AUTONOMOUS UNMANNED SURFACE VESSELS IN NAVAL WARFARE: SYSTEM SAFETY AND ETHICAL IMPLICATIONS IN CONGESTED AND LITTORAL WATERS

Ong Yi Sheng Michael; Sung Wei Pei

Monterey, CA; Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AUTONOMOUS UNMANNED SURFACE VESSELS IN NAVAL WARFARE: SYSTEM SAFETY AND ETHICAL IMPLICATIONS IN CONGESTED AND LITTORAL WATERS

by

Ong Yi Sheng Michael and Sung Wei Pei

March 2024

Thesis Advisors: Joshua A. Kroll
Bradley J. Strawser
Co-Advisor: Logan O. Mailloux

THIS PAGE INTENTIONALLY LEFT BLANK

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

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2024	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE AUTONOMOUS UNMANNED SURFACE VESSELS IN NAVAL WARFARE: SYSTEM SAFETY AND ETHICAL IMPLICATIONS IN CONGESTED AND LITTORAL WATERS	**5. FUNDING NUMBERS**
6. AUTHOR(S) Ong Yi Sheng Michael and Sung Wei Pei	
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)

The growing desire to deploy unmanned surface vessels (USVs) in naval operations is attributed to their effectiveness demonstrated in capability development trials. However, to remove human-in-the-loop in current USV operations presents significant ethical challenges, including attributing responsibility, complying with international law, operating safely, and distinguishing non-combatants in congested and littoral straits. This thesis adds to the understanding of key enablers for small and medium-sized autonomous USV deployment with a complementary system safety approach. First, we build a notional vignette based on ethical challenges and hazards identified using system theoretic process analysis (STPA). We then measure the utility and complexity of USVs conducting autonomous launch and recovery (LAR), navigation, intelligence, surveillance, and reconnaissance (ISR), and firing. Using our system-level model, our analysis finds humans must retain active control when resorting to firing, exert supervisory control during navigation and ISR, and allow zero human control in LAR operations. While human operators can be the moral agents to augment artificial intelligence (AI)'s lack of emotion and reasoning, the concept of meaningful human control is central to our recommendations. We recommend additional measures such as clarity in roles and responsibilities, along with training and certification, to prevent humans from being the moral crumple zone of AI.

14. SUBJECT TERMS artificial intelligence, AI, ethics, utilitarianism, unmanned surface vessels, USV, system safety, MIL-STD 882E, system theoretic process analysis, STPA, accountability, transparency, explainability, governance, meaning human control, ISR	**15. NUMBER OF PAGES** 147
	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.

AUTONOMOUS UNMANNED SURFACE VESSELS IN NAVAL WARFARE: SYSTEM SAFETY AND ETHICAL IMPLICATIONS IN CONGESTED AND LITTORAL WATERS

Ong Yi Sheng Michael
Military Expert 5, Republic of Singapore Navy
BSIT, Singapore University of Social Sciences, 2015

Sung Wei Pei
Major, Republic of Singapore Navy
BME, Nanyang Technological University, 2014

Submitted in partial fulfillment of the
requirements for the degrees of

MASTER OF SCIENCE IN COMPUTER SCIENCE

And

MASTER OF SCIENCE IN DEFENSE ANALYSIS (IRREGULAR WARFARE)

from the

NAVAL POSTGRADUATE SCHOOL
March 2024

Approved by: Joshua A. Kroll
Advisor

Bradley J. Strawser
Advisor

Logan O. Mailloux
Co-Advisor

Gurminder Singh
Chair, Department of Computer Science

Carter Malkasian
Chair, Department of Defense Analysis

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The growing desire to deploy unmanned surface vessels (USVs) in naval operations is attributed to their effectiveness demonstrated in capability development trials. However, to remove human-in-the-loop in current USV operations presents significant ethical challenges, including attributing responsibility, complying with international law, operating safely, and distinguishing non-combatants in congested and littoral straits. This thesis adds to the understanding of key enablers for small and medium-sized autonomous USV deployment with a complementary system safety approach. First, we build a notional vignette based on ethical challenges and hazards identified using system theoretic process analysis (STPA). We then measure the utility and complexity of USVs conducting autonomous launch and recovery (LAR), navigation, intelligence, surveillance, and reconnaissance (ISR), and firing. Using our system-level model, our analysis finds humans must retain active control when resorting to firing, exert supervisory control during navigation and ISR, and allow zero human control in LAR operations. While human operators can be the moral agents to augment artificial intelligence (AI)'s lack of emotion and reasoning, the concept of meaningful human control is central to our recommendations. We recommend additional measures such as clarity in roles and responsibilities, along with training and certification, to prevent humans from being the moral crumple zone of AI.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AGI	artificial general intelligence
AI	artificial intelligence
ASW	anti-submarine warfare
C2	command and control
C4ISR	command, control, communications, computers, intelligence, surveillance, and reconnaissance
CA	control actions
CDCA	collision detection/collision avoidance
CNO	Chief of Naval Operations
COLREG	Convention on the International Regulations for Preventing Collisions at Sea, 1972
CONOPS	concept of operations
EW	electronic warfare
FB	feedback
GNSS	Global Navigation Satellite System
HQ	headquarters
IEC	International Electrotechnical Commission
IMO	International Maritime Organization
ISIS	Islamic State of Iraq and Syria
ISO	International Organization for Standardization
ISR	intelligence, surveillance, and reconnaissance
IUU	illegal, unreported, and unregulated
JTC	Joint Technical Committee
LAR	launch and recovery
LARS	launch and recovery system
LLM	large language model
LOA	levels of autonomy

LOAC	Law of Armed Conflict
LUSV	large unmanned surface vessel
MIL-STD 882E	Military Standard 882E
MIT	Massachusetts Institute of Technology
ML	machine learning
ONR	U.S. Office of Naval Research
RAI	responsible AI
ROE	rules of engagement
ROV	remotely operated vehicle
SATCOM	satellite communication
SC	subcommittee
SLOC	sea lines of communication
SOI	system of interest
STAMP	systems-theoretic accident model and processes
STPA	system-theoretic process analysis
STPA-Sec	system-theoretic process analysis approach for security
SuW	surface warfare
UAV	unmanned aerial vehicle
UCA	unsafe control actions
UNCLOS	United Nations Convention on the Law of the Sea
U.S. DOD	United States Department of Defense
U.S. Navy SEAL	United States Navy Sea, Air, and Land
USV	unmanned surface vessel
UUV	unmanned underwater vehicle
VUCA	volatile, uncertain, complex, and ambiguous

EXECUTIVE SUMMARY

In this thesis, we examine the use of autonomous unmanned surface vessels (USVs) in congested and littoral waters. Our research is a collaborative effort between two Naval Postgraduate School departments (Defense Analysis and Computer Science), merging ethical considerations and system safety perspectives to ensure the safe and ethical deployment of autonomous USVs. Employing autonomous USVs in military operations entails both safety hazards and ethical considerations; our central research question is, how do these ethical considerations and system safety processes influence the implementation of meaningful human control in a littoral and congested maritime environment? Our work integrates and explores the relationship between ethical considerations and system safety concepts and how this relationship necessitates human oversight in autonomous USV operations. As autonomy changes the calculus of USV employment, including military ethics, and the system safety identifies risks, meaningful human control can mitigate the inherent distrust military practitioners have toward fully autonomous USVs. Our research advocates for an operating model in which human operators are not the "moral crumple zones" for AI shortcomings [4].

For our research design, we start with hypothetical scenario of an important maritime passage that serves as a testing ground for exploring the intersection of ethics and system safety within autonomous naval operations. We then delve into STPA method, traditionally used in civil industries, to systematically identify potential hazards and developing loss scenarios that incorporate ethical considerations [1]. Next, we explore the crucial aspects of accountability, transparency, and explainability that foster military practitioners' trust in AI systems [2]. Last, we propose a model determining the extent of human involvement necessary to reach ethical decisions and safe operation of autonomous USVs.

During our analysis, we identify three main categories that can result in losses (i.e., civilian and military casualties, damage to one's own USV or other vessels, and mission failure):

1. Faults related to the system—technical faults that occur independent from human control can result in the damage of the USV itself or the objects surrounding it.

2. Gray zone engagement concepts, where the blame is shifted to AI—in circumstances when direct confrontation or response is likely to trigger escalation, the USVs can be used as mere instruments to inflict intended harm that appear as accidents, allowing the nation to achieve its objectives while attributing responsibility to the AI, thereby minimizing the likelihood of escalation.

3. Pushing boundaries when unmanned systems are exploited to perform more dangerous maneuvers than manned ships—the lack of crew onboard the USVs enables commanders to test the capabilities of the USVs beyond limits that would constrain manned ships.

To address these faults, we measure the utility and complexity of USVs conducting launch and recovery (LAR); navigation; intelligence, surveillance, and reconnaissance (ISR); and firing autonomously. Our model, shown in Figure 1, found that humans should retain active control during firing, exert supervisory control during navigation and ISR, and allow zero human control in LAR operations.

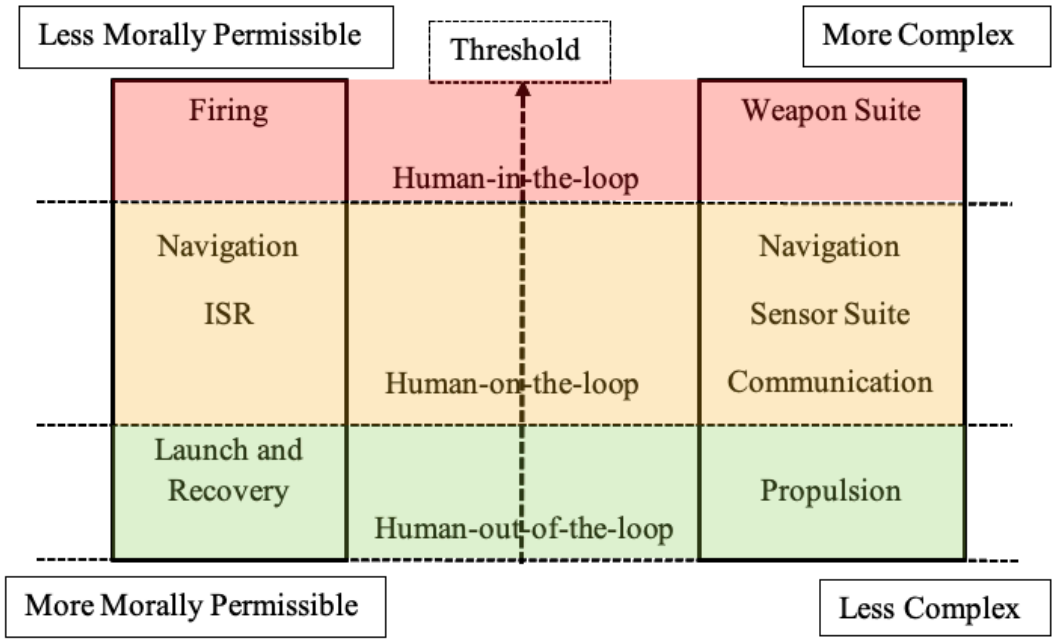

Figure 1. System-Level Model for Meaningful Human Control

This analysis goes beyond just assessing a system's capability to perform tasks safely but also emphasizes the importance of adhering to ethical boundaries. While the model provides a comprehensive engineering analysis, it primarily serves as a tool for navigating ethical dilemmas by connecting engineering decisions with moral evaluations and broader human values. From an engineering and ethical perspective rooted in utilitarianism, the model does not claim to solve these ethical challenges outright but rather offers guidance for thoughtful consideration and finding balance in designing and utilizing autonomous naval systems [3].

Human operators serve as a safety net, providing a necessary layer of active or supervisory control—human-in-the-loop or human-on-the-loop, respectively [5]. The human operator's role is crucial in filling the gaps that autonomous systems may have, providing context-based decision making and acting as a fail-safe mechanism to alter or halt operations, if necessary. The essence of this thesis lies in establishing a framework that ensures the safe and ethical operation of USVs, reflecting a blend of ethics and engineering pragmatism. By adapting STPA for the military domain and supplementing it with the current MIL-STD-882E safety processes, we pave the way for a robust approach to system

safety that accommodates ethical considerations and the nuances of human control. Ultimately, this research not only provides insights into the safe deployment of autonomous systems but also lays the groundwork for future studies to further refine the balance between autonomy and human oversight.

Recommendations

The key recommendations for ensuring the safe deployment of autonomous USVs, incorporating meaningful human control, are as follows:

1. Commanders should confine their usage of autonomous USVs based on each USV's intended design, so as to mitigate the risk of system faults caused by actions beyond operational parameters.

2. Commanders should not blame operators for all AI-related accidents. Instead, commanders should understand challenges operators face when exercising active or supervisory control of the autonomous USVs.

3. Designers should build autonomous USVs that offer options for human operators to be present onboard for active or supervisory control.

4. Commanders and operators need to have clarity in their roles and responsibilities, understand the context of the operating environment, know the acts of omission or commission and the consequences that results from the action, and be accountable for those actions taken.

5. Commanders should train and certify operators to perform active or supervisory control of autonomous USVs.

6. Programmers should ensure that autonomous USVs are programmed in a way that allows for a smooth transition of control from AI to operators, providing operators with ample time to address any AI-related faults that may arise.

References

[1] E. Jatho, "Finding and fixing fragility in machine learning," Ph.D. dissertation, Dept. of Comp. Sci., NPS, Monterey, CA, USA, 2023. Available: https://calhoun.nps.edu/handle/10945/72193

[2] R. Sparrow, "Killer robots," *J. Appl. Philos.*, vol. 24, no. 1, pp. 62–77, 2007. Available: https://www.jstor.org/stable/24355087

[3] J. Bentham, *An Introduction to the Principles of Morals and Legislation*. Oxford, England: Clarendon Press, 1907. Available: https://www.econlib.org/library/Bentham/bnthPML.html

[4] M. C. Elish, "Moral crumple zones: Cautionary tales in human-robot interaction," *Engag. Sci. Technol. Soc.*, vol. 5, pp. 40–60, Mar. 2019. Available: https://doi.org/10.17351/ests2019.260

[5] P. Scharre, "Centaur warfighting: The false choice of humans vs. automation," *Temple Int. Comp. Law J.*, vol. 30, no. 1 (Spring 2016), pp. 151–165. Available: https://sites.temple.edu/ticlj/files/2017/02/30.1.Scharre-TICLJ.pdf

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

My gratitude goes to the Chairs of the Computer Science and Defense Analysis departments, Dr. Singh and Dr. Malkasian, whose support for the collaboration allowed MAJ Sung and me to bring our thesis to fruition. I owe a debt of thanks to MAJ Sung, who has worked tirelessly alongside me, consistently bringing an exceptional level of dedication that has been fundamental to our success. I am deeply appreciative of IGPO for providing us the opportunity to brief Admiral Paparo, Commander, United States Pacific Fleet, regarding our thesis. Special gratitude is extended to my advisors. I am particularly indebted to Professor Kroll, whose insights have significantly shaped our research. Professor Mailloux deserves special recognition for my countless hours spent in his office discussing STPA and for meticulously reviewing our work, offering his expertise even post-retirement. Professor Strawser's insightful advice has significantly broadened our perspective and enriched our ideas. A special acknowledgment from my heart goes out to the love of my life, Wina Foo, whose unwavering love and support across the miles have been the bedrock of my resilience and drive. To my children, Aidan and Alyssa, your love has been my solace and strength.

—Michael Ong

I am profoundly grateful to Naval Postgraduate School's Defense Analysis Department Chairman, Dr. Carter Malkasian, and Computer Science Department Chairman, Dr. Gurminder Singh, for their support and collaboration on this academic venture. This jointness is the epitome of adopting systems thinking to address complex issues. A heartfelt appreciation to our advisors, Dr. Bradley J. Strawser and Dr. Joshua Kroll, and Lt. Col. Mailloux Logan, for their invaluable guidance throughout the thesis process. Their expertise was indispensable in refining our ideas. My appreciation extends to DA and IGPO staff for the opportunity to brief Admiral Paparo, Commander of the United States Pacific Fleet. ME5 Michael Ong, whom I only met in my second quarter, is efficient, responsible, and knowledgeable. Thank you for bringing new perspectives to

stimulate discussions and enduring the sleepless nights working hard together. Most importantly, I am immensely grateful to my wife, Wong Jie Hui. You are the core pillar of strength in our family, ensuring our son, Sung Chen Rui, and I are always mentally and physically well taken care of. Sincere appreciation to both sides of the family who give us support throughout this academic journey.

—Sung Wei Pei

I. INTRODUCTION

Unmanned operations in the maritime domain provide significant strategic and tactical advantages, especially when the risk to human life is reduced [1]. While technology is the key enabler for vessels to be unmanned, the innovative concept of unmanned vessels attacking manned ships has historical roots dating back to the battle between the Athenian and Syracusan navies. During the conflict, the Syracusan navy deliberately set an uncrewed merchant vessel on fire and let it drift toward the Athenian fleet [2]. Although the Athenians successfully evaded the threat through effective countermeasures, this historical event serves as evidence of military desires to exploit the potential of unmanned vessels for offensive operations.

In the past, various militaries experimented with unmanned technology for its advantage of enhanced efficiency at a reduced cost [1]. The development and deployment of Germany's unmanned explosive speedboats during World War I and II marked a significant advancement in naval warfare tactics, demonstrating early attempts at operationalizing remotely controlled unmanned technology for offensive purposes [1], [3]. In a similar period, the U.S. Navy also experimented with unmanned surface vessels (USVs) to remove operational risk toward human operators in hazardous missions such as sea-mine clearance.

In the present day, the Russia-Ukraine conflict highlights the strategic utility of remotely controlled USVs in modern naval warfare [4]. Various reports and video footage suggest the Russian Black Sea flagship vessel Admiral Makarov was attacked by agile and explosive-laden USVs [5], [6], [7]. Although the attack on the frigate at the Sevastopol naval base did not cause significant damage to the vessel itself, it has a notable impact on the morale of the Russia Navy and prompted additional security measures [8]. The ability of the USVs to penetrate the protected harbor demonstrated the valuable role USVs have in shaping the behaviors of major powers on the battlefield.

There is an inherent distrust by scholars and military practitioners when autonomous systems make and execute decisions on their own, however, particularly when

the actions impose unnecessary cost to non-combatants. As maritime military platforms become more autonomous, critics such as Robert Sparrow in his article "Killer Robots" claim that the attribution of responsibility becomes increasingly difficult, thus contributing to the public's doubts about autonomous systems behaving ethically [9]. In 2013, a non-governmental group produced "Killer Robots: UK Government Policy on Fully Autonomous Weapons" to encourage United Kingdom policymakers to thoroughly examine the use of fully autonomous weapons [10]. Aaron Johnson and Sidney Axinn expressed their unease with autonomous systems killing humans, stating that the decision to take a human life should be essentially human [11]. Military practitioners share concerns about unmanned aerial vehicles (UAVs) that are applicable to USVs, including whether drones make conflict more possible and whether deadly force employed by autonomous systems is ethically and legally justified [12].

The limited trust of scholars and military practitioners in fully autonomous USVs hampers these USVs' applicability during gray zone operations, especially in congested and littoral waters when a challenge exists in distinguishing between combatants and non-combatants. Allowing these autonomous USVs to operate without humans involved in the decision-making processes and their autonomous weapon systems to identify, select, and fire targets raises concerns [13]. Employing autonomous USVs in military operations entails both safety hazards and ethical considerations. Thus, the question remains: how do these ethical considerations and system safety processes influence the implementation of meaningful human control in a littoral and congested maritime environment? This research project undertakes a comprehensive evaluation of small and medium-sized autonomous USVs in naval warfare, considering both operational and ethical considerations, as well as safety at a policy and system level. Once these important considerations have been sufficiently discussed, we will analyze when and to what extent USVs should operate autonomously. We propose a model mapping operational tasks to levels of meaningful human control. As autonomy changes the calculus of USV employment, including military ethics and the system safety approaches at identifying risks, meaningful human control can mitigate the inherent distrust military practitioners have of fully autonomous USVs.

A. IMPORTANCE OF SYSTEM SAFETY ASSESSMENT FOR AUTONOMOUS USVS IN NAVAL WARFARE

The deployment of autonomous USVs in military operations in contested and congested waterways is highly desirable on many tactical and strategic grounds. Pursuing this goal, USVs are quickly becoming pervasive in naval warfare experimentation and future operations, but the U.S. Navy opined that "there's still more work to do before unmanned ships become a permanent fixture in fleet operations" [14]. Failure to properly address operations and safety concerns for advanced USVs can result in catastrophic consequences, such as loss of life, system failure, environmental damage, and collisions at sea. It is important to ensure that appropriate operational and safety processes are in place for USVs to mitigate, prevent, and reduce the probability of adverse events. A classic example is the USS Vincennes (CG 49) incident that occurred during the height of the protracted Iraq-Iran war (1980–88). In 1988, the Vincennes, a Ticonderoga-class guided missile cruiser fitted with the then state-of-the-art Aegis Combat System, shot down an Iranian commercial airplane (Iran Air Flight 655) [15]. While a full accounting of how this incident occurred will not be discussed in this work, this incident demonstrates that autonomous weapon systems, while functioning properly, can harm non-combatants, even though the U.S. military has been known for their ethical targeting practices. When such an incident happens, there is a "visceral human desire to find an individual accountable" [16]. As maritime military platforms become more autonomous, however, critiques such as Robert Sparrow's "Killer Robots" will argue that the assessment of responsibility is increasingly difficult, thus contributing to the public's distrust of autonomous systems [9].

To mitigate such risks, we explore a renowned system safety assessment framework for its applicability to the autonomous USVs: Leveson's STPA [17]. The STPA is used mainly in the civilian aviation industry, offering a safety analysis approach to understand complex behaviors, and control the system's operations to prevent it from entering hazardous systems state.

B. THESIS SCOPE, ASSUMPTIONS, AND LIMITATIONS

With the deployment of autonomous USVs involving ethical implications and compatibility of system safety, the purpose of this thesis is to analyze the effects of these two parameters in congested and littoral waters. To achieve the objectives of this study, a qualitative and analytical research design is adopted. This will involve a review of existing literature, including academic journals and government publications. As the research project covers two different fields (ethics and system safety), the literature review is embedded within the chapters that follow. The broad scope of this thesis covers three main parts. First, we create a hypothetical vignette under the constraints of the Law of Armed Conflict (LOAC) [18], United Nations Convention on the Law of the Sea (UNCLOS) [19], and Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREG) [20], [21] in a littoral and congested environment. Second, we identify the safety hazards using Leveson's STPA and explore its applicability to the autonomous USVs [17]. Lastly, we analyze the moral permissibility of the potential autonomous USVs' tasks and the complexity of the system to achieve those tasks. The ethical analysis adopts a consequentialist normative ethics, utilitarianism, to understand how it impacts non-combatants in the aforementioned operating environment [22].

The methodology for this research project entails the review of the literature on the use of USVs in naval operations using open and unrestricted sources. An overview of the concept of operations (CONOPS) for USVs is provided at the outset of the research. *U.S. Navy Employment Options for USVs* [1], *USV Master Plan* (version 2007) [23], and "Unmanned Systems Integrated Roadmap FY2011–2036" [24] provide a history of military USVs deployments as well as the envisioned deployment in future. As the in-depth details of any specific military USVs' capabilities are classified in nature, a notional reference USV platform is presented for the sake of the ethical and system safety discussion. We have also reviewed current literature on the levels of autonomy and decided to follow European policy makers' terminology—human-in-the-loop, human-on-the-loop, and human-out-of-the-loop—in this research project for its clarity on the levels of human intervention in our analysis [25]. With this methodology, we set the foundation for an operational vignette through hypothetical deployment of autonomous USVs in the 2030–

2035 period. This vignette takes into account the challenges associated with international law (LOAC, UNCLOS, and COLREG) briefly and imposes the constraints on the autonomous USVs in the contested maritime environment.

The vignette is completed after identifying the hazards using the STPA approach, extracted from *Engineering a Safer World: Systems Thinking Applied to Safety* [17]. These hazards can result in damages to the USVs as well as injury, loss of life, and property damages to others. The systems and subsystems selected for analysis are highly relevant to the autonomous nature of the USV such as collision detection/collision avoidance (CDCA) and supporting systems that enable the vessel to operate autonomously. Examining the suitability of employing STPA for hazard analysis on the current Department of Defense's Military Standard 882E (MIL-STD 882E) USVs calls for a more thorough methodology. This involves adopting a mixed-method approach, combining both quantitative and qualitative methods. A literature review is conducted to identify the advantages and limitations of applying STPA to autonomous systems such as USVs. Subject matter experts or practitioners in the field of system safety and autonomous systems will be consulted for their assessments on the matter.

The next phase of the methodology focuses on responsible artificial intelligence (RAI) for USVs. This includes a review of the need for RAI in naval operations, the U.S. Department of Defense's five principles of RAI, and designing RAI systems for USVs. Through the discussion of system safety and ethical implications of operational deployment, the hazards identified and the utility of deploying autonomous USVs aid in the implementation of RAI and safety engineering approaches in the real-world scenarios. On the assignment of responsibility, the research project draws the examples from autonomous cars and aircrafts fitted with autonomous systems due to the lack of operational data for autonomous USVs.

Lastly, the analysis covers the moral permissibility of each task conducted by the autonomous USVs. We aim to arrange the relative order of comparative moral permissibility of each task, recognizing that this order may differ depending on the political and military context. In addition, we assess the complexity of each system, with a greater emphasis on safety for more intricate systems. The analysis ends with the proposal of a

model that integrates ethical considerations and engineering complexity. This model seeks to define a threshold for human involvement, aiming to enhance commanders' trust in the autonomous actions of USVs.

There are four assumptions taken in this research. First, military personnel distrust autonomous systems and seek their improvement before the full adoption into the naval fighting force. Second, autonomous USVs do not have the same reliability issue as autonomous cars and aircraft systems. Third, the two different fields, system safety and ethics, are co-dependent in operational USV scenarios, thereby eligible for an integrated analysis. Lastly, the use of a conceptual system developed in this research project can be generalized to other classes of USVs or autonomous systems.

C. NAVAL CHALLENGES IN THE "GRAY ZONE"

Coastal states, regardless of their size and military might, face an immense responsibility in safeguarding their sovereignty and maritime resources. The maritime security landscape has shifted dramatically in recent times, however, presenting a multitude of challenges extending beyond conventional state-based threats. Today, naval challenges are not limited to traditional adversaries; they encompass a diverse array of actors ranging from states and non-state entities to paramilitary groups and terrorist organizations.

Many of these challenges emerge from the "gray zone," commonly referred to as the deployment of non-traditional forces to achieve national interest below the threshold of war [26], [27]. Oftentimes, these threats operate in an ambiguous and elusive realm [28]. One prominent example is the use of fishing militias by China to assert territorial claims and exert pressure for maritime dominance over neighboring nations. These militias, acting under the guise of civilian fishing vessels, engage in aggressive and provocative actions, leading to territorial disputes and increased tensions. While the inconsistent operation within the self-proclaimed "Nine Dash Line" can be attributed to differing interpretation of international maritime laws such as UNCLOS, a more dire concern arises from the blurred distinction between combatant and non-combatant giving rise to ethical dilemmas. Remarkably, even when the threatened neighbor is a U.S. ally, as is the case with Vietnam,

such questionable actions are "far less likely to trigger U.S. intervention," thus highlighting the complexities of such unconventional threats [29].

Non-state actors are as great a security concern as unconventional state-sponsored actors. Another example that demonstrates the challenge of maritime security in such waters is the infamous USS Cole (DDG 67) tragedy; a terrorist attack that left 17 U.S. Sailors dead and close to 40 other crew members injured in the port of Aden, Yemen during its preplanned fuel stop [30].

In response to the growing risks posed by these diverse threats, there is a potential drive toward the adoption of USVs to be the strategic alternative for U.S. Navy conventional forces to guard against aggressive unconventional forces and non-state actors [31]. Amid the backdrop of great power competition, the U.S. Navy is likely to prioritize their conventional forces to defend against conventional threats. This prioritization leaves the United States and its allies vulnerable to China's state-controlled unconventional maritime forces and maritime terrorism. Due to the absence of human operators and the long endurance of USVs, these USVs are suitable assets to track China's fishing militias and non-state actors should they perform any suspicious activities.

Deploying USVs in contested waters has encountered its own set of issues. For instance, the confiscation of and towing away of U.S. Navy unmanned systems by a Chinese rescue and salvage ship and the Iranian Revolutionary Guard Corps, in two cases, have raised concerns about the vulnerability of unmanned assets in hostile environments [28]. Furthermore, uncertainty on how UNCLOS and COLREG apply to USVs adds another complex layer to the multipolar security landscape [32].

Despite these issues, there are three compelling reasons to invest in unmanned systems for naval operations. First, USVs can perform the "dirty, dull, and dangerous" tasks to minimize risks to human personnel [23]. Second, a reliable USV is able to reduce navigational errors in straits or congested waters, such as the error that led to the USS John S. McCain (DDG 56) collision [33]. Third, the advancement of artificial intelligence (AI) and machine learning (ML) technologies enable USVs to adapt and improve their

capabilities over time, enhancing their effectiveness in countering diverse naval challenges [34].

D. USE CASE INTRODUCTION

To prime for an effective ethical and safety analysis, the conceptual military USV is described in this section. This initial setup is crucial for laying the foundation to understand and improve the effectiveness of autonomous USVs in the real world. Following this, the specific scenario to study USVs' operating environment and operational responses to the hypothetical fishing militias is developed in Chapter II, thus setting the parameters to glean any valuable insights into their behaviors and decision-making processes. Next, the system of interest (SOI) to examine the STPA approach is elaborated in Chapter III.

1. The Notional System of Interest

The use case SOI refers to a conceptual medium-sized USV, modelled on a 20-meter hull length. Comparable in size to sailing yachts and small commercial vessels, the USV operates in congested and littoral waters. Therefore, the scope of this paper limits the discussion to a range of small to medium USVs.[1] The function of these USVs is to maintain a continuous maritime patrol presence and respond promptly to threats, if necessary. For the purposes of the safety analysis, the notional USVs are assumed to be outfitted with communication technologies, autonomous navigation, radar, and collision detection and avoidance systems. Figure 1 illustrates the communication linkage between the notional USVs and the shore-based headquarters (HQ), facilitated by satellite communication (SATCOM) technology. The notional USVs are also equipped with a fully autonomous non-lethal and lethal weapon system to counter hostile threats.

[1] The US Navy organizes its USVs into four size-based categories: very small, small, medium, and large [35]. In congested waters, it is improbable for a mothership to stay stationary or linger around to control the very small USV, due to the risk of collision and adherence to international law. On the other hand, large USVs have limited application of intelligence, surveillance, and reconnaissance (ISR) in the straits. Therefore, the we have excluded large and very small USV in this work.

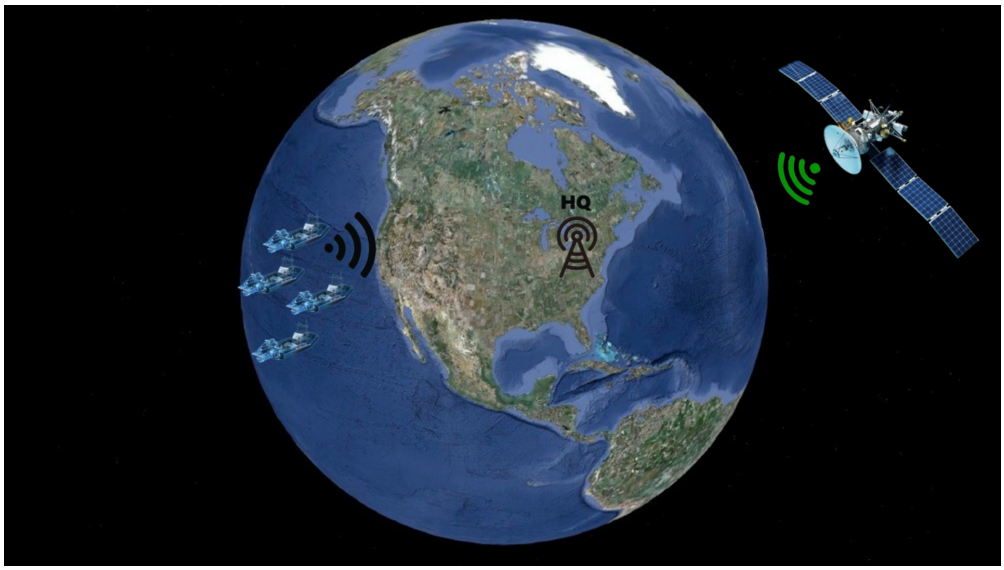

Figure 1. USV-Shore Connection Overview

The SOI is employed for mine sweeping, mine disposal, and patrol operations. During interdiction missions, USVs are remotely managed by operators at HQ. Operating autonomous USVs in congested and littoral maritime settings can present additional safety risks. For instance, the USV's default fail-to-safety response, such as remaining idle, could lead to a collision with other vessels. Unpredictable or unsafe maneuvers by other vessels could also provoke the USV into a hazardous condition. The USV's capabilities play a significant role in two primary controlled processes: autonomous navigation and communication.

a. Autonomous Navigation

Interactions with other vessels during operations can generate conflicts with the USV's autonomous programming, possibly leading to unsafe behaviors. This risk can be further escalated by the actions and responses of other vessels.

b. Communication

Link latency can impact real-time intervention from HQ, potentially affecting situational awareness of the operator or even the timeliness for autonomous USVs to

interdict hostile threats. Delayed situational awareness could result in less-than-optimal decision making during critical situations.

2. The Notional Operational Vignette

A brief introduction to the operational vignette is provided here with more detailed descriptions provided in Chapter II and III. First, in Chapter II, a hypothetical scenario is presented that explores the challenges faced by autonomous USVs in the busy waterways of a chokepoint, the distinguishability of combatants and non-combatants, and the legal dilemmas of USVs. Next, in Chapter III, hazards are identified with the STPA approach and later developed into different scenarios that the USVs may encounter. These scenarios are used to analyze ethical considerations and the application of system safety to draw valuable insights and understanding of autonomous USVs. The typical phases of operation for USVs are shown in Figure 2, which starts by launching the USV from the naval base; navigating to the disputed waters; conducting intelligence, surveillance, and reconnaissance (ISR); and returning to the base.

While these phases seem applicable to many autonomous vessels, the span of USV operations within phase 3 of Figure 2 is wide and distinct. According to *U.S. Navy Employment Options for USVs*, current USVs can execute the 16 unique mission types depicted in Figure 3 [1].

Figure 2. Phases of Operation for USV Deployment

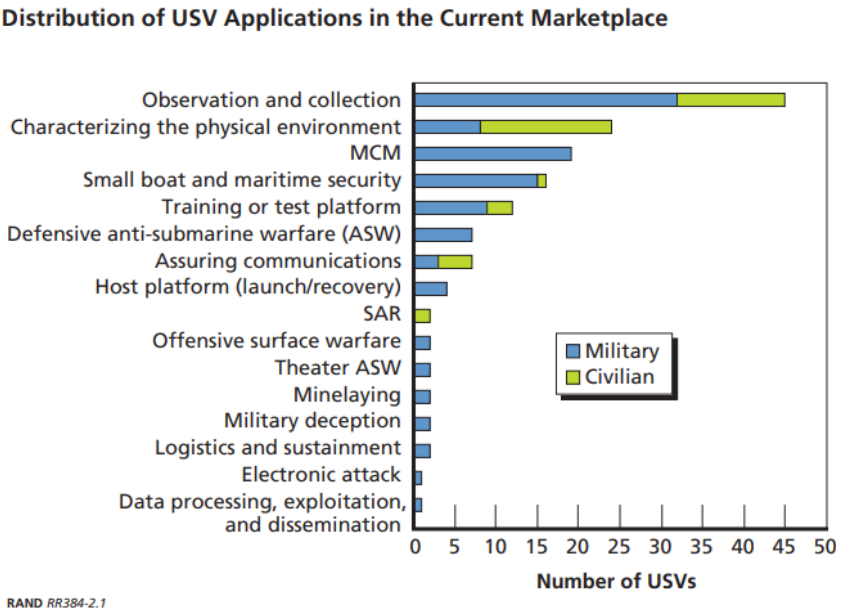

Figure 3. List of USV Operations. Source: [1].

For the purpose of the ethical analysis, we aim to generalize the vignette for two reasons. First, it provides a common-to-all type of scenario that allows for the exploration of fundamental ethical concepts. Second, the vignette allows for identification of common ethical challenges and the exploration of the ethical decision-making process. While the thesis recognizes the intricacies of the unique activities conducted within the mission envelope, these nuanced details are kept broadly applicable to discuss sufficiently the ethical dilemma but not dilute the intended research questions. The generalized phases of operation, shown in Figure 4, are appliable to the ethical discussion in this thesis. The notional USVs are capable of conducting these four tasks autonomously: launch and recovery (LAR), navigation, ISR, and firing, if required.

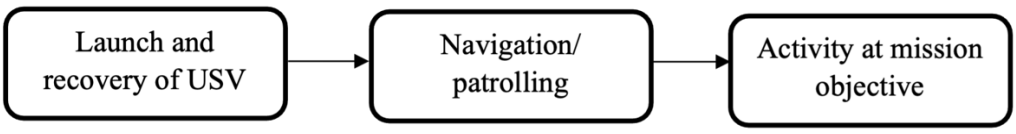

Figure 4. Phases of Operation Discussed in Ethical Aspect

For the system safety analysis, it is important to highlight that we continue the generalized vignette approach without any analysis in the activities conducted at the mission objective. This is attributed to the cross-domain specialization (e.g., weapons) required to provide a thorough and comprehensive analysis. It is also important, however, to acknowledge that specialized domain analysis is a crucial aspect that should be considered in future work. The generalized phases of operation, shown in Figure 5, are discussed in the system safety analysis.

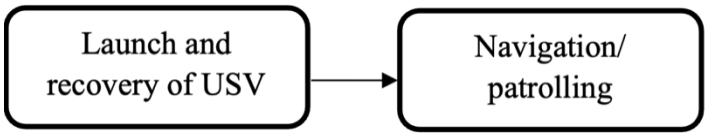

Figure 5. Phases of Operation Discussed in System Safety Aspect

E. RESEARCH QUESTION

Employing autonomous USVs in military operations entails both safety hazards and ethical considerations; the research question is, how do these ethical considerations and system safety processes influence the implementation of meaningful human control in a littoral and congested maritime environment?

F. THESIS ORGANIZATION

The thesis is composed of five chapters:

Chapter I begins by laying the foundation through an overview of conventional USV operations in military operations. This provides a macro view of ethical considerations faced by USV commanders and the key drivers behind the formulation of system safety standards. With the system of interest and notional autonomous USVs' deployment, a list of research questions is developed and presented in this chapter.

Chapter II creates the first part of the vignette using the potential challenges autonomous USVs may face in a congested and littoral operational environment. This chapter begins with a literature review of USV CONOPS, different levels of autonomy, and international maritime laws. These reviews highlight three potential challenges that

sets the operational boundary for the vignette. First, the limited navigable waters of a strait act as a chokepoint, thus increasing the risk of collision and the adverse effects of operating near a littoral space. Second, ethical considerations such as distinguishability and proportionality arise when employing autonomous USVs within the bounds of the LOAC principle. Lastly, we briefly cover portions of UNCLOS and COLREG that autonomous USVs may not be able to adhere.

Chapter III creates the rest of the vignette through STPA's six-step process. This chapter begins by conducting a literature review on the accident causality model and STPA. Next, it examines the STPA approach to develop the loss scenarios with ethical consideration. Given that the STPA approach is widely recognized in the aerospace and other civilian industries, this research aims to apply STPA to complement military USV systems, which are designed, built, and tested safe to MIL-STD 882E. This thesis attempts to study the advantages and limitations of this aviation standard on autonomous USVs. To ensure that the technology is used responsibly and ethically, this thesis further explores the impact ethical constraints have on the outcome of missions and safety-related issues of non-combatants in the environment.

Chapter IV delves into the meta-ethical questions associated with fully autonomous USV such as command, control, and responsibility. In order to guide responsible use of AI, this chapter examines the governance of AI policy by different stakeholders (e.g., U.S., European Commission, and UNESCO). It also examines how policies such as governability and traceability are important to close the responsibility gap. The research project also uses Robert Sparrow's argument in "Killer Robots" to expound on the "attribution of responsibility" in fighting a just war [9].

Chapter V uses a three-stage process to develop a systemic model that includes the analysis of ethical considerations and engineering complexity; it also aims to provide a threshold of when humans should become involved, thus increasing commanders' trust in autonomous USVs acting independently. In addition, we provide actionable recommendations to ensure the safe deployment of the autonomous USV system. Synergizing with the previous chapters on AI ethics, this chapter aims to aggregate the

findings and present the results of this project. This chapter also compiles the summary of findings and provides future research directions for the deployment of autonomous USVs.

II. SETTING THE STAGE: CHALLENGES OF AUTONOMOUS USV OPERATIONS

To comprehend the growing interest in USVs, particularly from the naval military standpoint, this chapter sheds light on their roles in naval warfare and distinctive advantages in the reduction of human risk, operational and maintenance costs, and long endurance [1]. During research and development incorporating AI and ML, however, concerns have arisen among policy makers and academics on the ethics of arming USVs. This is due to uncertainties and mistrust of USVs' ability to navigate and make appropriate decisions to fire autonomously [12]. This chapter delves into the potential concerns about fully autonomous USVs, using the principles of the LOAC and UNCLOS to provide critical insights. Lastly, the chapter concludes by creating a vignette to set the stage for the next few chapters to argue how USVs can avoid these moral and ethical dilemmas as well as recommending how to strengthen the safety standard through the integration of STPA into the military.

A. USVS' ROLES AND THE OPERATING ENVIRONMENT

USVs have earned the military's trust and dependence due to their successful contribution to dirty, dull, and dangerous tasks [23], [36]. For example, during the Able and Baker atomic bomb test in 1946, these USVs, documented as drone boats, collected water samples from the lagoon surrounding Bikini Atoll, which were later used to assess the radiological impact of the atomic bomb detonation. The USVs were remotely piloted using radio signals by operators on the drone control ship, USS Begor (APD 127), minimizing human risks and exposure to hazardous environments [36].

The value that military practitioners see in USVs did not immediately allow USVs to gain acceptance for an expanded role. Since their inception, USVs were primarily used for ISR missions or hazardous situations such as mine countermeasures [23]. Yet, the focus on USVs lagged behind that of their unmanned counterparts serving in the air, on the land, and beneath the sea, resulting in comparatively lesser emphasis overall [1]. Recognizing the USVs' potential in the maritime domain, the Department of Defense (DOD) sponsored

research on USVs' mission and functions for the U.S. Navy in 2013, which suggested 10 suitable mission sets, an expansion from the initial U.S. Navy *USV Master Plan*'s seven mission sets [1]. The research and development of USV employment options paved the way for autonomous USVs to be an important element in U.S. naval warfare.

Later in 2016, ADM Gary Roughead, Chief of Naval Operations (CNO) in the U.S. Navy, boldly reiterated at the Brookings Institution, "I represent the Navy, and it is imperative that unmanned systems encompass all the domains within which the Navy conducts its operations" [37]. ADM John M. Richardson, U.S. Navy CNO, provided further options for the integration of USVs with existing systems in his first of four key lines of effort: "strengthen naval power at and from sea" [38]. In this article, he proposed an integrated approach that "explore [s] alternative fleet designs, including kinetic and non-kinetic payloads and both manned and unmanned systems" [38]. Since then, the United States has firmly committed to further advancing its unmanned technology capabilities in the maritime surface domain.

1. Unique Challenges of USVs in Straits

The geographical constraint of a strait often revolves around the restricted navigable waters flanked by its shores, posing greater navigational and operational risks than in ocean passage. Since the definition of a strait is "a narrow passage of water connecting two large bodies of water," the limitation of the narrow passage amplifies when it serves as a bustling channel leading to major trading routes, sometimes referred to as shipping chokepoints [39]. These chokepoints restrict the transit capacity and lack viable alternatives, often entailing hefty costs and unnecessary delays when alternative routes are sought. Such circumstances contribute to a high volume of traffic, resulting in three possible high concern risks. First, the risk of collision is high. Second, if the law enforcement around the area is absent, cargos carrying high value goods are an easy prey for pirates [40]. Third, if the law enforcement is active but the jurisdiction of the area is shared by two or more countries, dispute between states may arise. Each state may assert military strength and tight restriction such as threatening a blockade for the purpose of

deterrence, disrupting sea lines of communication (SLOC), and even leading to disputes that persist until a formal agreement to cooperate has been signed [41], [42], [43].

Thus, navigating through straits and congested waters presents a unique set of challenges for USVs, be they remotely controlled or fully autonomous. One primary concern is the increased likelihood of encountering and potentially causing harm to non-combatant vessels. The confined nature of these waterways, often densely populated, raises the stakes for unintended interactions with civilian traffic. Moreover, the risk of collision in these congested areas significantly rises [44]. Multiple vessels, unpredictable currents, and narrow passages demand precise navigation and collision-avoidance capabilities. These challenges highlight the need for advanced navigational systems and algorithms to ensure safe and effective USV operations.

Beyond the increased risks of encountering non-combatants and collision, USVs operating in straits and congested waters face additional complexities. System errors, particularly radar clutter from the shoreline posing as contacts approaching USVs as well as loss of contact during high-speed maneuvers, can disrupt navigational decision-making processes and further complicate operations in these confined spaces [45]. Additionally, USVs are susceptible to adversaries' electronic warfare (EW) capabilities such as jamming or spoofing of the Global Navigation Satellite System (GNSS), which can disrupt or compromise their functionality, posing a significant operational risk in a GNSS-denied environment [46]. Lastly, there is the critical concern of inadvertently crossing into territorial boundaries, as many straits serve as international borders. Navigating these intricate territorial dynamics requires precise coordination from the military to avoid diplomatic and security complications.

2. Concept of Operation for USVs

The first breakthrough for USV development in military operations was the formalization of the U.S. Navy's *USV Master Plan* [23], guided by clear vision from Sea Power 21 [47] and Quadrennial Defense Review 2006 [48]. The objective of USVs then was to support homeland defense, the War on Terror/irregular warfare and conventional campaigns. With the U.S. Navy's emphasis on unmanned technology supporting a full

array of sea strike and sea shield options, the envisioned tasks for USVs expanded to include mine countermeasures, anti-submarine warfare (ASW), maritime security, surface warfare (SuW), special operations forces support, EW, and maritime interdiction operations support [47].

Despite acknowledging that fully autonomous USVs will not be achieved within 5 to 10 years from the promulgation of the *USV Master Plan* (2007 version), each mission set strategically leverages the benefit offered by USVs [23]. Two examples drawn from the mission sets are used to illustrate the role of the USVs and how the distinct advantages complement the execution. First, for ASW, the USVs will "patrol, detect, hand off, or engage adversary submarines" [23]. This establishes an additional layer of ASW defense for the manned surface group. It also allows the manned combatants to focus on their core tasks and minimizes potential risks that the manned platforms could face, should they be directly involved in the ASW missions [23]. Second, for maritime security, the USVs will conduct ISR and/or deter hostile actions over a prolonged period. The long endurance and sustainable nature of the USVs allows them to provide a conspicuous presence in the area of operation for deterrence and immediate intervention. In addition, undisrupted data collected over a long period can be transferred back, at near real-time, for further processing and sensemaking [23].

Subsequent investment in USV suitability study suggests that USVs can augment manned vessels to do more with higher payload and cross-domain integration [1]. This study, which outlines potential applications for the USVs, was caveated not to be viewed as a replacement or update of the *USV Master Plan* [23] or "The Unmanned Systems Integrated Roadmap FY2011–2036" [24]. Instead, it identifies 10 distinct mission sets for the U.S. Navy to consider when USVs can be employed more effectively owing to the growing capability to integrate better [49]. The mission sets include command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR); military deception/information operations/electronic warfare; SuW; mine warfare; ASW; logistics; ground attack; air and missile defense functions; and missions not currently being performed. Similarly, two examples are highlighted for simple illustration of the USVs' CONOPS. First, for mine warfare in confined waters, the USVs will detect and collect

relevant information for sensemaking. Compared to UAVs and unmanned underwater vehicles (UUVs), USVs are able to carry larger payloads such as sonar to perform effective mine hunting operations [1]. Second, as part of the C4ISR mission to conduct communications relay, the USVs are deployed within communication range to perform an intermediatory role between manned vessels, UAVs, USVs, and/or UUVs. The unique position of USVs sailing above water allows them to relay information for a robust cross-domain integration, bridging the air and the sea (underwater) interface. [1].

Moving forward from the theoretical discussion, the concept of unmanned systems performing coordinated movements and distinct tasks became reality a decade ago. In 2014, a demonstration conducted by the U.S. Office of Naval Research (ONR) on Virginia's James River witnessed five autonomous USVs escorting a simulated high-value target against a potential aggressor. While the nascent autonomous USV swarming capability was described as a "chaotic group of kids learning to play soccer for the first time," these USVs were able to break off from the protected ship and establish a defensive barrier between the intruding vessel and the protected ship autonomously [50], [51]. Not only did this event show their ability to coordinate a set of complex maneuvers, including maritime protection and interdiction, the autonomy to engage the enemy as a sacrificial shield for the high-value target is a testament that unmanned systems can be sacrificial.

Haunted by the tragic suicide bombing of USS Cole, ONR orchestrated yet another impressive USV swarming demonstration in 2016, capable of protecting a harbor from intruders "without direct human control" [50]. The autonomous USVs were built off of an existing rigid hull inflatable boat integrated with a software originally designed for the NASA Mars Rover [51]. In the demonstration, four USVs were seen patrolling within the grid of 4 nautical miles by 4 nautical miles. Once a potential threat was identified, the group of USVs collaboratively determined which among them should undertake the task of tracking and trailing the intruding vessel [50]. The CONOPS for swarming USVs, even without any weaponry onboard, opens up a chest of options for any military commander throughout the peace-to-war continuum. These options start from the most benign activity such as coordinated surveillance and reconnaissance to complex maneuvers like effective

interception or kamikaze-type operations [52]. These unmanned vessels continue to demonstrate reliability and versatility in naval operations.

Despite Rear Adm. Casey Moton's confidence in the development of unmanned systems in the U.S. Navy, touting progress as "fairly close on the autonomy," he further explained that "we are definitely going to have a requirement for crew support on large USVs (LUSVs), or a smaller crew, to handle those things that are just not quite there with maneuvering critical situations" [53], [54]. This statement further exemplifies the military practitioner's distrust of autonomous systems making and executing decisions without human intervention. In the next section, we explore what autonomous means and the potential distrust of autonomous USVs in the context of the LOAC and maritime laws.

3. Defining Levels of Autonomy

The concept of autonomy in unmanned systems has garnered substantial attention, with researchers and experts proposing a plethora of frameworks. Before going into depth on the existing literature, the terms autonomy and automated should be clarified. The term "autonomy" is derived from the Greek words "autonomia" and "autonomous," where "auto" signifies *self* and "nomos" stands for *law*. As a compound word, "autonomy/ autonomous" encapsulates the concept of being independent and self-governing [55]. It also connotes a system that has the freedom to make and execute decisions on its own. "Automation" on the other hand, relates to a system that operates precisely as programmed by the developer, devoid of any room for choice or the ability to deviate in response to varying circumstances [56]. Its actions are predetermined from the outset with no capacity to alter them in the future, regardless of the situation. Lower levels of autonomy (LOA) tend to be automated or predefined while higher levels assert much more independent decision making. Further, lower-LOA systems tend to include more harm intervention than higher levels.

A prevailing commonality across LOA frameworks is the effort to categorize varied nuances of interactions between the humans and machine, regardless of its usage on land, at sea, or in the air. One of the earliest and most significant frameworks developed by Thomas Sheridan and Bill Verplank [57] and Sheridan [58] presents 10 LOA,

encompassing an array of options for human-machine collaboration ranging from fully manual control to fully autonomous [56]. Between the two extremities, the decision lies with human from levels 2 to 4. Level 5 is the middle ground when the computer executes after the human approves. From level 6 to 9, human intervention is a minimal supervisory role to override, while the computer decides and executes on its own. This LOA was later modified by Parasuraman et al. into four stages of autonomy that consider the behavior of information processing by humans [59]. Even though the approach and the presentation of the LOAs are different, the principle of fully autonomous and sharing of controls by human and machine largely remains the same. This reinforces the misconception that we should not perceive autonomy in binary, rather "one system is more or less autonomous than another" [60].

Not all studies are bounded by the extremities of fully manual to fully autonomous, which is a tailored approach to specific application and industry. Mica Endsley's conceptualization introduces four LOA, prioritizing increased system autonomy over human control [61]. The taxonomy being rooted in the context of the advanced cockpit that may have automation such as autopilot mode. In this situation, it comes as no surprise that she eliminated the fully manual control of the airplane in her lower bound. Bernd Lorenz et al. likewise present a compact version with three LOA, accounting for a fault management level [62]. It is composed of the baseline, the automation support, and the automation support failure level. Lastly, the remotely piloted to fully autonomous taxonomy proposed by Bruce Clough illustrates some of the unconventional ways to present the LOA [63]. To expand the flexibility of controls, some theorists such as Celestine Ntuen and Eui H. Park [64] suggest the importance of fully manual option progressively throughout their book, building off a similar concept from Lorenz et al.'s work [62]. Victor Riley [65] and subsequently Ben Shneiderman [66], however, took a more innovative approach that deviates from the shared control between human and machine to a taxonomy that is entirely different. Both concepts suggest the co-existence of high human control and high autonomy.

Within the maritime industry, there is also a diverse way to present the taxonomy due to the multifaceted nature of the field, but similarities to the previous literatures were

observed. For example, Clough's taxonomy [63], spanning from remotely controlled to fully autonomous, is evident in documents such as U.S. Navy *USV Master Plan* [23], "Unmanned Maritime Systems Engineering Technologies and Applications for Unmanned Maritime Systems" [67], and *U.S. Navy Employment Options for Unmanned Surface Vehicles* [1]. The first document simplifies the taxonomy into three levels: manual, semi-autonomous, or autonomous/fully autonomous [23]. Manual is defined in an engineering term to have "man in the loop continuously or near continuously." Semi-autonomous recognizes the autonomy of the system itself but with operator input such as the authority to fire the weapon. Fully autonomous is the ability to self-govern from start to finish. The second document consists of five LOA shaped by the communication linkage, the nature of decisions made, and the risks involved [67]. They suggest a higher level of human control when the communication linkage is the most stable and during lower risk situations and more autonomy in adverse conditions. The third document with five LOA is primarily categorized based on the control of navigation between user input and machine-initiated actions, for example, plotting of waypoints by human or machine for collision avoidance [1].

Contrary to all the above LOA taxonomy that focuses on the different types of human and machine interaction, the International Maritime Organization (IMO) hinges its level of autonomy with the location of seafarers and how the ship is being controlled [68]. The four LOAs are uniquely presented with the first three levels remotely controlled but differentiated with the immediacy for human intervention during the ship's voyage and the fourth level fully autonomous. Critics have argued the inaccuracy of assumptions and proposed changes for the amendment of IMO Conventions [32]. From the perspective of European policy makers, we have noticed a parallel trend in terminology, prioritizing human-centric distinctions when categorizing robotic weapon functions. The three primary classifications based on the degree of human involvement are the human-controlled ("human-in-the-loop") system, the human-supervised ("human-on-the-loop") system, and the autonomous ("human-out-of-the-loop") system [25].

The trend of differing taxonomy highlighted the difficulty of imposing a single definition of LOA and thereby offers three insights. First, there is no one-size-fits-all

model. Second, despite the differences in taxonomy, there is "no good or bad taxonomy," only a framework that fits one application better than another [56]. Third, most of the literature, regardless of technological maturity, gravitates toward a fully autonomous USV, evidently designating it as the upper bound. As far as this thesis is concerned, we follow the terminology of human-in-the-loop, human-on-the-loop, and human-out-of-the-loop for its simplicity to examine the level of human intervention in our analysis at a later chapter (Chapter V).

Even after gleaning clarity on the levels of autonomy, current research for fully autonomous USVs has varied understanding. According to Vincent Boulanin, autonomy is a broad definition that covers many functions in a system [69]. Andrew William [70], Arnold Roberta [71], and Paul Scharre [72] shared the same sentiment that a more insightful approach is to focus on the autonomous functioning within a system rather than categorizing the entire system as autonomous. This is particularly useful when striving to assess the ethical and safety aspects of autonomy relevant to the USVs and to determine the extent to which they can operate independently.

B. AUTONOMY IN THE EYES OF INTERNATIONAL LAW

Throughout history, resistance to new technology is a recurring theme, often due to the uncertainties surrounding its legal implications [73]. The introduction of new technology will in one way or another change the status quo, but not all disruption faces equal resistance—some, such as electric lights and the advent of the telephone, are more well-received than others [74]. Generally, disruptions that cause instability to the younger generation, jeopardize competitive edge, infringe upon rights, or threaten regional peace are widely perceived as less favorable. In the case of autonomous USVs, their increasing autonomy introduces a host of military, civilian, and legal complexities that must be navigated. These complexities include the deployment and operation of autonomous USVs in relation to existing international laws, such as LOAC, UNCLOS, Just War Theory, and COLREG.

1. Law of Armed Conflict

The primary objective of the LOAC is to safeguard individuals who are unable to defend themselves during hostilities [75]. Two entities that may be the subject of concern are the combatants and noncombatants. Traditionalist Michael Walzer claims that combatants have lost their rights to life and liberty because they have agreed to be weaponized into lethal beings [76]. Thus, it is permissible to kill combatants in war. Noncombatants, on the other hand, enjoy those rights and cannot be the targets of military actions until absolutely necessary to prevent further catastrophes. Even in those sporadic situations to kill noncombatants unintentionally as collateral damage, it should be done in a necessary and proportionate means adhering to the objectives of the attack [76]. Therefore, killing of noncombatants intentionally is impermissible in normal circumstances of the armed conflict.

While the LOAC, also known as international humanitarian law, primarily defines the permissible targets and outlines the legal obligations and constraints concerning potential harm to civilians and their property during armed conflicts, its scope extends beyond traditional warfare. This legal framework serves as the applicable regime even in situations falling below the threshold of war [77], [78]. LOAC's provisions continue to be relevant in conflicts that might not meet the traditional criteria of full-scale armed confrontations but still involve the potential use of force or military capabilities. One such example is the non-international armed conflict where there are "protracted armed confrontations occurring between governmental armed forces and the forces of one or more armed groups, or between such groups arising on the territory of a State" [77], [79]. This provides valuable perspective on the relevance of the LOAC against terrorism and non-state actors, though the caveat in practice is the corporation of multiple groups to be treated as a single party to the conflict [80], [81]. For example, the United States uses the phrase, "al-Qaeda and its associated forces" to justify the legality of its War on Terror [82].

Discrimination, proportionality, and necessity are the bedrock principles of LOAC and these are core pieces of the Just War Theory that armed autonomous USVs may find challenging to follow [75]. While military necessity is bounded by the law, and thus easier for rule-following AI to obey, discrimination and proportionality require more attention.

Extreme discrimination is needed to distinguish between combatants and noncombatants engaged in hostilities, as well as military objectives and civilian objects [83]. Subtle differences in how an object is perceived can indeed have significant implications, especially in the context of an armed autonomous USV. One critical challenge that arises is the amount of data required for the system to consistently make correct distinctions. Armed autonomous USVs not only need to differentiate between various types of ships but also understand the different functions of these ships. This entails recognizing the broad category of civilian vessels apart from military warships and more subtle differences such as a warship from a neutral nation, an enemy warship repurposed as a hospital ship, or a warship that has surrendered. Additionally, armed autonomous USVs must possess the capability to assess the combat effectiveness of a ship accurately. For instance, it must be able to distinguish a fully operational enemy warship from one that has been seriously damaged, resulting in zero combat effectiveness (commonly known as *hors de combat*). When the target is an individual person, how is the machine trained to distinguish personnel from the armed forces or an innocent civilian before they display any hostile intent or actions? If it is through the identification of insignia or uniform, how effective will the recognition be in the absence of these items [84], [85]?

The advantage of USVs' long endurance can be a double-edged sword. While USVs can remain longer at sea than humans, the time factor plays a crucial role in a volatile, uncertain, complex, and ambiguous (VUCA) environment [85]. A warship initially identified as an enemy ship can become a neutral ship once a peace treaty has been signed. A state-sponsored fisherman can be a fishing militia once the executive order has been issued. A civilian port may temporarily be set up as a military forward staging area. These varying permutations pose significant challenges to distinguishability, with the context changing with time.

The principle of proportionality in armed conflict mandates that military objectives be targeted without causing incidental or collateral damage to civilians and civilian objects [86], [87]. Proportionality may be more abstract than it seems to be, however, because of the underlying judgement required. Should damages occur, it must not be "excessive in relation to the direct and concrete and direct military advantage anticipated" [86], [87]. In

additional to the challenges to omit civilians and civilian objects discussed above, armed autonomous USVs are faced with the challenge of deciding what is excessive to obtain military advantage. The process is not a numerical rigor or "a crude tit-for-tat" exercise for inflicting damage equal to that which it received. It involves comprehending the target's military significance, considering its current role in aiding the enemy, and evaluating the advantages gained by its neutralization or destruction in the given situation [85], [88]. Therefore, in the absence of an ethical framework for autonomous system, the question remains "can LOAC eloquently govern the usage of armed autonomous USVs?"

2. United Nations Convention on the Law of the Sea

The legal framework, often designed under the assumption of a static status quo, are constantly challenged by rapid technological advancements [32]. From new technology permitting drilling for offshore gas and oil to commercial fishing in distant waters, these advancements consequently allowed coastal states to assert rights and jurisdiction over natural resources beyond the territorial sea. UNCLOS went through multiple iterations before reaching a consensus in 1982, which provided states the basis for a rules-based order for all uses of the oceans [19]. UNCLOS establishes the principle of freedom of navigation on the high seas, ensuring that all states have the right to navigate, fly over, and other permissible activities. The principle underpinning UNCLOS that "vessels will always need a captain, a flag, and a crew" does not fit well with autonomous USVs, however [32]. This raises questions about how autonomous USVs are categorized legally, and what rights and responsibilities they have under UNCLOS.

As USVs become more prevalent, issues like collision avoidance and safe passage for other vessels in the vicinity of USV operations need to be addressed to prevent interference with freedom of navigation. Brendan Gogarty and Meredith Hagger [89] propose limitations on the activities USVs can engage in during innocent passage, while McLaughlin [90] highlights that USVs are unequivocally bound by COLREG. The USVs must demonstrate a high level of collision avoidance capability to fulfill the requirement of maintaining a vigilant lookout.

USVs present several legal issues when it comes to operation within the framework of the UNCLOS. Although we are not delving into the legal intricacy and debate, this section highlights two legal dilemmas posed by USVs in the maritime regulations.

Firstly, one of the fundamental legal challenges arising from the use of USVs is the terminology and classification of these vessels. UNCLOS employs terms like "ship" and "vessel" interchangeably but the third UN Conference on Laws of the Sea ratified in 1982 favors "vessel" due to its inclusivity of ships and other floating structures [19]. Yet, precise meanings can vary from one country to another. For instance, India's definition of a "vessel" differs from Korea's definition of a "ship." India, for example, does not include wing-in-ground craft potentially leading to differences in legal treatment that affect the application of admiralty laws. A separate example relevant to USVs is the seizure of an unmanned and remotely operated tethered submersible vehicle. The claimant supported the court with Canadian maritime law when a remotely operated vehicle (ROV) is considered a ship due to its broader definition [91]. The court gave the verdict that the ROV is not a ship under the Australian admiralty law, however, because the ROV (1) "does not fall within the general words of the statutory definition," (2) does not possess the characteristics of a ship, (3) does not achieve buoyancy through the displacement of water, (4) does not navigate on its own but instead is transported as cargo, and (5) is not registered as a ship [92]. This diversity in definitions underscores the pressing need for international consensus on how USVs should be categorized within the legal framework, especially when autonomous vessels may potentially sail to countries of different jurisdictions.

Military practitioners and scholars hold differing opinions regarding whether USVs qualify as warships and thus enjoy sovereign immunity, constituting the second legal issue under scrutiny [19]. According to UNCLOS 1982, the definition of warship means "a ship belonging to the armed forces of a State bearing the external marks distinguishing such ships of its nationality, under the command of an officer duly commissioned by the government of the State and whose name appears in the appropriate service list or its equivalent, and manned by a crew which is under regular armed forces discipline" [19]. According to the U.S. naval warfare publication "The Commander's Handbook on the Law of Naval Operations," USVs are allowed to engage in belligerent actions, such as capturing

and diverting adversarial or neutral vessels, even without a commanding officer onboard [93]. If these USVs are commanded by a commissioned officer or manned by a crew subject to regular armed forces discipline through remote or alternative means, they may be designated as warships. This classification grants them similar rights and freedoms as manned warships, including navigational rights and internationally authorized uses of the waters. Therefore, under the sovereign immunity accorded to the warship, USVs may not be searched, boarded, inspected, or seized without the permission of the flag state, regardless in foreign territorial waters or international waters. This handbook has faced criticism from China scholars, however, due to the perceived intention of the U.S. government to unilaterally influence international regulations concerning the status of unmanned systems at sea [94].

3. International Regulations for Preventing Collisions at Sea

COLREG, established by the IMO, serves the crucial purpose of optimizing navigational practices to avert potential collisions between two or more vessels at sea [95]. Paramount to this research project is the governance of the Traffic Separation Scheme and the conduct of vessels regardless of visibility [96]. In 1972, the Traffic Separation Scheme was included in COLREG as a vital maritime traffic management system. This scheme was first implemented in the Dover Strait "on a voluntary basis" in 1967 [95]. Its obligatory adherence came into effect after the IMO Assembly adopted the resolution in 1971. Since then, traffic in or near the Traffic Separation Zone must observe the regulation. This Traffic Separation Zone can be viewed as a highway of the sea where the flow of the traffic is separated. If the ship is on the correct side of the waters and wishes to merge into or out of the Traffic Separation Zone, it can "normally join or leave a traffic lane at the termination of the lane, but when joining or leaving from the side shall do so at as small an angle to the general direction of traffic flow as practicable" [96]. If it is on the opposite side of the waters and wishes to cross and join the traffic flow from the other direction, it "shall cross as nearly as practicable at right angles to the general direction of traffic flow." This minimizes ambiguity for other vessels regarding the intentions and course of the crossing ship, while simultaneously facilitating a swift crossing of the lane.

One may wonder whether USVs and the military need to obey these rules, but the general rule of COLREG states that "the rules apply to all vessels upon the high seas and all waters connected to the high seas and navigable by seagoing vessels" [96]. Even major powers such as United States and China have acknowledged that their navies will comply with COLREG, so as to "maintain a high level of safety at sea" [96], [97]. Therefore, it is without a doubt that USVs will need to adhere to COLREG for their safety as well as the safety of others.

C. VIGNETTE: DEFENDING THE STABILITY OF THE STRAIT TO MINIMIZE DISRUPTION

The comprehension and practical application of ethical principles hold more significance than the methods employed. Past and present scholars have adopted different methodologies to set up the ideal conditions for examining ethical dilemmas. Some scholars such as Jeff McMahan [98] skillfully narrated fictitious exemplars, while others like Cécile Fabre [99] built her argument based off historical events to analyze the principles of the Just War Theory. We are of the view that the accuracy of the historical events may result in bias through different political lenses, resulting in an additional layer of divergence to the already complicated ethics of war. Moreover, the ethical implications of employing fully autonomous technology have not been thoroughly assessed in any realistic conflicts. Therefore, in this study, we use a notional vignette to compartmentalize actions and evaluate their impact.

The following vignette, which bears the characteristics of littoral and congested waters, serves as an introduction to the scenario that is used to explore ethical implications and system safety in the subsequent chapters.

1. Geography

Figure 6 shows X Strait, a body of water located between the southeastern coast of Country 1 and West of island Country 2. The strait extends approximately 180 kilometers (97 nautical miles) in length and the width varies, with its narrowest point being around 130 kilometers (70 nautical miles). The maritime chokepoint serves as a vital passage for its diverse maritime activities, including shipping, naval operations, and fishing. The strait

functions as an international trade route through which commercial shipping from multiple countries passes, linking the Eastern hemisphere to the Southeastern hemisphere.

Strait X is known for its rough sea conditions. Over the past three decades, a rise in the occurrence of tropical storms and typhoons has led to increased wind speed and wave height. Maximum wind speed has been recorded at 12m/s (27mph) during winter; primarily attributed to the northeast monsoon and cold wave. The maximum wave height recorded was around 3 meters (~10 feet). The wind and waves are from the southwest during summer and northeast at other times. The strong winds and rough seas pose a navigational risk and are known to affect the accuracy of small arms weapons during training exercises.

The United States, not shown in Figure 6, a neutral country that promotes peace and stability around the region, has a satellite naval base, shown in Figure 6, 1000 nautical miles away from X Strait. The naval base docks a mixture of U.S. Navy warships, ranging from manned ships to fully autonomous USVs.

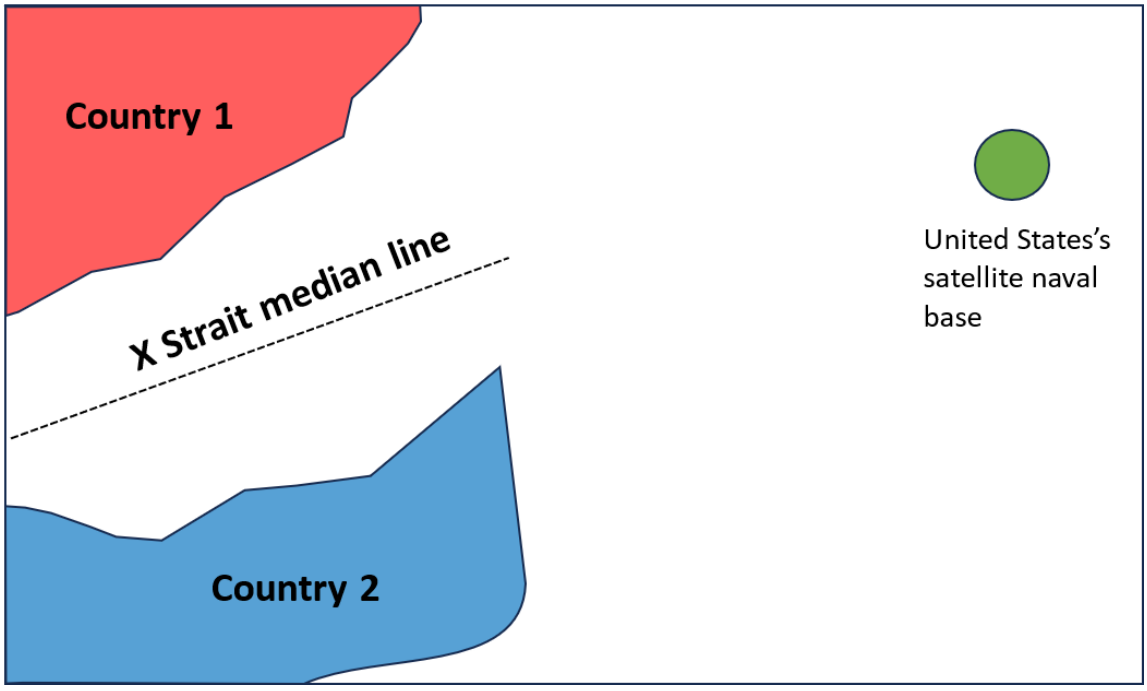

Figure 6. Geographic Location of the Countries Involved

2. Situation

In the near future from 2030 onwards, there has been a rising tension between Country 1 and 2 over a territorial dispute. Country 1 strives for victory without engaging in armed conflict. Recently, it was observed that multiple fishing militias carrying plainclothes coast guard officers from Country 1 were operating outside their jurisdiction, infringing territorial boundaries demarcated by the median line. An estimate of over 8 million fishing militia personnel were reported in 2010 with their modus operandi including establishing a prolonged presence around disputed islands, maintaining close distance from naval ships and coast guards, and harassing the neighboring country's fishing vessels. These militias consist of civilian fishing vessels linked to Country 1's government and military. Although Country 2 expelled around 6,000 of Country 1's vessels and impounded around 300 of Country 1's ships over the past five years, Country 2 has seen a continuous pattern of deployment of the gray zone aggressor's fishing militias in advancing her maritime strategy. These fishing militias are also speculated to operate in X Strait for ISR. As they are employed as full-time fisherman, it is sometimes hard to differentiate whether they belong to the fishing militias or are normal fishing vessels without connections to the government or military. This allows them to disguise their actions as well as maintain a certain level of plausible deniability to Country 1.

3. U.S. Navy's Response to Unconventional Forces

At the macro level, amid the backdrop of great power competition, conducting ISR using strategic assets such as manned warships and SATCOM on fishing militias is disproportionate and ineffective. Given the limited sphere of influence of fishing militias, deploying manned warships to monitor their activities represents a misallocation of valuable resources. In addition, the satellite orbits around the earth, which is not suitable for ISR that requires persistent surveillance.

Military strategy drives technology, but technology also reciprocally influences military strategy. The need for low-risk solutions in hazardous environments has driven the technological pursuit of unmanned technology. For example, the U.S. Navy's USV acquisition initiatives were structured around four size-based divisions based on the length

of the vessel: large, medium, small, and very small [35]. For the purpose of this thesis, the LUSVs are excluded because of their limited deployment within the straits. On the other hand, very small USVs are excluded due to their reliance on a mothership to tether and control them in congested waters, potentially leading to an increased risk of collision and violating international maritime laws.

In response to Country 1's coercion using fishing militias, the U.S. Navy launches a swarm of autonomous USVs to patrol in X Strait. This allows the United States to demonstrate deterrence around the region and conduct ISR such as facial recognition and record any egregious activities performed by the armed fishing militias. The U.S. Navy's USVs do not have any personnel onboard and are capable of conducting the following tasks autonomously: launch and recovery, navigation to X Strait, patrolling, ISR, and firing if necessary. The autonomous USVs' mission is to maintain peace and stability of the region. Should the armed fishing militias conduct any suspicious or egregious activities that undermine stability, the USVs can capture, divert, or conduct time-critical strikes on the armed fishing militias.

After AI detects abnormal behaviors, Figure 7 shows the sailing formation of autonomous USVs when approaching vessels of interest such as the fishing militias. During the approach, the autonomous USVs establish communication either through loud hailer or communication channels that aims to slow down or stop the vessel of interest. During interdiction, the autonomous USVs maintain a safe distance of more than 1 cable away from the vessel of interest. This distance is necessary to avoid collision between the USVs and fishing militias and prevent the fishing militias from boarding the unmanned vessel. The dotted USVs shown in Figure 7 are the force disposition of USVs during interdiction, which allows them to contain the vessel of interest and concurrently divert neutral vessels away from the incident site.

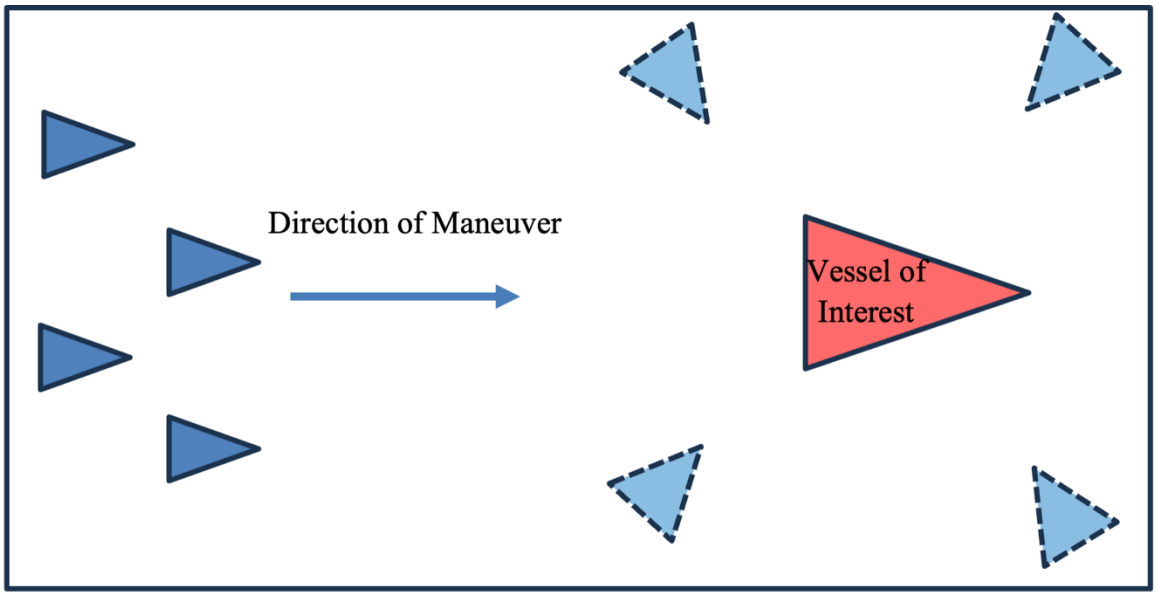

Figure 7. Fully Autonomous USVs Transitioning and Patrolling Formation, Leading to Eventual Interdiction Position. Adapted from [51].

To support the USVs, a fuel supply vessel and manned vessel are 100 nm away from the autonomous USVs' surveillance grid. Although the support ship is capable of refueling four USVs and recovering up to two faulty USVs, this thesis neither discusses the ethical implications nor identifies the hazards of these tasks. The omission of these tasks is due to the complications of man-machine teaming and human errors.

Behind the theater of operation, the autonomous USVs are under supervisory control of the human operator. The operator may choose to intervene in the navigational route, communication lines for verbal warning, or employment of weapons. The employment of weapons includes non-lethal and lethal weapons as well as the precise location where the munitions should hit. As Country 1 is notorious in their disinformation campaign, the commander's intent for this operation is to prevent escalation and limit the use of direct force against the fishing militias, so as to reduce reputational cost.

After examining the potential challenges in the operating environment, we have provided an overview of the operational setting in which the autonomous USVs are designed operate. Next, we use the STPA methodology to identify risks and hazards, so

that we can further develop the situation into operational scenarios that may damage or result in damages caused by the autonomous USVs.

III. SETTING THE STAGE: HAZARD IDENTIFICATION USING SYSTEM THEORETIC PROCESS ANALYSIS METHODOLOGY

In this chapter, we delve into the system theoretic process analysis (STPA) approach by using our hypothetical SOI. It is worth noting that since USVs are military vessels, they are designed to adhere to the safety standards defined by MIL-STD 882E. These standards ensure a certain level of safety in their design through comprehensive safety evaluations. The key difference between STPA and MIL-STD 882E is that the former focuses on "hazard analysis technique" while the latter leans toward "management process" [100]. Our research aims to explore how STPA can complement or enhance these existing safety measures for USV systems that regularly undergo MIL-STD 882E evaluations.

A. WHY ACCIDENT CAUSALITY MODELS ARE USEFUL IN CONGESTED AND LITTORAL WATERS

Unimpeded SLOCs, especially in strategically key straits, remain vital for globalization. Singapore, being one of the busiest waterways in the world, has seen an average of 100,000 ships passing through its strait annually [101]. This figure accounts for a fourth of all goods transported globally, indicating how connected the waterway is and the global reliance on the narrow channel. Another strategic chokepoint is the Strait of Hormuz, which allows passage of one fifth of the world's oil and liquefied natural gas supply [102]. Hence, any disruption to these waterways will directly or indirectly impact the global supply chain of essential goods and the world's economy.

Maritime accidents, in particular ship collisions, occur more frequently in these congested waterways than open waters. In the effort to prevent accidents, literature varies from accident probability studies to accident causality models. An accident probability study is an empirical analysis to identify navigational risks. It maps the maritime traffic pattern and traffic density, based on the Automatic Identification System and collision diameter of the vessel, to warn seafarers off collision hotspots and reduce the probability of maritime accidents [103]. The other approach within our area of interest is the accident causality model. From the earliest publication of Heinrich's Domino Model in 1931 to the

latest addition of STPA, the underlying objective of the accident causality model is to "impose patterns on accidents and influence the factors considered in any safety analysis" [17]. In other words, identify the hazards and mitigate the risks. This prevents unsafe environments and reduces workplace accidents through the engineering of safety systematically.

The evolution of the accident causality model was driven by the pursuit of enhanced safety and risk management across various industries [104]. These models have transitioned from simplistic, single-factor explanations to complex, multifaceted frameworks that account for the factors leading to accidents. For example, the initial Domino Model focuses on identifying a single root cause that topples the first domino, which begins a chain of dominos falling until an accident happens [105]. As the understanding of accidents sharpens, however, accident causality models such as Bird and Loftus's Domino Theory and Swiss Cheese model recognize the multifaceted and multilayered issues that involve management, basic causes (personal and job factors), and immediate causes (bad practices) [106], [107]. With the advent of new technology, digitalization, automation, and higher order decision making, the nature of accidents has changed considerably due to new pathways to loss. The six-step STPA approach, a more robust hazard analysis technique, witnessed growing influence across multiple sectors including our proposed inclusion of autonomous USVs [17], [104].

B. STPA APPROACH TO SYSTEM SAFETY ANALYSIS

This section provides a detailed exploration of employing STPA for assessing and enhancing the safety of autonomous systems such as USVs. Building upon the systems theoretic accident model and processes (STAMP) safety approach, STPA is an innovative hazard analysis method developed by Dr. Leveson at the Massachusetts Institute of Technology (MIT) [17]. By leveraging systems theory, this chapter aims to identify USV hazards and uncover root causes within these complex modern systems [17].

As this study investigates whether STPA can further improve safety for USVs, we will use the STPA Handbook as a primary reference to step through the STPA methodology. This handbook, developed by Leveson and John Thomas, serves as a

comprehensive guide for the practical application of STPA in various complex systems [108]. It is important to emphasize that our examination of STPA should not be seen as criticism toward MIL-STD 882E. Rather, it stems from recognizing that USVs operating independently in complex littoral waters present unique challenges, requiring a thorough analysis of system safety.

At its core, the STPA methodology centers on modelling system interactions through a functional control structure; it allows the user to gain greater insight into potential risks that may be overlooked by traditional safety techniques, such as failure mode and effect analysis [109]. A key distinction of STPA is its emphasis on fundamental, adaptable safety principles rather than mere compliance with prescriptive checklists. This versatile approach makes STPA widely applicable across diverse domains. Importantly, STPA encompasses both software and hardware considerations to enable holistic safety assessment of socio-technical systems. For autonomous vessels like USVs, this means STPA can shed light on potential flaws in the human-machine teaming decision-making process as well as physical reliability issues. STPA's merits have been demonstrated through widespread adoption in various safety-critical fields like aerospace, automotive, and nuclear power [110].

Given its growing track record with evaluating autonomous systems, STPA shows considerable promise as a means of augmenting existing safety protocols for autonomous USVs. By applying STPA's systematic techniques for modeling and analyzing control structures, potential hazards can be identified early and addressed proactively before fielding systems. This section first provides an overview of the STPA methodology and a specialized six-step approach tailored for greater clarity in safety analysis. In this section, we incorporate the information from Chapter II's vignette into our STPA analysis with the intent to align the safety analysis with the operational environment (congested and littoral waters) and challenges that are specific to USVs patrolling the X Strait.

C. THE SIX-STEP STPA APPROACH

While conventional STPA is based on four key steps, this analysis employs an expanded six-step variant, shown in Figure 8. Selecting the expanded variant in this thesis

enables greater thoroughness, granularity, and explicit emphasis on hazard mitigation recommendations throughout the safety assessment process [111].

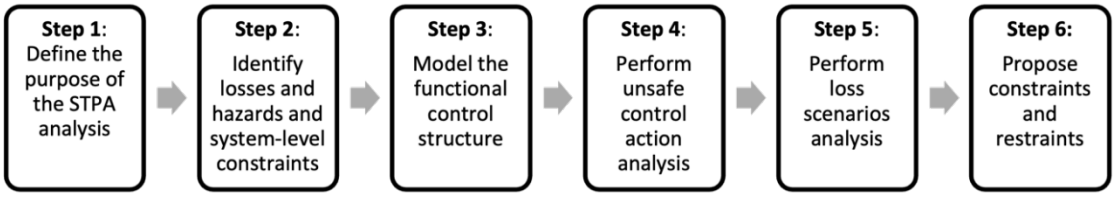

Figure 8. Six-Step STPA Safety Approach

From the six-step STPA safety approach, starting with defining the purpose of the analysis and potential losses upfront provides important direction and bounds the scope. Next, identifying foreseeable hazards throughout the entire system meticulously aids in comprehensive risk capture during the early design phase. The functional control structure is then modelled to gain valuable insight into the complex interactions between system components and illustrates risk pathways. By thoroughly cataloguing the system's potential control actions (CAs) and associated feedback mechanisms, it is possible to gain insight into how unsafe conditions and states can arise. When identified risks are subjected to unsafe control actions (UCA) analysis, explicit traceability between hazards and their potential causes is established. Finally, by directly addressing hazardous scenarios with appropriate constraints and control measures, targeted mitigations close the loop. These six analysis steps, when taken together, allow for a thorough examination of the system's behavior, tendencies for unsafe control, and pathways to increased safety. The sections that follow demonstrate these risk-based techniques for a notional, medium-sized USV case study.

1. Step 1: Define the Purpose

The primary objective of conducting an STPA analysis for autonomous USVs patrolling in the X Strait is to understand the desired capabilities and operational environment in which these USVs are expected to operate. This step also entails capturing

the rationale behind carrying out this safety analysis (i.e., the escalating hostilities resulting from territorial disputes between Country 1 and Country 2). By defining these elements, we purposefully limit the scope of the safety analysis effort in addition to setting it.

This step's primary goal is to recognize and efficiently handle risks and hazards related to USV operations in the demanding maritime environment of the X Strait. These hazards include things that have to do with USVs directly as well as how they interact with changing and crowded marine environments. Unpredictability, resulting from variables like traffic patterns and constantly shifting environmental conditions, characterizes these environments. Moreover, improving the reliability and safety of USVs is a major emphasis of this analysis. These USVs are not only supposed to fulfil their mission requirements, but also to do so in a way that prioritizes safety. It is critical that these safety issues are addressed because of the congested environment in which they operate, where even the smallest error can have serious repercussions. The USVs' CDCA system and communication systems are the two main areas of focus for this STPA safety analysis. These systems are selected with consideration for the complex web of AI algorithms that control their behavior. The STPA approach explores these areas and offers a risk-based view of safety that could significantly improve the current compliance-driven safety frameworks for sea vessels. A well-defined analysis effort at the outset of the STPA is crucial for laying the groundwork for a contextually appropriate safety assessment.

STPA step 1 involves following a systematic approach:

a. Defining Desired Abilities

The main focus is on automated USVs, which are expected to play crucial roles in modern maritime operations. It is essential to define their abilities accurately from the start of the study, ensuring there is a clear and comprehensive understanding of these systems.

b. Establishing the Operational Environment

USVs are specifically designed for operation in crowded and littoral waters. These environments are highly dynamic and complex, presenting unique challenges that need to be addressed.

c. Outlining the Justifications for STPA Analysis

The STPA analysis goes beyond being just a theoretical exercise. There are specific goals driving this effort. Not only does it serve as a potential complement to established safety standards like MIL-STD 882E, but it also provides a structured approach for exploring and explaining core concepts underlying complementary safety methodologies.

d. Determining Scope and Boundaries

It is crucial to set limits on the scope of safety analysis. This includes identifying specific operational scenarios for USVs such as docking and launching procedures, as well as sustained patrolling in designated waters. These constraints ensure that the analysis remains focused and produces practical results.

When it comes to managing risks and hazards, it is important to note that this process aims to identify and handle the various risks involved in autonomous operations of USVs. Improving safety and reliability is a core objective in addition to identifying potential risks. As we move forward with Steps 2–6, our focus is on developing effective strategies to enhance the performance of USVs, especially in challenging congested and littoral maritime regions. STPA offers a valuable risk-based perspective that complements existing safety models for autonomous maritime systems, going beyond mere compliance centric approaches.

2. Step 2: Identify Potential Losses and Hazards

In line with the details outlined in the vignette and the STPA methodology, the second step seeks to meticulously unearth potential losses and associated hazards that the system could confront.

a. Potential Losses

Drawing from the STPA handbook, "A loss involves something of value to stakeholders. Losses may include a loss of human life or human injury, property damage, environmental pollution, loss of mission, loss of reputation, loss or leak of sensitive information, or any other loss that is unacceptable to the stakeholders" [108]. It delineates

circumstances where "stakeholders" or the organization experience a deprivation of something they value. Explicitly, losses could manifest as human casualties or injuries, damage to assets, environmental harm, mission failures, reputational setbacks, breaches leading to the leakage of sensitive information, or any adverse outcome deemed unacceptable by the involved stakeholders. This multifaceted understanding of "loss" forms the foundation for our analysis, directing our focus toward aspects that are of paramount significance to stakeholders in the maritime domain.

To comprehensively identify potential losses and hazards, a diverse team of stakeholders is essential. The ideal team for such analysis would encompass the operators, capability owner, system manager (bridging both engineering and maintenance), project manager, and representatives from the platform and combat systems OEMs. Although resource limitations may make it difficult to put together such a comprehensive team, expert consultations can help make up for any shortfalls. Through harnessing available resources like expert consultations, the depth and scope of the study remain uncompromised.

In the current STPA exploration, from the maritime safety lens, the identified potential losses are as follows:

L-1: Loss of human life or significant injury: This includes both direct and ancillary human casualties. Despite the unmanned nature of USVs, potential collisions with other manned vessels or inadvertent disruptions in operations are a significant concern. Considerations should also incorporate situations involving naval technicians or contractors onboard during maintenance or sea trials.

L-2: Loss or damage to the USVs: Representing both a financial and strategic asset, the USVs' severe damage or complete loss would have a significant impact on the mission success.

L-3: Loss or damage to external entities: This represents the loss and/or damage that the USVs might inflict on other vessels.

L-4: Loss of operational effectiveness: This loss pertains to the failure of USVs to accomplish designated operational tasks due to reasons other than physical damage. This includes software malfunctions, communication disruptions, or inadequate responses to dynamic environmental factors that may render the USV unable to complete its mission effectively. Although the USV remains intact, its operational purpose is compromised, resulting in a failure to deliver the mission's strategic or tactical outcomes.

L-5: Loss of sensitive information: This refers to unauthorized access or compromised information about the USV operations, tactics, and technological data. In this fast-changing cyber threat landscape, ensuring the integrity of such information is important. An information breach gives adversaries an advantage.

L-6: Environmental Loss: This refers to the USVs' impact on the marine environment. Losses may include oil spills or debris from unintended incidents or accidents. Such losses harm the environment and are damaging to the Navy's reputation.

L-5 and L-6, initially identified as part of the losses, were later streamlined out, refocusing the analysis on our core questions on safety and ethics. Specifically, L-5, pertaining to the loss of sensitive information, can potentially find its place in a cybersecurity-centric study, making it a potential avenue for future research.

b. Potential Hazards

After identifying the losses, our next endeavor is to delve into the hazards. The STPA handbook furnishes us with "three basic criteria for defining system-level hazards" [108]:

1. Hazards are system states or conditions (not component-level causes or environmental states).

2. Hazards can lead to a loss in worst-case conditions.

3. Hazards must describe states or conditions to be prevented.

It is necessary to determine the systems to be studied, as well as the system boundary in order to identify the mitigatable system-level hazards. This approach to hazards requires recognizing states or conditions within our control that could lead to a mishap. For instance, while rough sea conditions or high maritime traffic are not inherently hazards, the inability of USVs to maintain a safe distance from other maritime entities could signify the inception of a hazard, culminating in an unfortunate event. Consequently, after demarcating the system and its boundaries, the subsequent task is to categorize system-level hazards. The relationship between systems, system boundary, and the environment is shown in Figure 9.

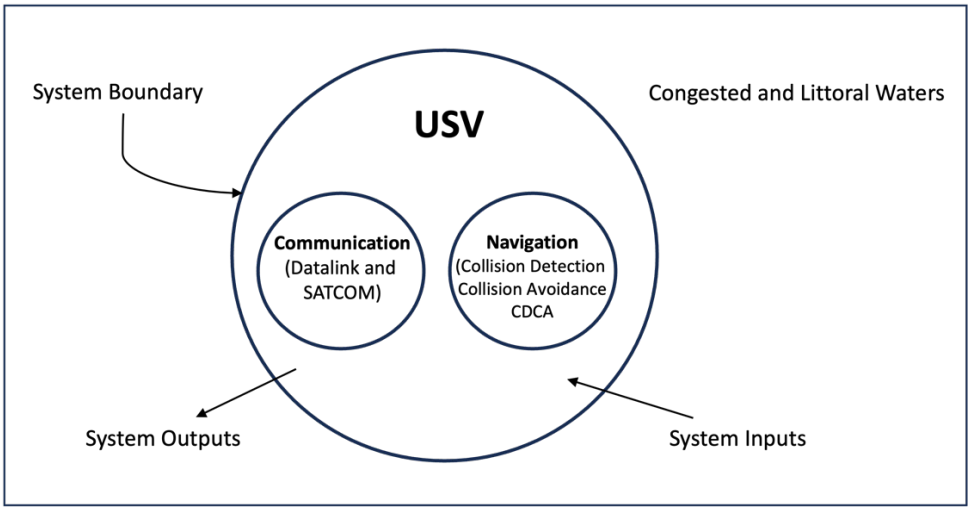

Figure 9. Relationship between Systems, System Boundary, and the Environment

Understanding the relationship between the systems, boundary, and operating environment allows us to recognize system states or situations that, under the most adverse environmental conditions, might result in a loss. The following are the determined hazards:

H-1: USV leaves patrolling sector, runaway, or stalls [L-1, L-2, L-3, L-4]

H-2: USV does not maintain safe distance from vessels and other obstacles [L-1, L-2, L-3, L-4]

H-3: USV performs unexpected operations [L-1, L-2, L-3, L-4]

H-4: USV experiences disorientation in situational awareness [L-1, L-2, L-3, L-4]

3. Step 3: Model the Functional Control Structure

Building upon the insights gained from the vignette and the identification of potential losses and hazards, the next step involves modeling the control structure of the USV systems patrolling in the X Strait. We use this control structure analysis to gain valuable insight into the complex interactions between system components and to identify potential risk pathways within this intricate environment. In this context, the control structure encompasses the various CAs, feedback mechanisms, and autonomous decision-making processes that govern the behavior of the USVs. By cataloguing these components comprehensively, we understand how they contribute to system safety and identify potential unsafe conditions and states. With the unique challenges presented by the maritime environment of the X Strait, where congested waters and unpredictable maritime traffic patterns are prevalent, the control structure modeling should also consider such complexities. We do this by analyzing the specific mission activities linked to the identified potential hazards, which are carried out by the USVs during operations.

1. Launch and Recovery: It is crucial to make sure that USVs have reliable systems in place for both slipping off and docking at the wharf in their operational base. Mishandling, system malfunctions, or unfavorable weather conditions all pose risks here.

2. Navigation/Patrolling: USVs must ensure their collision detection and avoidance systems are operational while they navigate, especially in congested and littoral waters. These waters are inherently unpredictable, due to variables like traffic patterns and shifting weather, so careful hazard anticipation and management are necessary.

3. Activity at Mission Objective: After reaching their patrolling sectors, which involves transiting through this challenging environment, USVs must carry out assigned operations, which may entail gathering intelligence, conducting reconnaissance, or handling other specialized duties. It is imperative that they function safely and effectively during these missions.

Additionally, the control structure must reflect the role of AI systems and algorithms in facilitating autonomous decision making, as outlined in the vignette. The modelling of the control structure provides a clear visualization of how the USV system operates, how it interacts with its surroundings, and how it responds to various inputs and conditions. This step is pivotal in uncovering potential risk pathways and understanding how unsafe conditions and states can arise, ultimately contributing to the development of targeted hazard mitigation measures. Taking into consideration the vignette and AI-driven autonomous decision-making, in Figures 10 and 11, we derived the following functional control structure diagrams:

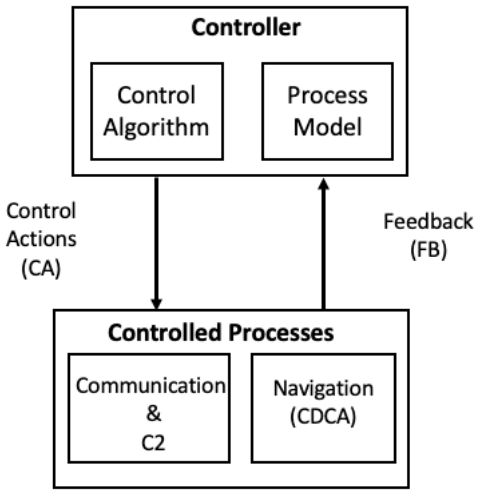

Figure 10. High-Level Functional Control Loop

Figure 10 provides a visual representation of the USV's functional control structure, illustrating the feedback loop that enables continuous adaptation and safe operation in the challenging (unpredictable) environment of the X Strait. This model forms

the backbone of our STPA analysis, highlighting the interplay between autonomous decision-making processes and human oversight in ensuring mission success. At the top in Figure 10, we have the "controller." The controller processes the operations of the USVs within the X Strait. The controller consists of two important elements, and both these elements work together to direct and coordinate actions undertaken by the USV. They take feedback to adapt and respond effectively to real time situations.

1. Control algorithm. This element represents a set of decision-making protocols that guide the USV's autonomous actions. It includes rules and computations that help ensure that it aligns with its mission objectives and safety requirements.

2. Process model. This element contains operational parameters and environmental models that provide information to the control algorithm about both the USV itself and its surroundings. This includes dynamic maritime conditions within the X Strait.

At the bottom in Figure 10, we have the "controlled processes." They represent the execution of CAs and are divided into two main processes.

1. Communication & C2 (command and control). This process manages the transmission and reception of data and commands, ensuring continuous communication between the USV and its operational base for command and control purposes.

2. Navigation (CDCA): This process manages navigation information while also focusing on collision avoidance. This is particularly important when navigating through complex and congested waters in the X Strait.

There are two lines interacting between the controller and controlled processes.

1. Control actions. The controller issues CA to influence the controlled processes. These CAs involve the necessary operational commands for the USV to carry out its mission, such as maneuvering, adjusting speed, and following engagement protocols.

2. Feedback (FB). This is the data that controlled processes send back to the controller. This feedback loop is crucial for enabling real time control and changes in algorithms and process models by allowing them to adapt to dynamic operational environments.

Figure 11 is a detailed representation of the different levels of control and oversight involved in managing the navigation and communication functions of the USV. This figure provides a more in-depth view of the high-level functional control loop shown in Figure 10. For example, the controller shows the HQ Command Center, which holds the highest level of control authority. This is where information about global tracks is compiled. It serves as a decision-making hub, taking into account both the USV's status and the broader operational environment. The USV Automation is the autonomous capabilities of the USV, which includes automated navigation controls crucial to its operations, as well as processing sensor data. The USV Automation combines information on USV tracks and sensor data with external data streams to make autonomous decisions. Located on the left side, perpendicular to the flow from the controller to the controlled processes, are what we term "manual navigation controls." These interventions are driven by humans and serve as a mitigation measure. They are not analyzed as CA due to their specific use in interdiction scenarios, which is beyond the scope of this thesis.

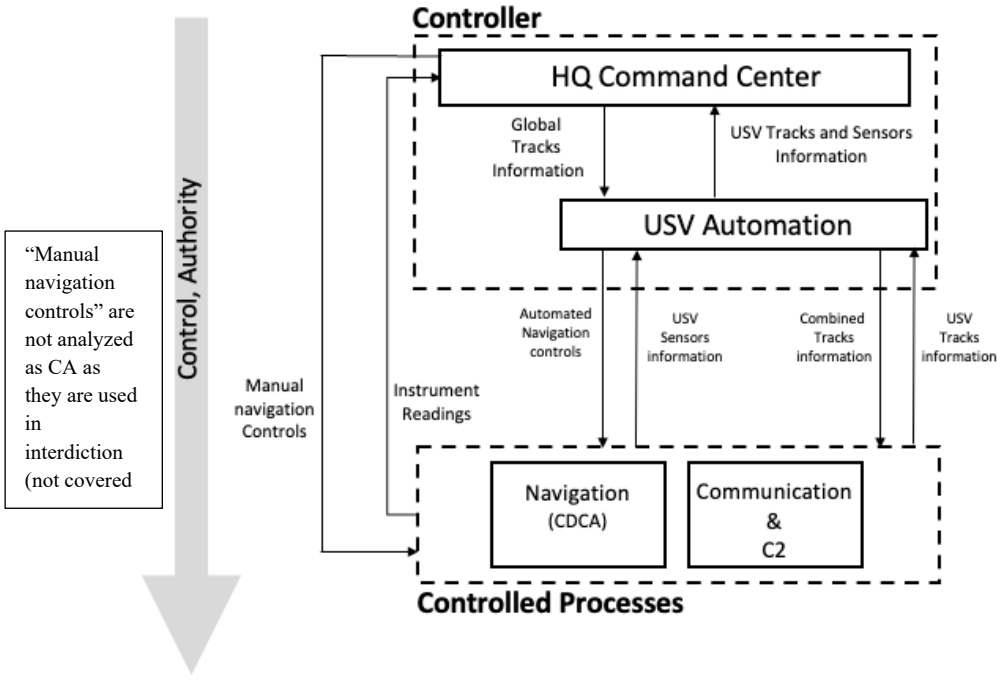

Figure 11. Notional USV Hierarchical Control Structure (Navigation and Communication)

4. Step 4: Perform Unsafe Control Action Analysis

Using the control structure shown in Figure 11, all possible CAs related to each activity were listed during the UCA analysis. Following that, scenarios were found for every single action, where the action leads to unsafe conditions and is also traceable to hazards. The STPA handbook prescribes the four ways a CA can be unsafe, which are shown in Table 1 [108].

1. Not providing the CA leads to a hazard
2. Providing the CA leads to a hazard
3. Providing a potentially safe CA but too early, too late, or in the wrong order
4. The CA lasts too long or is stopped too soon

Table 1. Unsafe Control Action Analysis

Control Actions	**Not providing causes hazards**	**Providing causes hazards**	**Too early, too late, out of order**	**Stopped too soon, applied too long**
Provide global tracks information (no imminent threat as organic components would have performed their functions)	UCA-1: Failure to provide global tracks information compromises situational awareness, increasing the risk of navigation errors [H-2]	UCA-2: Incorrect global tracks information could misdirect USV actions, leading to unsafe maneuvers or engagements [H-2, H-4]	UCA-3: Delayed or out-of-order tracks information prevents timely response to dynamic conditions [H-2]	UCA-4: Incomplete information may cause prolonged uncertainty, affecting decision making [H-2, H-4]
Provide combined tracks information (imminent threats exist)	UCA-1: Lack of combined situational data could lead to a misunderstanding of the operational environment [H-1, H-2, H-3, H-4]	UCA-2: Wrong combined information may trigger inappropriate USV actions [H-1, H-2, H4]	UCA-3: USV provides combined tracks information too early with no update; target changes course overriding previous safe condition [H-2, H-4] UCA-4: USV provides combined tracks information too late [H-2]	UCA-5: Incomplete information may cause prolonged uncertainty, affecting decision making [H-1, H-2, H-4]

Control Actions	Not providing causes hazards	Providing causes hazards	Too early, too late, out of order	Stopped too soon, applied too long
Execute automated navigation controls (steering and speed regulation)	UCA-1: Failure to execute automated navigation controls [H1, H2, H3]	UCA-2: Executing automated navigation controls with suboptimal level of steering adjustment and speed regulation [H1, H2, H3]	UCA-3: Executing automated navigation controls too early [H1, H2, H3] UCA-4: HQ Executing automated navigation controls too late [H1, H2, H3]	UCA-5: Executing navigation controls for an extended period can lead to persistent unsafe states [H1, H2, H3]
Acquire and process sensor data	UCA-1: Failure to acquire sensor data leads to blind spots in environmental awareness [H-2]	UCA-2: Acquiring flawed sensor data misinforms the operational picture [H-2, H-4]	UCA-3: Acquiring sensor data out of necessary sequence impairs real-time response [H-2]	UCA-4: Halting the acquisition of sensor data too soon leaves the USV without current updates [H-2, H-4]

In STPA, control actions are the commands that the system's control algorithms issue to control the behavior of the controlled process. In our analysis, we focused on four CAs: "provide global tracks information," "provide combined tracks information," "execute automated navigation controls," and "acquire and process sensor data." These were selected due to their direct impact on the USV's operational capabilities in patrolling, navigation, and decision making in response to imminent threats. We walk through

"provide global tracks information" as an example to explain the rationale of each unsafe condition.

1. Not providing causes hazards (UCA-1): If the USV does not provide global tracks information during patrolling, it might miss critical situational awareness cues. This could lead to navigational errors or collision risks, thus linked to Hazard H-2, which pertains to maintaining a safe distance from obstacles.

2. Providing causes hazards (UCA-3): If the USV provides incorrect global tracks information, it could misrepresent the location of other vessels or obstacles, leading to erroneous navigational decisions and increasing the risk of collision (Hazards H-2 and H-4).

3. Too early, too late, out of order (UCA-2): If the USV provides global tracks information too late during patrolling, it may not allow enough time for the USV to take corrective actions to avoid hazards, again relating to Hazard H-2.

4. Stopped too soon, applied too long (UCA-4): Providing incomplete global tracks information could result in an incomplete picture of the operational environment for the USV, potentially leading to unsafe operational conditions (Hazards H-2 and H-4).

5. Step 5: Perform Loss Scenarios Analysis

To identify situations in which the autonomy of the USV may result in unsafe conditions or contribute to hazardous events, a thorough understanding of UCAs and the USV's operational conditions described in the vignette is crucial when creating loss scenarios. When these two factors are combined, we produce a comprehensive analysis that captures the nuances and hazards related to USVs. Through careful examination of the CA and consideration of its consequences when such CA is not provided, provided incorrectly, or performed inappropriately, we are able to predict potentially dangerous states that may emerge while the USV is operating. The vignette gives particulars about

the surroundings, the goals of the mission, the USV's capabilities, and the ways in which the navigation and communication systems interact. Using the vignette and its operation conditions helps to contextualize and provides loss scenarios that are practically relevant, by taking into account the real-world circumstances and potential difficulties the USVs may encounter.

While ethical considerations are not part of the steps for STPA approach, we have included the ethical consideration as part of the loss scenarios. To ensure that the technology is used responsibly and ethically, ethical considerations provide insight into accountability, responsibility, and the proper way to deploy autonomous USVs. It is critical to uphold ethical integrity in the military because it affects both operational and strategic interests. A breach of ethical standards can have negative repercussions, including harming people and communities, damaging the military's reputation, and eroding public trust. The military unit must act with a strong moral compass to make sure that the autonomy given to USVs does not violate moral standards or human values. Therefore, it is advantageous to include ethical issues in the loss scenario analysis in addition to technical and operational parameters. The ethical considerations allow us to reduce risks, improve safety, and make a lead into Chapter IV on the moral application of autonomous systems in military operations.

Three broad categories have been identified when the fully autonomous USVs may encounter loss scenarios. First, in the category 1 of system related faults, with the increasing autonomy of USVs through rapid progress of AI, the USVs may execute tasks that deviate from the norm. These technical fault that occur independent from human control can result in the damage of the USV itself or the objects surrounding it [112]. Second, in the category 2 of gray-zone engagement concepts, in circumstances when direct confrontation or response is likely to trigger escalation, the USVs can be used as a mere instrument to inflict intended harm. A programmer may devise actions that appear as accidents, allowing the nation to achieve its objectives while attributing responsibility to the AI, thereby minimizing the likelihood of escalation. Third, in the category 3 of pushing the boundary, the lack of crew onboard the USVs enables commanders to test the capabilities of the USVs beyond limits that would constrain manned ships.

In the first scenario of category 1, shown in Table 2, we examine "drop anchor during launch and recovery," which is one phase of USV operation. This scenario highlights a situation where the autonomous navigation controls fail to accurately execute launch or recovery maneuvers due to conflicting tasks. The USV may erroneously initiate an anchor drop while still in motion, potentially due to miscommunication or misinterpretation of sensor data. This unsafe act not only risks damage to the USV and its equipment but also poses a hazard to nearby personnel who may be within the operational vicinity.

We have factored in ethical concerns based on the principle that the autonomous system should not cause harm through its actions. For each loss scenarios, we have identified a specific normative claim which violates the loss scenarios. This scenario is identified under Category 1: System Related Fault, as it demonstrates the tangible consequences of a system's inability to navigate the nuances of complex operations, despite the sophisticated interplay of algorithms and ML models designed to automate such tasks. In this scenario, the ethical imperative is clear: to ensure that the system's autonomy does not lead to outcomes that could undermine the ethical principles inherent in military operations.

Within each category, the loss scenarios are shown in Table 2.

Table 2. Loss Scenarios and Categories

Loss Scenario	Description	Loss and Ethical Considerations
Category 1: System Related Fault		
1.1 Drop anchor during launch and recovery (conflicting tasks)	Automated navigation controls fail to provide proper steering and speed regulation during recovery, leading to the unsafe condition where a USV drops anchor while still in motion. This could lead to equipment damage and potential harm to nearby personnel.	Ensuring the safety of personnel and equipment is paramount. Proper fail-safes and manual override capabilities should be in place to prevent such scenarios.
1.2 Speed up instead of slow down during imminent navigation risk	USV's automated navigation controls incorrectly interpret sensor data, causing it to speed up when an imminent navigation risk is detected, instead of slowing down.	The USV's decision-making algorithms must be transparent and reliable. Accounts of mishap and near misses must be made available to aid investigation without any intention to deceive.
1.3 Crossing territorial water	USV's automated navigation controls do not recognize or respond (avoid crossing over) to territorial water boundaries, leading to potential international incidents.	Respecting territorial boundaries is crucial to avoid escalating international tensions. The USV's algorithm must ensure adherence to international laws and rules of the road. If faced with a "trolley car"-type problem, visual recording, backed with electronic chart, must commence and be made available to justify the means to the ends.
1.4 Wrongly identifies vessel of interest	USV provides wrong combined tracks information, misidentifying a vessel of interest, potentially leading to wrong engagement.	Ensuring correct target identification is vital to prevent harm to innocent vessels and to maintain trust in autonomous systems. Physical or hardware constraints must be in place for any further actions to be made.

Loss Scenario	Description	Loss and Ethical Considerations
1.5 Fires at target ship or person without valid reason	USV takes engagement action without valid reason due to a misinterpretation of sensor data or a malfunction in its decision-making algorithm.	The use of force must be justified and captured through video recordings. Systems that can initiate force must have stringent checks and balances to prevent misuse or errors.
Category 2: Gray-Zone Engagement Concepts (Pushing the Blame to AI)		
2.1 Collided with enemy coastal structures/ warship/ merchant/ civilian vessel causing navigational disruption	USV's automated navigation controls (CDCA) fail to avoid collision, causing navigational mishaps.	Ethical question of whether this should this be allowed so as to gain military advantage. If the system is used as a mere means and as an instrument of harm, the responsibility must be clearly defined. This is to ensure that responsibility is not placed solely on the "shoulders" of AI (pushing the blame).
2.2 Swarming formation in a disoriented manner causing navigational hazard	USVs' command and control algorithms malfunction, causing them to swarm in a disorderly manner instead of in a formation, which leads to navigational hazards.	Ethical question of whether this should this be allowed so as to gain military advantage. Understand the advantage of mobility and make trade off on safety. Ensuring the predictability and safety of autonomous swarm behaviors is crucial to prevent mishaps and maintain safe navigation.
2.3 Anchor landed on enemy submarine cable/ electricity grid	USV's automated navigation controls fail to recognize underwater infrastructure, causing damage due to anchor landing or dragging underwater cables.	Ethical question of whether this should this be allowed so as to gain military advantage. USV should be programmed to minimize environmental and infrastructural impact, ensuring responsible navigation and operations.

Loss Scenario	Description	Loss and Ethical Considerations
2.4 Warning shot hitting fishing militias due to strong wind/high sea state	USV misjudges environmental conditions, leading to a wrong target hit, escalating tensions between countries.	Ethical question of whether this should this be allowed to happen so as to gain military advantage. USV must be able to accurately assess environmental conditions to ensure the accuracy of weapons.
Category 3: Pushing the Boundary		
3.1 More aggressive interdiction	USV takes an overly aggressive stance during interdiction, escalating the situation.	Ensuring the military action is justified, else, de-escalation and proportionality in engagements is vital to prevent unnecessary conflict or tension.
3.2 Persistent surveillance in strategic locations	USV engages in persistent surveillance, potentially escalating tensions.	Balancing the need for surveillance and respect for international norms is necessary to maintain a stable international environment.
3.3 Coordinated swarming to overwhelm enemy defenses	USVs participate in a coordinated swarming that overwhelms enemy defenses, potentially leading to an escalated conflict.	Ensuring that USVs do not contribute to unintentional escalation of conflict is crucial.
3.4 Work in tandem with helicopter and manned submarines	Poor coordination with manned assets, leading to potential friendly fire or mission failure.	Ensuring tight integration and clear communication between USV and manned systems is vital.

6. Step 6: Propose Constraints and Restraints

Leveraging the loss scenario analysis carried out in Step 5, this section covers the step of formulating constraints and restraints. This step presents an opportunity to exert control over UCAs through various strategies, including design modifications, the introduction of new controls, procedural updates, or the implementation of audits. These measures are important for guaranteeing the autonomous USVs operate securely,

effectively reducing the risks tied to previously pinpointed UCAs and potential loss scenarios. The constraints and restraints serve as protective mechanisms, delineating the operational parameters of the system and mandating specific behaviors to avert mishaps. The adoption of these safeguards is instrumental in enhancing the safety of autonomous USVs, directly addressing the vulnerabilities identified in preceding steps and fortifying the system's overall resilience and integrity.

Building upon the analysis of potential losses in Step 5, this section explores the development of constraints and restraints. It does so with a broad approach, however, avoiding getting too implementation specific. This allows for a comprehensive overview, encompassing various safety measures while acknowledging that the specific solution details may differ based on mission requirements, technological advancements, or unexpected operational challenges. By focusing on two main categories: (1) system-level constraints and (2) operational limitations, we aim to establish clarity on the protective measures needed to enhance the safety and reliability of autonomous USVs.

When establishing system-level constraints, the goal is to strengthen the reliability and accuracy of the navigation and control systems in USVs. Reflecting on the hazards identified, such as the potential for collision during complex maneuvers (H-2), our constraints aim to reinforce safe operational practices. By stipulating a safe operational distance at all times, particularly during launch, recovery, and mission-critical tasks, we address the UCAs related to automated navigation controls, mitigating risks similar to the scenarios where a USV might erroneously drop anchor while still in motion (Loss Scenario 1.1). In terms of communication, it is crucial to establish an uninterrupted connection between the USVs and the command center. This involves addressing any UCAs related to providing global and combined track information, as well as mitigating the risks of communication failures. In addition, it is important to implement remote monitoring capabilities for the USVs' system health. This includes automated alerts and fail-safes that provide warnings in advance and enhance the resilience of the system.

On the operational side, it is essential for the USVs to strictly follow assigned patrol boundaries and avoid entering unauthorized territorial waters. By doing so, we can directly address any UCAs related to crossing territorial boundaries and prevent potential

international incidents, a direct response to hazards like crossing territorial waters (Loss Scenario 1.3). In extreme situations where the USV finds itself in an unsafe state, designed restraints will activate and transit into a safe mode; our design incorporates fail-safes to transition into a secure mode, echoing the need for manual overrides to prevent incidents such as misidentification of vessels of interest (Loss Scenario 1.4). For example, it can navigate toward the nearest secure location and maintain its position or return to base. To ensure the possibility of immediate human intervention when needed, it is vital to incorporate an emergency override or, if required, a kill function. Such measures highlight the critical role of human oversight and intervention in safety-critical systems. This allows us to intervene promptly if the USV is about to engage in an unsafe action or has entered a hazardous state. This is especially relevant in scenarios like Loss Scenario 1.5, where rapid corrective action is imperative in the event of unjustified engagement.

The broad discussion on constraints and restraints serves as an additional step toward creating a safer operational environment for autonomous USVs. It sets the foundation for developing more detailed safety protocols tailored to specific contexts in the operating environment. From the phases of operation and notional operational vignette developed in Chapters I and II, we have stepped through STPA and identified the hazards based on operating scenarios. These steps allowed us to develop the events specific to the vignette. The nuances of human oversight and the ethical implications of intervention are further explored and discussed in Chapter IV.

IV. RESPONSIBLE ARTIFICIAL INTELLIGENCE FOR USVS

This section explores the meta-ethical questions surrounding the usage of autonomous systems, specifically in the areas of command, control, and responsibility. We also explore the applications of accountability, transparency and explainability in the need for reliable AI systems and the current efforts in AI governance. As the deployment of autonomous systems in the Navy has not yet proliferated, there are limited examples to draw upon. Therefore, we use other autonomous systems such as aircraft and cars as a reference to illustrate our points.

A. WHY ADOPT A REALIST APPROACH TO ETHICS MATTERS?

To have a meaningful discussion on ethical issues, we must first understand the difference between ethical "relativism" and "realism" before we lay out the reasons for taking a realist approach for this thesis. Most philosophers supporting ethical relativism credit Herodotus's *The History* for its descriptive illustration of different cultures and their different moral codes [113]. Herodotus recounted a story by the king of ancient Persia, Darius, and how he called upon the Greeks, who traditionally practiced cremation. He asked them how much would it take to eat their deceased father. Feeling surprised, the Greeks replied no amount of cash would make them do it. Subsequently, Darius, called upon the Callatiae, one of the Indian tribes, who traditionally practiced cannibalizing their deceased parents, and asked them how much would it cost to burn their deceased father. Similarly, they were caught by surprise and exclaimed that it was a terrible thing to mention. This anecdote is one of many examples relativists would argue illustrates that the judgement of ethics is relative, meaning that ethical principles and values can vary depending on cultural, societal, or individual perspectives. In relativistic ethics, there is neither an ultimate truth nor right or wrong; it is entirely culturally bound [114].

Ethical relativism tolerates different belief systems and rejects moral truth or justifications [115]. Critics such as James Rachels argues that this thinking leads to three consequences [116]. First, it is impossible to evaluate the moral status (superiority or inferiority) of our society with another society because these moral codes will be perceived

merely as different [116]. Second, right or wrong is determined by social standards and we cannot criticize these standards even though they are our own because—a relativist argues—from a moral standpoint who is to say which standard is better [116]. Third, moral progress cannot exist because the concept connotes that there is an advancement in social changes [116].

A realist, on the other hand, often adopts a more assertive view on ethical issues and judges the matter as right or wrong. Their judgements are grounded in an objective and reasoned approach, with claims supported by empirical data. A realist strives to remove biases and personal beliefs, basing their arguments on observable facts and rational deductions. This approach provides clear and universal answers to ethical questions, which hold true across various contexts and cultural differences [114]. More importantly, the argument can be "verified" and "withstand rational criticism" [114].

This thesis desires to reach an absolute position on ethical issues related to autonomous USVs. Although the task of weighing the intangibles such as moral cost and benefits may seem insurmountable, we hope to shed light on who should be responsible and what is the right way to use autonomous USVs ethically. One may criticize that taking a stance on ethical judgement implies having superior knowledge of all the facts before weighing each viewpoint methodically to reach a single argumentative position [114]; the point here is not having superior knowledge, however, but the process to analyze the issue rationally with defensible evidence to support the argument.

Utilitarianism is a concept in realists' normative ethics to evaluate whether actions are right or wrong based on the maximum happiness they yield. British philosopher Francis Hutcheson argues that "happiness" is the overarching objective of life; thus, actions can be judged by their effectiveness in attaining that goal [117]. He wrote "that Action is best, which procures the greatest Happiness for the greatest Numbers; and that, worst, which, in like manner, occasions Misery" [117]. Using this concept, Jeremy Bentham coined what we now know as utility or utilitarianism [22]. He theorized that any action will consequentially produce either pain or pleasure to concerned parties. In the context of morality, it differentiates the action as either right or wrong. It is also known as the

consequentialist approach because it is a type of normative ethical theory that assesses the morality of an action based on the resulting consequences.

B. MILITARY INTEREST IN MAINTAINING ETHICAL INTEGRITY

Maintaining ethical integrity within the military is essential for both strategic and operational interests; failing to do so not only affect the individual but compromises the entire military outfit. A prime example is the trial of then-Navy Sea, Air, and Land (SEAL) Chief Petty Officer, Eddie Gallagher. He was accused of stabbing a 17-year-old Islamic State of Iraq and Syria (ISIS) fighter and circulating a picture of himself and the corpse to his friends. Even though he was acquitted of murder and charged with "the violation of posing for photographs with a dead war casualty," it greatly influenced public perception and the international community [118]. To that end, Navy leadership publicly acknowledged the U.S. Navy elite SEALs has problems with "character and ethics" and called for the comprehensive review on these aspects [119]. The failure to maintain moral foundation universally on the battlefield erodes the public's trust toward the military, even if they are an elite fighting unit who have already earned their respect.

Technology not only shapes our ethical responses but also influences the degree of accountability and responsibility individuals feel for their actions. The "trolley car" thought experiment is a simplistic idea that shows people are willing to sacrifice one person to save five men by pulling the track changing handle, because this outcome is not only morally permissible, but a morally right action than letting the trolley car continue its path to kill five men by doing nothing. When faced with a separate moral dilemma, now the trolley is charging towards five people. The only way to save the five men is to push the fat man off the bridge effectively stopping the trolley before the trolley reaches the five men. While the concept of sacrificing one to save five men is the same, people are less willing to push a man off the bridge than to pull the track changing handle to save five men from imminent danger. This provides evidence that people are more likely to act ethically when they are "intimately connected to the decisions they are making" [120]. In the first trolley car problem, no harm is intended, people are more likely to choose pulling the handle because sacrificing one is the unintended consequence of saving the five men, whereas actually

pushing a man off a bridge is when harm is intended and thus a more accountable act. Even if the action leads to a good outcome of saving five men, the act is morally wrong when deliberate harm is intended. To further illustrate this argument in the military context, a soldier, out in the field, will likely think twice before pulling the trigger on a young child, armed with a gun, exploited for terrorism activities. The act of killing the young fighter, no matter what the cause is, is an accountable act when harm is intended. No matter what course of action the soldier chooses, he must bear the responsibility for the decision and live with its consequences.

Similar to pushing the man off the bridge, the action of killing the young fighter evokes emotion, and the ethical implications are immediate and personal [120]. In contrast, a soldier, in the command center, operating an armed aerial drone or missile system remotely may experience a degree of detachment from the consequences of his actions. The physical separation created by technology can reduce the soldier's perception of personal responsibility, such as hitting a button to fire a missile, potentially making it easier to make difficult decisions [121].

C. ETHICS IN A TECHNOLOGICAL ERA

As technology advances, a higher level of autonomy is introduced into decision-making processes. Autonomous systems, including USVs, may operate with varying degrees of independence, from basic human oversight to fully autonomous systems. These autonomous systems may reduce human error caused by impaired judgement. For example, studies have shown that fatigue, stress, and anxiety increase the probability of human error and are often the root cause of many accidents [122]. Likewise, such consequences can be detrimental if human error violates the principle of harm, especially in congested and littoral waters when non-combatants are injured. Even with this distinct advantage, ethical challenges do arise and even tend to intensify as the level of autonomy increases. For instance, armed autonomous USVs can deviate from traditional warfare dynamics, as they can be designed to navigate and engage targets without direct human control. This raises ethical questions about the delegation of lethal force to machines. Scholars such as Aaron

Johnson and Sidney Axinn [11], Noel Sharkey [123], and Shannon Vallor [124] are some of the strong proponents of human involvement to end a life rather than machines.

Apart from armed autonomous USVs causing commanders to contend with ethical issues, drivers in autonomous vehicles navigating through day-to-day situations share similar concerns. These situations can be as simple as approaching a demarcated area where pedestrians may cross, executing a turn amid oncoming traffic, or maneuvering through bustling intersections. Johannes Himmelreich argues that unlike humans, autonomous vehicles cannot make intuitive choices [125]. The algorithm programmed to these vehicles does not allow flexibility to changing traffic conditions. The first ethical question at the policy level revolves around the safety standards autonomous vehicles should adhere to. Should the government prioritize imposing minimum safety standards on manufacturing companies, encourage companies to share safety features with competitors, or allow a consumer-driven initiative where market forces determine the standards? Second, the interaction between the safety algorithm of the vehicle and the ecosystem of the other road users has ethical implications at the social and operational level. To what extent should safety be prioritized so that it will not impede traffic flow? These situations may not present imminent danger and evoke emotions like the "trolley car" thought experiment, but they highlight the intricacies of navigating ethical considerations in the realm of autonomous vehicles in the technological era. The lack of transparency and pushing the blame to drivers are the root causes of the distrust of autonomous vehicles [126], [127].

In today's context, autonomous vehicles are no longer theoretical thought experiments but publicly assessable transportation with direct societal impact. This is exemplified by the industry leader, Tesla. The threat caused by self-driving cars is clear and present. After the widespread implementation of Tesla's "Full Self-Driving" software from 12,000 vehicles to approximately 400,000 vehicles within a year, fatalities in the United States alone have increased from 3 to 17 and authorities have recorded an astonishing 736 crashes [128]. Various U.S. and overseas sources report on the erroneous decisions autonomous vehicles make when faced with day-to-day situations [129], [130], [131]. For example, during instances that require the autonomous vehicle to stop, the

vehicle, instead, chooses to accelerate. Furthermore, these numbers did not take into consideration Teslas exported globally, where such accidents may not have recorded. Tesla publicly denied earlier allegations of unintended acceleration, however [132]. Inquiries into accidents and incidents involving Tesla's autonomous function, carried out by the National Highway Traffic Safety Administration, consistently attributed fault to the drivers rather than the automotive manufacturing giant [126], [133], [134], [135]. While these tragedies are a stark reminder that autonomous systems have not reached full maturity, the surviving family members not only face a deep sense of grief but have to contend with no compensation. The ethical question of who is responsible for accidents involving autonomous systems has far-reaching legal consequences that also need to be addressed.

The assignment of responsibility becomes complex in accidents involving autonomous systems when there is no negligence on the part of the human operator. This ambiguity increases the difficulties of conducting ethical analysis when accidents occur without clear agency to be held responsible, assuming all accounts are factual. In the case of Air France 447, the pitot tube froze, causing the autopilot to be disengaged [136]. The control was handed back to the pilot. Around five minutes later, the plane fell into the Atlantic Ocean, resulting in the death of all 228 personnel onboard. The freezing of the pitot tube seems to be a case of *force majeure* leading to a series of unfortunate events. The pilots, represented by Air France, and autonomous system, represented by Airbus, were both found to be not guilty [137]. Legal issues, compensation, and emotions aside, the ethical question of who is to be responsible for this accident remains highly debatable. This accident raises several interesting questions that puts Airbus, Air France, Thales (French manufacturer of the pitot tube), and pilots liable. If Airbus's autonomous system were to be responsible, there are three queries that need to be answered. First, should Thales be faulted for the defective pitot tube? Second, should the designer be blamed for the stall alarm originally designed to warn, but which instead became a source of confusion? Third, should Airbus be faulted for not demanding Air France change the defective pitot tube? If Air France was to be responsible, depriving the pilots from manual control training at cruise altitude and procrastinating to change the pitot tube are the two most plausible reasons [138].

During the trial, both Airbus lawyers and prosecutors blamed the deceased pilots for the crash. We argue that it is the pilot's responsibility to ensure safety of the aircraft, but it is not their fault as the outcome was unforeseeable [136]. The pilots performed their job duly; actively salvaging the situation even though they had less than 5 minutes of reaction time. The autonomous system in autopilot mode handed over control to the pilots when the readings of the outside air pressure became abnormal. The pilots may not have fully comprehended the sequence of events that led to the chaotic situation, however. Some may ask, how long does it take for the autonomous system to hand over control to the human? When is the pilot really "responsible"—is it 1 second, 1 minute, or other units of measurement? Madeleine Clare Elish termed this as moral crumple zone [139]. A crumple zone of the car is a sacrificial part that increases the survivability of the human driver during a crash. The pilot, in this case, is a moral crumple zone of the AI, when the "responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system" [139].

Objectively, we believe that tasks can be delegated to the autonomous system but the overall responsibility ultimately lies with the pilots; as Lisanne Bainbridge points out, however, one of the "ironies of automation" is that skillsets are perishable [140]. The automation that supposedly makes tasks easier for the operator in turn increase their inexperience. This is evident in the case of pilots who heavily depend on autopilot during cruising altitude. The system's reliability may explain why Air France does not prioritize training pilots to manually fly at cruising altitude, as it is not necessary when flying with a working autopilot system.

While we have shed light on the potential responsibility gap autonomous systems face, the choice lies with us whether we want to live in a world there is a system with fault or a world without autonomous systems. If we choose the former, where fault is accepted, we need tangible constraints to reduce accident rates. The examples we have drawn upon using autonomous vehicles and automated systems are carefully chosen to display the flaws of the autonomous systems. We should also look at the consequences of not using reliable and responsible autonomous USVs to save human lives, however. To reiterate, these USVs, once activated, can launch and recover, navigate, conduct ISR, and select and engage

targets. These activities can even include executing contingency plans in the absence of human intervention. Additionally, safety mechanisms can be in place to override the individual functions as well as the entire USV. While it seems unsettling for an autonomous USV to take action against a human enemy, we argue that there is great utility to deploy any autonomous USV in dirty, dull, and dangerous tasks. For the sake of argument, if we take the technophobic stand of restricting autonomous USVs in contested waters seriously, here are some possible consequences:

1. We Cannot Deploy Autonomous USVs Even Though They Minimize Unnecessary Risks

A typical ethical objection armed autonomous USVs receive is the asymmetric advantage autonomous USVs possess, perceived as unfair when machines fight against humans. Consider an armed conflict when the commander needs to decide between sending a manned warship or an autonomous USV into hostile waters. Due to the "dishonorable" scene of machines fighting men, the commander is forced to send the manned warship to engage hostile targets. This ethical conundrum contradicts Bradley J. Strawser's Principle of Unnecessary Risk, which asserts that commanders are morally obligated to minimize risks to their own forces [141]. While the context of Principle of Unnecessary Risk is depicted in the context of unmanned aerial vehicles, it is without any reasonable doubt that this principle applies to other autonomous machines such as autonomous USVs. Even though autonomous USVs can mitigate unnecessary risks in this armed conflict scenario, if we take technophobia seriously, we have to reject autonomous USVs as a mere means and continue to deploy our manned ships in all hazardous situations, inadvertently putting our sailors in harm's way.

2. We Make Conscious Decisions to Accept Human Errors over Reliable Machines

Contesting the use of autonomous USVs implies embracing human deficiency over the reliability of these advanced machines. For example, the human attention span is limited and can at times be distracted during an ISR mission, resulting in oversights at critical moments [142]. Ronald C. Arkin points out that unmanned systems have the

capacity to collect, store, and sense-make data faster and more accurately than human [143]. If we take technophobia seriously, we have to accept the fact that we are not optimizing our resources appropriately. Instead, we are choosing a lesser option that may jeopardize the mission.

In short, these two consequences highlight the advantage of autonomous USV deployment during conflict. Taking a holistic lens from mistakes made by Tesla's Full Self-Driving and Air France 447 autopilot systems to the advantages of autonomous systems, we recognize the efficiency of AI and areas where humans perform better than AI. So, it depends on what is being done and what the risk tolerance is for the specific task. To elaborate, humans are unwilling to accept an error that evokes emotions such as human casualties due to mislabeled intelligence collection. In other areas, however, we may be willing to accept all kinds of errors and risks to prevent the needless loss of a life. In the next section, we explore how responsible use of AI, such as accountability and explainability, can potentially close the distrust gap in naval operations.

D. CONSIDERATIONS FOR IMPLEMENTING RESPONSIBLE AI IN NAVAL OPERATIONS

The rapid rise of AI has generated much attention, especially so with the explosion of interest in large language models (LLM) such as ChatGPT from OpenAI with its abilities to engage humans as a chatbot [144]. The surge in interest surrounding ML such as reinforcement learning-based autonomous navigation for USVs has also brought about great anticipation. There are great concerns, however, regarding the lack of understanding of the inner working of the system as well as explaining how the decisions are made [145], [146].

The military often operates in VUCA environments, which necessitates trading off transparent AI with higher performance [147], [148], [149]. Ensuring trustworthy AI, particularly in USVs, requires not only upholding ethical principles but also translating such principles into practical governance. Governance, through its framework of rules and principles, provides a structured approach to transform values into actionable policies. These values should form the foundational basis of all such policies. Differentiating

between system errors and system component errors is a crucial first step. System errors are related to hardware or mechanical malfunctions, such as malfunctioning sensors, broken antennas, or propulsion system failures. On the other hand, system component errors such as faults from AI, data, algorithm, or ML model limitations can have a mixture of algorithmic fairness problems, overlooked considerations, lack of data, or human biases [150]. For instance, an incorrect target engagement could result from a sensor perception algorithm erroneously classifying radar signatures. Robust algorithms, thorough validation, and careful data curation are necessary to prevent system component errors. If an AI programmer uses faulty training data, it creates the condition under which human decision in the system can be made faulty and thus lead to mistakes in engagement decisions. These mistakes are a result of flawed system design. Recognizing this distinction makes it possible to choose the right precautions such as designing-in redundancy, quality control, and reliability engineering that can all be used to address system safety and ethical considerations. After system deployment, stakeholders must keep an eye out for any emerging problems throughout the system life cycle and fix them with patches and updates. The legal requirements and accountability systems controlling these procedures, however, have not kept up with the rapid development of technology [151]. There is an urgent need to close the gap between the legal and accountability frameworks that are outdated and the widespread use of AI in order to ensure governability and accountability.

The U.S. DOD's five AI principles—responsible, equitable, traceable, reliable, and governable systems—offer a baseline foundation for oversight frameworks [152]. Ethically sound AI complies with rules and regulations while carrying out task. Systems that are equitable do not discriminate on the basis of unsuitable characteristics like gender or race [153]. Traceable AI makes it possible to audit procedures and identify the causes of errors, facilitating accountability. Reliable AI operates as intended across contexts, delivering consistent results without malfunctions or deterioration. Finally, as systems increase autonomy, governable AI through auditing framework allows for continued organizational and human oversight [154]. Converting these into requirements for USVs might entail making sure that the inner workings of the AI are accurate, reliable, transparent, and equitable, resulting in more accountable, explainable, and governable autonomous USVs.

The next part highlights the need for accountability, explainability, governance, and essential parameters to advance beyond current governance.

1. Accountability

USVs exemplify the remarkable advancements in technology, bringing forth greater efficiency and autonomy in naval operations. As USVs become more integrated into naval use, however, important questions arise concerning accountability and the need for reliable AI systems. These important queries are a part of a larger conversation about AI systems and are not specific to USVs. The increasing use of USVs raises debates about accountability and responsible AI, and it is crucial to recognize that USVs share the same accountability concerns that have been raised in the broader context of AI. These are not novel questions; these concerns have long dominated discussions about the responsible creation and application of AI systems in a variety of domains and contexts.

Accountability serves as a fundamental concept in both governance and AI, playing a significant role in shaping the responsible development and deployment of USVs. Mark Bovens expressed the overarching notion of accountability as a "golden concept" that universally embodies transparency and trustworthiness [155]. This principle is fundamental in contemporary governance and holds profound relevance in the responsible implementation of technologies such as USVs and AI systems. In the context of AI systems, accountability signifies the ability to assess whether decisions align with established procedural and substantive standards, with the capacity to hold individuals or entities responsible when these standards are not met [156]. This notion bears similarity to USVs since autonomous USVs too are entrusted with making safety-critical decisions. Therefore, it becomes crucial to determine who should be held accountable when these safety-critical decisions fall short of optimal outcomes.

Thompson wrote about the "many hands problem," which is a familiar topic in political governance discussions [157]. He describes the problem, "Because many different officials contribute in many ways to decisions and policies of government, it is difficult even in principle to identify who is morally responsible for political outcomes" and it reflects the difficulty of identifying individual responsibilities within the context of USVs.

Consequently, assigning blame becomes challenging when things go wrong, a phenomenon known as the "responsibility gap" or "accountability gap" [9], [158]. Addressing these accountability issues become imperative when striving for responsible and trustworthy AI in USVs. Helen Nissenbaum contends that accountability has uses beyond merely placing blame and is essential in encouraging the adoption of better practices [159]. In this context, the capability owner, USV program members, and system designers are motivated to aim for excellence in the safe deployment of USVs in operations when they are held accountable for the potential harms or risks they may contribute to. This is similar to the concept of how important accountability is in governance, where it motivates attempts to reduce risks and maximize the benefits of technology. Accountability emerges as a universal concept that resonates both in USVs and reliable AI systems. Accountability turns into a crucial instrument for building more stable and trustworthy systems. Thus, successfully navigating the complex challenges posed by the accountability gap and the involvement of multiple parties requires a steadfast commitment to transparency and promoting best practices.

2. Transparency and Explainability

The role of AI in shaping decision-making processes has a significant impact, as it can enhance or even replace human judgment with faster and more accurate outcomes in controlled scenarios. The widespread adoption of this transformative technology is hindered by a lack of reliability that persists among many, however. This lack of confidence is partly due to concerns about the opacity of AI systems, which have been examined in the study, "Transparency and Trust in Artificial Intelligence Systems" [160]. Additionally, the Human Rights Watch report named "Losing Humanity - The Case Against Killer Robots" offers insights into understanding this apprehension, highlighting how trust issues arise from fears about potential misuse of lethal autonomous systems [161].

The apprehension surrounding AI primarily focuses on fears: the emergence of a technological singularity where artificial general intelligence (AGI) becomes uncontrollable and the possibility that "killer robots" could fall into malicious hands [9].

These concerns were particularly heightened when there were calls to halt what was perceived as an experiment with generative AI after just one year after launch of a popular LLM [162]. The demand for a moratorium received support from various signatories, including notable AI researchers and individuals, such as Elon Musk, who co-founded OpenAI [162]. On this apprehension, perspectives vary greatly, and there are views that posit that the fears surrounding AI and autonomous weaponry are overblown, advocating for a nuanced view that takes into account the myriad of safeguards and applications of AI [163]. These voices advocate for a nuanced perspective that considers the various safeguards and applications of AI [164], [165], [166]. To address these issues, Robert Sparrow's work on "Killer Robots" discusses the complexities of assigning responsibility in autonomous weapon systems [9]. The essence of the dilemma often revolves around the enigmatic nature of AI systems, which emphasizes the critical need of explainability.

Explainability or explainable AI is a commonly used term in the domain of RAI, signifying the importance of transparency and understandability in AI systems and decisions [167], [168]. The concept of explainability in AI brings up an important dilemma: finding the right balance between making models transparent and safeguarding proprietary algorithms crucial for advancing technology. The financial industry is a prime example, as lenders find themselves torn between regulators who seek insights for fairness and the need to protect their proprietary models to stay competitive [169]. We posit that the same "competitive edge" issue is also seen in the military, where maintaining tactical advantage requires secrecy but oversight and accountability are equally important. Different industries will have different expectations when it comes to explainability; this raises the question of how we can ensure safety and accountability in USV actions without revealing classified or sensitive technological expertise.

Google's document "Perspectives on Issues in AI Governance" offers some guidance on establishing minimum acceptable standards across various sectors and applications [170]. The guidance recommends that while it is impractical for governments and civil society to define exhaustive explanation standards for every potential use of AI, this guidance can offer insight through illustrative scenarios. These scenarios help industries gauge the necessary balance between AI system performance and varying levels

of explainability. Another proposed solution is the creation of a graded scale of explanations that can serve as a benchmark for setting minimum acceptable standards tailored to different sectors and contexts. As a result, policymakers and military contractors must carefully strike a balance between regulatory transparency in USV operations and protecting vital technological secrets for national security. Practically speaking, this means carefully crafting guidelines that allow for sufficient regulatory oversight of USV AI systems without necessitating the disclosure of sensitive or classified data. Additionally, it is important to make sure that these systems can function in a wide range of situations, which can be quite challenging considering how complex their operational environments are.

Furthermore, the integration of legacy systems with AI-driven designs adds another level of complexity. With reference to the U.S. Government Accountability Office's AI life cycle phases, it is generally more practical to incorporate AI principles from the beginning rather than trying to adapt them to existing systems [171]. When it comes to implementing AI systems in real world scenarios, integrating AI principles during the design phase is more efficient and cost effective than making changes to systems after deployment. This approach is particularly relevant when dealing with outdated technology in legacy systems, as it may be more practical to replace them entirely rather than attempting extensive modifications as part of system upgrades. The balance between confidentiality and transparency remains a significant challenge, especially as international cooperation becomes increasingly important. The development and use of USVs and their AI systems take place on a global scale, where harmonized regulations, ethical considerations, and safety standards are crucial. As nations strive to maintain their competitive edge, they have to work together toward common objectives that will foster stability, safety, and trust in the rapidly evolving landscape of modern naval warfare, particularly for the safety in the maritime environment. In addressing these complexities, we also acknowledge that certain uncertainties persist, and solutions may only emerge from a globally cohesive regulatory framework.

3. Regulatory Efforts in Artificial Intelligence Governance

Establishing robust and widely accepted guidelines for AI, especially in the realm of military operations, is important [172]. The ethical considerations surrounding the use of AI in warfare demand a governance structure that ensures AI acts in a reliable and justifiable way. This common governance framework should align with human values and adhere to international laws governing warfare. To address this issue, different agencies and nations are working separately to develop standards and guidelines, albeit in a silo manner [173], [174], [175].

The International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) Joint Technical Committee (JTC) 1/Subcommittee (SC) 42 is an international standard and provides a focal point for the standardization of AI and guidance to different communities developing AI applications, fulfilling its aim to serve as the proponent for JTC 1's standardization program on AI [173]. Similarly, the EU's Ethics Guidelines for trustworthy AI, crafted by the high-level expert group on AI appointed by the European Commission, set out seven key requirements for AI systems to be deemed trustworthy, with an emphasis on human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, societal and environmental well-being, and accountability [174]. In the United States, the U.S. DOD updated Directive 3000.09, reflecting a commitment to developing and employing autonomous weapon systems in a responsible and lawful manner. The updated directive aims to minimize failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements, focusing on appropriate levels of human judgment over the use of force, demonstrating a cautious approach toward autonomous weapon systems. The requirements in the directive include "the design, development, deployment, and use of systems incorporating AI capabilities is consistent with the U.S. DOD AI Ethical Principles and the U.S. DOD Responsible AI (RAI) Strategy and Implementation Pathway" [176]. Other global efforts include UNESCO's Recommendation on the Ethics of Artificial Intelligence. In November 2021, all 193 member states unanimously adopted the recommendation and this global standard focuses on AI ethics and puts human rights at the forefront. The standard highlights important principles such as proportionality, safety and

security, privacy and data protection, multi-stakeholder governance, and accountability [175]. Instead of being just a statement of principle, UNESCO's recommendation provides actionable steps by offering comprehensive policy action areas that translate the highlighted core values and principles into practical strategies.

The need to establish regulations for AI is driven by the understanding that AI governance goes beyond individual national boundaries. This is particularly evident when autonomous USVs maneuver transboundary through the sea. Due to the extensive efforts in AI regulation and ethics, particularly in the defense sector, the United States may be well placed to lead in the governance of AI as the first mover. We should reassess, however, if the governance of AI in the maritime environment needs to be ratified by all nation states akin to UNCLOS. Recently, the first U.S. AI executive order issued by President Biden exemplifies the nation's evolving policies as it mandates safety assessments for AI, indicating a shift toward enforceable regulations [177]. In this emerging AI governance field, distinguishing between conceptual development and actionable policies is important. While the former sets the foundation and guides constructive discussions, the latter provides a clear roadmap for implementation. The progression of the U.S. standpoint from conceptual frameworks to actionable policies can be observed through recent updates in U.S. DOD's directive and the executive order on AI. It demonstrates a big step forward in regulatory approaches, transitioning from theoretical debates to practical implementations.

4. Advancing Governance

Other than the five principles of AI, the advancement of governance in AI involves trustworthy AI, including international coordination, transparent practices, accountability, and prioritizing safety. In improving international coordination, it is important to recognize the nexus of international collaboration, transparency, and accountability that underpin the responsible deployment of USVs. As articulated by Horowitz, the U.S. military, through its codification of AI principles, propels a wave of RAI policy formulation among its international partners, emphasizing the need for an interconnected approach in military AI norms and controls [178]. The United States and her allies should continue to cooperate

with international partners to strengthen the governance of AI, so as to push autonomy forward in a "safe and responsible way" [178].

Openness and explainable AI is essential for facilitating improvements in system design and deployment strategies through transparent practices. One should see transparency in AI governance as not merely a tenet but rather a practice, as demonstrated by U.S. DOD's openly accessible policies on AI and emerging technologies. Kroll argues that such openness is indispensable for building an environment where stakeholders can trust and critically assess AI systems, ensuring that systems are not inscrutably complex but are built and operated with transparent and verifiable practices [179]. Stakeholder participation is vital in the oversight processes, ensuring that governance is not a monolith but a responsive construct that evolves with technological advancements and operational needs. This inclusive approach is crucial as nations strive to balance confidentiality with transparency, particularly in the complex and often sensitive realm of military operations. Further, methodologies like red teaming, comprehensive simulation testing, a certain level of human oversight, and the advancement of explainable AI are instrumental in fortifying the governance framework. These strategies ensure that USVs not only comply with military doctrine and the laws of war but also meet the highest standards of operational safety and effectiveness.

Auditing serves as a means to govern accountability effectively. Auditing capabilities and accountability mechanisms, such as detailed logging and post-incident analyses, are cornerstones of a governance structure that values responsibility. These audit functions allow for a transparent review process and facilitate improvements in system design and deployment strategies. Performance metrics and continuous monitoring are also integral, enabling human oversight to intervene when necessary and ensuring that AI systems function within their operational parameters. Standardization in documentation and reporting further streamlines the governance process, allowing for consistent and clear communication of USV operations.

Trustworthiness in AI must work in tandem with the safe usage of systems, which can be learned from autonomous technology pioneers. In the pursuit of trustworthiness in AI, best practices are the guiding stars; these practices include safety designs, exhaustive

testing, and validation protocols akin to those adopted by pioneers in autonomous vehicle technology, like Tesla and Cruise [180], [181]. These companies prioritize safety, a core principle that the military can emulate to enhance the reliability of USVs. Learning from these companies, with their vast user feedback and incident management experiences, offers valuable insight into the resilience and reliability of autonomous navigation systems. By analyzing these experiences and lessons learned, particularly in accident investigations, one can better understand the strengths and weaknesses of autonomous navigation technology, ultimately aiding in its improvement and development.

The advancement of governance in AI is a multidimensional endeavor, demanding international coordination, adherence to transparent practices, and a commitment to accountability. It is a journey that must harmonize relentless innovation in technology with the imperatives of ethics and operational excellence. As we stand at the crossroads where the paths of AI and ethics intersect, we are faced with a choice, a choice to sculpt a future that upholds our moral compass alongside technological prowess, ensuring safety and trust in the rapidly evolving landscape of naval warfare. In the next chapter, we synergize our discussion on ethics and the identified loss scenarios from STPA to analyze the findings and present the results and recommendations of this thesis project.

V. RESULTS AND ANALYSIS

This chapter proposes a model that recommends when and to what extent human involvement is necessary for safe deployment of autonomous USVs in congested and littoral waters, and which provides recommendations to keep humans from being the moral scapegoat of the AI. This model alleviates the inherent distrust military practitioners have toward autonomous USVs and it is conceptualized through a three-stage deduction process. First, we measure the moral permissibility of each task conducted by the autonomous USVs. By employing this approach, the extent of the claim and the significance of consequences are used to gauge the moral permissibility of the corresponding autonomous USV tasks: firing, ISR, navigation, and launch and recovery. While the type of tasks and outcome of these measurements vary based on political and military contexts, this model offers a sample to illustrate our points relating to the ethical implication of deploying autonomous USVs in a busy strait. Second, we investigate the complexity of each component of the autonomous USV system independently from the ethical discussion. The more intricate the system, the more safety needs to be emphasized. Lastly, a model is built using ethical considerations and engineering complexity to determine the thresholds when humans should have an active control, a supervisory control, or zero human control. As a result, the level of human intervention required to mitigate the complexity and ethical considerations of each task enhances commanders' confidence in the autonomous USVs acting independently.

While a human operator can be the moral agent to augment AI's lack of emotion and reasoning ability, the operators are not a panacea for all AI-related faults, particularly not as the moral crumple zone of AI [72], [139]. The chapter aims to clarify what it means to have meaningful human control—beyond simply pressing a button as instructed—in navigation and ISR tasks. In addition, this chapter provides recommendations for safe USV deployment, which include designing autonomous systems with human intervention in mind, clarifying the operator's role and responsibilities, and advocating for relevant training and certification, be it for active or supervisory control of the autonomous USVs. Therefore, human operators are only responsible for decisions within the scope of their

work and actions within their control, eliminating the risk of becoming a moral crumple zone of the AI.

A. SUMMARY OF CHAPTERS I THROUGH IV

In the preceding chapters, we have explored the intricate details of autonomous naval operations, with a specific focus on USVs operating in a challenging environment. Chapter II lays the groundwork by painting a hypothetical geopolitical vignette in the X Strait, an important maritime passage marked by volatility and strategic tensions within the next decade. Our notional USVs played a central role in this scenario, operating in this dynamic environment. Equipped with advanced navigation and communication systems, granting a significant level of autonomy, the USVs can perform various complex yet routine tasks in man-out-of-the-loop conditions. These tasks include patrolling and ISR as well as interactions with different maritime vessels, including friendly, neutral, and adversarial ships in a littoral and congested waterway. This sets the stage to examine the ethical and safety implications of autonomous USVs.

Chapter III delves into the STPA methodology, which we were able to apply to our notional USVs to identify potential hazards and thoroughly analyze them. We employ an enhanced six-step version of STPA to define the purpose of our analysis, identify hazards accurately, and model the control structure. By thoroughly examining the potential CAs and feedback mechanisms of the notional USVs, we uncover the conditions that can lead to unsafe operations or situations. Starting from system malfunctions to strategic exploitation of AI capabilities, these categories encompass a wide range of risks associated with deploying autonomous naval technology. During our analysis, we identify three main categories that can result in losses: (1) faults related to the system; (2) gray zone engagement concepts, where the blame is shifted to AI; and (3) pushing boundaries when unmanned systems are exploited to perform more dangerous maneuvers than manned ships. Each category encompasses various loss scenarios where the autonomy of USVs may intersect with ethical dilemmas and challenges related to system safety.

Chapter IV delves into the meta-ethical questions that arise with the deployment of autonomous systems in military contexts, focusing on responsibility and blame. This

chapter dissects how accountability, transparency, and explainability are critical to foster reliable AI systems and discusses ongoing efforts in AI governance. Given the nascent stage of autonomous system deployment in naval operations, the chapter draws parallels from other domains, such as autonomous systems in civil aviation and autonomous vehicles in the automotive industry, to highlight the challenges and considerations pertinent to the military use of autonomous technology. These examples highlight the potential ethical issues and governance needs of future naval autonomy.

B. STAGE 1: MORAL PERMISSIBILITY OF THE GENERAL TASKS CONDUCTED BY THE AUTONOMOUS USV (ETHICAL DISCUSSION)

Ethical discussion on autonomous USVs' tasks is necessary to establish moral boundaries at both the policy and operational levels. More importantly, these discussions enable states to hold the moral high ground in the face of adversaries' nefarious actions. Such actions, often carried out illegally by aggressors, may involve practices like illegal, unreported, and unregulated (IUU) fishing, leading to the illicit exploitation of marine resources as well as the assertion of territorial claims. These activities not only risk escalating tensions but also threaten to compromise ethical principles. The following ethical discussions aim to provide a basis for establishing norms and guidelines in order to foster a responsible approach to deploying autonomous USVs in maritime operations.

As we have mentioned in Chapter IV, not all ethical issues present themselves as an urgent and unavoidable "trolley car" thought experiment. Day-to-day situations of normal USV operation such as ISR, navigation, and launch and recovery may put forward critical ethical dilemmas worth discussing. Linking back to the vignette in Chapter II, the United States is sending armed autonomous USVs to provide a military presence and to maintain regional peace between Countries 1 and 2. The routine operation involves launching a swarm of autonomous USVs from an allied naval base in Country 3, subsequently navigating to X Strait. Thereafter, the autonomous USVs patrol and conduct ISR around the region to detect any egregious activities engaged by the fishing militias.

Recognizing the diversity in naval tactics, techniques and procedures among different countries, this thesis avoids being too prescriptive about the tactical maneuvers

that the USVs can perform in different scenarios. Instead, we will generalize the mission tasks to allow for adaptable and context-specific applications across various naval contexts and operational environments. Specifically, autonomous USVs are equipped to operate for extended duration at sea, so as to deter any IUU fishing activities, and repel them when absolutely necessary [182].

Among the normative ethics, utilitarianism is used to measure the moral permissibility of the four tasks (navigation, ISR, launch and recovery, and firing) conducted by the autonomous USVs in the notional vignette. Observable behaviors in strategic dynamics and consequences of those actions carry more weight in the international arena than the declared motivation. For instance, if a country deploys military forces near its borders or territorial waters, neighboring nations and global observers may react based on the perceived threat posed by the action, regardless of the nation's stated intention. This phenomenon is true for Russia's intention to pull back troops before the Russia-Ukraine conflict and is equally relevant in China's construction of an aircraft carrier under the guise of scientific research [183], [184]. According to the consequentialist approach, particularly utilitarianism, the act is morally permissible when its consequences contribute to the overall maximization of pleasure or well-being and minimize pain or suffering [22]. Here, the evaluation of the morality is centered on the utility in terms of promoting happiness and minimizing harm. The measurements of moral permissibility of the four tasks are as follows:

1. Is It Morally Permissible for USVs to Conduct Navigation Autonomously?

We argue that it is morally permissible for autonomous navigation for two reasons and raise two challenges that may surface, but not affect, the moral permissibility of this task.

Firstly, in the military context, autonomous USVs minimize unnecessary risks to sailors in a "dirty, dull, and dangerous" mission. The absence of personnel onboard to drive the ship reduces the potential for injury, trauma, or death when a collision occurs.

Therefore, navigating autonomously aligns with the utilitarianism approach that prioritizes minimizing harm.

Second, a robust and reliable navigation system in autonomous USVs holds the potential to save lives by mitigating risks associated with human errors and fatigue. Navigating through narrow channels or challenging maritime environments, an autonomous system can offer precision and consistency, outperforming human capabilities in certain scenarios. This reliability contributes to the moral permissibility of autonomous navigation by emphasizing the potential for increased safety and accident prevention.

A flexible policy review is needed to allow autonomous USVs these deviations in order to safely navigate and avoid collision, if necessary. Drawing parallels with an autonomous car, the rule-following vehicle has difficulty crossing the double yellow line in a two-directional single lane when a cyclist impedes its way. The traffic rules should allow waivers for the autonomous car to cross the double yellow line when there is no oncoming vehicle [185]. This temporal violation of traffic law allows the vehicle to avoid hitting the cyclist and prevents it from hogging the road from other road users. In the maritime context, the challenge arises when autonomous USVs navigate within the Traffic Separation Scheme. The Traffic Separation Scheme may be seen as the highway of the seas where traffic lanes are separated between vessels travelling in different directions. Breaking COLREG becomes necessary in collision avoidance scenarios. The autonomous USV may cross the separation line temporarily to avoid collision before returning back to its lane. In this special circumstance, international law should allow such provision for ethical and safety reasons. Just as autonomous cars may require adjustments to navigate complex traffic scenarios, autonomous USVs may need the flexibility to navigate ethical challenges in their operational environments. This flexibility emphasizes the importance of integrating ethical considerations into the development and deployment of autonomous systems, ensuring that their actions align with broader principles of safety.

The second challenge confronting autonomous USVs is that when temporary deviations are required, AI are limited by their rule-following structure [185]. AI developers need to factor temporary deviations within the algorithm, particularly in critical instances like the cyclist example mentioned above, to prioritize safety. Similar dilemmas

in the maritime domain may warrant such deviation. Suppose an autonomous USV needs to cross territorial waters to avoid collisions or adhere to navigational safety protocols. The rule-following AI should deviate under this special circumstance and overwrite its initial set of programs, so as to avoid imminent danger to its surrounding mariners.

We claim that it is morally permissible for USVs to navigate autonomously as the USVs minimize unnecessary risks to sailors and overcome human errors. To ensure USVs are reliable, flexible policy structures and the adaptable AI must be in place for robust decisions to be made autonomously.

2. Is It Morally Permissible for USVs to Conduct ISR Autonomously?

To address this question, we argue two political utilities, two operational utilities, and an area of ethical concern for autonomous USVs to conduct ISR in X Strait. The purpose of ISR is to achieve information superiority in the VUCA environment presented by the aggressor's gray zone coercion. By obtaining ample advance warning, coercion signal, and evidences of the egregious gray zone activities, neither the U.S. nor Country 2 will be caught off guard by the strategic or operational surprise [31].

The first political utility, amid the gray zone scenario, ISR conducted by the autonomous USVs can play a pivotal role in shaping Country 1's behavior and acting as a deterrent. The proactive surveillance makes it clear that any aggressive moves will be closely observed and responded to effectively. This concept is akin to Jeremy Bentham's unrealized prison architecture of the panopticon [186]. The prison is designed to make the prisoners think that they are constantly surveilled but can never actually confirm if they are being observed. This unsettling feeling from being watched is Bentham's way of socio-psychological control, in hope that prisoners work harder for the sake of evading punishments. In a similar vein, autonomous USVs' ISR capabilities may influence Country 1 to adopt a more restrained approach, so as to avoid an escalating conflict with another major power.

Second, ISR is a non-kinetic task with "unattributable effects" that minimizes the risk of escalation [31]. In the intricate landscape of international relations and conflicts, where the potential for unintended consequences looms large, ISR offers a valuable

alternative to kinetic actions. With its passive nature, ISR can exert influence discreetly, making it challenging for a gray zone aggressor to pinpoint the intent behind a USV's routine actions. This strategic approach not only reduces the likelihood of unintended escalation but also allows for nuanced and calibrated responses in dynamic geopolitical environments.

Third, the autonomous USVs can operate longer and better than manned ships in a harsh maritime environment. To collect valuable information in a contested water, persistence surveillance at sea and processing of large data are two crucial factors. Autonomous USVs eliminate the need for replenishments such as food and water; thus, if the USVs can be refueled at sea, the length of deployment can be significantly extended. The long deployment ensures a comprehensive situational awareness in the strait. Next, AI can collect accurate results and analyze large data faster than humans. Persistent surveillance requires undivided attention to ensure important events are recorded, which may cause human to miss key information when distracted or tired. AI is able to process a large amount of data, analyze patterns, and warn users of any "abnormal or suspicious activities" [187], [188]. This is especially useful when collecting and analyzing a large amount of data such as facial recognition, types of fishing vessels, and the fishermen's pattern of lives. Therefore, the deployment of autonomous USVs increases the operational utility with a lower manpower and maintenance cost as well as a more reliable set of data in near real time.

Lastly, with the rise of great power competition, the U.S. Navy's conventional forces should be reserved against Country 1's conventional naval forces; autonomous USVs are a suitable alternative to deter unconventional forces [31]. Although the true strength of a country's military might cannot be judged solely based on the number of assets, it provides a good indication of relative combat power. U.S. DOD's annual report to Congress shows Country 1 has approximately 440 ships compared with the U.S. Navy's 300 ships. The unequal balance of power may force the U.S. Navy's conventional assets to deprioritize Country 1's unconventional ships and focus on dealing with conventional threats. Autonomous USVs need to be the strategic alternative against Country 1's unconventional ships such as fishing militias, which operates as fishermen during peace

and transform into militias upon state orders. Therefore, autonomous USVs have a significant role in gray zone operation amid the backdrop of great power competition.

Despite numerous advantages to conducting ISR autonomously, using facial recognition technology in ISR operation raises privacy concerns [189]. The data of the AI Global Surveillance Index shows that at least 64 countries are using facial recognition systems [190]. While this technology actively prevents illegal activities, the archiving and usage of these data are often deployed with opacity, limited public oversight, or no internationally recognized legal framework [191]. As facial recognition technology is integrated in our daily lives such as immigration, bank accounts, and other security protections, intentional or unintentional leakage of these data may damage international relations and foreign cooperation.

We claim that it is morally permissible for USVs to autonomously conduct ISR, particularly in the context of intense geopolitical competition, as autonomous USVs are a strategic option to counter fishing militias. Stringent measures must be implemented to safeguard against the leakage of sensitive data collected by these USVs, however.

3. Is It Morally Permissible for USVs to Conduct Launch and Recovery Operation Autonomously?

This thesis assumes launch and recovery to be conducted within the naval base using slipways, before transiting toward the strait in the company of a mother ship. We argue that the conduct of launch and recovery autonomously is morally permissible for two reasons and a concern that can be mitigated:

First, autonomous launch and recovery eliminates operational risk to human operators. Ensuring the safety of the personnel engaged in the handling of USVs is the primary objective of successful launch and recovery operations [192]. Launching and recovering autonomously brings about a substantial reduction in risk by eliminating the human factor from these critical maritime operations. The traditional approach, which involves humans in launch and recovery procedures, introduces a range of potential risks. Human errors, fatigue, and unpredictable conditions can contribute to accidents, injuries,

and even fatalities. By transitioning to autonomous systems, these risks associated with human involvement are significantly mitigated.

Secondly, autonomous launch and recovery procedures alleviate considerable stress placed on the coxswain responsible for ensuring the safety of autonomous USVs. Recovering the USVs from a slipway poses inherent challenges compared to the relatively simpler task of launching [193]. Coxswains require extensive training, a patient understanding of the surrounding waters, and a deep knowledge of seamanship to navigate these complexities [194]. Hanyok and Smith highlight Ullman's enhanced interface, featuring a more ergonomic seat, an intuitive button layout, and a more responsive handling mechanism. These design improvements aim to "reduce the mental strain on coxswains and prevent task saturation" [193]. The introduction of autonomous launch and recovery offers a transformative solution to these challenges, however. The autonomy of the launch and recovery processes eliminates the need for constant human supervision, thereby mitigating the stress on coxswains and streamlining the overall operational efficiency.

Lastly, a potential ethical concern of job displacement may arise, which creates resentment toward the organization. Existing USV coxswains may not be able to transition to a newer autonomous platform once it is introduced. There are two mitigating measures. One, existing manned USVs can undergo a thorough maintenance regime and refit to tailor to operations that requires a coxswain to perform context-based tasks. The USV's coxswains can be assigned another naval platform based on his expertise. Two, the operator can go through retraining and reskilling programs to work in an alternate portfolio. By doing so, the ethical concern of job displacement can be mitigated.

From the points above, we claim that it is morally permissible for USVs to launch and recovery autonomously, so as to eliminate operational risks to humans and mitigate stress on coxswains.

4. Is It Morally Permissible for USVs to Fire Autonomously?

The traditional mentality of deploying soldiers on the ground is to ensure their force preservation. In a soldier-to-soldier combat scenario, once the rules of engagement (ROE) have been approved, soldiers in high-threat conditions will be on high alert to detect

specific signs of hostile intention and actions that satisfy the ROE. This action of firing to disable or to kill ahead of the enemy soldier saves lives. The same ROE may apply to an attack using crew-served weapons by the terrorists who are civilians but soon change the moral status to combatants with clear hostile intentions and actions. Lionel K. McPherson describes how U.S. Lieutenant General James Dubik's perception on command responsibility is to ensure the safety of the command's personnel. According to McPherson, "a commander is not only responsible for protecting the rights of [another state's] civilians, but also for protecting the rights of [his or her] soldiers, to ensure that [his or her soldiers] are only exposed to due risk" [195].

In the context of an autonomous USV, the conventional imperative of preemptively firing on the enemies before they pose a threat is not necessary [73]. The traditional mindset of force preservation, crucial in scenarios involving human lives, takes a different form when deploying autonomous systems. With no immediate risk to human life, autonomous USVs can withstand not just one but multiple hits before responding to the situation and distinguishing between combatants and non-combatants. Even if evidence shows the autonomous USVs being shot by fishing militias, however, the prospect of allowing a machine to autonomously retaliate against a human threat raises ethical concerns. The lingering discomfort of machines killing humans is expressed by Aaron Johnson and Sidney Axinn that "the decision to take a human life must be an inherently human decision…it would be unethical to allow a machine to make such a critical choice" [11].

To avoid the unfair scenario of man versus machine, there are two alternatives for autonomous firing to be morally permissible. First, instead of firing to cause harm, the autonomous USV can be programmed to fire with the sole purpose of immobilizing the vessel. Since the role of fishing militias is to harass Country 2's fishermen, establish prolonged presence around disputed islands, and maintain close distance to naval ships and coast guards, immobilizing their vessels away from the contested area can serve as an effective deterrent. Second, instead of firing lethal munitions that may result in unintended harm, the autonomous USV can be equipped with non-lethal weapons "with the intention of incapacitating personnel or equipment while minimizing the risk of fatalities, permanent injury or damage to materiel and the environment" [31]. This approach disrupts the fishing

militias' ability to carry out their job, at the same time demonstrating a commitment to reducing the risk of severe consequences while still addressing security concerns effectively.

It can be presumed that autonomous USVs firing on humans is morally impermissible, primarily due to the inherent unfairness of such scenario. We claim, however, that it is morally permissible for autonomous USVs to fire non-lethal weapons autonomously with the effect of immobilizing or impeding the fishing militias so as to demonstrate resolve in a controlled manner.

While it is difficult to measure the moral permissibility of different tasks, we propose to approach this complicated issue by understanding how stringent the claims to each task are and its resulting consequence. If one task has a narrower scope and heavier consequence than the other, it is analyzed to be less morally permissible. This idea is not novel; we do this cost-benefit analysis in our daily lives before taking an action. An analogy to illustrate this point is questioning whether lying is right or wrong. The morality police can argue that it is not virtuous and that is not how one wants to be treated; therefore, a person should not lie. In very stringent circumstances, however, lying is morally permissible. A utilitarian can argue that actions that make the most people happy are good and morally permissible. Suppose someone close to the family is injured badly in a car accident; it is arguable that most people will not break the bad news to their grandmother who is terminally ill and instead give a comforting but convincing excuse to explain their absence. The claim is now narrowly scoped to include telling a white lie to avoid the grandmother's disappointment or grief. On the other hand, should the truth be told, the consequence may be worsening of the grandmother's health condition because of grief. Telling the truth, in this case, may not have any tangible and short-term benefit as well, and thus, it should be avoided.

Comparing the moral permissibility of difficult tasks requires careful deliberation of the arguments. For example, comparing telling lies with doing a push-up exercise, we can immediately judge the moral permissibility of the latter far exceeds former. The metric of measurement is the scope of claim and the consequence. Doing push-ups has significant health benefits but one must avoid doing so when their arms are injured. While there are

caveats to both claims, the scope of claim is less restrictive in the push-up example than telling lies. One good way to understand this concept of consequence is to constantly ask three questions: First, what is the worst possible outcome that can happen if the assigned task is not performed as planned? Second, should the undesirable situation happen, who will be implicated and how many stakeholders are involved? Third, what is the consequence if an incident makes a headline and draws negative public perception?

Returning to the ethical discussion of the autonomous USV's task, it is easy to point out that firing autonomously has the narrowest scope among the other three tasks. When a live munition, regardless of the caliber of the round, is fired within the busy waterway, three consequences can be postulated. First, if the round hits Country 1's ship unintentionally, the action may be seen as an act of provocation. This provocation can result in the tipping point to justify initiating war against Country 2 and the United States. Second, if the rounds hit a civilian ship, the United States may be criticized for testing a technologically immature product that causes harm to civilian property. Lastly, if the round causes loss of life to Country 1's fishing militia or civilian ship crew, the United States may be accused of not valuing life and violating human rights. This accusation can result in the loss of confidence toward the U.S. Navy domestically as well as the public in the affected country.

On the other end of the spectrum, autonomous launch and recovery operation at allied bases has minimal impact compared to the other three tasks. While there are risks involved, the benefits of saving time and manpower far outweigh the cost.

Conducting ISR autonomously is more morally permissible than navigating autonomously because ISR is a non-kinetic task with effects that cannot be attributed to the United States. The task has minimal impact on the combatants and non-combatants in the vicinity. Chances of the autonomous USV injuring civilians or damaging civilian property are lower in the ISR operation than the navigation task. Therefore, prioritizing autonomous ISR over autonomous navigation aligns with ethical principles by minimizing potential harm and unintended consequences.

Based on the aforementioned analysis, we can infer that the sequence of comparative moral permissibility between these tasks is as follows: launch and recovery operations, followed by ISR, navigation, and firing, with launch and recovery being identified as the more morally permissible task.

C. STAGE 2: INVESTIGATION ON THE COMPLEXITY OF SYSTEMS (SAFETY DISCUSSION)

The fabric of modern systems presents a duality of both capability and potential risks in autonomous USVs. Charles Perrow, in his book *Normal Accidents*, characterizes complex systems through a sociological lens, transcending the boundaries of mere technical or engineering scrutiny [196]. He illustrates this concept through the Three Mile Island incident, a catastrophic confluence of minor faults coming together in unforeseen ways, culminating in a systemic failure. Perrow postulates that such accidents in complex systems should be considered normal behaviors and these occurrences were not aberrations but rather intrinsic attributes of complex systems, where increased layers of control might paradoxically amplify the risk by introducing unanticipated paths to failure. This perspective allows us to evaluate the enigmatic nature of complexity within systems like USVs, acknowledging that while complexity can inadvertently create paths to unforeseen accidents, it may not be synonymous with high accident rates, relative to the safety of commercial and space aviation.

Amid the discourse on system safety, a fundamental divide and disagreement exists between advocates of Perrow's Normal Accidents theory and proponents of reliable systems. There are inherent differences between these two types of thinking: Perrow's adherents argue that the pursuit of high reliability may inadvertently sow the seeds for more systemic accidents through complex, unforeseen paths. This debate echoes through the corridors of ethical responsibility within complex systems, a concept we previously explored while discussing Robert Sparrow's "responsibility gap" [9]. The contention arises over who bears the brunt of responsibility when system accidents occur. Bridging this dichotomy, Leveson's model, particularly the STPA applied in Chapter III, offers a reconciliatory perspective on the divergent views between accidents in complex systems and the drive toward high reliability systems.

The opacity of a system, in terms of its inner workings and the resulting explainability, sits on a continuum with transparency on one end and opacity on the other. The more transparent and explainable a system is, the lower the opacity. Conversely, a system shrouded in opacity may signal complexity. A key factor contributing to this complexity is the level of interconnectivity and interactions within the system's architecture. As interconnectivity and interactions increase, so too does the complexity, escalating the challenge of comprehending the scope of the system's potential behaviors and failure.

To evaluate and rank the complexity of the system components identified in our notional USVs, we employ the STPA approach. This approach allows us to examine the functional control structure of each component, taking into account their interconnectivity and interactions between the controllers and controlled processes. By leveraging STPA, we categorize the components into "baskets of systems" based on the intricacy of control and oversight to manage them effectively. Such classification, grounded in STPA's comprehensive assessment of potential control flaws and failure conditions, ensures that the complexity ranking reflects the real-world operational demands.

In Table 3, we outline the complexity levels of components within our notional USV systems, alongside their operational capabilities. This structured representation provides a clear overview that not only enumerates the complexities but also specifies the degree of human involvement required for each task. It becomes a crucial reference in clarifying the roles and responsibilities of the personnel involved with the USV system. The table articulates a clear delineation of the system, dissecting its components, functionality, and inherent complexity, thus offering a detailed perspective on the operational intricacies of USVs. This effort complements our system-level model introduced in Figure 12, serving as a foundational element in our comprehensive examination of autonomous naval operations.

Table 3. System Components and Human Oversight in USVs

S/N	System Component	Functionality	Complexity	Explainability	Human Control Threshold
1	Propulsion	Movement and maneuverability (float/move)	Low	High	Human-out-of-the-loop
2	Communications (datalink/ SATCOM)	Voice/data transmission and reception	Medium	Medium	Human-on-the-loop
3	Sensor suite	Tactical picture compilation	Medium	Medium	Human-on-the-loop
4	Navigation (CD/ CA)	Pathfinding and obstacle avoidance (sensor + software)	Medium	Low	Human-on-the-loop
5	Weapon suite	Target engagement and defense	High	Low	Human-in-the-loop

D. STAGE 3: SYSTEM-LEVEL MODEL FOR MEANINGFUL HUMAN CONTROL (ETHICAL AND SAFETY ANALYSIS)

From the ethical discussion in the previous section, we can attempt to arrange the moral standing of each task. When we juxtapose the moral standing and the complexity of a system through a safety lens, we can effectively decide different degrees of human intervention the system requires in order to operate morally. This interaction is illustrated in Figure 12. The threshold will be further explained in the analysis. While the full analysis of the threshold will be elaborated later in this chapter, the level of human intervention is not definitive. States and commanders may make adjustments based on their risk appetite or a stronger moral argument. We are presenting this approach as an outcome of the ethical considerations and the system safety standards imposed on the autonomous USVs. Even if

one disagrees with this approach, one can see that it is the correct approach when we can define the level of meaningful human control based on the ethical and safety perspective.

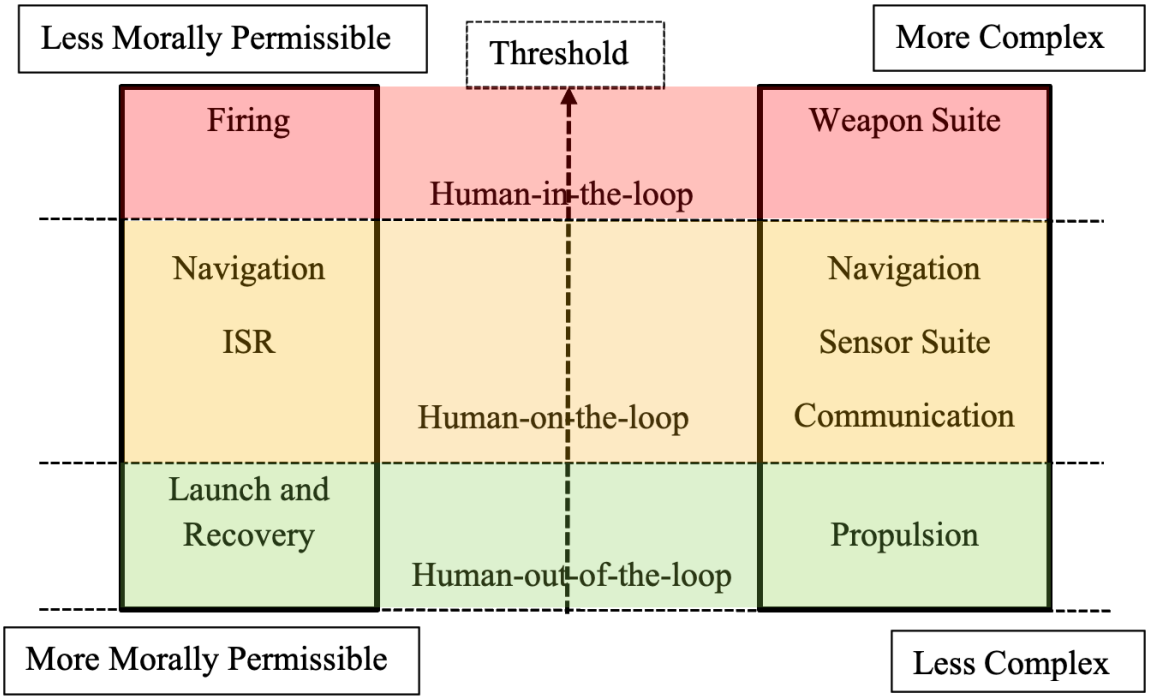

Figure 12. System-Level Model for Meaningful Human Control

Here we present a system-level model, which should be read from the bottom to the top. These tasks are conducted by our notional USVs in the maritime environment of the X Strait, characterized by its congested waters and unpredictable traffic patterns. The model is visually structured into two primary boxes, showing the moral permissibility and complexity, respectively. The level of human intervention and its threshold is depicted by the horizontal lines cutting across the two boxes illustrated using the dashed line. An explanation of Figure 12 is as follows:

Mission Tasks and Moral Permissibility (Depicted by the Left Box): From this chapter's earlier discussion, we arrange the relative order of comparative moral permissibility of each task: launch and recovery, ISR, navigation, and firing. Progressively, we start from more morally permissible to less morally permissible. The principle of the ordering is based on how stringent the claims are and the weight of the consequences. The

more stringent the conditions to execute and the greater the consequences are, the less morally permissive the task is. Therefore, in the diagram, these tasks are sequenced as such with launch and recovery at the bottom, followed by ISR and navigation in the middle and finally firing at the top.

Systems and Complexity (Depicted by the Right Box): On the right box, we derive the order of the systems from the safety discussion in Part C of this chapter. This process is conducted independently from the moral permissibility from Part B1 to B4. Correspondingly, we notice the launch and recovery system (LARS) positioned at the bottom, followed by communication, sensor suite, and navigation in the middle, and ending with the positioning of the weapons suite at the top. This organization is based on the system complexity and interaction range, starting from systems that are considered to be less complex at the bottom to the more intricate and complex system on top.

Level of Human Intervention (Depicted by the Dashed Line): From the independent discussion on moral permissibility and system complexity, we derive the two boxes on the left and right. Next, we can determine the level of human intervention based on the two independent parameters. The task that is more morally permissible and less complex in nature can be carried out autonomously. From the first line from the bottom to the center line (row highlighted in green) lies what we refer to as the man-out-of-the-loop zone where activities like launch and recovery take place alongside systems like LARS, which can be performed without human intervention.

Above the upper line is what we term the "human-in-the-loop" zone. This zone consists of tasks related to firing and systems like the weapons suite. This zone involves tasks that are considered less morally permissible, and the systems are more complex than the rest. Here, our recommendation is for these activities to remain non-autonomous, necessitating direct human control. In the intermediate zone between these two thresholds is the human-on-the-loop area, encompassing ISR and navigation as well as the systems that enable those tasks. These activities call for a degree of human oversight, albeit less direct than in the human-in-the-loop threshold.

The level of human intervention that can be considered meaningful goes beyond pressing a button as instructed [197]. In 2013, a non-governmental organization published *Killer Robots: UK Government Policy on Fully Autonomous Weapons* to seek the United Kingdom policymakers' deeper consideration for the usage of fully autonomous weapons [10]. One of the key highlights of this document is a proposal for meaningful human control, which proposes three minimum standards. First, on information, the operator needs to have sufficient contextual knowledge of the reason to attack, the selection of target, and the consequences of the action. Second, on action, the operator needs to be intentional and perform deliberate actions. Third, on accountability, the operator needs to do his or her due diligence to process any information and be held accountable for the "outcome of the attack" [10]. While this document is written for fully autonomous weapons, the concept is also relevant to an autonomous USV performing its routine tasks. Since launch and recovery is considered permissible for man-out-of-the-loop and the standards surrounding meaningful human control of autonomous weapon systems are discussed in Article 36, we explore how such meaningful human control can be implemented on navigation and ISR. Human-on-the-loop covers a wide spectrum of human oversight that is mainly supervisory, from the most abstract form of determining the permissible rules of engagement for a particular military action by the commander, to another commander providing precise instructions for a strike on a target deemed legal under international law. In this analysis, we cover the operator-level supervision, sufficient to accomplish the mission:

1. Navigation (Human-on-the-Loop)

Information: the operator needs to (a) understand the navigational rules such as UNCLOS and COLREG; (b) be briefed on the mission, navigational route, and communication, contingency, and support plan; and (c) know how to react diplomatically and non-escalatory when faced with issues outside the contingency plan.

Action: the operator needs to have sufficient rest to have the cognitive ability to participate in pre-planned and contingency actions. These actions include but are not limited to navigation according to the designated route as well as overwriting steering control to avoid imminent dangers.

Accountability: the operator is responsible for decisions within the scope of his work and actions within his control.

2. ISR (Human-on-the-Loop)

Information: the operator needs to (a) understand the area of operation, navigable waters, and suitable area to conduct the operation; (b) be briefed on the mission, as well as communication, contingency, and support plan; and (c) know how to respond when faced with issues outside the contingency plan, be it civilian or governmental agencies.

Action: the operator needs to have sufficient rest to have the cognitive ability to participate in pre-planned and contingency actions. These actions include but are not limited to conducting proper surveillance as well as overwriting steering control to avoid imminent dangers.

Accountability: the operator is responsible for decisions within the scope of his work and actions within his controls.

The human operator involved in autonomous USV operations may act as an additional line of defense when an autonomous USV does not perform according to its pre-planned assigned tasks. According to Paul Scharre, human operators, whether in a human-in-the-loop or human-on-the-loop situation, can be effective in three key roles [72]. First, the "human as essential operator" fills up the potential gaps autonomous USVs may have [72]. The operator can value-add by providing context-based scenarios and deliberate decision that accountability needs to be traced. For example, last week a particular fisherman performed something out of the norm. This week he is doing it again. The human operator can answer the question of what we should do with him. The solution has to be traced back to the operator and also to the chain of command. Second, the "human as moral agent" supplements AI's inability to process emotion and reasoning [72]. The human operator can complement this technological gap by weighing the consequences of certain actions and assessing the short-term and long-term effects on the civilian before becoming victims of potential collateral damage. Third, the "human as fail-safe" discerns when to alter course, change the mission scope, or stop the operation entirely [72]. Therefore, humans can mitigate system faults and enhance the operational envelope, if necessary.

E. RECOMMENDATIONS FOR SAFE USV DEPLOYMENT

To ensure the responsible deployment of USVs, we begin with the fundamental assumption that these systems are designed with a high level of commitment to system safety. This assumption involves the implementation of well-designed redundancy measures for system robustness and reliable remote-control capabilities. Here, the key recommendations for ensuring the safe deployment of USVs pertain to the scope of usage and a shift in the mental model for human intervention, shown in the following recommendations:

1. Adherence to Intended Scope of Use

The operation of USVs within the confines of their intended design is paramount to their safe utilization. It is also essential to avoid being too mission-oriented, as the narrow focus on objectives can compromise safety protocols. This is particularly important during operational test & evaluation, where testing the system's limits must occur within a controlled, fail-safe environment. This approach helps mitigate the risk of system faults caused by actions beyond operational parameters.

2. Human Intervention Not Seen as Moral Crumple Zone of AI

Instead of blaming the operators for all accidents, military leaders need to understand the challenges of the operator when they perform active or supervisory controls on the autonomous USVs. Often times, operators step up as a safety net for human oversight to ensure the safe operation of autonomous systems. While meaningful human control can help mitigate potential distrust of autonomous USVs, humans should not be viewed as a panacea for all AI-related faults, particularly not as the moral crumple zone of AI. To prevent humans from being the moral scapegoat of AI, our recommendations are as follows:

a. Operator as a Mitigating Measure

Recognizing the challenges of deploying the operator, it is recommended that military leadership consider a comprehensive list of factors before assigning individuals to these critical roles. These considerations should encompass the full spectrum of

operational, technical, and environmental factors that impact the efficacy of the operators and to deploy them meaningfully.

b. Design Autonomous Systems with Human Intervention in Mind

Systems not originally designed for an additional layer of human control, such as automated fire suppression systems onboard ships like FM-200, may need modification to accommodate human presence during operations.

c. Clarity in Roles and Responsibilities

A clear definition of roles and responsibilities is vital for the effective deployment of operators, who must understand their duties within the system's safety framework. As autonomous systems become more advanced, the lines between "human-in-the-loop" and "human-on-the-loop" may blur, necessitating a reassessment of human oversight roles.

1. Human-in-the-loop: For critical safety systems, human operators are essential for active monitoring and control.

2. Human-on-the-loop: Constant human oversight is necessary, with the capacity for immediate intervention. The positioning of personnel, either at a remote command center or on the USV, is crucial for prompt and effective response.

d. Training and Certification

Competency is a cornerstone of system safety. To mitigate skills atrophy, periodic certification is essential to maintain proficiency. Drawing lessons learned from incidents such as the Air France Flight 447 disaster, the interval between certifications must be carefully determined to ensure skills remain current.

e. Timing for AI to Hand Over Control to Operator

The time at which the controls are handed over to the operator should allow sufficient time for them to fulfill their responsibility effectively. Rushed or poorly timed

transitions can jeopardize system safety and the operators' ability to manage emerging situations.

F. CONCLUSION

This system-level model for meaningful human control demonstrates how mission activities and system complexity interact, highlighting the important role of human control in ensuring ethical and safe use of autonomous USV systems. The analysis goes beyond just assessing a system's capability to perform tasks safely and emphasizes the importance of adhering to ethical boundaries. While the model provides a comprehensive engineering analysis, it primarily serves as a tool for navigating ethical dilemmas by connecting engineering decisions with moral evaluations and broader human values. From an engineering and ethical perspective rooted in consequentialism, the model does not claim to solve these ethical challenges outright but rather offers guidance for thoughtful consideration and finding balance in designing and utilizing autonomous naval systems.

The technological adoption for autonomous USVs can be embraced through alleviating the distrust humans have of them. The identification of the five principles of AI and updates in Directive 3000.09 by the U.S. DOD are steering autonomous weapon systems toward a more responsible and effective governance of AI. Despite these efforts, ethical and legal concerns surrounding the use of autonomous USVs persist. While this paper aims to provide more clarity on the level of human intervention required from autonomous USVs' tasks, including routine ones, we caution military leadership against making humans the crumple zone of AI.

G. RECOMMENDATIONS FOR FUTURE STUDY

Throughout our research project, the application of STPA has been pivotal in navigating the operational intricacies of USVs within the waters of our hypothetical X Strait. By delving deep into specified capabilities and the challenging operational environment, we have identified an array of risks and hazards intrinsic to USV missions, ranging from direct system-related issues to complex interactions with a dynamic maritime environment. Our focus has been to bolster the reliability and safety of USV operations, ensuring mission success is inextricably linked to stringent safety measures. One critical

facet that remains unexplored within our analysis is security, however. Here, a system-theoretic process analysis approach for security (STPA-Sec) emerges as a necessary extension, bridging the gap left by STPA. It offers a methodical approach to adding security considerations into the safety-centric analysis, as demonstrated by Mailloux et al. in their application to notional space systems. By integrating STPA-Sec, we could significantly enhance our operational vignette, aligning safety and security analysis within the unique operational context of the congested and littoral environments of the X Strait. This integrated approach promises a comprehensive safety-security paradigm that is acutely tuned to the nuances of USV operations.

To further refine our paper, we propose adopting a holistic strategy that synergizes STPA with the MIL-STD 882E. The MIL-STD 882E, an established approach dedicated to the design and operational safety of military systems, provides a systematic procedure for mitigating safety risks. By conducting a comparative assessment of STPA and MIL-STD 882E, we can evaluate their efficacy in hazard identification, analysis, and mitigation within military USV operations. The insights gained from such a comparative study could pave the way for a hybrid framework, one that amalgamates the strengths of STPA or STPA-Sec and MIL-STD 882E, to forge a comprehensive safety and security schema.

Our current safety analysis is based on a notional vignette approach, analyzing specific mission tasks such as launch and recovery operations and navigation. This safety analysis deliberately omits detailed analysis of the mission objective, specifically firing of weapons. This deliberate omission stems from the specialized cross-domain knowledge required to conduct an authoritative and exhaustive analysis, a proficiency beyond our current expertise. Future study should aim to address this gap, incorporating domain-specific insights to extend safety analysis to the domain of weapons. This effort requires collaboration with domain experts, thereby ensuring a holistic examination of USV operations across all operational dimensions.

Continued reassessment of aggregated utilitarianism is imperative with the commissioning and deployment of fully autonomous USVs in the straits. This paper has considered the ethical implications based on a notional vignette and conceptual USV systems. Ethical deployment should be reassessed, throughout the peace-to-war

continuum, due to the growing maturity of autonomous technology, increase in deployment of such systems, and consequential impact to the mariners on real-life interactions. This ongoing analysis should encompass short-term considerations such as reliability and safety of collision avoidance systems. For long-term impacts, evaluation of the environmental consequences concerning climate change and measurement of effectiveness in deterrence capabilities must be sought. To complete the analysis, the political and economic impact should be analyzed when the data becomes available for evaluation.

The deontological approach is another normative ethics that can be used to evaluate the morality of autonomous USVs' actions when the technology matures [198]. In the current stage of technological infancy for autonomous systems, AI lacks the ability to have consciousness, intention, emotion, and the ability to reason [199], [200]. If, in the future, technology has made it possible for autonomous USVs to achieve human-like traits, the application of the deontological approach onto autonomous USVs' actions can be reevaluated [201]. When that day comes, the integration of mature AI and the safe deployment of autonomous USVs could potentially revolutionize the landscape of naval warfare, posing disruptive changes for scholars and military practitioners alike to observe.

LIST OF REFERENCES

[1] S. Savitz *et al.*, *U.S. Navy Employment Options for Unmanned Surface Vehicles (USVs)*, Santa Monica, CA, USA: RAND Corp., 2013. Available: https://www.jstor.org/stable/10.7249/j.ctt5vjw3v

[2] R. Rawles, "Lysimeleia (Thucydides 7.53, Theocritus 16.84): what Thucydides does not tell us about the Sicilian Expedition," *J. Hellenic Stud.*, vol. 135, pp. 132–146, 2015. Available: https://doi.org/10.1017/S0075426915000105

[3] G. Galdorisi, "The war in Ukraine and its impact on the future of naval warfare: The USV dimension," Second Line of Defense, Dec. 22, 2022. Available: https://sldinfo.com/2022/12/the-war-in-ukraine-and-its-impact-on-the-future-of-naval-warfare-the-usv-dimension/

[4] H. I. Sutter, "Suspected Ukrainian explosive sea drone made from recreational watercraft parts," *USNI News*, Oct. 11, 2022. Available: https://news.usni.org/2022/10/11/suspected-ukrainian-explosive-sea-drone-made-from-jet-ski-parts

[5] L. Harding and I. Koshiw, "Russia's Black Sea flagship damaged in Crimea drone attack, video suggests," *The Guardian*, Oct. 30, 2022. Available: https://www.theguardian.com/world/2022/oct/30/russias-black-sea-flagship-damaged-in-crimea-drone-attack-video-suggests

[6] The Economist, "Ukrainian ingenuity is ushering in a new form of warfare at sea," Dec. 07, 2022. Available: https://www.economist.com/science-and-technology/2022/12/07/ukrainian-ingenuity-is-ushering-in-a-new-form-of-warfare-at-sea

[7] H. Bachega and J. Gregory, "'Massive' drone attack on Black Sea Fleet - Russia," BBC, Oct. 29, 2022. Available: https://www.bbc.com/news/world-europe-63437212

[8] H. I. Sutton, "Why Ukraine's remarkable attack on Sevastopol will go down in history," Naval News, Nov. 17, 2022. Available: https://www.navalnews.com/naval-news/2022/11/why-ukraines-remarkable-attack-on-sevastopol-will-go-down-in-history/

[9] R. Sparrow, "Killer robots," *J. Appl. Philos.*, vol. 24, no. 1, pp. 62–77, 2007. Available: https://www.jstor.org/stable/24355087

[10] Article 36, "Killer robots: UK government policy on fully autonomous weapons," Apr. 2013. Available: https://article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf

[11] A. M. Johnson and S. Axinn, "The morality of autonomous robots," *J. Mil. Ethics*, vol. 12, no. 2, pp. 129–141, Jul. 2013. Available: https://doi.org/10.1080/15027570.2013.818399

[12] J. E. Jackson, Ed., *One Nation Under Drones: Legality, Morality, and Utility of Unmanned Combat Systems*. Annapolis, Maryland, USA: Naval Institute Press, 2018.

[13] M. Roorda, "NATO's targeting process: Ensuring human control over (and lawful use of) 'autonomous' weapons," in *Autonomous Systems: Issues for Defence Policymakers*, W. Andrew and P. Scharre, Eds., Norfolk, Virginia, USA: NATO Allied Command, 2015, pp. 152–168.

[14] M. Eckstein, "US Navy's four unmanned ships return from Pacific deployment," Defense News, Jan. 16, 2024. Available: https://www.defensenews.com/naval/2024/01/16/us-navys-four-unmanned-ships-return-from-pacific-deployment/

[15] V. Vincze, "The USS Vincennes incident: A case study involving autonomous weapon systems," *Honvédségi Szemle – Hungarian Defence Review*, vol. 148, no. 2, pp. 92–101, Aug. 2021,. Available: https://doi.org/10.35926/HDR.2020.2.6

[16] D. M. Stewart, "New technology and the law of armed conflict," *Int. Law Stud.*, vol. 87, no. 1, pp. 271–298, 2011. Available: https://digital-commons.usnwc.edu/ils/vol87/iss1/12/

[17] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, MA, USA: The MIT Press, 2012. Available: https://doi.org/10.7551/mitpress/8179.001.0001

[18] H. M. Hensel, Ed., *The Law of Armed Conflict: Constraints on the Contemporary Use of Military Force* (Global interdisciplinary studies series). Aldershot, Hants, England; Burlington, VT, USA: Ashgate Pub. Co, 2005.

[19] M. H. Nordquist, *United Nations Convention on the Law of the Sea, 1982: A Commentary*, vol. 2. Martinus Nijhoff Publishers, 1985. Available: https://books.google.com/books?id=LBpWsUYTO7QC&printsec=copyright#v=onepage&q&f=false

[20] J. Zhuang, J. Luo, Y. Liu, R. Bucknall, H. Sun, and C. Huang, "Collision Avoidance for Unmanned Surface Vehicles based on COLREGS," in *2019 5th Int. Conf. Transp. Inf. Safety*, Jul. 2019, pp. 1418–1425. Available: https://ieeexplore.ieee.org/document/8883829/citations

[21] L. Hu, H. Hu, W. Naeem, and Z. Wang, "A review on COLREGs-compliant navigation of autonomous surface vehicles: From traditional to learning-based approaches," *J. Autom. and Intell.*, vol. 1, no. 1, pp. 1–11, Dec. 2022. Available: https://doi.org/10.1016/j.jai.2022.100003

[22] J. Bentham, *An Introduction to the Principles of Morals and Legislation*. Oxford, England: Clarendon Press, 1907. Available: https://www.econlib.org/library/Bentham/bnthPML.html

[23] U. S. Navy, *The Navy Unmanned Surface Vehicle (USV) Master Plan*. 2007. Department of Defense, Washington, D.C., USA. Available: https://apps.dtic.mil/sti/citations/ADA504867

[24] Office of the Secretary of Defense (Acquisition Technology and Logistics), "Unmanned Systems Integrated Roadmap FY2011-2036," Washington, DC, USA, Oct. 2011. Available: https://apps.dtic.mil/sti/citations/ADA558615

[25] N. Melzer, "Human rights implications of the usage of drones and unmanned robots in warfare," Geneva Centre for Security Policy, Brussels, Belgium, EXPO/B/DROI/2012/12, May 2013. Available: https://data.europa.eu/doi/10.2861/213

[26] S. Luo and J. G. Panter, "China's maritime militia and fishing fleets," *Mil. Rev. 101*, vol. 1, pp. 6–21, Jan. 2021. Available: https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/January-February-2021/Panter-Maritime-Militia/

[27] L. J. Morris, M. J. Mazarr, J. W. Hornung, S. Pezard, A. Binnendijk, and M. Kepe, "Gaining competitive advantage in the gray zone: Response options for coercive aggression below the threshold of major war," RAND Corporation, Santa Monica, CA, USA, RR2942, Jun. 2019. Available: https://www.rand.org/pubs/research_reports/RR2942.html

[28] C. Robinson, "Protect unmanned surface vessels in the gray zone," *Proceedings*, vol. 149, no. 1, p. 1439, Jan. 2023. Available: https://www.usni.org/magazines/proceedings/2023/january/protect-unmanned-surface-vessels-gray-zone#:~:text=The%20pressure%20for%20the%20Navy,rather%20than%20expand%20USV%20employment.

[29] D. Grossman and L. Ma, "A short history of China's fishing militia and what it may tell us," The RAND Blog, blog. Available: https://www.rand.org/pubs/commentary/2020/04/a-short-history-of-chinas-fishing-militia-and-what.html

[30] Federal Bureau of Investigation, "USS Cole bombing." Accessed: Feb. 14, 2024. Available: https://www.fbi.gov/history/famous-cases/uss-cole-bombing

[31] B. Ang, "Deterring maritime gray zone aggression ethically with emerging technologies," Naval War College, Newport, Rhode Island, USA, 2019. Available: https://apps.dtic.mil/sti/citations/AD1083850

[32] D. M. Coello, "Is UNCLOS ready for the era of seafaring autonomous vessels?," *J. Territ. Marit. Stud.*, vol. 10, no. 1, pp. 21–37, 2023. Available: https://www.jstor.org/stable/48713706

[33] USNI, "NTSB accident report on fatal 2017 USS John McCain collision off Singapore," Aug. 06, 2019. Available: https://news.usni.org/2019/08/06/ntsb-accident-report-on-fatal-2017-uss-john-mccain-collision-off-singapore

[34] S. Campbell, W. Naeem, and G. W. Irwin, "A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres," *Annual Reviews in Control*, vol. 36, no. 2, pp. 267–283, Dec. 2012. Available: https://doi.org/10.1016/j.arcontrol.2012.09.008

[35] United States Navy, "Navy, Marine Corps release unmanned campaign plan," Mar. 2021. Available: https://www.navy.mil/Press-Office/Press-Releases/display-pressreleases/Article/2538616/navy-marine-corps-release-unmanned-campaign-plan/

[36] L. Berkhouse, "Historical Report Atomic Bomb Tests ABLE and BAKER (Operation Crossroads)," AD0473909, Jun. 1984. Available: https://apps.dtic.mil/sti/citations/AD0473909

[37] Brookings, *CNO Admiral Gary Roughead: The Future of Unmanned Naval Technologies*. 2009. [Mp3]. Available: https://www.brookings.edu/events/the-future-of-unmanned-naval-technologies/

[38] J. M. Richardson, "A Design for Maintaining Maritime Superiority," Washington, DC, USA, Jan. 2016. Available: https://apps.dtic.mil/sti/citations/AD1002755

[39] "Strait," *Dictionary*. Aug. 28, 2023. Available: https://www.dictionary.com/browse/strait

[40] J. Buschmann *et al.*, "Maritime domain protection in the Straits of Malacca," M.S. thesis, Dept. of Syst. Eng. Anal., NPS, Monterey, CA, USA, 2005. Available: https://calhoun.nps.edu/handle/10945/6910

[41] W. Angkasari, "Indonesia-Malaysia dispute over maritime boundaries in the northern region of the Malacca Straits: Implication to fisheries management regime," *J. Critical Rev.*, vol. 7, no. 03, 2020. Available: https://www.academia.edu/68017505/Indonesia_Malaysia_Dispute_Over_Maritime_Boundaries_in_the_Northern_Region_of_the_Malacca_Straits_Implication_to_Fisheries_Management_Regime

[42] A. Smoltczyk, "Show of force in Strait of Hormuz: Risk of 'accidental' Gulf War on the rise," *Der Spiegel*, Jan. 31, 2012. Available: https://www.spiegel.de/international/world/show-of-force-in-strait-of-hormuz-risk-of-accidental-gulf-war-on-the-rise-a-812199.html

[43] AP News, "Malaysia, Indonesia end 18-year sea border disputes, vow to cooperate in defending palm oil industry," Jun. 08, 2023. Available: https://apnews.com/article/malaysia-indonesia-sea-dispute-palm-oil-3704cdddad393425a1cdf94055607e6e

[44] S. Bhattacharjee, "How to handle a ship in congested (high-traffic) waters?," Marine Insight, Aug. 28, 2019. Available: https://www.marineinsight.com/marine-navigation/how-to-handle-a-ship-in-congested-high-traffic-waters/

[45] J. Larson, M. Bruch, and J. Ebken, "Autonomous navigation and obstacle avoidance for unmanned surface vehicles," in *Unmanned Syst. Technol. VIII*, SPIE, May 2006, pp. 53–64. Available: https://doi.org/10.1117/12.663798

[46] I. Baumann, "GNSS cybersecurity threats: An international law perspective," *Inside GNSS Eng. Pol. Des.*, Jun. 2019. Available: https://insidegnss.com/gnss-cybersecurity-threats-an-international-law-perspective/

[47] V. Clark, "Sea Power 21: Projecting decisive joint capabilities," *Proceedings*, vol. 128, no. 10, p. 1196, Oct. 2002. Available: https://www.usni.org/magazines/proceedings/2002/october/sea-power-21-projecting-decisive-joint-capabilities

[48] Office of the Secretary of Defense, "Quadrennial defense review report," Washington, DC, USA, Feb. 2006. Available: https://apps.dtic.mil/sti/citations/ADA442905

[49] P. J. Winstead, "Implementation of unmanned surface vehicles in the distributed maritime operations concept," M.S. thesis, Dept. of Systems Engineering., NPS, Monterey, CA, USA, 2018. Available: https://calhoun.nps.edu/handle/10945/61301

[50] J. Hsu, "U.S. Navy's drone boat swarm practices harbor defense," *IEEE Spectrum*, Dec. 2016. Available: https://spectrum.ieee.org/navy-drone-boat-swarm-practices-harbor-defense

[51] J. Hsu, "U.S. Navy tests robot boat swarm to overwhelm enemies," *IEEE Spectrum*, Oct. 2014. Available: https://spectrum.ieee.org/us-navy-robot-boat-swarm

[52] "Turkey unveils locally designed kamikaze USV," *Janes*, Jul. 26, 2023. Available: https://www.janes.com/defence-news/news-detail/idef-2023-turkey-unveils-locally-designed-kamikaze-usv

[53] R. O. Rourke, "Navy large unmanned surface and undersea vehicles: Background and issues for Congress," Congressional Research Service, Washington, DC, USA, CRS Report No. R45757, Apr. 2023. Available: https://crsreports.congress.gov/product/details?prodcode=R45757

[54] S. LaGrone, "CNO: Navy to finalize large unmanned surface vessel requirements later this year," *USNI News*, Apr. 05, 2023. Available: https://news.usni.org/2023/04/05/cno-navy-to-finalize-large-unmanned-surface-vessel-requirements-later-this-year

[55] "Autonomy," *Dictionary*. Aug. 23, 2023. Available: https://www.dictionary.com/browse/autonomy

[56] M. Vagia, A. A. Transeth, and S. A. Fjerdingen, "A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed?," *Appl. Ergonom.*, vol. 53, no. Part A, pp. 190–202, Mar. 2016,. Available: https://www.sciencedirect.com/science/article/abs/pii/S0003687015300855

[57] T. B. Sheridan, W. L. Verplank, and T. L. Brooks, "Human/computer control of undersea teleoperators," in *14th Annu. Conf. Manu. Control*, Nov. 1978. Available: https://ntrs.nasa.gov/citations/19790007441

[58] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA, USA: MIT Press, 1992. Available: https://books.google.com/books/about/Telerobotics_Automation_and_Human_Superv.html?id=eu41_M2Do9oC

[59] R. Parasuraman, "Designing automation for human use: Empirical studies and quantitative models," *Ergonomics*, vol. 43, no. 7, pp. 931–951, Jul. 2000. Available: https://doi.org/10.1080/001401300409125

[60] J. Chapa, "The ethics of remote weapons," in *One Nation Under Drones: Legality, Morality, and Utility of Unmanned Combat Systems*, J. E. Jackson, Ed., Annapolis, Maryland, USA: Naval Institute Press, 2018, pp. 176–193.

[61] M. R. Endsley, "The application of human factors to the development of expert systems for advanced cockpits," *Proc. Human Factors Soc. Annu. Meeting*, vol. 31, no. 12, pp. 1388–1392, Sep. 1987. Available: https://doi.org/10.1177/154193128703101219

[62] B. Lorenz, F. Di Nocera, S. Röttger, and R. Parasuraman, "The effects of level of automation on the out-of-the-loop unfamiliarity in a complex dynamic fault-management task during simulated spaceflight operations," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 45, no. 2, pp. 44–48, Oct. 2001. Available: http://doi.org/10.1177/154193120104500209

[63] B. T. Clough, "Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway?," Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio, USA, Aug. 2002. Available: https://apps.dtic.mil/sti/citations/ADA515926

[64] C. A. Ntuen and E. H. Park, "Human factor issues in teleoperated systems," in *Proc. First Int. Conf. Ergonom. Hybrid Autom. Syst. I*, NLD: Elsevier Science Publishers B. V., Oct. 1988, pp. 203–210.

[65] V. Riley, "A general model of mixed-initiative human-machine systems," *Proc. Human Factors Soc. Annu. Meeting*, vol. 33, no. 2, pp. 124–128, Oct. 1989, doi: 10.1177/154193128903300227.

[66] B. Shneiderman, "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," arXiv, University of Maryland, College Park, MD, USA, UMD HCIL-2020-01, Feb. 2020. Available: https://doi.org/10.48550/arXiv.2002.04087

[67] S. Mias, "Unmanned maritime systems engineering technologies and applications for unmanned maritime systems," *Int. J. Unmanned Syst. Eng.*, vol. 1, no. 4, pp. 23–30, 2013. Available: https://www.proquest.com/docview/1519059598/abstract/793F19BDF8444D26PQ/1

[68] IMO, "Outcome of the regulatory scoping exercise for the use of maritime autonomous surface ships (MASS)," presented at the Maritime Safety Committee 103 session, Jun. 2021, pp. 1–105. Available: https://wwwcdn.imo.org/localresources/en/MediaCentre/PressBriefings/Documents/MSC.1-Circ.1638%20-%20Outcome%20Of%20The%20Regulatory%20Scoping%20ExerciseFor%20The%20Use%20Of%20Maritime%20Autonomous%20Surface%20Ships...%20(Secretariat).pdf

[69] V. Boulanin and M. Verbruggen, "What are the technological foundations of autonomy?," in *Mapping the Development of Autonomy in Weapon Systems*, Stockholm, Sweden: Stockholm International Peace Research Institute, 2017, pp. 5–18. Available: https://www.jstor.org/stable/resrep24528.8

[70] A. Williams, "Defining autonomy in systems: Challenges and solutions," in *Autonomous Systems: Issues for Defence Policymakers*, A. Williams and P. Scharre, Eds., Norfolk, Virginia, USA: NATO Allied Command, 2015, pp. 27–64.

[71] R. Arnold, "The legal implications of the use of system with autonomous capabilities in military operations," in *Autonomous Systems: Issues for Defence Policymakers*, W. Andrew and P. Scharre, Eds., Norfolk, Virginia, USA: NATO Allied Command, 2015, pp. 83–97.

[72] P. Scharre, "Centaur warfighting: The false choice of humans vs. automation," *Temple Int. Comp. Law J.*, vol. 30, no. Number 1 (Spring 2016), pp. 151–165. Available: https://sites.temple.edu/ticlj/files/2017/02/30.1.Scharre-TICLJ.pdf

[73] J. Thurnher, "No one at the controls: Legal implications of fully autonomous targeting," *Joint Force Quart.*, no. 67, pp. 77–84, Dec. 2012. Available: https://papers.ssrn.com/abstract=2296346

[74] G. Watson, "Resistance to change," *American Behavioral Scientist*, vol. 14, no. 5, May 1971. Available: https://doi.org/10.1177/000276427101400507

[75] L. R. Blank and G. P. Noone, *International Law and Armed Conflict: Fundamental Principles and Contemporary Challenges in the Law of War*, Second edition. in Aspen casebook series. New York: Wolters Kluwer, 2019.

[76] M. Walzer, *Just and Unjust Wars - Moral Argument Historical Illustrations*, 5th ed. New York, NY, USA: Basic Books, 2015. Available: https://books.google.com/books/about/Just_and_Unjust_Wars.html?id=EuTQCQAAQBAJ

[77] L. Doswald-Beck, *Human Rights in Times of Conflict and Terrorism*. New York, NY, USA: Oxford University Press, 2011. Available: https://books.google.com/books?id=fcraLq0NBesC&printsec=copyright#v=onepage&q&f=false

[78] M. N. Schmitt, "Narrowing the international law divide," in *One Nation Under Drones: Legality, Morality, and Utility of Unmanned Combat Systems*, J. E. Jackson, Ed., Annapolis, Maryland, USA: Naval Institute Press, 2018, pp. 133–149.

[79] United Nations Office for Disaster Risk Reduction, "Non-International Armed Conflict (NIAC)," Jun. 2023. Available: http://www.undrr.org/understanding-disaster-risk/terminology/hips/so0002

[80] U.S. House, 107th Congress Public Law 40. (2001, Sep. 18), *Authorization for Use of Military Force*. Available: http://www.congress.gov/bill/107th-congress/senate-joint-resolution/23/text

[81] B. Obama, "Remarks by the President at the National Defense University," The White House. Available: https://obamawhitehouse.archives.gov/the-press-office/2013/05/23/remarks-president-national-defense-university

[82] J. C. Johnson, "National security law, lawyers, and lawyering in the Obama administration," *Yale Law & Policy Rev.*, pp. 141–150, 2012. Available: https://yalelawandpolicy.org/national-security-law-lawyers-and-lawyering-obama-administration

[83] International Humanitarian Law Databases, "Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977." Available: https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/

[84] R. Sparrow and G. R. Lucas Jr., "When robots rule the waves?," in *One Nation Under Drones: Legality, Morality, and Utility of Unmanned Combat Systems*, J. E. Jackson, Ed., Annapolis, Maryland, USA: Naval Institute Press, 2018, pp. 75–98.

[85] C. M. Ford, "Autonomous weapons and the law," in *One Nation Under Drones: Legality, Morality, and Utility of Unmanned Combat Systems*, J. E. Jackson, Ed., Annapolis, Maryland, USA: Naval Institute Press, 2018, pp. 150–164.

[86] International Humanitarian Law Databases, "Article 51 - Protection of the civilian population."Available: https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-51

[87] International Humanitarian Law Databases, "Article 57 - Precautions in attack." Available: https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-57

[88] "Assessment, proportionality, and justice in war," in *Assessing War: The Challenge of Measuring Success and Failure*, Georgetown University Press, 2015, pp. 255–265. Available: https://www.jstor.org/stable/j.ctt19qgffn

[89] B. Gogarty and M. C. Hagger, "The laws of man over vehicles unmanned: The legal response to robotic revolution on sea, land and air," *J. Law, Inf., Sci.*, vol. 19, p. 73, 2008. Available: https://papers.ssrn.com/abstract=1796486

[90] R. McLaughlin, "Unmanned naval vehicles at sea: USVs, UUVs, and the adequacy of the law," *J. Law, Inf., Sci*, vol. 21, no. 2, p. [100]-115, Jan. 2011. Available: http://www5.austlii.edu.au/au/journals/JlLawInfoSci/2012/6.html

[91] M. Suri, "Autonomous vessels as ships – the definition conundrum," in *IOP Conf. Series: Mater. Sci. Eng.*, IOP Publishing, Nov. 2020. Available: https://dx.doi.org/10.1088/1757-899X/929/1/012005

[92] Jade. *Guardian Offshore AU Pty Ltd v Saab Seaeye Leopard 1702 Remotely Operated Vehicle Lately on Board the Ship "Offshore Guardian" [2020]*. 2020. Available: https://jade.io/article/717961

[93] Stockton Center for International Law, *The Commander's Handbook on the Law of Naval Operations*. Washington, DC, USA, 2022. Available: https://usnwc.libguides.com/ld.php?content_id=66281931

[94] M. J. Valencia, "The U.S. is trying to unilaterally shape the international rules governing the use of drones," Peking University Institute of Ocean Research, Haidian, Beijing, China, May 2022. Available: http://www.scspi.org/en/dtfx/us-trying-unilaterally-shape-international-rules-governing-use-drones

[95] International Maritime Organization, "Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs)." Available: https://www.imo.org/en/About/Conventions/Pages/COLREG.aspx

[96] Admiralty and Maritime Law Guide: International Conventions, *Convention on the International Regulations for Preventing Collisions at Sea (London 1972)*. 1972. Available: http://www.admiraltylawguide.com/conven/collisions1972.html

[97] United States Department of Defense, "Memorandum of Understanding Between the Department of Defense of the United States of America and the Ministry of National Defense of the People's Republic of China Regarding the Rules of Behavior for Safety of Air and Maritime Encounters," UNT Digital Library. Available: https://digital.library.unt.edu/ark:/67531/metadc949788/

[98] J. McMahan, "Innocence, self-defense and killing in war," *J. Polit. Philos.*, vol. 2, no. 3, pp. 193–221, 1994, Accessed: Sep. 06, 2023. Available: https://philosophy.rutgers.edu/joomlatools-files/docman-files/Innocence_Self-Defense_&_Killing_in_War.pdf

[99] C. Fabre, *Cosmopolitan War*. Oxford University Press, 2012.

[100] N. Leveson, "STPA (System-Theoretic Process Analysis) compliance with MIL-STD-882E and other army safety standards," Massachusetts Institute of Technology, Cambridge, MA, USA. Accessed: Feb. 08, 2024. Available: http://sunnyday.mit.edu/compliance-with-882.pdf

[101] Today, "What a Singapore Strait traffic jam says about the world economy," Mar. 06, 2019. Available: https://www.todayonline.com/world/what-singapore-strait-traffic-jam-says-about-world-economy?cid=internal_inarticlelinks_web_22012024_tdy

[102] Statista Research Department, "Oil flows through the Strait of Hormuz between 2014 and 2020." Accessed: Oct. 27, 2023. Available: https://www.statista.com/statistics/277157/key-figures-for-the-strait-of-hormuz/

[103] Y. C. Altan and E. N. Otay, "Spatial mapping of encounter probability in congested waterways using AIS," *Ocean Engineering*, vol. 164, pp. 263–271, Sep. 2018. Available: https://doi.org/10.1016/j.oceaneng.2018.06.049

[104] Reykjavik University, "What is STAMP/STPA?" Accessed: Oct. 26, 2023. Available: https://en.ru.is/stamp/what-is-stamp/

[105] E. Marsden, "Heinrich's domino model of accident causation," Risk Engineering. Available: https://risk-engineering.org/concept/Heinrich-dominos

[106] V. V. Khanzode, J. Maiti, and P. K. Ray, "Occupational injury and accident research: A comprehensive review," *Safety Science*, vol. 50, no. 5, pp. 1355–1367, Jun. 2012. Available: https://doi.org/10.1016/j.ssci.2011.12.015

[107] J. Reason, "The contribution of latent human failures to the breakdown of complex systems," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 327, no. 1241, pp. 475–484, Apr. 1990. Available: https://www.jstor.org/stable/55319

[108] N. Leveson and J. Thomas, *STPA Handbook*. USA, 2018. Available: https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf

[109] S. M. Sulaman, A. Beer, M. Felderer, and M. Höst, "Comparison of the FMEA and STPA safety analysis methods–A case study," *Software Quality J.*, vol. 27, no. 1, pp. 349–387, Mar. 2019. Available: http://doi.org/10.1007/s11219-017-9396-0

[110] MIT Partnership for Systems Approaches to Safety and Security (PSASS), "STAMP workshop." Accessed: Oct. 16, 2023. Available: https://psas.scripts.mit.edu/home/stamp-workshops/

[111] J. Edgar, "Finding and fixing fragility in machine learning," Ph.D. dissertation, Dept. of Comp. Sci., NPS, Monterey, CA, USA, 2023. Available: https://calhoun.nps.edu/handle/10945/72193

[112] M. Issa, A. Ilinca, H. Ibrahim, and P. Rizk, "Maritime autonomous surface ships: problems and challenges facing the regulatory process," *Sustainability*, vol. 14, no. 23, Art. no. 23, Nov. 2022. Available: https://doi.org/10.3390/su142315630

[113] Herodotus, *The Histories*. Translated by A.D. Godley, Cambridge, MA, USA: Harvard University Press, 1920. Available: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:abo:tlg,0016,001:3:38

[114] H. M. Curtler, *Ethical Argument: Critical Thinking in Ethics*, 2nd edition. New York: Oxford University Press, 2004.

[115] C. Gowans, "Moral relativism," in *The Stanford Encyclopedia of Philosophy*, Spring 2021., E. N. Zalta, Ed., Metaphysics Research Lab, Stanford University, 2021. Available: https://plato.stanford.edu/archives/spr2021/entries/moral-relativism/

[116] J. Rachels and S. Rachels, "The challenge of cultural relativism," in *The Elements of Moral Philosophy*, 7th ed., pp. 14–31., New York: McGraw-Hill, 2012. Available: https://books.google.com/books/about/The_Elements_of_Moral_Philosophy_7e.html?id=P6RvEAAAQBAJ&printsec=frontcover&source=kp_read_button&hl=en&newbks=1&newbks_redir=0#v=onepage&q&f=false

[117] F. Hutcheson, *An Inquiry into the Original of Our Ideas of Beauty and Virtue (1726, 2004)*. Indianapolis, Indiana, USA: Liberty Fund, 1726. Available: https://oll.libertyfund.org/title/leidhold-an-inquiry-into-the-original-of-our-ideas-of-beauty-and-virtue-1726-2004

[118] E. Fieldstadt, P. Helsel, and S. Dong, "Navy SEAL Edward Gallagher acquitted of murder in ISIS fighter case," NBC News, Jul. 02, 2019. Available: https://www.nbcnews.com/news/us-news/navy-seal-edward-gallagher-acquitted-murder-isis-fighter-case-n1025226

[119] K. Toropin, "Navy's top admiral said SEALs had 'character and ethics' issues," Military.com, Aug. 03, 2021. Available: https://www.military.com/daily-news/2021/08/02/navys-top-admiral-questions-seals-character-and-ethics.html

[120] C. Wicker, "Ethics and consciousness," in *Thinking Ethics: How Ethical Values and Standards are Changing*. London, England: Profile Books, 2005, pp. 5–29.

[121] L. Johansson, "Ethical aspects of military maritime and aerial autonomous systems," *J. Mil. Ethics*, vol. 17, no. 2–3, pp. 140–155, Jul. 2018. Available: https://doi.org/10.1080/15027570.2018.1552512

[122] A. Russo, L. Vojković, F. Bojic, and R. Mulić, "The conditional probability for human error caused by fatigue, stress and anxiety in seafaring," *J. Marine Sci. Eng.*, vol. 10, no. 11, pp. 1–17, Oct. 2022. Available: https://doi.org/10.3390/jmse10111576

[123] N. Sharkey, "Saying 'No!' to lethal autonomous targeting," *J. Mil. Ethics*, vol. 9, no. 4, pp. 369–383, Dec. 2010. Available: https://doi.org/10.1080/15027570.2010.537903

[124] S. Vallor, "The future of military virtue: Autonomous systems and the moral deskilling of the military," in *2013 5th Int. Conf. Cyber Conflict*, Jul. 2013. Available: https://ieeexplore.ieee.org/document/6568393

[125] J. Himmelreich, "Never mind the trolley: The ethics of autonomous vehicles in mundane situations," *Ethical Theory and Moral Practice*, vol. 21, no. 3, pp. 669–684, May 2018,. Available: http://doi.org/10.1007/s10677-018-9896-4

[126] I. Duncan, "Bursts of acceleration in Tesla vehicles caused by drivers mistaking accelerators for brakes, feds conclude," *Washington Post*, Jan. 08, 2021. Available: https://www.washingtonpost.com/transportation/2021/01/08/tesla-brakes/

[127] D. Hull and K. Naughton, "Tesla blames driver in fatal crash as victim's family lawyers up," *Bloomberg*, Apr. 11, 2018. Available: https://www.bloomberg.com/news/articles/2018-04-11/tesla-says-inattentive-driver-to-blame-for-fatal-model-x-crash

[128] S. Blanco, "Report: Tesla autopilot involved in 736 crashes since 2019," *Car and Driver*, Jun. 13, 2023. Available: https://www.caranddriver.com/news/a44185487/report-tesla-autopilot-crashes-since-2019/

[129] S. Alvarez, "Revel driver claims Tesla 'unintended acceleration' in lawsuit: Report," *Teslarati*, May 19, 2023. Available: https://www.teslarati.com/revel-driver-tesla-unintended-acceleration-lawsuit/

[130] C. Agatie, "Another Tesla crashes in China after high-speed driving, Tesla pledges full cooperation," *Autoevolution*, Feb. 20, 2023. Available: https://www.autoevolution.com/news/another-tesla-crashes-in-china-after-high-speed-driving-tesla-pledges-full-cooperation-210604.html

[131] S. Alvarez, "Tesla China pledges full cooperation with investigators after fatal crash," *Teslarati*, Feb. 19, 2023. Available: https://www.teslarati.com/tesla-china-cooperation-investigators-fatal-crash/

[132] Tesla, "There is no 'unintended acceleration' in Tesla vehicles." Accessed: Nov. 15, 2023. Available: https://www.tesla.com/blog/no-unintended-acceleration-tesla-vehicles

[133] National Transportation Safety Board, "Electric Vehicle Run-off-Road Crash and Postcrash Fire," Washington, DC, USA, Feb. 2023. Available: https://www.ntsb.gov/investigations/Pages/HWY21FH007.aspx

[134] National Transportation Safety Board, "Tesla Crash Investigation Yields 9 NTSB Safety Recommendations," Washington, DC, USA, Feb. 2020. Available: https://www.ntsb.gov/news/press-releases/Pages/NR20200225.aspx

[135] A. Roy, D. Levine, and H. Jin, "Tesla wins bellwether trial over autopilot car crash," *Reuters*, Apr. 22, 2023. Available: https://www.reuters.com/legal/us-jury-set-decide-test-case-tesla-autopilot-crash-2023-04-21/

[136] Bureau d'Enquêtes et d'Analyses, "Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris," Le Bourget, France, f-cp090601, Jul. 2012. Available: https://bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf

[137] E. Cobbe, "Air France and Airbus acquitted of involuntary manslaughter in 2009 crash of Flight 447 from Brazil to Paris," *CBS News*, Apr. 17, 2023. Available: https://www.cbsnews.com/news/air-france-flight-447-crash-airbus-and-airline-acquitted-involuntary-manslaughter/

[138] C. Irving, "Air France Flight 447: Who's fault was it?," *The Daily Beast*, Jul. 13, 2017. Available: https://www.thedailybeast.com/articles/2011/07/29/air-france-flight-447-who-s-fault-was-it

[139] M. C. Elish, "Moral crumple zones: Cautionary tales in human-robot interaction," *Engag. Sci. Technol. Soc.*, vol. 5, pp. 40–60, Mar. 2019. Available: https://doi.org/10.17351/ests2019.260

[140] L. Bainbridge, "Ironies of automation," in *Analysis, Design and Evaluation of Man–Machine Systems*, G. Johannsen and J. E. Rijnsdorp, Eds., Pergamon, 1983, pp. 129–135. Available: https://doi.org/10.1016/B978-0-08-029348-6.50026-9

[141] B. J. Strawser, "Moral predators: The duty to employ uninhabited aerial vehicles," *J. Mil. Ethics*, vol. 9, no. 4, pp. 342–368, Dec. 2010. Available: https://doi.org/10.1080/15027570.2010.536403

[142] G. R. Lucas, "Military ethics and unmanned systems," in *Military Ethics: What Everyone Needs to Know*, New York, NY, USA: Oxford University Press, 2016, pp. 167–188.

[143] R. C. Arkin, "The case for ethical autonomy in unmanned systems," *J. Mil. Ethics*, vol. 9, no. 4, pp. 332–341, Dec. 2010. Available: https://doi.org/10.1080/15027570.2010.536402

[144] E. Rosenbaum, "ChatGPT AI hype cycle is peaking, but even tech skeptics doubt a bust," CNBC. Available: https://www.cnbc.com/2023/02/11/chatgpt-ai-hype-cycle-is-peaking-but-even-tech-skeptics-doubt-a-bust.html

[145] N. Yan, S. Huang, and C. Kong, "Reinforcement learning-based autonomous navigation and obstacle avoidance for USVs under partially observable conditions," *Math. Prob. Eng.*, vol. 2021, May 2021. Available: https://doi.org/10.1155/2021/5519033

[146] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI 2017 Workshop Explainable Artificial Int. (XAI)*, 2017, p. 6. Available: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf

[147] H. F. Barber, "Developing strategic leadership: The U.S. Army War College experience," *J. Manage. Develop.*, vol. 11, no. 6, p. 4, 1992. Available: https://doi.org/10.1108/02621719210018208

[148] C. Molnar, *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Bookdown, 2022. Available: https://christophm.github.io/interpretable-ml-book/index.html

[149] J. Kwik and T. van Engers, "Algorithmic fog of war: when lack of transparency violates the law of armed conflict," *J. Future Robot Life*, vol. 2, no. 1–2, pp. 43–46, 2021. Available: https://papers.ssrn.com/abstract=4126001

[150] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions," *Annu. Rev. Stat. Appl.*, vol. 8, no. 1, pp. 141–163, Mar. 2021, doi: 10.1146/annurev-statistics-042720-125902.

[151] J. A. Kroll *et al.*, "Accountable algorithms," *University of Pennsylvania Law Review*, vol. 165, pp. 633–705. Available: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

[152] C. T. Lopez, "DOD adopts 5 principles of artificial intelligence ethics," U.S. Department of Defense. Available: https://www.defense.gov/News/News-Stories/article/article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/

[153] A. Datta, A. Datta, J. Makagon, D. K. Mulligan, and M. C. Tschantz, "Discrimination in online advertising a multidisciplinary inquiry," *Proc. 1st Conf. Fair., Account., Transpar.*, vol. 81, pp. 20–34, 2018. Available: https://proceedings.mlr.press/v81/datta18a

[154] I. D. Raji *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. 2020 Conf. Fair., Account., and Transpar.*, in FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 33–44. doi: 10.1145/3351095.3372873.

[155] M. Bovens, "Analysing and assessing accountability: A conceptual framework," *European Law J.*, vol. 13, no. 4, pp. 447–468, Jun. 2007. Available: https://doi.org/10.1111/j.1468-0386.2007.00378.x

[156] F. Doshi-Velez *et al.*, "Accountability of AI under the law: The role of explanation." Dec. 20, 2019. Available: https://doi.org/10.48550/arXiv.1711.01134

[157] D. F. Thompson, "Moral responsibility of public officials: The problem of many hands," *Amer. Political Sci. Rev.*, vol. 74, no. 4, pp. 905–916, Dec. 1980. Available: https://doi.org/10.2307/1954312

[158] T. Chengeta, "Accountability gap, autonomous weapon systems and modes of responsibility in international law," *J. Int. Law & Policy*, vol. 45, no. 1 Fall, Jan. 2016. Available: https://doi.org/10.2139/ssrn.2755211

[159] H. Nissenbaum, "Accountability in a computerized society," *Sci. Eng. Ethics*, vol. 2, no. 1, pp. 25–42, Mar. 1996. Available: https://doi.org/10.1007/BF02639315

[160] P. Schmidt, F. Biessmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," *Journal of Decision Systems*, vol. 29, no. 4, pp. 260–278, Sep. 2020. Available: https://doi.org/10.1080/12460125.2020.1819094

[161] B. Docherty, "Losing humanity: The case against killer robots," Human Rights Watch, USA, Nov. 2012. Available: https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

[162] J. Vincent, "Elon Musk and top AI researchers call for pause on 'giant AI experiments,'" The Verge. Available: https://www.theverge.com/2023/3/29/23661374/elon-musk-ai-researchers-pause-research-open-letter

[163] P. Scharre, "Why you shouldn't fear 'slaughterbots,'" IEEE Spectrum. Available: https://spectrum.ieee.org/why-you-shouldnt-fear-slaughterbots

[164] E. Jatho and J. A. Kroll, "Artificial intelligence: Too fragile to fight?," *Proceedings*, vol. 148/2/1, p. 428, Feb. 2022,. Available: https://www.usni.org/magazines/proceedings/2022/february/artificial-intelligence-too-fragile-fight

[165] E. M. Bender, "On NYT magazine on AI: Resist the urge to be impressed," Medium. Available: https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd

[166] I. Van Rooij, O. Guest, F. G. Adolfi, R. De Haan, A. Kolokolova, and P. Rich, "Reclaiming AI as a theoretical tool for cognitive science." PsyArXiv, Aug. 01, 2023. Available: https://doi.org/10.31234/osf.io/4cbuv

[167] Microsoft, "Responsible and trusted AI," Microsoft Azure. Available: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai

[168] IBM, "What is explainable AI?" Accessed: Nov. 27, 2023. Available: https://www.ibm.com/topics/explainable-ai

[169] R. Pace, "Dark skies ahead: The CFPB's brewing algorithmic storm," Pace Analytics Consulting LLC. Available: https://www.paceanalyticsllc.com/post/cfpb-brewing-algorithmic-storm

[170] Google AI, "Responsibility: policy perspectives." Accessed: Nov. 09, 2023. Available: https://ai.google/responsibility/public-policy-perspectives/

[171] T. Ariga, "AI: Key practices to help ensure accountability in Federal use," Government Accountability Office, Washington, DC, USA, Statement of Taka Ariga, Chief Data Scientist, Science, Technology Assessment and Analytics GAO Report No. GAO-23-106811, 2023. Available: https://www.gao.gov/assets/gao-23-106811.pdf

[172] D. Vergun, "U.S. endorses responsible AI measures for global militaries," *Defense Department News*, Nov. 22, 2023. Available: https://www.defense.gov/News/News-Stories/Article/Article/3597093/us-endorses-responsible-ai-measures-for-global-militaries/

[173] ISO/IEC JTC 1/SC 42, "Artificial intelligence," ISO. Accessed: Nov. 19, 2023. Available: https://www.iso.org/committee/6794475.html

[174] European Commission, "Ethics guidelines for trustworthy AI," Apr. 2019. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[175] United Nations Educational, Scientific and Cultural Organization (UNESCO), "Recommendation on the ethics of artificial intelligence." Available: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

[176] U.S. Department of Defense, "DOD announces update to DOD Directive 3000.09, 'autonomy in weapon systems.'" Available: https://www.defense.gov/News/Releases/Release/Article/3278076/dod-announces-update-to-dod-directive-300009-autonomy-in-weapon-systems/

[177] The White House, "Fact sheet: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence," Washington, DC, USA, Oct. 2023. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[178] J. Clark, "DOD committed to ethical use of artificial intelligence," *U.S. Department of Defense News*, Jun. 15, 2023. Available: https://www.defense.gov/News/News-Stories/Article/Article/3429864/dod-committed-to-ethical-use-of-artificial-intelligence/

[179] J. A. Kroll, "The fallacy of inscrutability," *Philos. Trans. Roy. Soc. A: Math., Physical, Eng. Sci*, vol. 376, no. 2133, Nov. 2018, doi: doi.org/10.1098/rsta.2018.0084.

[180] Tesla, "Tesla vehicle safety report," Jan. 2023. Available: https://www.tesla.com/VehicleSafetyReport

[181] Cruise Automation, "Cruise safety report," 2022. Available: https://assets.ctfassets.net/95kuvdv8zn1v/zKJHD7X22fNzpAJztpd5K/ac6cd2419f2665000e4eac3b7d16ad1c/Cruise_Safety_Report_2022_sm-optimized.pdf

[182] A. C. Min, "USVs could deter IUU fishing," *Proceedings*, vol. 149, no. 8, Aug. 2023. Available: https://www.usni.org/magazines/proceedings/2023/august/usvs-could-deter-iuu-fishing

[183] P. Wintour, "Russia has amassed up to 190,000 troops on Ukraine borders, U.S. warns," *The Guardian*, Feb. 18, 2022. Available: https://www.theguardian.com/world/2022/feb/18/russia-has-amassed-up-to-190000-troops-on-ukraine-borders-us-warns

[184] E. Y. Kong and K.-L. Yu, "Deciphering Chinese strategic deception: The Middle Kingdom's first aircraft carrier," M.S. thesis, Dept. of Def. Anal., NPS, Monterey, CA, USA, 2013. Available: https://calhoun.nps.edu/handle/10945/34690

[185] Stanford University, "How the 'trolley problem' applies to self-driving cars," *Futurity*, Jan. 26, 2023. Available: https://www.futurity.org/autonomous-vehicles-av-ethics-trolley-problem-2863992-2/

[186] UCL, "The Panopticon," Bentham Project. Accessed: Jan. 28, 2024. Available: https://www.ucl.ac.uk/bentham-project/about-jeremy-bentham/panopticon

[187] C. Davenport, "Future wars may depend as much on algorithms as on ammunition, report says.," *Washington Post*, Apr. 09, 2023. Available: https://www.washingtonpost.com/business/economy/future-wars-may-depend-as-much-on-algorithms-as-on-ammunition-report-says/2017/12/03/4fa51f38-d6b7-11e7-b62d-d9345ced896d_story.html

[188] D. M. West and J. R. Allen, "How artificial intelligence is transforming the world," Brookings, Washington, DC, USA, Apr. 2018. Available: https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/

[189] R. Talbot, "Automating occupation: International humanitarian and human rights law implications of the deployment of facial recognition technologies in the occupied Palestinian territory," *Cambridge University Press*, vol. 102, no. 914, pp. 823–849, Dec. 2021. Available: https://doi.org/10.1017/S1816383121000746

[190] S. Feldstein, "The global expansion of AI surveillance," Carnegie Endowment for International Peace, Washington, DC, USA, Sep. 2019. Available: https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847

[191] A. Toh, "Opinion: Worried about how facial recognition technology is being used? You should be," *Los Angeles Times*, Nov. 18, 2019. Available: https://www.latimes.com/opinion/story/2019-11-18/facial-recognition-artificial-intelligence-watching-you

[192] C. L. Strouse, "A launch and recovery system for integrating unmanned ocean vehicles onto surface platforms," M.S. thesis, Dept. of Mech. Eng., NPS, Monterey, CA, USA, 2019. Available: https://calhoun.nps.edu/handle/10945/63189

[193] L. W. Hanyok and T. C. Smith, "Launch and recovery system literature review," Naval Surface Warfare Center, West Bethesda, Maryland, USA, NSWCCD-50-TR-2010/071, Dec. 2010. Available: https://apps.dtic.mil/sti/citations/ADA590153

[194] Naval History Heritage Command, "Chapter VI. The coxswain takes over." Accessed: Feb. 15, 2024. Available: https://www.history.navy.mil/research/library/online-reading-room/title-list-alphabetically/s/skill-in-the-surf-a-landing-boat-manual/chapter-vi-the-coxswain-takes-over.html

[195] J. M. Dubik, "Human rights, command responsibility, and Walzer's Just War Theory," *Philos. & Public Affairs*, vol. 11, no. 4, pp. 354–371, 1982. Available: https://www.jstor.org/stable/2265156

[196] C. Perrow, *Normal Accidents: Living with High-Risk Technologies*. New York, NY: Basic Books, 1984.

[197] M. C. Horowitz and P. Scharre, “Meaningful human control in weapon systems: A primer,” Center for a New American Security, Washington, DC, USA, Mar. 2015. Available: https://www.cnas.org/publications/reports/meaningful-human-control-in-weapon-systems-a-primer

[198] L. Alexander and M. Moore, “Deontological ethics,” Stanford Encyclopedia of Philosophy. Available: https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/

[199] E. Hildt, “Artificial intelligence: Does consciousness matter?,” *Frontiers in Psychology*, vol. 10, Jul. 2019,. Available: https://doi.org/10.3389/fpsyg.2019.01535

[200] J. P. Sullins, “When is a robot a moral agent,” *Int. Rev. Inf. Ethics*, vol. 6, no. 12, pp. 23–30, 2006. Available: https://philarchive.org/rec/SULWIA-2

[201] B. Brożek and B. Janik, “Can artificial intelligences be moral agents?,” *New Ideas in Psychology*, vol. 54, pp. 101–106, Aug. 2019,. Available: https://doi.org/10.1016/j.newideapsych.2018.12.002

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