability—has been slow. Most DoD deployment of unmanned systems stem from pressing needs of conflicts, which means that the systems are rushed to combat without the necessary support, resources, and training (Defense Science Board, 2012). These factors, combined with lack of trust that an unmanned system will perform as intended, makes it hard for DoD to realize the full potential of autonomous USVs (Defense Science Board, 2012). To understand more specifically how trust may correlate to adoption rates of USVs, the next section defines trust.

B. DEFINING TRUST

Most trust definitions in the context of autonomous systems date back to the 1960s and 1980s. More recent definitions of trust include "willingness to place oneself in a relationship that establishes or increases vulnerability with the reliance upon someone or something to perform as expected" (Johns, 1996, p. 81). Moorman, Deshpande, and Zaltman (1993, p. 82) explain trust as "willingness to rely on an exchange partner in whom one has confidence." Mayer, Davis, and Schoorman (1995) contend that trust entails 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 party" (p. 712). This thesis uses the Johns definition as the most applicable and measurable in the context of military USVs.

C. DIFFERENCE BETWEEN AN AUTOMATED SYSTEM, AUTONOMOUS SYSTEM, AND ARTIFICIAL INTELLIGENCE (AI)

Automated systems are machines that have predetermined tasks based on programming. Robots that manufacture automobiles are one example. Autonomous systems, on the other hand, are built upon the concept of self-governance. For instance, many cars today are able to automatically decrease their speed based on the distance to other cars without human intervention because they are programmed to do so. AI, on the other hand, has the ability to gather information and then use that information to make decisions or take actions beyond their original programming. AI is sometimes mistaken with ML, but they are fundamentally different. For instance, "machine learning often relies on pattern detection made possible by ingesting massive amounts of data rather than the inferential reasoning that defines human intellect" (Hartig & Vanhoose, 2019, p. 2). AI on the other hand reaches far deeper in mimicking aspects of human intelligence to include perception, emotion, judgment, recognition, understanding, creating, and thinking (Li & Du, 2017, p. 1). Adoption of autonomous systems with aspects of ML are the focus of this research since these systems are more widespread in the DoD (Defense Science Board, 2012). AI systems are still in their infancy and not discussed in this research, though the impact of AI will likely be the greatest (Horowitz, 2018).

D. THE NEED FOR AUTONOMOUS TECHNOLOGY AND LIMITATIONS OF TESTING

The September 11th terrorist attacks became a pivotal event for the deployment of unmanned aerial vehicle because the United States realized that the terrorist tactics required a different style of warfare. 9/11 warranted the deployment of weaponized UAV because, by so doing, the United States could quickly respond to an unknown method of warfare (Bowden, 2013). UAVs were also cheaper, stealthier, and required no risk to American lives as compared to manned military air combatants. In retaliating for the attacks, the United States emphasized the use of

weapon-equipped UAVs in taking down the leadership of Al Qaeda, hoping to both spare further American lives and cause less collateral damage (Bowden, 2013). Although there have been civilian casualties due to UAV employment, these are rare circumstances (Byman, 2013). UAVs provided an extension of human warfighting capability, but these systems still required some type of human interaction. For instance, the Scan Eagle has some autonomous capabilities, such as flying in waypoints provided by the flight crew, but the system still requires a human in the loop when it comes to takeoff and landing of the aircraft.

Although there was a significant increase in the adoption of UAVs after 9/11, the adoption of fully autonomous UAVs has been slow especially when the UAV is equipped with weapons such as the Predator, which is capable of releasing weapons on human targets or infrastructure. Similarly, DoD has only slowly adopted autonomous USVs. Commanding officers are reluctant to turn over control to artificial systems that lack transparency in decision-making and reporting. However, the need for autonomous systems arises from mission priorities and how complex technology has evolved today (Palmer et al., 2016). The management of this complexity is what drives the issue of trust. Intelligent systems will face the same deficiencies as humans mainly because of the complexity of the software required for a machine to be fully autonomous.

The decision-making required by autonomous systems involves complex algorithms that are bound by the limited amount of time for software testing. Full testing would require that the system is confronted with every scenario and combination of inputs that it may encounter, resulting in an infinite number of testing scenarios. According to Atkinson (2015), the autonomous systems are limited by the amount of available information, computational resources, and the limited amount of time available to research a conclusion. Since infinite testing is not reasonable, the software testing is limited to reasonable scenarios that the developers are able to envision at the time. The goal of the autonomous systems, then, is to make them act independently of humans and have the ability to make independent decisions.

E. CONTROVERSY THAT SURROUNDS HUMAN-MACHINE TRUST

Such a large and crucial goal understandably engenders controversy; one crucial argument surrounding human-machine trust is whether we need a "human in the loop." Some critics argue that "everything we know about dangerous machines under the control of complex software

systems says we still want people there to mitigate the risks to human life" (Hsu, 2015). As an example, Mindell in Hsu article, mentions that human involvement is critical in self-driving vehicles and how these cars should not be fully autonomous. However, we know humans get tired and fall asleep at the wheel, drink and drive, text and drive, and have many other operator failures that make automobile accidents one of the leading causes of death. Fully autonomous self-driving cars, creating a safe and convenient environment, could potentially mitigate those operator errors; after all, machines do not get tired nor have we seen very many drinking alcohol. Many newer cars are already equipped with sensors and radar that allows these cars to automatically brake when they sense danger ahead or behind. Currently, these cars are playing supportive roles, but Tesla and other car companies are taking these technologies to the next level of full automation, performative roles. While self-driving cars still have flaws, they point to an example of where autonomous systems, systems without humans "in the loop," may become safer than the alternative.

In the context of the DoD, not having humans in the loop certainly makes sense for time critical systems such as close-in weapon systems (CIWS). CIWS is a ship's weapon defense system designed to search, track, and engage targets within milliseconds without any human involvement. Other critiques like Lin (2016) suggest that autonomous systems should not be given the power to take human lives because, no matter the technological advancement, these systems will never be able to truly deliberate and appreciate the weight of taking a human life.

The physical removal of the human from the battlefield should not alter the ethics of warfare. Autonomous/unmanned systems should complement military operations in hostile environment or terrain that are impossible for humans to navigate. Some of these systems are fully autonomous, meaning they do not have humans in the loop. They operate independently and can engage a target without any human interaction. An example of such an autonomous system is the SGR-A1, Sentry Guard Robot, used in South Korea to patrol its borders with North Korea (Ross, 2013). With technological advancements in the past decades, the way we fight wars has shifted from man-to-man to man-to-machine and sometimes machine-to-machine. Ethical considerations are crucial but also made more difficult because of the abstract nature of trust. Were researchers better able to grasp specifics that influence trust, we could possibly increase human-machine trust beginning with the testing phase of autonomous systems. Trust, as stated earlier, is controversial

in definition and assigning a trust number to an autonomous system with potentially lethal capability could raise ethical questions (Ross, 2013). While questions of what level of trust in a system is needed to allow it to autonomously conduct lethal actions and who would be held accountable when it makes the wrong decision are extremely valid, they are beyond the scope of this thesis. Before those can be considered, first, we need to determine if measuring trust is possible. Possibly, by building trust into the system, autonomous systems could reach their full potential.

F. TRUST FACTORS THAT COULD IMPACT ADOPTION OF AUTONOMOUS SYSTEMS

If autonomous systems can save lives yet decreased trust prevents or slows DoD's adoption, which specific factors impact trust? This thesis is not the first to attempt to measure trust. Current literature provides many avenues for defining and building trust. For example, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust" by Hoff and Bashir analyzes qualitative trust attributes. Taylor, Mittu, Sibley, and Coyne in 2016 attempt to define individual factors to evaluate their impact on trust. According to the Trust V paper by Palmer, Selwyn, & Zwillinger (2016), nine automation trust attributes and eight autonomous characteristics must be addressed in the design process to engender trust in an autonomous system. Schaefer, Chen, Szalma, & Hancock (2016) also addressed some of these attributes and characteristics. Both articles mention how humans trust automation based on perceived competence or predictability of the system, meaning that humans follow Bayes' Theorem, the operator's expectation that the "probability that something is going to happen [is] based on something that has already happened in the past." Both articles also identify reliability or dependability. Other notable attributes of automation are validity, where the system is able to solve the correct problems, and false-alarm rates, the known and acceptable error rate. Excessive alarms, especially an increase in false alarms, **can negatively affect human trust in such systems and inadvertently affect performance. As to** characteristics, this thesis considers four main autonomous characteristics from the Trust V paper relevant to our research: Human Interaction, Self-directed, Self-governed, and Adaptive/Learning, all described in our methodological framework.

G. A QUANTITATIVE MEASURE OF TRUST MAY FILL A GAP IN EXISTING LITERATURE AROUND DOD ADOPTION OF AUTONOMOUS SYSTEMS

Our goal in this research is to build a quantitative model of trust that encompasses the Trust V paper factors and assign a numerical value for trust in a given situation. Our model attempts to go further than previously identified trust factors so that our quantitative model allows systems engineers to design for specific areas of trust early in development. Statistically, we can show a system has been adequately tested up to a given level, which is represented by confidence, based on Bayes' Theorem. Obviously, no one wants to discover problems with autonomous systems during combat situations any more than DoD wants to risk untrained humans. Combat is not the time to find out your system needs to be recalled due to malfunctions, and it may be too late too due to potential casualties. Testing of autonomous unmanned systems, therefore, needs to be done in a controlled garrison or training environment first in order to fix the bugs. There should be a minimum level of reliability and predictability before an autonomous system is deployed in a combat environment because, generally, technologies are designed according to specifications. The minimum level should not change for a training or deployed environment. As the saying goes, "train like you fight or fight like you train." The baseline needs to be established during the testing phase, ensuring the systems' reliability and predictability before combat deployment. This research, therefore, focuses on a quantifiable way of testing and evaluating trust in autonomous USVs in the design and acquisition phase.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

In addition to decreasing risk to human lives, in order to keep up with peer competitors like China and Russia when it comes to the arms race, the United States must invest in autonomous technology, but integrating them into the military will require investing in trust (Hartig & Vanhoose, 2019). While Chapter II identified the need for a quantitative measure of trust in the design phase of autonomous USVs, Chapter III describes how we quantified and modeled trust attributes and characteristics to apply them in a military environment. The chapter first describes why we chose to address the design and operation of autonomous USVs in terms of human-machine teaming. Second, the chapter describes the trust factors adopted from the Trust V paper and then the trust factors we added. Last, the chapter describes the basic model and the possible outcomes expected.

A. METHODOLOGICAL FRAMEWORK FOR TRUST

In our model, autonomous systems complement the mission by integrating as a team member to extend human capabilities (Scharre, 2016), and our reasoning is based on Scharre's further point that humans may perform three roles in a human-machine team. When a human is playing the role as an essential operator, the autonomous system requires a human in the loop to effectively complete an engagement (Scharre, 2016). This teaming is applicable to weapons systems onboard naval ships that require human interaction during the flight of the missiles in order to change course. Even though the waypoints are predetermined, a human in the loop ensures each weapon proceeds to its designated target. The human can also add new waypoints as needed. When a human is acting as a moral agent, the human operator is making a decision on proportionality of force to be used (Scharre, 2016). This model is applicable to systems such as Predator, a potentially armed unmanned aerial vehicle. Predators are controlled by a human who can decide to release the payload based on collateral damage. For instance, if the target is at a gas station and seems to be alone, an autonomous system may engage such target. A human in the loop may realize that dropping munitions on a gas station will result in larger explosion than required and will wait until the target leaves the gas station. Lastly, when a human fills in as a fail-safe agent, the human can intervene and/or stop the operation if the system begins to fail or if the

situation does not warrant an engagement (Scharre, 2016). Weapons systems like the close-in weapon system can engage a target automatically but can be disengaged by a human when necessary. We posit that autonomous systems, even those without offensive capability, will require these same human roles.

It is clear that at the present time, and perhaps for some time to come, autonomous systems (especially lethal ones) will require humans in supervisory or controlling roles. A model may be used to show the interactions of the various concerns related to the adoption of autonomous systems. The purpose of this model is to show various contributions that affect adoption rate as a surrogate for trust.

B. TYPE OF DATA COLLECTED: TRUST V MODEL

The Trust V model summarizes many of the early findings as regards trust in autonomous systems and consolidates the findings into nine factors:

1. *Benevolence*—system is supporting the mission and operator (not in opposition)
2. *Directability*—system can be re-directed by the operator
3. *False-alarm rate*—certain error rates are known and acceptable
4. *Perceived competence*—the operator believes the system can perform a task
5. *Reliability*—the system has only a small chance of failing during a mission
6. *Robustness*—the system can appropriately handle perturbations
7. *Understandability*—the conclusions a system reaches can be understood
8. *Utility*—the system adds value
9. *Validity*—the system is solving the correct problems (Palmer et al., 2016, p. 62).

Trust in autonomous systems is guided by eight additional factors:

1. *Adaptive/Learning*—system can acquire information and then beneficially leverage that information

2. *Adversarial*—system can complete its mission when subject to opposing efforts

3. *Dynamic*—system can complete its mission in a changing environment

4. *Human interaction*—system can interact with humans, and humans with the system, for mutual benefit

5. *Self-directed*—system can direct itself (e.g., can determine what to do)

6. *Self-governed*—system can control itself (e.g., can do what needs to be done)

7. *Uncertain*—system can complete its mission in environments that are different from expectations

8. *Unstructured*—system can complete its mission in environments that are difficult to describe (Palmer et al., 2016, p. 63).

Each of the trust attributes and characteristics from the Trust V paper that we included in the simplified initial model are defined and described as follows.

Adaptive/Learning—This factor is analogous to ML; according to Mitchell (1997, p. 2), ML is where "a computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E."*

Adversarial—The system is capable of continuing its mission even when an adversary is actively trying to stop it. Examples could be a blockade trying to stop the USV from reaching its final destination or adversaries trying to flood the USV with bad radar returns.

Dynamic—When the environment changes, if visibility decreases or shipping traffic suddenly increases, for example, the USV can still continue its mission. In these cases, the USV should not go dead in the water simply because of a changed environment.

Uncertain—If the mission parameters are not clearly defined, the USV can still continue without having to be reprogrammed. For example, if the mission is to find a ship that is moving

erratically and has a peculiar automatic identification system (AIS) signature,[1] the USV can continue the mission despite ill-defined parameters.

Unstructured—This is the ability to receive data in many formats and incorporate them into the mission. Humans can receive data in nearly unlimited formats—hand written, typed, technical data, voice, and more—and still decipher the meaning. The USV should be able to continue its mission even if the data formats are not uniform.

Perceived competence—Though a system may be technically reliable and predictable, the operator may still believe it is malfunctioning. The USS *McCain* collision provides an example of a crew not understanding a new system's operation and believing it to be malfunctioning even though it proved as reliable during the acquisition process (LaGrone, 2019b). Another example is when a computer hangs; though it may be processing and operating exactly as specified, the user may believe it is malfunctioning. If, however, a hanging computer shows a graphic representing progress, the user will believe in the competence of the machine even though it may be unwarranted.

Directability—Once the mission is programmed, the user can change the parameters and even override the system.

Self-directed—Self-directed systems can decide on a course of action given a general set of guidelines, but without explicit direction for the specific situation (Mitchell, 1997). For example, in a nautical "rules of the road" situation, a system can use international regulations to determine a specific course of action even though a multi-ship interaction is not explicitly defined.

Self-governed—Self-governed systems can decide when to impose a limit on their operations due to any combination of technical and non-technical factors. For example, if an unmanned aerial vehicle is flying in a high-temperature environment with data feedback loops that show engine temperatures that continually violate operating guidelines, the system could decide to use lower speed or fly at different altitude to maintain the correct temperature.

[1] AIS is used for tracking ships at sea in order to avoid collision. Typically, merchant vessels are known to maintain a steady course and certain speed in order to save fuel and time. An observer of such AIS signature on their radar system can easily assume this is a merchant ship even while the vessel is still out of sight.

Reliability—This is the probability that something is going to be successful based on the fact the it has been successful in the past (Better Explained Discussions, 2017).

Understandability—When the system reaches a conclusion or takes an action, the exact steps and considerations that went into the system can be recalled and weighted to show why the system reached the conclusion it did. This ties strongly into the human interface because decision-making information must be presented in a clear manner with sufficient technical depth while avoiding information overload that would prohibit an end user from comprehending what the data points are telling them. Our initial model uses understandability as a stand-alone variable. The final model has understandability as a combination of subjective norms, human interaction, and robustness.

Human Interaction—This is the ability for a human to direct and glean information from the system. In our final model, it includes Perceived Competence, Directability, and Robustness. A good human interface provides only the relevant information in an easily digestible format (Taylor, Mittu, Sibley, & Coyne, 2016). The constraints are that sufficient information must be provided while not overloading the operator. A human interface that simply presents the decision lacks depth and is not understandable. Conversely, an interface that provides hundreds of sensor values and algorithmic outputs will likely overload the human operator. Once the information is presented, the operator needs an intuitive way to direct the system. A graphical user interface with simple commands such as "Fire" or "Turn Left" could be useful whereas a system requiring multiple terminal commands to achieve the end state decreases the usefulness of the system. If an autonomous system is to extend human capabilities, that goal is completely defeated if a system requires that a human must do most of the work, a situation which can decrease trust.

Robustness—This is a measure of the autonomous systems that we define as the combination of Reliability, False-alarm Rate, Adversarial, Dynamic, Uncertain, Self-governed, Self-directed, and Unstructured. It is the totality of systems characteristics that deal with perturbations.

C. TYPE OF DATA COLLECTED: ADDED ATTRIBUTES

We consider the following to be essential trust attributes for DoD testing of autonomous systems.

Failure Rate— Understandability is inherently tied in to the failure rate. For instance, when an autonomous system's actions are not being understood by its human counterpart, the system is perceived to be malfunctioning.

Regulations and Policies—This attribute relates to authoritative documentation from any level that can limit the operations of a system. Government regulations are currently imposed on consumer grade drones, and we believe this will expand in the future. At lower levels, organizations such as the Navy may impose different or tighter regulations than those legally required to maintain safety and public opinion. Regulations and policies are an inevitable byproduct of widespread use, but they are also influenced by public opinion, competition, and failure rate. Public opinion changes based on catastrophic failures. As public opinion of any component wanes, increased regulation can follow. Take, for example, the Boeing 737 Max crashes that recently occurred. Though only one model was faulted, all of Boeing's line as well as all aircraft with highly automated functions are being questioned, leading to further scrutiny and regulations for the entire airline industry (Federal Aviation Administration, 2019).

Risk—In risky situations, trust tends to be of significant importance. Citing Deutsch's data, Mayer explains that risk, or having something invested, is requisite to trust (Mayer et al., 1995). Our initial model had one risk factor that encompassed all levels of risk. For our final model, we break risk into three levels: strategic risk, risk to mission, and risk to force.

War—For the purposes of this model, war is on a continuous scale from 20% to 80%. This represents a state where U.S. forces are never fully withdrawn from conflicts around the world, but are also not involved in unrestricted warfare. The percentage of war that the nation wages will directly affect adoption of autonomous systems as necessary to reduce the risk of lives lost during any conflict. Autonomous systems also can give us a competitive advantage by complicating an adversary's targeting problem. A higher level of war leads to increased funding and, therefore, quicker adoption, which can drive competition to produce better systems, further accelerating trust and adoption. The converse is true during peacetime. War can also lead to less restrictive policies and regulations adding to adoption.

Competition—The ability for competition, whether civilian or military, to produce better systems than the current arsenal provides an impetus for military system improvements. If, for

instance, a civilian corporation produces fully autonomous systems with the ability to make life or death decisions, then the military will be questioned as to why they could not produce a similar system. With an uptick in civilian adoption comes lower cost, greater user bases, and an overall higher acceptance of the role of autonomous systems, which can lead to greater military adoption rates. On the military side, a foreign military's possession of superior capabilities may drive funding and development to maintain military superiority.

Strategic Risk—As defined by DoD Joint Risk Analysis (Chairman of the Joint Chief of Staff Manual [CJCSM], 2016, p. C-4), "strategic risk is the potential impact upon the United States—including the U.S. population, territory, civil society, critical infrastructure, and interests—of current and contingency events given their estimated consequences and probabilities (e.g., the security of the United States and its citizens)."

Risk to Mission—DoD Joint Risk Analysis states that risk to mission consists of *operational risk* and *future challenges risk* and defines them as follows. "Operational Risk reflects the current force's ability to attain current military objectives called for by the current National Military Strategy, within acceptable human, material, and financial costs" (Chairman of the Joint Chief of Staff Manual [CJCSM], 2016, p. C-8). "Future Challenges Risk reflects the future force's ability to achieve future mission objectives over the near and mid-term (0-7 years)" (Chairman of the Joint Chief of Staff Manual [CJCSM], 2016, p. C-9).

Risk to Force— DoD Joint Risk Analysis states that Risk to Force consists of *institutional risk* and *force management risk* and defines them as follows. "Institutional Risk (Risk-to-Force) reflects the ability of organization, command, management, and force development processes and infrastructure to plan for, enable, and improve national defense" (Chairman of the Joint Chief of Staff Manual [CJCSM], 2016, p. C-9). "Force Management Risk (Risk-to-Force) reflects a Service and/or Joint Force Provider's ability to generate trained and ready forces within established rotation ratios and surge capacities to meet current campaign and contingency mission requirements" (Chairman of the Joint Chief of Staff Manual [CJCSM], 2016, p. C-9).

Catastrophic Failure—Failures, which due to their severity or effect on public opinion, can cause a negative effect on the adoption of autonomous technology.

Subjective Norm—In 2016, Ho, Ocasio-Velázquez, and Booth defined subjective norms "as perceived social influences or pressure, determined by the total set of accessible normative beliefs." Subjective norms can be considered social norms from peers, informal circles, supervisors and senior leaders (Ho et al, 2016). For our research, we take Subjective Norms to mean the level of autonomous systems that the military member's peer group would find acceptable.

Time Sensitivity—Rice and Keller showed in 2009 that use of automation increased under time pressure by having multiple participants try to pick out a number of aerial vehicles in a picture with or without the help of automation. As the allowed time to complete the task decreased, participants tended to rely more heavily on the automation to decide how many vehicles were in the picture. We posit that the same will happen with autonomous systems. As the time window for a human to act becomes smaller, an increase in reliance on autonomous systems will occur.

Policy Effect—Policy effect is an attribute that looks at the number and complexity of policies and regulations to determine if they will speed up or slow down adoption. Though policies and regulations can help to further the adoption of autonomous vehicles, there is a tipping point where they will result in a slowdown (Adjerid, Acquisti, Telang, Padman, & Adler-Milstein, 2015).

D. METHOD IMPLEMENTATION

Our methodological framework uses Stella software to build a trust model as shown in Figure 1. The model depicts how an increase in positive trust would lead to a significant increase in the adoption of autonomous USVs. Consequently, an increase in negative trust would lead to a slow increase in the adoption of autonomous unmanned surface vehicle. We consider our methodological framework to be valid because the results from Chapter IV of this study were within expected parameters. Our method also builds upon the components of the Trust V model by quantifying some specific attributes of automated systems, "reliability and understandability" (Palmer et al., 2016, p. 62), and some specific characteristics of autonomous systems, "Adaptive/Learning, Human Interaction, Self-directed, Self-governed" (Palmer et al., 2016, p. 62). Additionally, our initial model includes War, Competition, Regulations/policies, risk, and failure

rate, allowing for quantitative comparison of future systems in terms of trust to aid in make/buy decisions.

Figure 1 represents our initial trust model, derived from *Business Dynamics: Systems Thinking and Modeling for Complex World* by John D. Sterman. His book describes how complex world problems could be solved via simple models. The squares in Figure 1 are known as "stocks" and are analogous to tanks in a water system. The valves represent "flow rate" and control the inflow and outflow of the stock. This controls the inflow of water into the tank, or a drainage, which controls the outflow of water from the tank. The circles represent "auxiliary variables," sets of values assigned in order to simplify the mathematical equation that needs to be entered in the flow rate. The auxiliary variables can be omitted, leaving the model with only of stocks and flows (Sterman, 2000). However, the use of auxiliary variables makes modeling easier to understand because the audience can easily view the variables going into an equation.

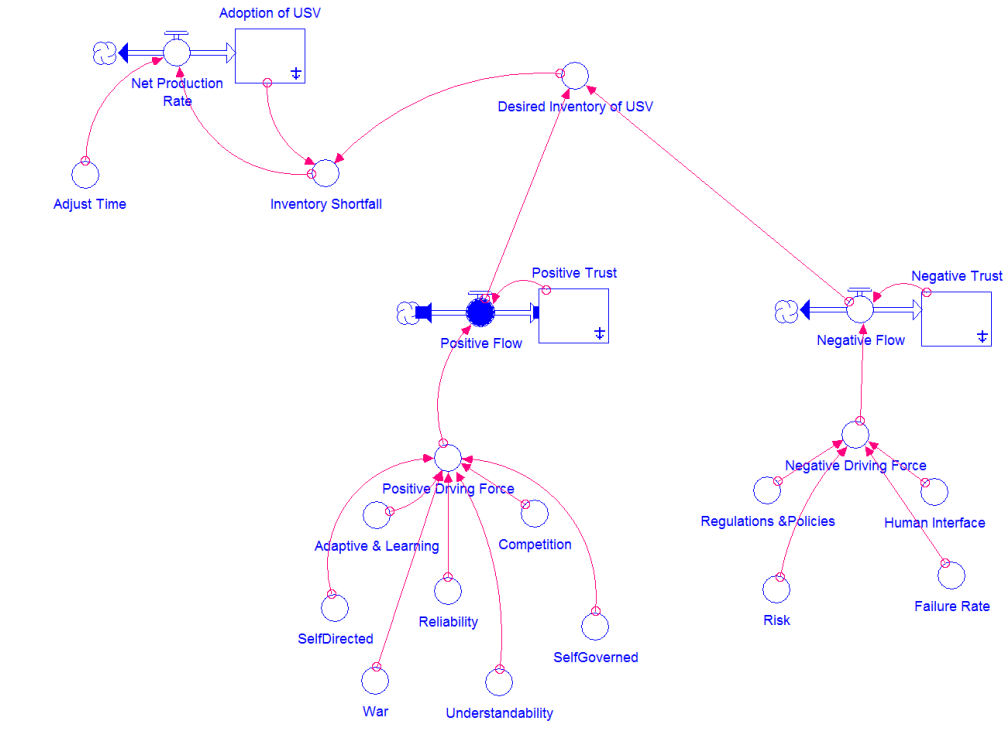


Figure 1. Initial Trust Model. Adapted from Sterman (2000).

Figure 1 depicts how an increase in Positive Driving Force will lead to Positive Flow of Trust while an increase in Negative Driving Force will lead to an increase in Negative Flow of Trust. Desired inventory is the number of USVs desired based on the positive and negative driving forces at the time. When Positive Flow is greater than Negative Flow, desired inventory equals 20. If Positive Flow is less than Negative Flow, desired inventory equals 5. This is based on the premise that an increase in Positive Flow of Trust in an autonomous system will generate a higher number of Desired Inventory of USVs whereas an increase in Negative Flow of Trust in autonomous system will lead to low number of Desired Inventory of USV as shown in Figure 2. According to LaGrone (2019a), the Navy plans to build two large unmanned surface vehicle (LUSV) every year (2020-2024) for a total of ten LUSV in five years. Based on that timeline and adoption rate, we anticipated that the Navy would build 20 LUSV in ten years but we do not anticipate this trend to continue indefinitely. We also anticipated that due to budget constraints and/or lack of trust based on high failure rate, regulations and policies, , and so on, that the number may be significantly reduced to five LUSV in ten years. As of this writing, only one LUSV, the Sea Hunter, exist and has been successfully demonstrated.

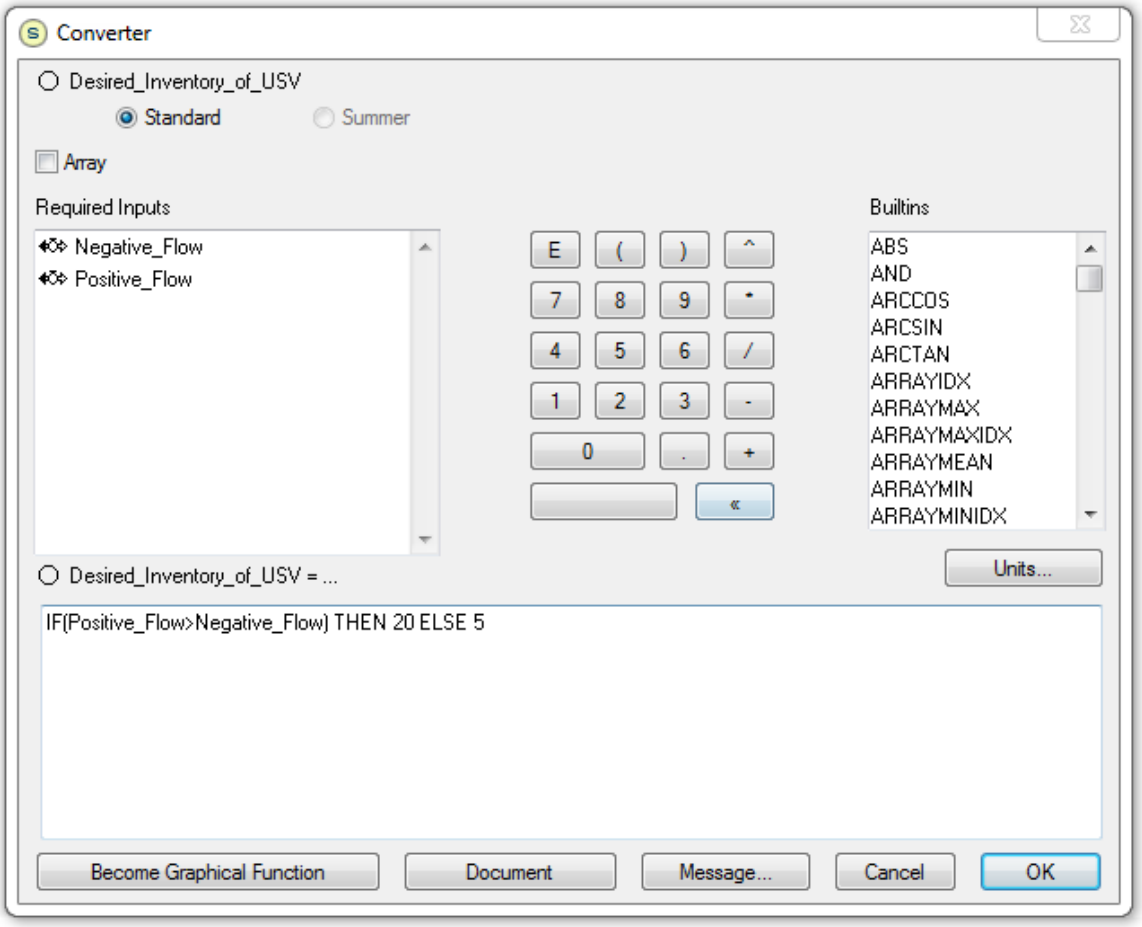


Figure 2. Desired Inventory of USVs Equation. Adapted from Sterman (2000).

Figure 3 shows an equation that comprises of all the formulas we used to derive our graphical output, detailed in Chapter V. The overall goal of our equation is to show how the difference between Positive Flow and Negative Flow drives Desired Inventory of USVs based on Inventory Shortfall (difference between Desired Inventory of USVs and Adoption of USVs) and Adjustment Time (set to 5 years). This will in turn lead to Adoption of USVs either increasing significantly or at a slow pace, all based on a key variable: War.

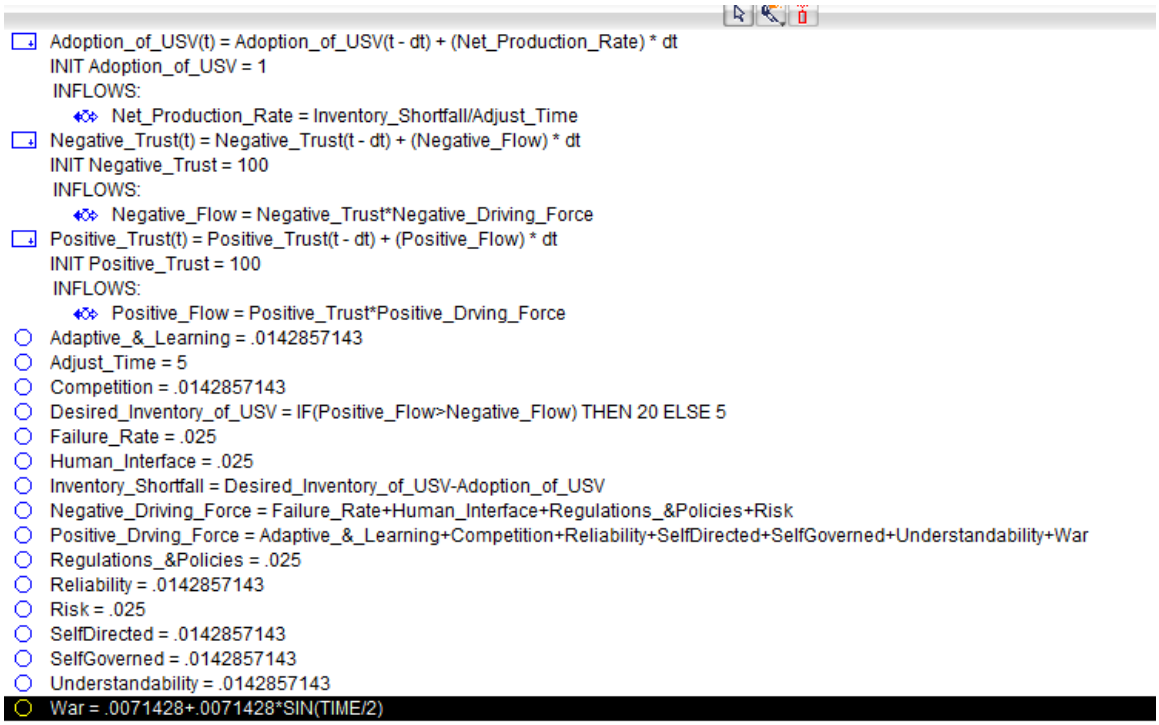


Figure 3. Trust Equation. Adapted from Sterman (2000).

E. POSSIBLE OUTCOMES EXPECTED

The outcome of our research will be a more robust software model of the factors affecting trust in autonomous USVs. The final model builds upon the basic model presented in this chapter by adding in new factors and making more complex models of the existing factors, demonstrating how the interplay may affect planned adoption of autonomous systems.

IV. DESIGN, ANALYSIS, AND RESULTS

Chapter IV describes the results of the thesis's modeling research. Using the Stella software, we show how adoption of autonomous unmanned surface vehicle (USV) is affected by trust attributes. We built two trust models for this research. The initial model presented in Chapter III includes some of the attributes and characteristics of the Trust V paper plus a few of our own attributes. The second trust model includes additional attributes from the Trust V paper plus a few additional attributes we considered relevant as we researched further into the relationship between trust and the adoption of USVs. Chapter IV summarizes and specifies the data, presents the final trust model in stages, and considers the negative and positive growth modules as well as adoption modules. Last, the chapter summarizes key findings, contributions to the research gaps, and implications for research and practice.

A. SUMMARY OF DATA COLLECTED

The data collected focused on the attributes of trust as described in Chapter III. We chose attributes from the Trust V paper since it succinctly categorizes factors of automated systems into nine attributes and characteristics of autonomous systems into eight attributes. The Trust V paper also consolidates early findings of other publications, including Schaefer, Chen, Szalma, and Hancock's (2016) article "A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications of Understanding Autonomy in Future Systems."

Although the Trust V paper described nine automation attributes and eight autonomous characteristics that must be part of the design and development process to engender trust, our final model includes only two out of the nine factors of automation—understandability and robustness. In addition, it includes only two out of the eight autonomous characteristics—human-interaction and adaptive/learning. In our model, other factors of automation and autonomy are combined into these four while three factors, benevolence, validity, and utility were assumed to exist in DoD systems, and therefore not analyzed. Lastly, we included ten of our own attributes to be considered in the design and development process of autonomous USVs. Table 1 depicts the different attributes and characteristics of the Trust V paper and the additional attributes we contributed to this thesis in order to build trust into the design process.

Table 1. Automation Attributes and Autonomy Characteristics. Adapted from Palmer, Selwyn, and Zwillinger (2016).

9 Trust V Automation Attributes	8 Trust V Autonomy Characteristics	10 New Attributes (our study)
Perceived Competence	Adversarial	Catastrophic Failure
Benevolence	Dynamic	Regulations/Policies
Understandability	Human Interaction	Risk to Force
Direct-ability	Self-directed	Risk to Mission
Reliability	Uncertain	Strategic Risk
Validity	Self-governed	War
Utility	Unstructured	Competition
Robustness	Adaptive/Learning	Subjective Norm
False Alarm Rate		Time Sensitivity
		Policy Effect

The yellow highlights in Table 1 are the Trust V paper attributes and characteristics that we explicitly adopted. Green highlights denote factors that were implicit in other factors. Blue highlights are additional factors that we believe are necessary to build trust in autonomous systems. In our first trust model, we assigned a specific number to each attribute and characteristics of trust. We kept all of the positive driving forces' values the same and all of negative driving forces' values the same, with the exception of one attribute, "war."

B. TRUST MODEL

In our final model, certain factors are static and user-defined while others are based on equations to show the interactions between the attributes.

1. S-curve for Simplified Adoption Model

To more accurately model adoption rates seen from videocassette recorders (VCRs) to railroads (Hall and Khan 2002), we use the S-shape curve. We believe that this is an accurate base to begin modeling since "it seems natural to imagine adoption proceeding slowly at first, accelerating as it spreads throughout the potential adopters, and then slowing down as the relevant population becomes saturated" (Hall & Khan, 2002). Figure 4 illustrates the technology adoption of the telephone, radio, television, and internet, demonstrating this familiar S-curve.

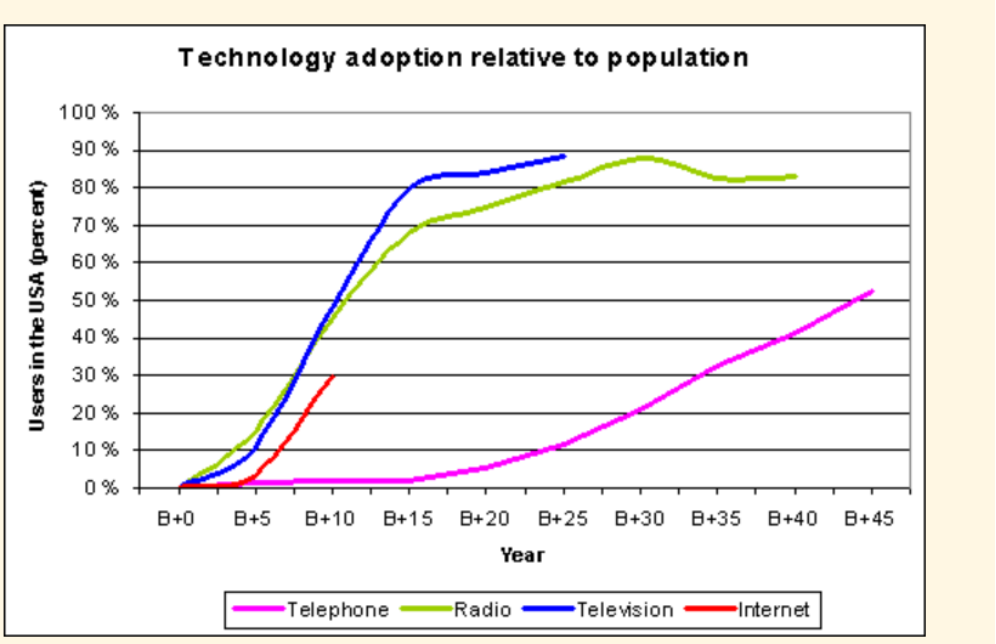


Figure 4. Technology Adoption Curves. Source: Hannemyr (2003).

In order to model this effect, we use a population growth equation,

$$\frac{dP}{dt} = r*P(1 - \frac{P}{K})$$

where P is population, *r* is the growth rate, and *K* is the carrying capacity. We change this equation to a time rate change of adoption,

$$\frac{dA}{dt} = r*A(1 - \frac{A}{M}) \quad ,$$

where A is the adoption in terms of percentage of the fleet that is autonomous, r is the rate of adoption, and M is the maximum percentage of the fleet that will be converted to autonomous. The M factor is important because we believe that, regardless of the sophistication of autonomous systems, some percentage of the fleet will remain manned. This could be, for example, a portion of the nuclear forces where we want a human to be in complete control to prevent any automation errors or hacking.

We set r = .2, resulting in maximum adoption within 50 years, which we feel is a reasonable timeframe for DoD adoption of USV, the Sea Hunter, out of 289 deployable battle force ships: 1/289 = .34% (U.S. Navy, 2019). The maximum adoption is set to 95 because we posit that Navy

will turn a high percentage of the fleet to autonomous while some percentage, in this case 5, will remain manned. This produces a simple model in Figure 5 that predicts adoption.

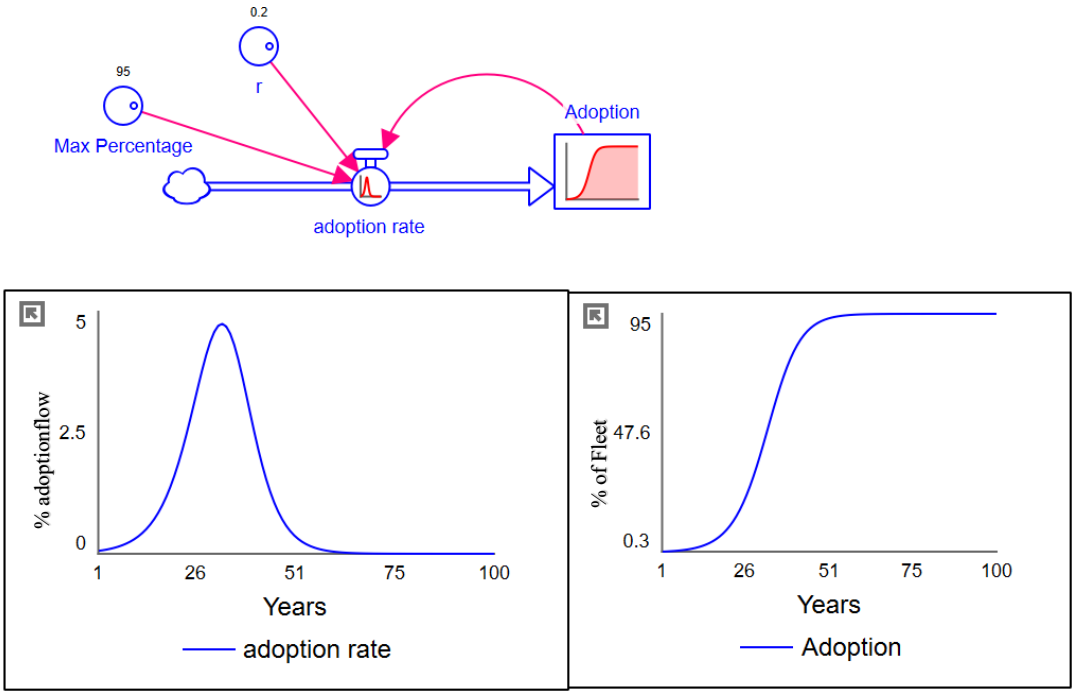
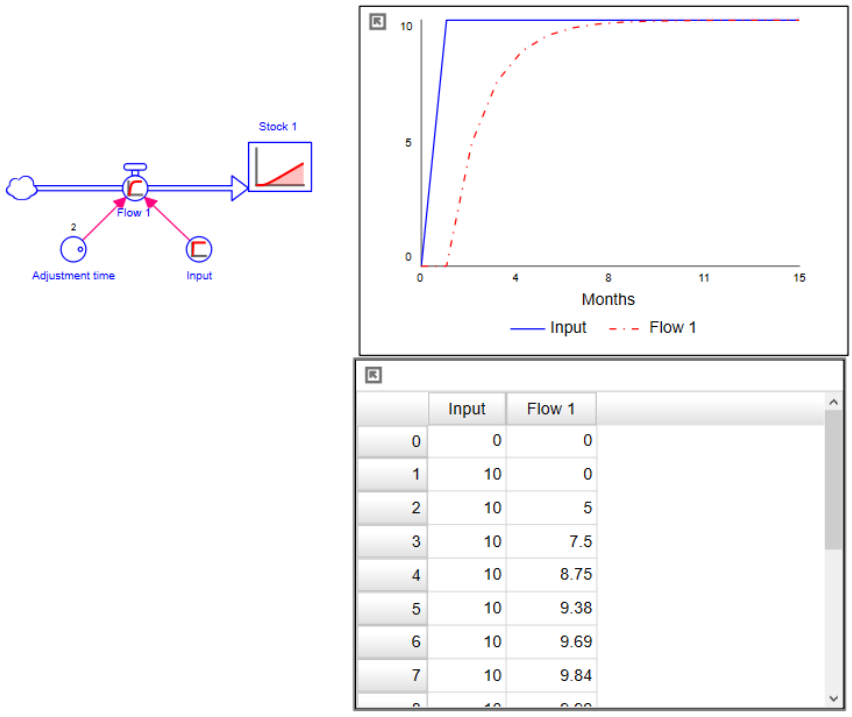


Figure 5. Simplified Adoption Model. Adapted from Sterman (2000).

2. Simple Smoothing Model

Here we add in a smoothing time that is constant throughout the model. The smoothing time causes the model to behave in more realistic manner since it allows changes to occur over time as shown in Figure 6.



	Input	Flow 1
0	0	0
1	10	0
2	10	5
3	10	7.5
4	10	8.75
5	10	9.38
6	10	9.69
7	10	9.84

Figure 6. Simple Smoothing Model. Adapted from Sterman (2000).

An example of smoothing in everyday life is adjusting the water in a garden hose. One may adjust the spigot to its maximum setting quickly, but the water flow at the end of the hose does not immediately respond. Rather, it reaches the new higher flow gradually. Figure 6 shows a simple example of smoothing where the input jumps to 10 at Time 1. The flow takes time to respond and will cut difference between the input and the flow down using the equation

$$\Delta \text{Flow}_{t+1} = \frac{input_t - flow_t}{Adjustment\ time} \text{ (Sterman, 2000)}$$

The chart shows that, with the adjustment time of 2, the difference between the input and the flow is cut in half each cycle.

3. Trust Model

We then added in modules to account for positive and negative growth, as shown in Figure 7. Additionally, we made the adoption rate a bi-flow. This means that adoption can also flow out based on negative growth factors. This is in contrast to our initial model where the flow could drop to 0 but could not reverse. Figure 7 and Table 2 show the final model along with associated

equations. As previously discussed, the adoption follows a population growth model using an adjusted growth rate. The adjusted growth rate is the difference between positive growth and negative growth modules, smoothed over time. The internals of the positive and negative modules are discussed in Section 4.

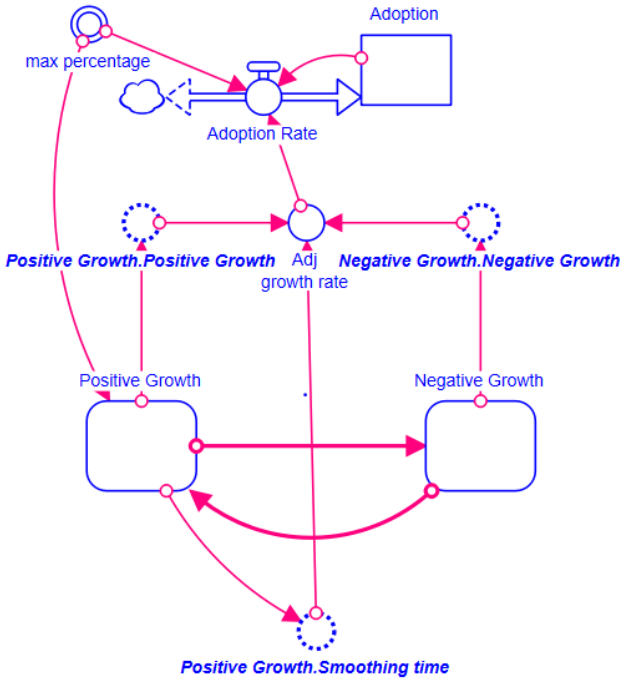


Figure 7. Final Trust Model. Adapted from Sterman (2000).

Table 2. Final Trust Model Equations

Attribute	Equation
Adoption	Initially 1/289 = .34%
Adoption Flow	Adj_growth_rate*Adoption*(1-Adoption/max_percentage)
Adj Growth Rate	SMTH1(Positive_Growth-Negative_Growth, 3)
Positive Growth Smoothing time	3 years
Max Percentage	95

4. Negative Growth Module

We next look into the negative growth module, shown in Figure 8. Negative growth factors include Catastrophic Failures, Time between Failures, Severity, Severity Overweight, Strategic Risk, and Risk to Mission, shown in Table 3.

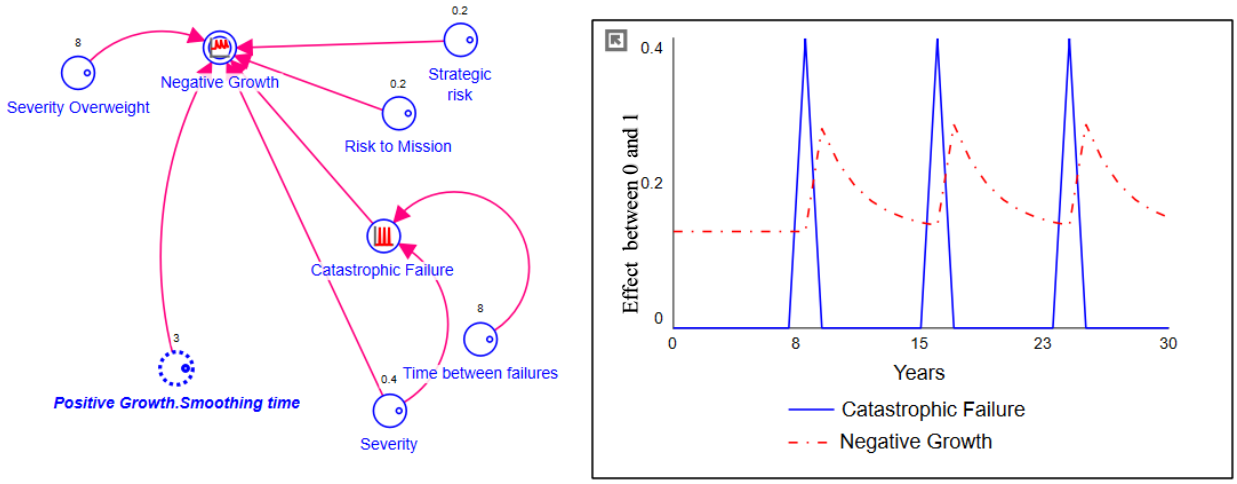


Figure 8. Negative Growth Module. Adapted from Sterman (2000).

We posit that negative growth is driven by catastrophic failures. Catastrophic failures are modeled using a pulse function as shown in Figure 8.

Table 3. Negative Growth Factors

Attribute	Equation
Negative Growth	SMTH1((Risk_to_Mission+Strategic_risk+(Severity_Overweight*Severity)*Catastrophic_Failure)/3, Positive_Growth.Smoothing_time)
Catastrophic Failure	PULSE(Severity, Time_between_failures, Time_between_failures)
Time Between Failures	User input in years
Severity	User input between 0 and 1
Severity Overweight	User input
Strategic Risk	User input between 0 and 1
Risk to Mission	User input between 0 and 1

Negative growth is a function equal to

$$\frac{SMTH1(Risk\ to\ Mission+Strategic\ Risk+Severity*Severity\ Overweight*Catastrophic\ Failure)}{3}$$

smoothed over three years. Though most factors are normalized to be between 0 and 1, negative growth is an exception. Effects of catastrophic failures depend on their severity. We measure Severity as a user input between 0 and 1 with sample values given in Table 4. While severity measures can be determined from past experience, they require calibration for the model to be accurate.

Table 4. Severity Index for Catastrophic Events

Event	Severity
Loss of control of USV	.1
Destruction of USV	.3
Destruction of unintended targets	.5
Loss of foreign civilians	.7
Loss of American military	.9
Loss of American civilians	1

In Figure 9, the Negative Growth Module, Run 547 uses Risk to Mission, Strategic Risk, Catastrophic Failures, and Time between Failures set to 1. This results in a high impact failure happening once a year and makes the negative growth rate reach a high of 3.3. Run 546 sets the values to .3 and 8 years between failures and reached a high of .28, which is more in line with reality. Additionally, as shown in Figure 9, the severity overweight factor causes a small increase in the severity of a catastrophic failure to result in a large increase in the negative growth.

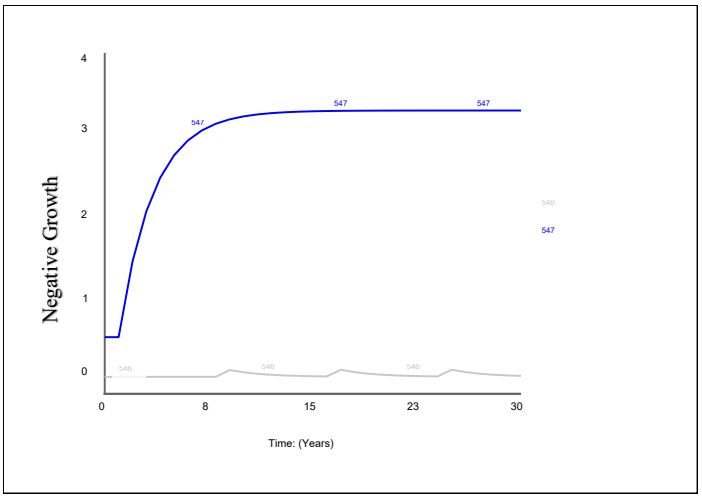


Figure 9. Severity Overweight Effect on Negative Growth

Time between failures is a user input corresponding to the years between catastrophic failures. Strategic Risk and Risk to Mission are user inputs between 0 and 1 as defined in Chapter III. Strategic Risk and Risk to Mission are two user-inputted variables that depend on the specific USV mission. A surveillance mission with very little risk to the American people or to the geographical combatant commander's mission will likely have low scores in both. On the other hand, a USV strike on a peer competitor could yield very high values. As Strategic Risk and Risk to Mission increase, we posit that adoption will decrease as trust will need to be very high to turn these missions over to autonomous control.

5. Positive Growth Module

We next examine the positive growth module as shown in Figure 10 and Table 5. Positive growth is the smoothed average of seven factors: Policy Effect, War, Risk to Force, Time Sensitivity, Understandability, Robustness, and Competition.

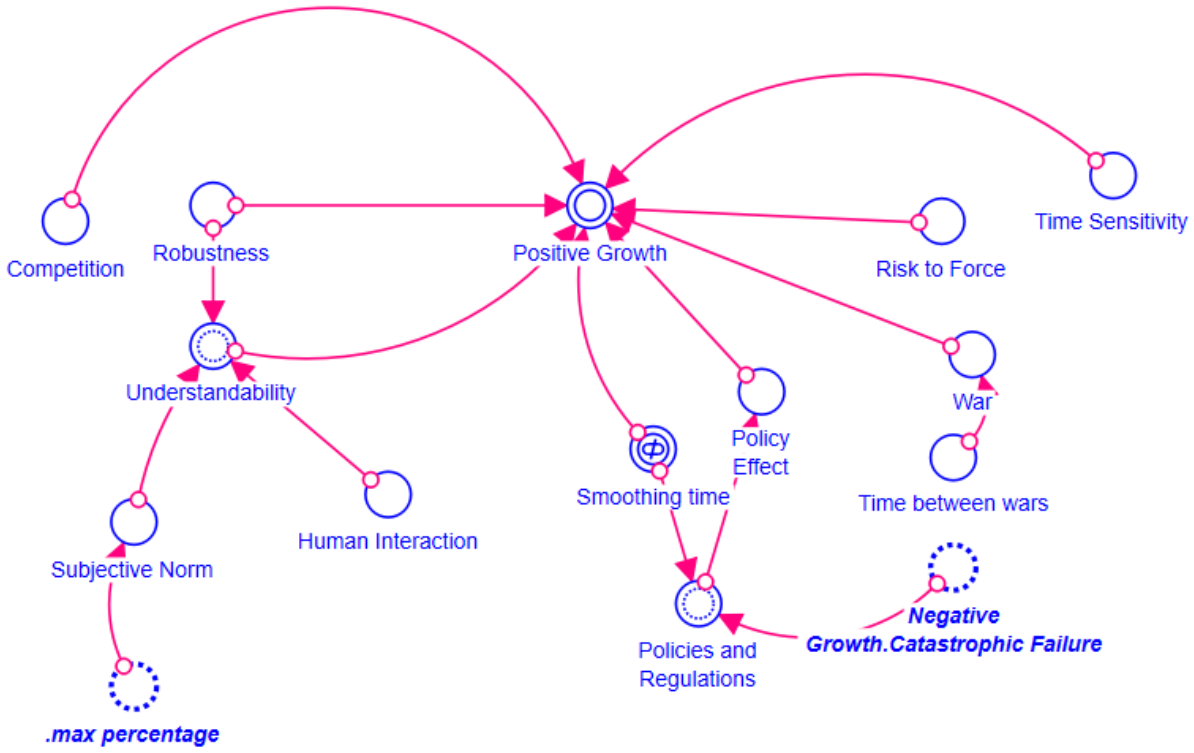


Figure 10. Positive Growth Module. Adapted from Sterman (2000).

Table 5. Positive Growth Factors

Attribute	Equation
Positive Growth	(SMTH1(Policy_Effect+War+Risk_to_Force+Time_Sensitivity+Understandability+Robustness+Competition, Smoothing_time))/7
Policies_and_Regulations	Smoothing_time*SMTH1(Negative_Growth.Catastrophic_Failure, Smoothing_time)
Policy_Effect	IF Policies_and_Regulations >1 THEN 1-Policies_and_Regulations ELSE Policies_and_Regulations
War	.5+SINWAVE(.3, Time_between_wars)
Understandability	IF Subjective_Norm>Robustness THEN (Subjective_Norm+Human_Interaction)/2 ELSE (Human_Interaction+Subjective_Norm-(Robustness-Subjective_Norm))/3
Subjective_Norm	.max_percentage/100-(.max_percentage/100)^((TIME+1))
Competition	User input between 0 and 1
Risk to Force	User input between 0 and 1
Robustness	User input between 0 and 1

Attribute	Equation
Time_Sensitivity	User input between 0 and 1
Time_between_wars	User input between 0 and 1
Smoothing_time	User input between 0 and 1
Human_Interaction	User input between 0 and 1

Our initial model had policies and regulations as a negative driving force with regards to adoption of autonomous systems. Upon further refinement, we posit that policies and regulations help to further technology within DoD. We believe that, although the initial catastrophic event will curb the adoption of new technology, the policies and regulations produced will provide DoD with a framework to build autonomous vehicles, thereby increasing adoption, illustrated in Figure 11. Policies and regulations can go above 1 if the severity and frequency of catastrophic events is high enough. This is dealt with using the policy effect variable.

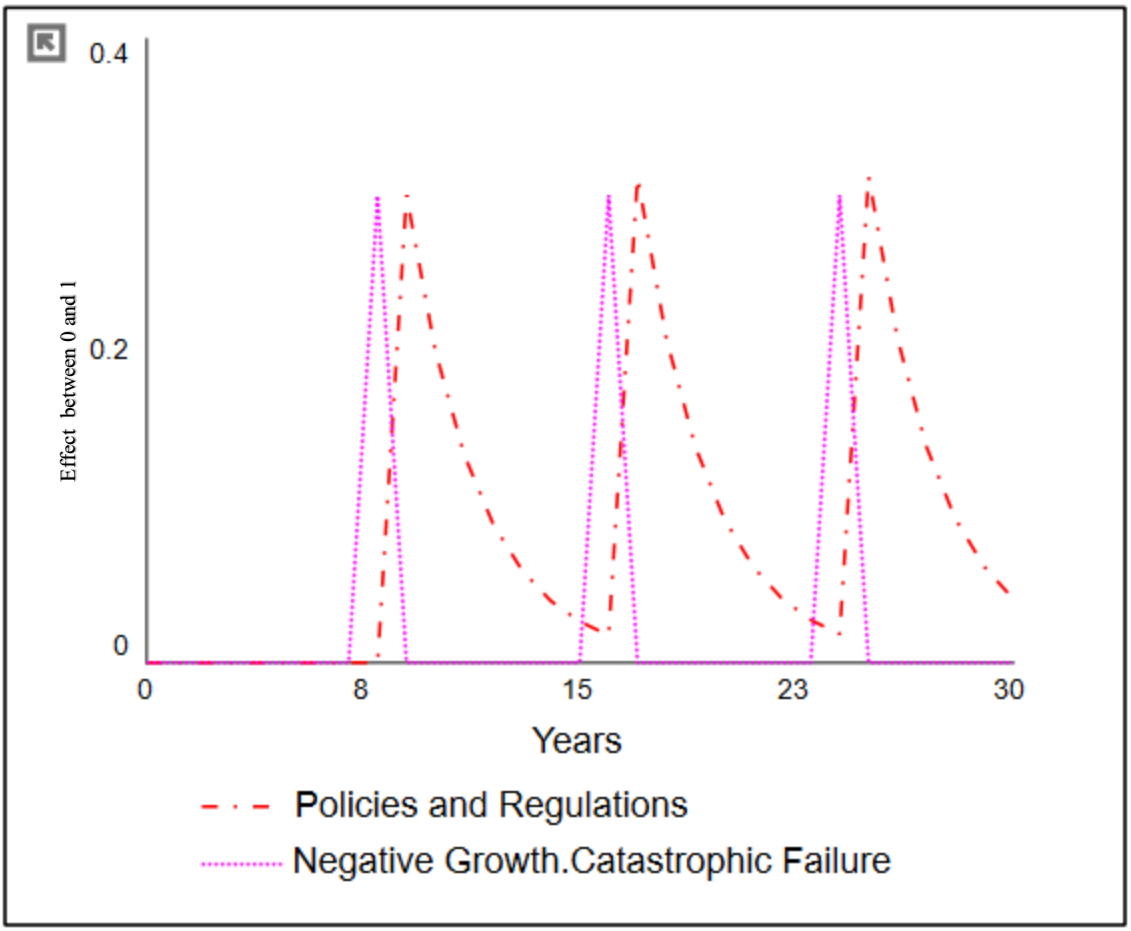


Figure 11. Policies and Regulation Results

As defined in Chapter III, the Policy Effect variable equals policies and regulations as long as they are less than 1. When policies and regulations go above 1, the equation changes to (1 – policies and regulations) as shown in Figures 12 and 13. This models the negative effect of having too many policies and regulations as described in Chapter III.

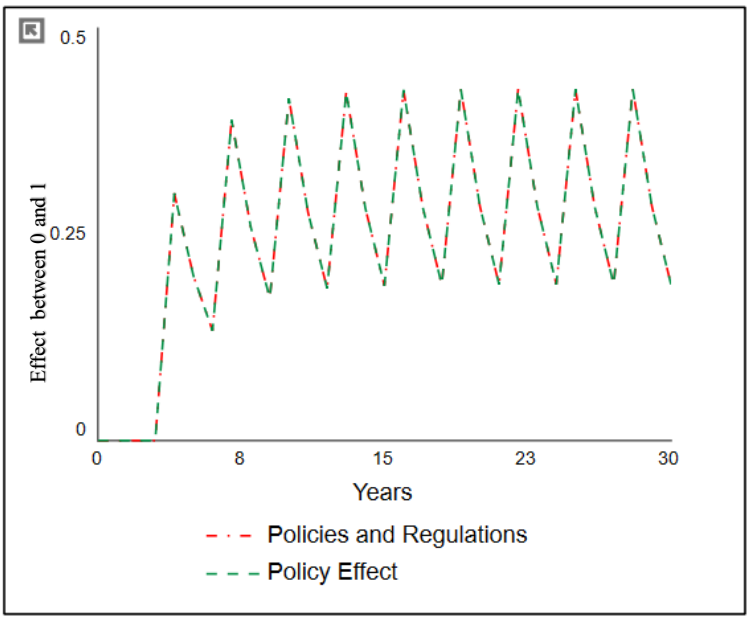


Figure 12. Policy Effect with Policies and Regulations < 1

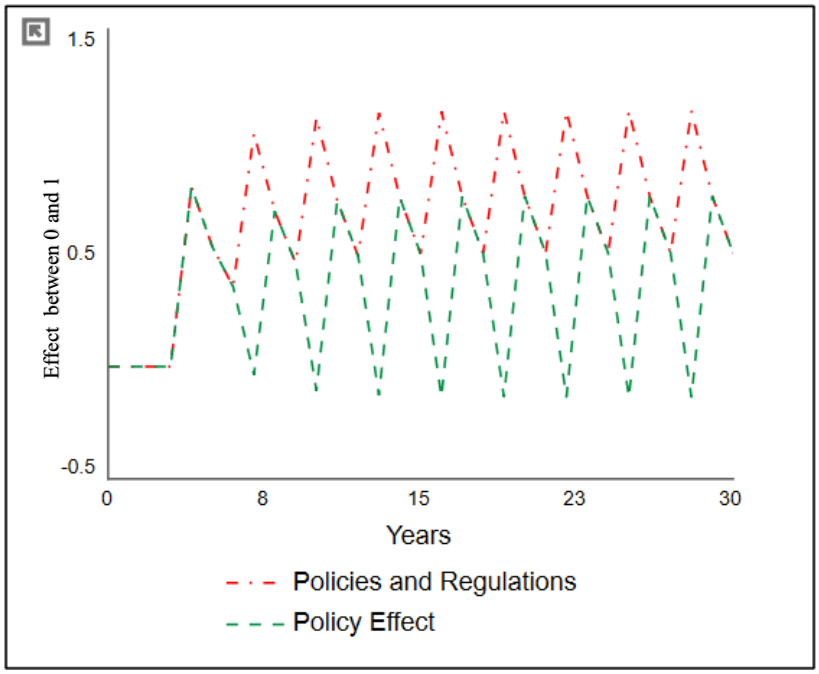


Figure 13. Policy Effect with Policies and Regulations > 1

Robustness is a measure of the system's perceived competence. It includes Reliability, False-alarm Rate, Adaptive/learning Capabilities, and an autonomous USV's ability to operate in adversarial, dynamic, uncertain, and unstructured environments while remaining self-governed and self-directed.

War is modeled using a sine wave alternating between 20% (.2) and 80% (.8) as shown in Figure 14. This represents a spectrum where the U.S. military is never fully withdrawn but also never fully deployed in unrestricted warfare. Input to this function is time between wars, which defines the time between the function's peaks in years.

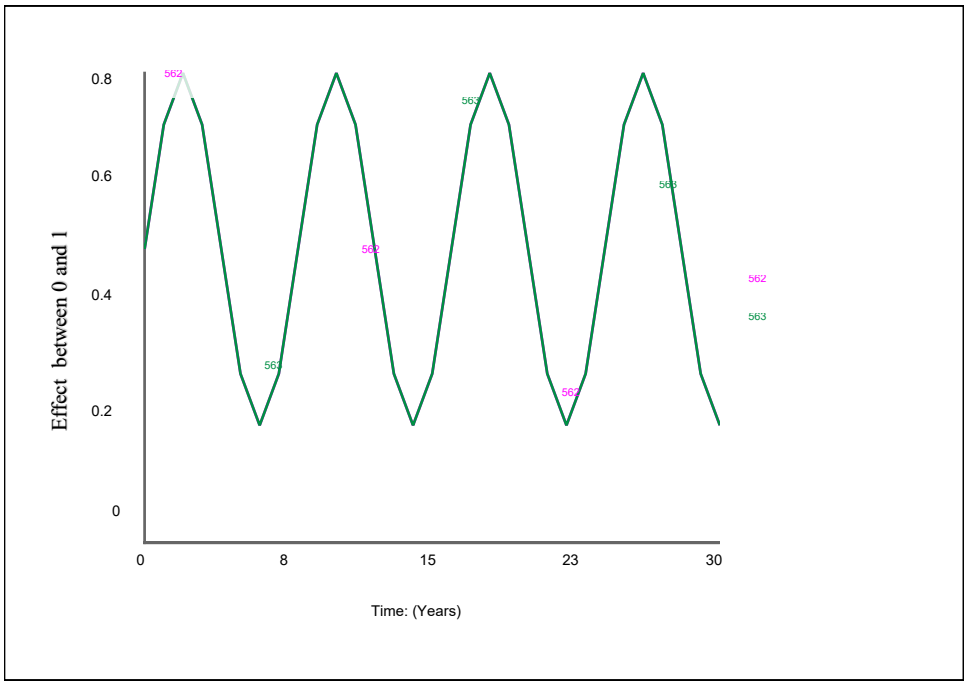


Figure 14. War Model

We posit that Subjective Norms will increase with regard to the acceptance of autonomous systems, as shown in Figure 15, eventually reaching the same maximum rate as is acceptable within the military. The difference here is that the subjective norm may proceed at a different rate for civilians versus the military. Military members will be influenced by the subjective norms of both groups because, at work, they use one standard while, in their personal lives, they use another. Take, for instance, the rise of social media. From a military perspective, sharing unnecessary

information is frowned upon. In contrast, in a military member's personal life, the sharing of personal information on social media has become quite acceptable. We posit that this disconnect led to increased comfort with military-only social media, such as Milsuite, in which people are willing to share information that is viewable to the entire DoD. We believe a similar crossover may occur in autonomous systems. As people become more comfortable with autonomous systems making higher risk decisions, such as autonomous driving, military members will become more comfortable using them in the field.

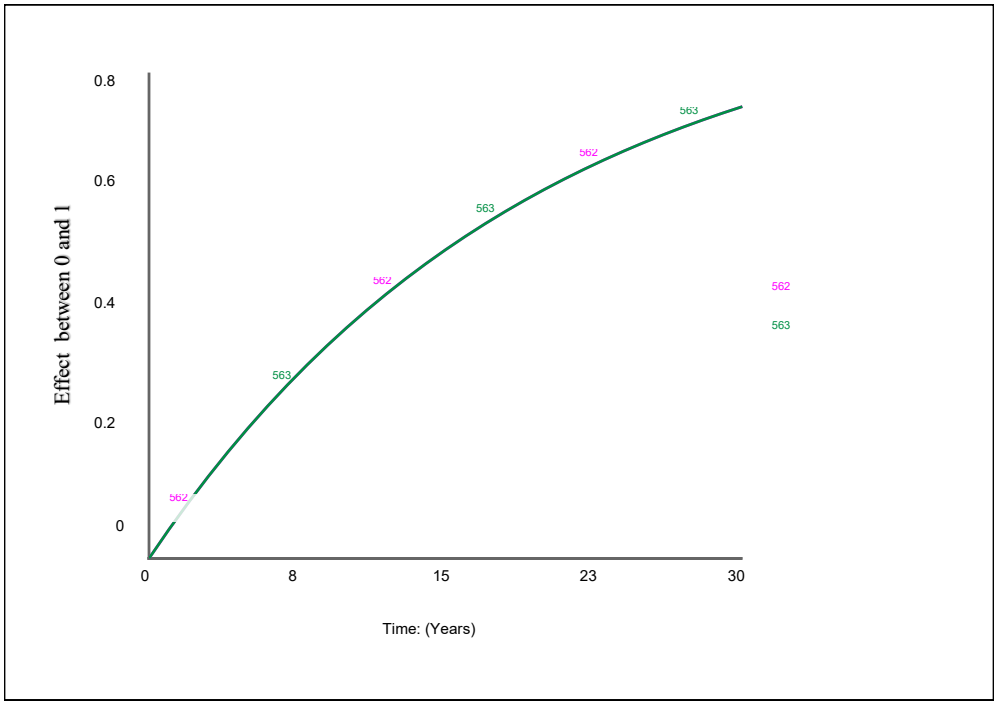


Figure 15. Subjective Norms Model

The Understandability component is made of three subcomponents: Subjective Norms, Human interaction, and Robustness. When the robustness of the system is greater than the subjective norm, we predict it will be less understandable. This model is defined by

$$\frac{Human\ interaction + Subjective\ Norm - (Robustness - Subjective\ Norm)}{3}$$

which creates a penalty, Robustness minus Subjective norm, for the robustness being greater than the subjective norm. When subjective norm overtakes the robustness of the autonomous system, the equation

$$\frac{Human\ interaction + Subjective\ Norm}{2}$$

is used, causing the jump in understandability shown in Figure 16. We posit that, as the subjective norms regarding use of autonomous systems increase, trust in autonomous systems in military settings will increase. This models the effect that Ho, Ocasio-Velázquez, and Booth described in 2017, in which familiarity with a system causes a sense of security.

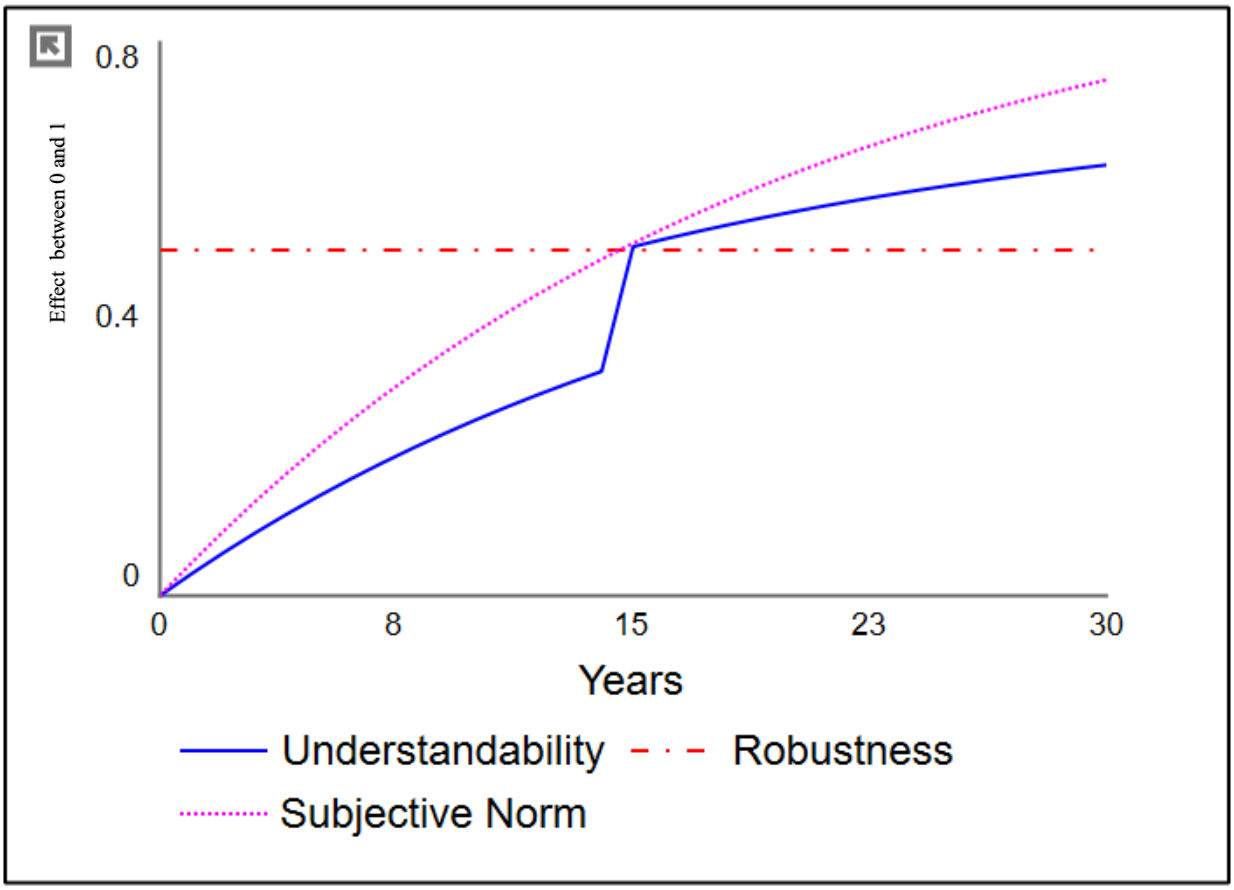


Figure 16. Understandability Model

The Time Sensitivity, Human Interaction, Competition, and Risk to Force factors, defined in Chapter III, are user inputs between 0 and 1 while Time between Wars is the time between the heights of war in years.

6. Normal Adoption

Figure 17 shows the output of the model showing normal adoption.

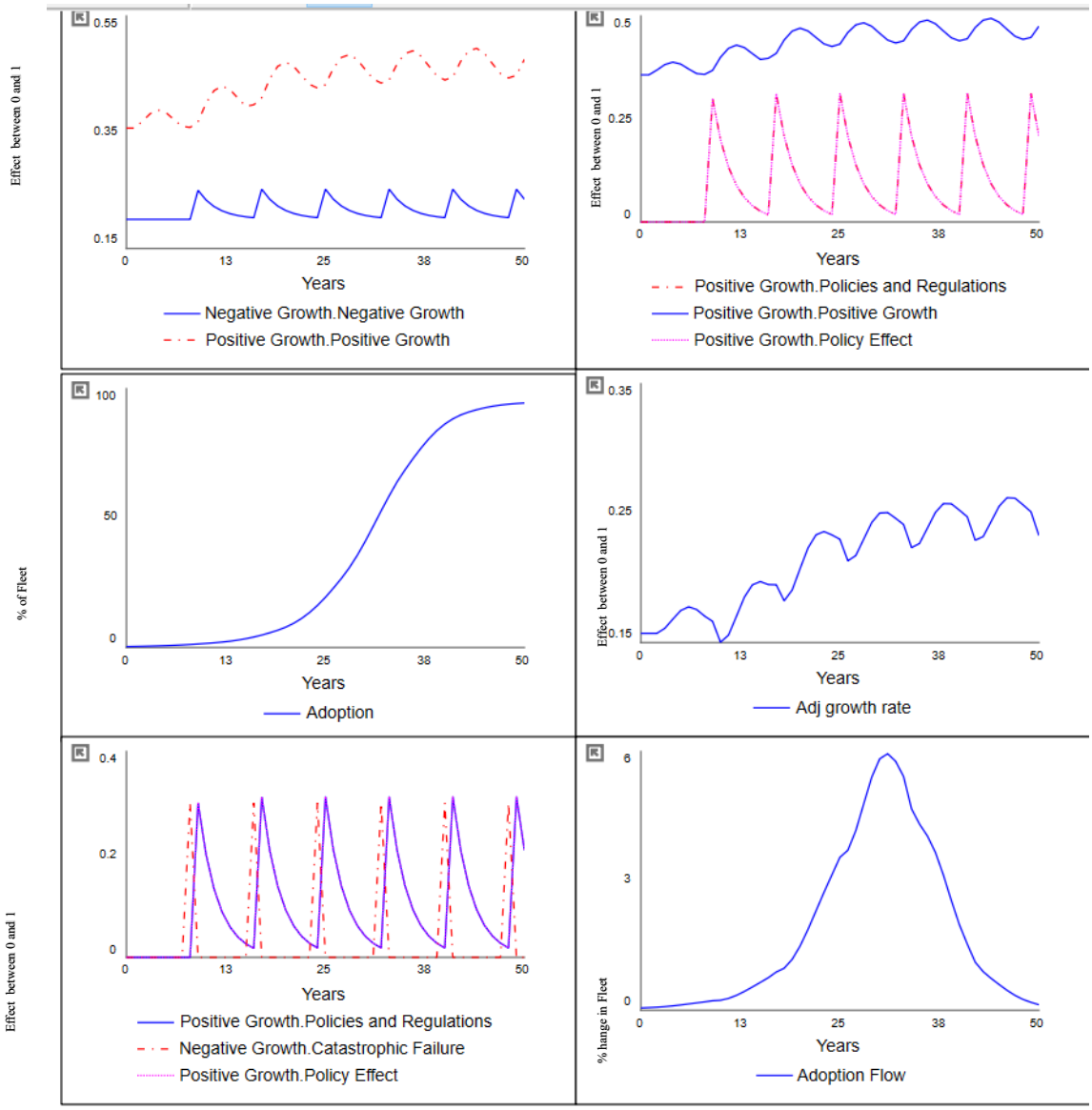


Figure 17. Normal Adoption

The top left graph shows positive growth rates are consistently above negative growth rate. The upper right graph shows that policies and regulations are consistently positive, therefore adding to positive growth. The central left graph shows the overall adoption with adoption reaching over 94% by year 50. The central right graph shows the adjusted growth rate, which is trending upward primarily due to increases in the subjective norm. The bottom left graph shows instances of catastrophic failures and how they drive policies and regulations and the policy effect. Finally, the bottom right graph shows overall adoption flow. It initially starts slow, reaching a peak as adoption spreads quickly, and then reverses approaching 0 as market saturation occurs.

7. Adoption Collapse

Figure 18 shows a model of adoption collapse. In this case, negative growth rate is consistently above the positive growth rate as shown by the upper left graph. The upper right graph demonstrate how policies and regulations can slow the adoption of new technology. Due to the frequency and severity of catastrophic failures, policies and regulations go above 1. This causes the policy effect to go negative, which, in turn, produces positive growth. The central left graph shows adoption initially increasing until the first catastrophic. The center right graph shows the adjusted growth rate initially positive, but going negative after the first catastrophic failure and remaining there. The bottom left graph shows the effect of the catastrophic failure of both policy and regulations and the policy effect. The bottom right graph shows the adoption flow exhibiting similar characteristics as the adjusted growth rate.

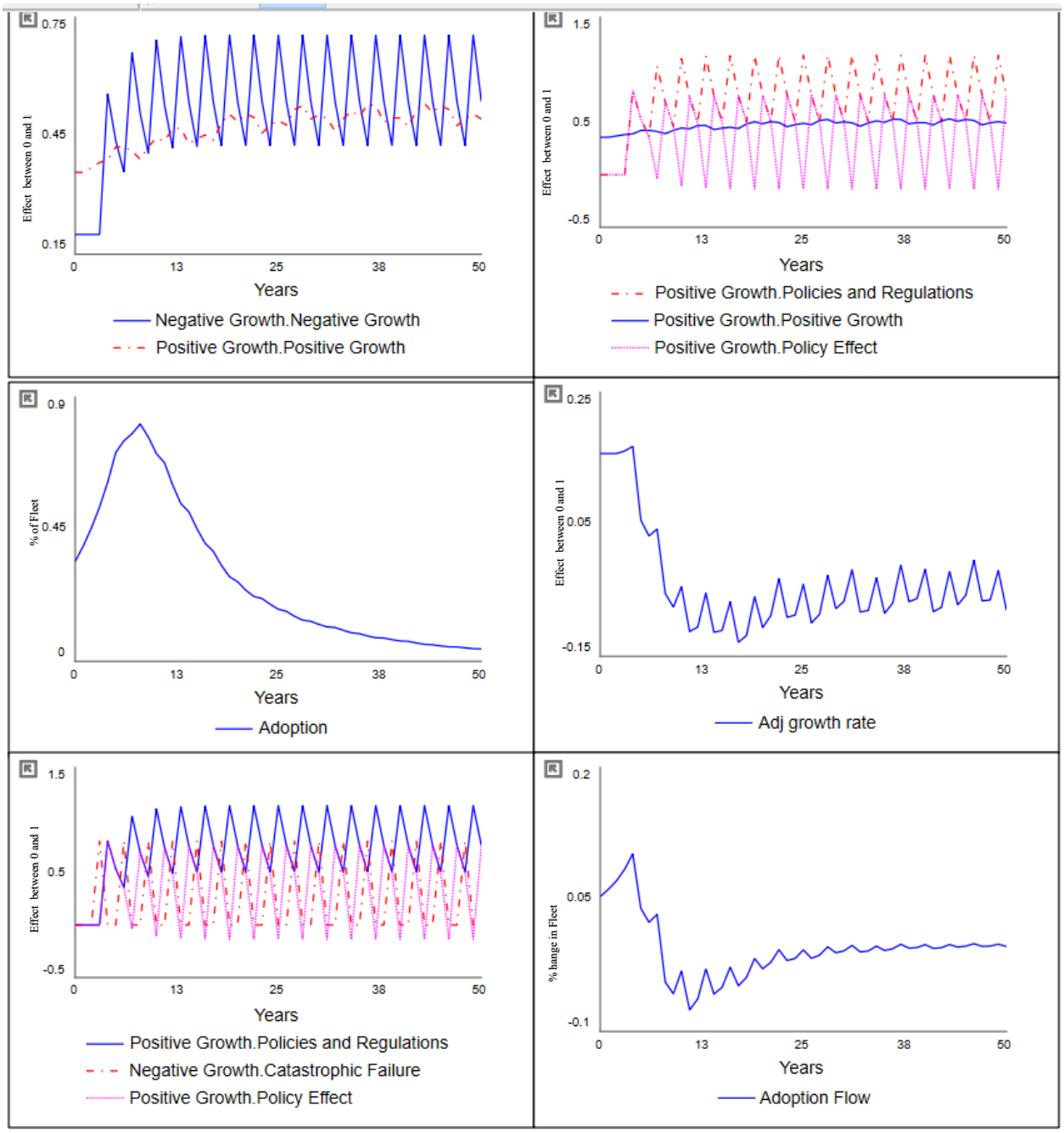


Figure 18. Adoption Collapse

8. Collapse and Rebound of Adoption

Figure 19 shows a collapse and rebound of adoption. Initially, the negative growth is higher than the positive growth rate as shown in the upper right graph. The policy effect remains positive adding to the positive growth rate after the initial catastrophic event. The positive growth increases

over time mainly driven by the increase in the subjective norm, which drives understandability, finally resulting in an increase in positive growth.

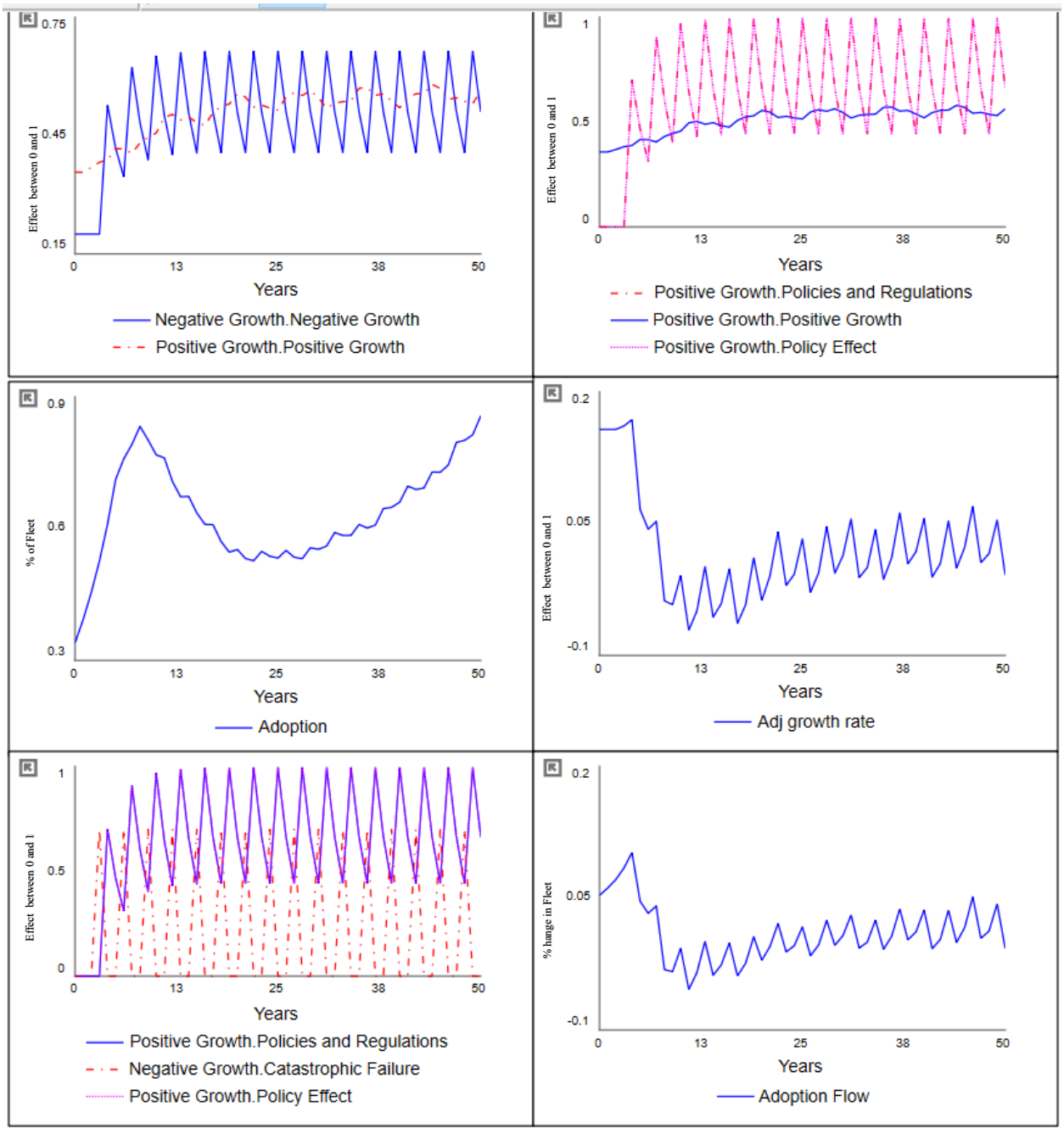


Figure 19. Collapse and Rebound

C. SUMMARY OF KEY FINDINGS

We derived two key findings from analyzing the results of our trust model. First, we were able to quantify the trust relationship between human and machine and to show how it leads to the increased adoption of autonomous USVs as shown in Figure 5. When trust is built-in from the beginning stage of the design and acquisition process, adoption of autonomous USV is likely to reach the desired state or at least very close. In order to build trust into the design and acquisition process, we defined trust by various attributes. We then assigned values to the attributes and plugged them into the trust model. The weight of the values assigned to each variable or characteristics should depend on the impact it had on the overall results of the model. Second, we were able to conclude that there is a relationship between understandability and adoption. Understandability is when an autonomous USV is able to convey its actions to its human counterparts via some kind of feedback mechanism, i.e., Common Operating Picture (COP) reporting. These reports need to be understood by the operators so that the operators thereby trust the autonomous USV to continue to perform its mission as tasked, rather than feeling the need to intervene due to lack of trust. Understandability has been built-in as part of our trust model, and then assigned a value like all other attributes and characteristics of trust to quantify how trust correlates with the adoption of autonomous USVs.

D. HOW RESULTS FILL KNOWLEDGE GAPS

Based on our quantitative trust model, we anticipate that trust can be better defined as either Positive Trust or Negative Trust. By assigning a value to trust, it makes trust measurable and comparable to other factors. From our analysis of the relationship between trust and the adoption of USVs, based on the mathematical calculations from our trust model, we conclude that trust plays a significant role. We anticipate that our trust model can be applied to future autonomous systems to mitigate the trust issue and increase human-machine teaming.

E. IMPLICATIONS FOR RESEARCH AND PRACTICE

Trust has been quantified into our trust model, and the result shows a significant increase in the adoption of autonomous USV during a time of war. However, a change in variable of one or more of the attributes of trust can change the outcome. For example, a significant increase in Failure Rate can lead to a decline in the adoption of autonomous USVs. The best way to fully

validate our results would be for DoD to test a trust model in the design and acquisition phase process for autonomous systems.

V. CONCLUSION AND RECOMMENDATIONS

This thesis showed that it is feasible to model trust into the Department of Defense (DoD) design and acquisition process in order to increase the adoption of autonomous unmanned surface vehicles (USV). To build a quantitative trust model, we researched and presented a greater understanding of various attributes and characteristics of trust. Understanding these attributes and characteristics was crucial to our research because it helped narrow down the critical trust factors to be included in our trust model. The use of quantitative trust models was paramount in the evaluation of the relationship between trust and the adoption of autonomous USVs.

After analyzing the graphical results from our trust models, we conclude that multiple factors go into a sharp increase or decrease in the future rate of USV adoption. Many variables that create trust can also have secondary effects that slow adoption. For example, competition surpassing our abilities in the USV realm can lead to increased adoption of USVs, which may potentially lead to more catastrophic failures, which, in turn, may decrease trust and dampen the adoption rate. Though we modeled the adoption of autonomous USVs with an S-curve, we also showed certain scenarios in which the curve could collapse based on large influxes of Negative Trust. With catastrophic failures set to 1 every 8 years, we saw that the model approached the maximum adoption within 50 years. Setting parameters to simulate catastrophic failures at yearly intervals resulted a collapse in trust, leading to a significantly longer adoption period with the adoption rate staying near 0 over the next 50 years. We modeled how subjective norms will progress and their effect on the adoption of USVs. Additionally, we modeled polices and regulations to initially have a dampening effect on adoption, but to eventually lead to greater trust and adoption as they create legal and procedural frameworks on which to build.

A. KEY CONTRIBUTIONS TO NEW KNOWLEDGE AND PRACTICE

This thesis provides a framework for building trust interfaces and understandability into the design and acquisition process. A quantitative approach to how trust relates to the adoption of autonomous USV brings trust from an abstract view to something measurable and comparable. We also furthered research, filling in a gap, by expanding on previous qualitative and quantitative

approaches to trust to specifically model how to build trust into the design and acquisition phase for autonomous USVs so as to examine trust's impact on the adoption rate.

B. KEY LIMITATIONS TO STUDY

One of the biggest challenges we faced during the experimentation of our trust model was how to properly assign values to trust attributes and characteristics and how to determine what makes one attribute or characteristic of trust weigh more than another. Trust, as most know it, is an immeasurable abstract. To bridge this challenge, we researched what others have done to quantify trust and attempted several mathematical calculations using our trust model. In short, while trust and its quantification remains subjective, our model offers a framework for researchers or developers to determine how they weigh each attribute and characteristic of trust.

C. RECOMMENDATIONS FOR AREAS OF FUTURE STUDY

For future work, we recommend that trust attributes and characteristics be implemented as part of the DoD design and acquisition process. Implementing trust factors into the testing process will not only allow for a measurable outcome as shown in this thesis, but it will also allow for predictability of how future systems may be adopted based on the given trust level. Further research could focus on three areas to refine the model.

1. The Connection between Subjective Norms and Adoption

Subjective norms can be tested by addressing them in two areas: subjective norms from professional and from personal life. For professional life, we propose a test during a war game scenario in which participants must pick their fleet assets based on the mission. They should be allowed to choose manned and unmanned assets with a limitation on the number of resources they can choose. For this example, assume that each participant can choose 50 fleet assets. They could then be shown a hypothetical breakdown of the choices of their peers and superiors. If, for example, the participant picked no unmanned assets, but the peer group selected an average 20% of the fleet to be unmanned, would that change their choices for the next round? Testing during war gaming may help to determine when the influence by the peer group changes the participant's subjective norm, resulting in a change in trust. A second test could compare subjective norms in civilian life versus hypothetical use in a battlefield scenario. The data gathered could determine

whether a correlation exists with market penetration of a technology and adoption in the battlefield. This can determine the "tipping point" where market penetration and familiarity with a technology result in increased battlefield use. Example questions could include the following.

a. *Would you use commercial language translation software when an interpreter is not available?*

b. *Would you use commercial agricultural drones for ISR when other assets are unavailable?*

c. *Would you use autonomous robots to accompany ground troops to carry supplies (similar to Boston Scientific robots)?*

d. *Would you use autonomous ground vehicles to deliver supplies between bases?*

e. *Would you use autonomous ground vehicles to shuttle troops between bases?*

f. *Would you use autonomous air vehicles to deliver supplies?*

g. *Would you use autonomous air vehicles to deliver troops?*

h. *Would you use a small unmanned vehicle (human controlled RC car) to deliver munitions?*

i. *Would you use a small unmanned autonomous vehicle (software controlled RC car) to deliver munitions?*

2. The Connection between Catastrophic Failures and Adoption

During a war game, future researchers could allow for the use of autonomous vehicles. Researchers could program in a failure rate and gather data on the time delay between a failure and the re-use of that technology. The data could be used as an analogy for re-establishing trust in a technology after a failure. This could also illustrate the organizational memory of failures as shown in Figure 20 and 21. In Figure 20, the catastrophic failures effect goes to 0 and has a negligible effect on negative growth as long as the combination of the severity and time between failures is low. In Figure 21, the effect of catastrophic failures never goes to 0 and continues having a negative growth because the combinations of severity and time between failures is relatively high.

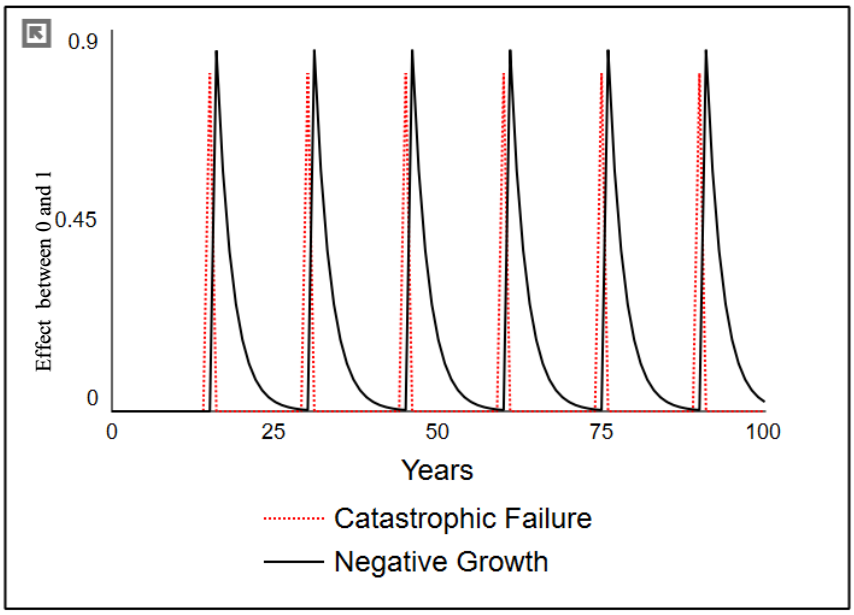


Figure 20. Catastrophic Failure effect on Negative Growth Returning to 0

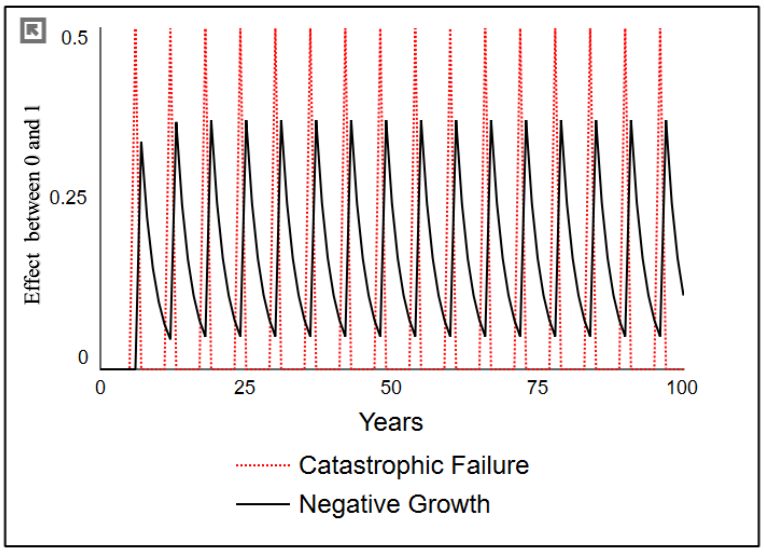


Figure 21. Catastrophic Failure with Organizational Memory

3. The Connection between Policies and Regulations and Adoption

During a war game, future researchers could initially allow use of the autonomous vehicles without any policies or regulations. Later, researchers could introduce basic guidelines from employment to see if adoption increases with a basic framework for use. Finally, researchers could increase the regulations to see if a complex framework stifles use. Figure 22 illustrates our model output for this scenario. Initially, the policies and regulations add positively to growth. After the quantity of policies and regulations reach a critical point, they subtract from positive growth.

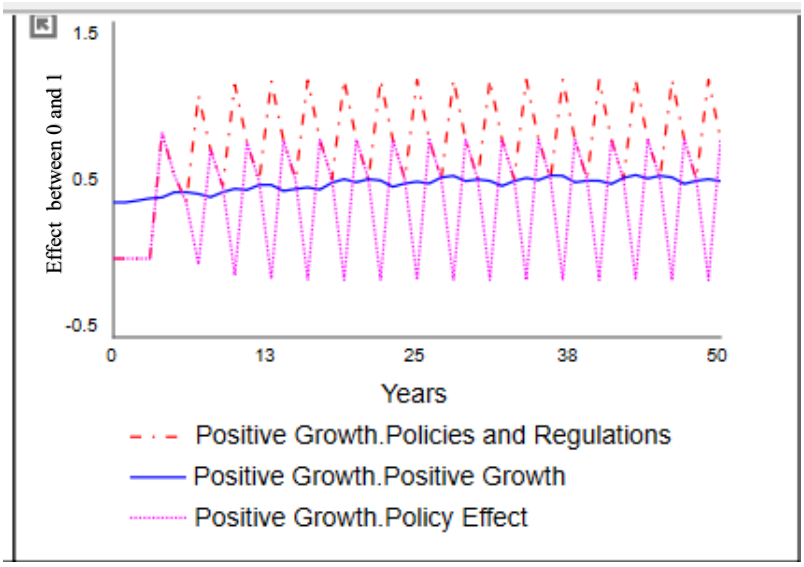


Figure 22. Policy Effect Going Negative due to Excessive Policies and Regulations

D. CONCLUSION

Quantifying trust may help military members understand the benefits of autonomous USVs as well as limitations rather than suffering fears or excessive expectations. By lessening the abstract nature of trust, we may be able to positively impact the human-machine relationship. We recommend using the terms Positive Trust and Negative Trust as a basic way to start quantifying trust as relates to DoD adoption of autonomous USVs, and we highly recommend implementing trust modeling in the DoD design and acquisitions phase.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Adjerid, I., Acquisti, A., Telang, R., Padman, R., & Adler-Milstein, J. (2015). The impact of privacy regulation and technology incentives: The case of health information exchanges. *Management Science*, *62*(4), 1042-1063.

Atkinson, D. J. (2015). Emerging cyber-security issues of autonomy and the psychopathology of intelligent machines. *Association for the Advanced of Artificial Intelligence,* 6–13.

BBC News. (2019). Boeing 737 Max: What went wrong? Retrieved from https://www.bbc.com/news/world-africa-47553174

Better Explained Discussions. (2017). An intuitive (and short) explanation of Bayes' Theorem. *Communications Standard Dictionary*. Retrieved from https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/

Brown, J. (2019). Tesla autopilot malfunction caused crash that killed Apple engineer, lawsuit alleges. Gizmodo. Retrieved from https://gizmodo.com/tesla-autopilot-malfunction-caused-crash-that-killed-ap-1834453661

Bowden, M. (2013, November). How the predator drone changed the character of war. *Smithsonian Magazine.* Retrieved from https://www.smithsonianmag.com/history/how-the-predator-drone-changed-the-character-of-war-3794671/

Byman, D. L. (2013). Why drones work: The case for Washington's weapon of choice. Brookings. Retrieved from https://www.brookings.edu/articles/why-drones-work-the-case-for-washingtons-weapon-of-choice/

Defense Science Board. (2012). Task force report: The role of autonomy in DoD systems. *Office Of The Under Secretary Of Defense For Acquisition, Technology and Logistics.* Retrieved from https://fas.org/irp/agency/dod/dsb/autonomy.pdf

Federal Aviation Administration. (2019). Operators of Boeing Company Model 73 7-8 and Boeing Company Model 73 7-9 Airplanes. Washington, D.C: U.S. Government Printing Office. Retrieved from https://www.govinfo.gov/content/pkg/FR-2019-03-18/pdf/2019-05067.pdf

Hall, B. H., & Khan, B. (2003). Adoption of new Technology (No. w9730). National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w9730

Hannemyr, G. (2003). The Internet as hyperbole: A critical examination of adoption rates. *The Information Society*, *19*(2), 111-121.

Hartig, L., & Vanhoose, K. (2019). Solving one of the hardest problems of military AI: Trust. *Defense One, 4*(1), 1-3. Retrieved from https://www.defenseone.com/ideas/2019/04/solving-one-hardest-problems-military-ai-trust/155959/?fbclid=IwAR2olPo4MVSniNNvN_kcI2No3QehD1KVwzAo2rp7BNAAdfRfa8KmCG8dZss

Ho, Shuyuan Mary, Ocasio-Velázquez, Mónica, & Booth, Cheryl. Trust or Consequences? Causal Effects of Perceived Risk and Subjective Norms on Cloud Technology Adoption. *Computers & Security 70* (2017): 581–595. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167404817301591

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434. Retrieved from https://journals.sagepub.com/doi/full/10.1177/0018720814547570

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf

Hsu, J. (2015). Q&A: Why fully autonomous robot cars hail from the 20th century. *IEEE Spectrum,* Retrieved from https://spectrum.ieee.org/cars-that-think/transportation/self-driving/why-fully-autonomous-robot-cars-hail-from-the-20th-century

Johns, J. L. (1996). A concept analysis of trust. *Journal of Advanced Nursing*, 24, 76–83.

Klein, D. (2018). Unmanned systems & robotics in the FY2019 defense budget, *AUVSI,* Retrieved from https://www.auvsi.org/sites/default/files/DoD%20FY19%20Budget%20Report_FINAL%20DRAFT_WITH%20SENATE%20NDAA.pdf

LaGrone, S. (2019a). Navy wants 10-Ship unmanned 'Ghost Fleet' to supplement manned force. *USNI NEWS,* Retrieved from https://news.usni.org/2019/03/13/navy-wants-ten-ship-3b-unmanned-experimental-ghost-fleet

LaGrone, S. (2019b). NSTB. Lack of Navy oversight, training were primary causes of fetal McCain collision. *USNI NEWS,* Retrieved from https://news.usni.org/2019/08/06/ntsb-lack-of-navy-oversight-training-were-primary-causes-of-fatal-mccain-collision

Li, D., & Du, Y. (2017). *Artificial intelligence with uncertainty*. Boca Raton, CRC Press. Retrieved https://www.taylorfrancis.com/books/9781315366951

Lin, P. (2016). Killer robots: New reasons to worry about ethics. *Forbes*. Retrieved from http://onforb.es/1muUSnm

Mayville JR., W. (October 14, 2016). *Joint risk analysis* [Memorandum]. Washington, DC: Chairman of the Joint Chief of Staff. Retrieved from https://www.jcs.mil/Portals/36/Documents/Library/Manuals/CJCSM%203105.01%C2%A0.pdf?ver=2017-02-15-105309-907

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20, 709–734.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market-research relationships. *Journal of Marketing*, *57*(1), 81–101.

Palmer, G., Selwyn, A., & Zwillinger, D. (2016). The "Trust V": Building and measuring trust in autonomous systems. *Robust Intelligence and Trust in Autonomous Systems*, 55-77. https://doi.org/10.1007/978-1-4899-7668-0_4

Rice, S., & Keller, D. (2009). Automation reliance under time pressure. *Cognitive Technology*, *14*(1), 36-39.

Ross, A. (2013). Cyber and drone attacks may change warfare more than the machine gun. *The Atlantic*. Retrieved from https://www.theatlantic.com/technology/archive/2012/03/cyber-and-drone-attacks-may-change-warfare-more-than-the-machine-gun/254540/

Schaefer, K. E., Chen J. 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

Scharre, P. (2016). Centaur Warfighting: The false choice of humans vs. automation. *Temple Int'l & Comp. L.J.,* 30(1), 153–159

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. New York, NY: Shelstad.

Taylor G., Mittu R., Sibley C., Coyne J. (2016). Introduction. In: Mittu R., Sofge D., Wagner A., Lawless W. (Eds.), *Towards modeling the behavior of autonomous systems and humans for trusted operations* (pp. 78-79). Boston, MA: Springer

U.S. Navy. (n.d.). U.S. Navy demographic data. Navy.mil Home Page. Retrieved May 15, 2019, from https://www.navy.mil/navydata/nav_legacy.asp?id=146

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