Using Reinforcement Learning to Provide Decision Support in Multi-Domain Mass Evacuation Operations

Mark Rempel and Nicholi Shiell
Centre for Operational Research and Analysis
Defence Research and Development Canada
60 Moodie Drive, Ottawa, CANADA
K1A 0K2

mark.rempel@forces.gc.ca, nicholi.shiell@forces.gc.ca

ABSTRACT

In this paper, we study a scenario in which a large number of individuals in various levels of medical distress are stranded at a remote location, such as in the Arctic, and must be evacuated. Set within this context, we examine a multi-domain operation in which the evacuation of individuals occurs via one of two ways, either by helicopter or by ship, each with their own capacity constraints. The aim of this research is to determine a decision policy whose objective is to maximize the number of survivors. This is achieved by seeking a policy that throughout the operation effectively coordinates the selection of those individuals to be evacuated via helicopter and those to be evacuated via ship. Our contributions are twofold. First, we formulate the multi-domain mass evacuation operation as a Markov Decision Process. Second, due to the fact that the curse of dimensionality renders exact methods not applicable, we employ an Artificial Intelligence framework, namely, Reinforcement Learning (RL), also known as Approximate Dynamic Programming (ADP) within operations research, to learn a near-optimal policy. Using a value function approximation based on state aggregation, we design an ADP algorithm to learn a policy within the context of a representative planning scenario. We then apply this policy across a range of test scenarios and compare the outcomes to those achieved using non-coordinated benchmark policies. Although our learned policy does not outperform all benchmarks, our results demonstrate how Artificial Intelligence may be used to evaluate candidate policies and provide decision support in multi-domain operations.

1.0 INTRODUCTION

Over the last few decades, the decrease in Arctic sea ice has been substantial, particularly during the summer seasons [1]. Although further reduction depends on many factors, such as climate change [2], activity in the region is expected to increase [3], [4]. In particular, if the ability to navigate through the Northwest Passage, Northern Sea Route, and Transpolar Sea Route (see the left panel of Figure 1) become commonplace, then their use for both trade and the transport of individuals, specifically cruise ships, will follow – although not without its challenges [5]. While the Polar Code sets out goals and functional requirements for ships operating in the region [6], when disaster strikes it remains the responsibility of government departments and agencies to conduct the necessary operations, including Search and Rescue (SAR). This is evidenced by recent exercises conducted by Arctic nations, such as the SARex exercise conducted near Spitzbergen, Norway in 2016, which aimed to assess the effectiveness of safety equipment during a mass evacuation after a cruise ship sinking [7], and NANOOK-TATIGIIT 21 conducted by the Canadian Armed Forces (CAF), which aimed to test an "interagency response to a major maritime incident requiring a Mass Rescue Operation along the eastern coast of Baffin Island" ([8], p. 30).

In such Major Maritime Disaster (MAJMAR) evacuation operations, the number of individuals that survive is influenced by a range of factors [9], including, but not limited to: response time; environmental conditions [10], [11]; the number of passengers, crew, and their medical conditions [12]; infrastructure [13], [14], [15]; and the decision policies used during the operation [16]. In particular, the latter study highlighted that when

considering an individual's medical condition, the decision policy – "a rule (or function) that determines a decision given the information available" (Ref. [17], p. 221) – used to determine the order in which individuals are evacuated, may have a significant impact on the number of survivors. With this in mind, this study examines a Multi-Domain Operation (MDO) in response to a MAJMAR scenario, in particular a cruise ship carrying up to 2,000 individuals, that occurs in a remote location and explores the effectiveness of evacuation decision policies given limited resources and uncertainty regarding an individual's medical condition over time.[1] It seeks to find a decision policy whose objective is to maximize the number of survivors by coordinating the selection of individuals to be evacuated across multiple domains.

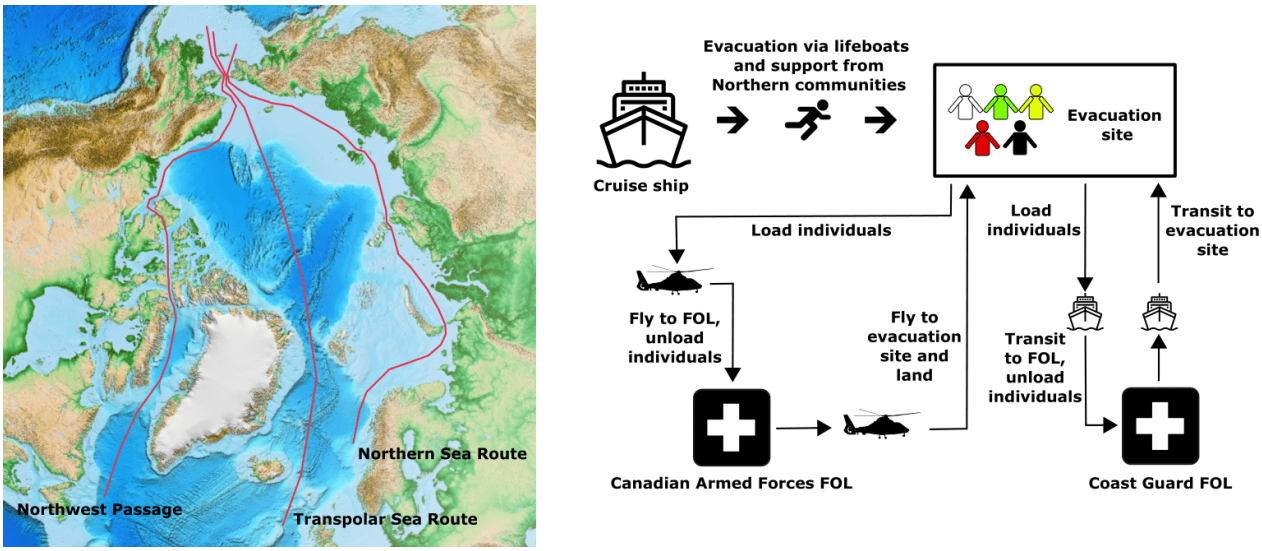


Figure 1: Arctic multi-domain evacuation. Left panel: Example Arctic Sea routes (actual routes may vary). Source: Basemap is taken from https://armap.org/web-services/. Right Panel: Multi-domain evacuation via air and sea.

Beyond the defence domain, the allocation of scarce resources to maximize the number of survivors of a mass evacuation scenario has been studied in an array of contexts, including blunt trauma [19], mass casualty events [20]-[25], healthcare facility evacuation [26], and aeromedical evacuation of an international traveller [27]. Many of these studies tend to focus on the evacuation of a small number of individuals, such as 12 in Ref. [25] and 100 in Ref. [21]; assume that each individual requires the same resources (e.g., a stretcher in an ambulance); and conduct the evacuation in a single domain (e.g., land or air). When compared to a MAJMAR scenario such as described above, these scenarios are relatively small. While scenarios that involve a very large number of individuals have been studied, such as several thousand during the evacuation of the New Orleans Airport during Hurricane Katrina [28] and over 100,000 during a major earthquake in Istanbul [29], these tend to be set in populated areas with access to a large number of resources. Thus, outside the defence domain there appear to be few, if any, studies that focus on finding decision policies in large-scale evacuation operations in which resources, in particular transport, are limited.

Within the defence domain, several studies have examined the medical evacuation of individuals. For example, Keneally et al. [30] studied aeromedical helicopter dispatch policies in a combat environment, constructed a Markov Decision Process to describe a scenario, and used a value iteration algorithm to develop an optimal

[1] Within the NATO context, as of 29 July 2022, the working definition of a Multi-Domain Operation was given as "the orchestration of military activities, across all domains and environments, synchronized with non-military activities, to enable the Alliance to deliver converging effects at the speed of relevance" [18].

policy to maximize the number of soldiers saved. Jenkins et al. [31] examined the military medical evacuation location-allocation problem and designed an integer programming model to determine the location and number of assets required over multiple phases of a military deployment where the model's objectives were to maximize the total expected demand covered, minimize the maximum number of locations in any phase, and minimize the number of times locations needed to be moved between phases. Robbins et al. [32] also investigated how best to dispatch aeromedical assets in response to calls for service, but employed an Approximate Dynamic Programming (ADP) based approach to seek a near-optimal policy. While such studies are related to the aforementioned scenario, to the best of our knowledge within the defence-related open literature, only Rempel et al. [16] seek a policy that determines the order in which individuals are to be evacuated. In addition, as Rempel et al. [16] and Robbins et al. [32] demonstrate, when seeking policies in medical evacuation scenarios, Artificial Intelligence (AI) frameworks such as ADP, known as Reinforcement Learning (RL) in computer science, may be useful to determine near-optimal decision policies as exact methods are not applicable due to the curse of dimensionality.

The application of these AI frameworks within MDOs is sparse. A query of the Web of Science on 11 May 2022 using the search term "multi-domain operation" and ("artificial intelligence" or "reinforcement learning" or "approximate dynamic programming") yielded three relevant results that focus on the security and trustworthiness of AI [33], how human decision-making can be integrated with AI [34], and predicting who is responsible for terrorist incidents [35]. A similar query of Google Scholar returned 62 results, including those that focus on the need to share data [36], the need to integrate AI into the battle management process [37], and how RL may be used for autonomous strategic manoeuvre and disruption [38]. Given these results, and the recent review of ADP being applied for decision support in a military context [39], to the best of our knowledge using RL/ADP to seek near-optimal decision policies in MDOs is either not widespread or non-existent in the open literature.

Our contributions are twofold. First, we formulate the multi-domain mass evacuation operation as a Markov Decision Process (MDP). Second, due to the problem's size and that the curse of dimensionality renders exact methods not applicable, we employ ADP to learn a near-optimal evacuation decision policy with the objective of maximizing the number of survivors. Using a Value Function Approximation (VFA) based on state-aggregation, we design an ADP-based algorithm to learn a policy within the context of a representative planning scenario. We then apply this policy across a range of test scenarios and compare the outcomes to those achieved using non-coordinated benchmark policies. Although our learned policy does not outperform all benchmarks, our results demonstrate how Artificial Intelligence may be used to evaluate candidate policies and provide decision support in multi-domain operations.

In addition to these contributions, this article extends our previous study [16] in three ways. First, it includes an MDO response to the situation rather than only an air domain response. Second, it expands the uncertainty considered by capturing the possibility of an individual's health degrading multiple triage categories between sequential decision epochs. Lastly, rather than using a one-to-one lookup table within the VFA, this study employs a state-aggregation approach similar to that found in Ref. [25].

The remainder of this article is organized as follows. The problem definition section describes the mass evacuation problem considered and the MDO that occurs in response. The methodology section defines the MDP formulation and the ADP formulation employed to search for a near-optimal decision policy. The computational results section presents computational results for a series of experiments, including the comparison of the ADP-generated policy to benchmark policies. Lastly, a conclusion and directions for future research are provided.

2.0 PROBLEM DEFINITION

The scenario considered in this article is based on the MAJMAR scenario described in Ref. [15]. The scenario described in Ref. [15] (pp. 6-18) requires the CAF to:

1) Deploy personnel, materiel, and multiple aircraft to a Forward Operating Location (FOL);
2) Transport individuals by helicopter from an evacuation site to an FOL; and
3) Transport individuals via fixed-wing aircraft from the FOL to a southern location.

In a recent paper, Rempel et al. [16] considered a portion of this scenario, in particular the movement of individuals from the evacuation site to the FOL, and searched for a near-optimal decision policy regarding evacuation via helicopter. This paper extends these two previous works by considering the response as an MDO. Specifically, the scenario considered which, like [16], focuses on the movement of individuals from the evacuation site to the FOL, allows for individuals to be evacuated through one of two ways, either by helicopter (such as those operated by a nation's military) or by ship (such as those operated by a nation's coast guard). Within this context, the aim is to determine a decision policy that coordinates the efforts of these two evacuation routes to collectively maximize the expected total number of survivors. The remainder of this section describes the representative scenario used in this article.

A cruise ship carrying 2,000 individuals, including passengers and crew hereafter described collectively as individuals, is transiting the Northwest Passage during August (such as the Crystal Serenity did in 2016 with approximately 1,000 passengers and 600 crew [39]). An accident, such as hitting an iceberg or engine failure, occurs at a location along the route that puts the ship in a category where the operational focus is on evacuation [40]. Within much of the Northwest Passage route the Canadian Coast Guard (CCG) is the lead agency to respond to this type of incident; however, while one of their ships is in the area, they do not have the capacity to evacuate 2,000 individuals. Given this situation, the CAF are requested to support the evacuation operation.

Within the first few hours, the individuals move from the accident location to an evacuation site on a nearby shore via the cruise ship's lifeboats and, if possible, with the assistance of those living in nearby Northern communities.[2] Concurrently, among other resources and personnel, the CAF deploy one helicopter, and set up an FOL, labelled the CAF FOL, that has a suitable airfield and is within the helicopter's operational range of the evacuation site. Likewise, the CCG divert their ship in the area to the evacuation site and identify a second FOL, labelled the CCG FOL, to which their ship transports individuals. Lastly, SAR technicians are deployed to the evacuation site to provide immediate medical care. The flow of the operation is depicted in the right panel of Figure 1.

Once the helicopter has arrived at the CAF FOL, SAR technicians have arrived at the evacuation site, and the CAF FOL is ready to receive individuals, the helicopter begins the following cycle: fly to the evacuation site; land and load living individuals (those deceased are left on site and will be recovered later); fly to the CAF FOL to unload individuals; fly to the evacuation site; and so on. A similar cycle occurs in parallel with the CCG ship: transit to the evacuation site; load living individuals; transit to the CCG FOL and unload individuals; return to the evacuation site; and so on. These cycles continue until all living individuals have arrived at either of the FOLs. In addition, the following assumptions are made: one helicopter may land at a time at the evacuation site (when multiple helicopters are considered in Section 4.0); the weather is clear for the duration of the operation; there are no significant aircraft or ship breakdowns; sufficient fuel is available at the CAF FOL to refuel the helicopters; and the ship has sufficient fuel to operate for the duration of the operation. These

[2] Based on Safety of Life at Sea guidelines [41], the maximum time for all survival craft to be launched with all individuals is 30 minutes from the time the abandon ship signal is given. This study assumes that the signal to abandon is not immediately given and includes the transit time to shore and the time to disembark the lifeboats. Thus, a few hours to reach the evacuation site is deemed reasonable.

assumptions are unlikely to be realized in practice; however, they result in an idealized scenario that may be used as a baseline on which to compare the impact of changes in the assumptions, equipment, number of individuals, and so on, in future studies.

At the evacuation site, the cruise ship's crew initially classify each of the individuals based on their medical condition, to one of the five triage categories – white, green, yellow, red, and black – as defined in Table 1. The classification of live individuals into four triage categories (black represents those deceased) based on medical need, as opposed to two categories based on capacity requirements, is consistent with the scenario described in previous work [15], [16]. The initial distribution of individuals is inspired by the Costa Concordia disaster, which resulted in 32 deaths and 64 injuries. Given this information, an initial count of 100 injured was deemed reasonable given a cold, remote, and isolated environment. While at the evacuation site, each individual's health, and hence their triage category, deteriorates over time (i.e., white → green, green → yellow, etc.). Given no intervention, all individuals will transition to the black triage category, that is, they perish, as depicted in Figure 2.

Table 1: Properties of the triage categories and initial count. See Ref. [17] for details.

Category	Treatment	Initial Count	Stretcher?	Mean Time [h]
White	None	1900	No	120
Green	Routine	40	No	48
Yellow	Early	30	Yes	8
Red	Immediate	30	Yes	1.5
Black	Deceased	0	Yes	-

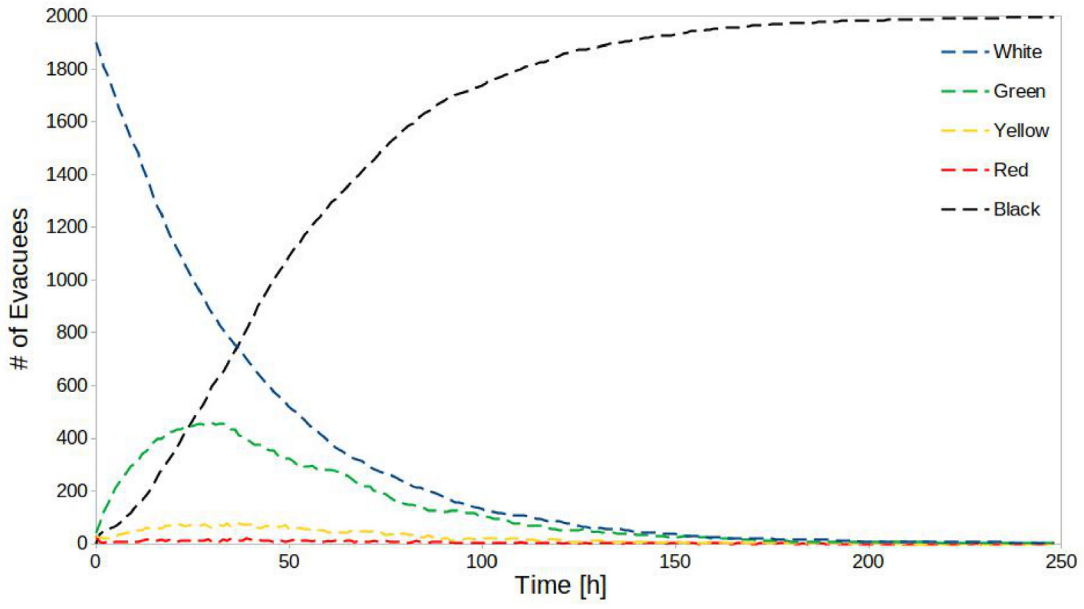


Figure 2: Number of individuals in each triage category as a function of time with no intervention. Initial distribution is 1,900 (white), 40 (green), 30 (yellow), 30 (red), and zero (black). Hour zero is defined as when the individuals have arrived at the evacuation site and the ship's crew have assessed the individuals. Category transition times follow an exponential distribution with mean times listed in Table 1.

Given that an individual's health may be influenced by many factors (e.g., cause of injury, age, medical history, etc.), neither their conditions nor their transitions between categories are modelled in detail. Rather, the transition time between two adjacent triage categories is assumed to follow an exponential distribution with mean times listed in the far right column of Table 1, consistent with previous work [15], [16].[3] These values are based on consultations with a medical professional and are to be interpreted as "rule-of-thumb estimates of how long a person would stay in a particular triage state under various environmental conditions and levels of care" ([15], p. 16). For example, the transition time from the red to black triage category is modelled after an individual having a heart attack (i.e., the American Heart Association recommends fewer than 90 minutes from the symptoms begin to the opening of the blocked artery). The transition times are the greatest source of uncertainty within the scenario, and their modification may have significant impact on the expected number of lives saved in the results presented herein. Further details may be found on pp. 16-17 of Ref. [15].

Given this triage model and mean transition times, the evacuation of individuals is a time-critical problem. The total number of individuals evacuated is non-deterministic as a result of an individual's transition between triage categories.[4] Given the scenario, its inherent uncertainties, and the decision policy sought – which individuals are loaded onto the helicopter or ship for transport to the respective FOL – this problem is aptly described as a sequential decision-making problem under uncertainty. Section 3.0 models this sequential decision-making problem as an MDP, which is "sufficiently broad to allow modelling most realistic sequential decision problems" (Ref. [42], p. 2). Specifically, we employ the modelling framework and notation suggested by Powell (Ref.[11], Ch. 6), as this allows greater flexibility and clearer representation of the uncertainty than the notation presented in Ref. [42]. Following the description of this model, Section 3.2 proposes an ADP formulation to learn a near-optimal decision policy that aims to maximize the expected total number of survivors.

3.0 METHODOLOGY

3.1 Markov Decision Process Formulation

3.1.1 State Space

Let the set of triage categories be represented by $\mathcal{T} = \{w, g, y, r\}$ – the black triage category is not included, since the deceased are recovered at a later time. Let the set of locations $\mathcal{L} = \{e, f^H, f^S\}$ represent the evacuation site, the FOL that receives individuals via helicopter, and the FOL that receives individuals via ship, respectively. Lastly, let the set $\mathcal{A} = \mathcal{A}^{\mathrm{H}} \cup \mathcal{A}^{\mathrm{S}}$ represent the helicopters and ships involved in the scenario, where $\mathcal{A}^{\mathrm{H}}$ is the set of helicopters that travel between the evacuation site and f^H and $\mathcal{A}^{\mathrm{S}}$ is the set of ships that travel between the evacuation site and f^S. It is assumed that all helicopters have identical attributes, as do the ships (when multiple of each are considered as in Section 4.0). The initial state S_0 is listed in Table 2 and contains parameters that are constant throughout the scenario. The MDP has a finite horizon, where the final decision epoch is the event in which no living individuals are remaining at the evacuation site. This event is labelled as K, and given the stochastic nature of the scenario the time at which it occurs, is variable.

The state variable is then defined as

$$S_k = (\tau_k, e_k, \rho_k), \tag{1}$$

[3] Using an exponential distribution to model the deterioration of individuals is effective for our problem for two reasons. First, it has been previously used in studies of priority assignment in emergency response [21] and, second, it requires no knowledge of how long an individual has been in a given triage category, just how long since their health was last sampled.

[4] While other uncertainties exist, such as breakdowns, refuelling and maintenance time, etc., these are outside the scope of this study.

where τ_k is the system time recorded during the k^{th} event, e_k is an integer event code that corresponds to Table 3, and $\rho_k = (\rho_{k,t})\ t \in \mathcal{T}$ where $\rho_{k,t}$ is an integer representing the number of individuals in each triage category t at the evacuation site. The inter-transition time listed in Table 3 includes the time to load individuals, transit to the FOL, unload individuals, and return to the evacuation site.

Table 2: Initial state S_0 are static parameters that are constant throughout the scenario.

$\mathcal{L}$			$\mathcal{A}$		
Evac (e)	**CAF FOL (f^H)**	**CCG FOL (f^S)**	$\mathcal{A}^H$	$\mathcal{A}^S$	**Description**
m^e	-	-	-	-	Vector of mean time (hrs) for an individual to deteriorate from a triage category
-	-	-	h^H	h^S	Total capacity of individuals
-	-	-	Δ^H	Δ^S	Vector of capacity consumed by each triage category
η^e	η^{fH}	η^{fS}	-	η^S	Initial location
-	-	-	φ^H	-	Maximum one-way transit time

Table 3: Transitory events that trigger a state transition.

Event (e_k)	Description	Inter-Transition Time
1	Helicopter arrives at evacuation site	Return Time
2	Ship arrives at evacuation site	Return Time

3.1.2 Decision Space

When a decision is made, it is made in response to a helicopter or ship arriving at the evacuation site and triggering event k. Let $x_k = (x_{kt})_{\forall t \in \mathcal{T}}$ where x_{kt} is an integer representing how many people are loaded from category t at the evacuation site onto the helicopter or ship that is onsite. A loading decision, be it for the helicopter or ship, is constrained by the space available onboard and the number of individuals at the evacuation site. For a helicopter, these constraints are stated as

$$\sum_{t \in \mathcal{T}} x_{kt} \Delta_t^H \leq h^H \tag{2}$$

and

$$x_{kt} \leq \rho_{kt}, \forall t \in \mathcal{T} \tag{3}$$

Similar constraints exist for a ship.

3.1.3 State Transition

The state transition function is defined as $S_{k+1} = S^M(S_k; x_k; W_{k+1})$, where W_{k+1} is exogenous information (i.e., the uncertainty within the decision problem), that arrives after decision x_k is made. W_{t+1} is given as

$$W_{k+1} = \{\hat{\delta}_{k+1,t}, \hat{e}_{k+1}, \hat{\tau}_{k+1}\} \tag{4}$$

The vector $\hat{\delta}_{k+1,t}$ represents the change in the number of individuals in each triage category t that follows the dynamics described in the previous section, $\hat{e}_{k+1}$ is the event that triggers the transition to state S_{k+1}, and $\hat{\tau}_{k+1}$ is the inter-transition time to the next event $k+1$.

3.1.4 Contribution

At each event a contribution is received, which is the number of individuals loaded onto the helicopter or ship. The contribution function is given as

$$C(x_k) = \sum_{t \in \mathcal{T}} x_{kt} \tag{5}$$

3.1.5 Objective Function

Given this MDP, the problem is to find the best decision policy. Let $X^{\pi}(S_k)$ be a decision policy that determines how many individuals from each triage category are loaded onto either a helicopter or ship for each state $S_k \in S$. To determine the optimal policy π^* from the class of policies $(X^{\pi}(S_k))_{\pi \in \Pi}$, the objective function is defined as

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\sum_{k=0}^{\mathcal{K}} C\left(X^{\pi}(S_k)\right) | S_0 \right] \tag{6}$$

where $\mathbb{E}^{\pi}$ indicates the expectation over the chosen policy. The optimal policy is found via Bellman's equation

$$V(S_k) = \max(C(X^{\pi}(S_k) + \mathbb{E}[V(S_{k+1}) | S_k, X^{\pi}(S_k)]) \tag{7}$$

using the decision function

$$X^{\pi}(S_k) = \underset{x_k \in \mathcal{X}_k(S_k)}{\operatorname{argmax}} \left(C(x_k) + \mathbb{E}[V(S_{k+1}) | S_k, x_k] \right) \tag{8}$$

where $\mathcal{X}_k(S_k)$ captures the constraints that define the feasible region.

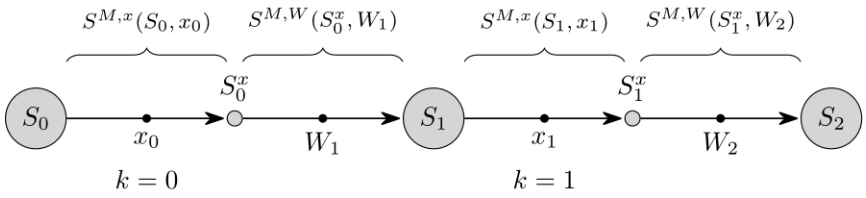


Figure 3: Example of a sequence of decisions as an MDP with the Post-Decision State Variable (PDSV). At each decision epoch k, a vector of decisions x_k is selected via a decision policy $X^{\pi}(S_k)$.

3.2 Approximate Dynamic Programming Formulation

The classic approach to finding an optimal decision policy is to solve Bellman's optimality equation via a Dynamic Programming algorithm, such as value iteration or policy iteration [43]. However, there are three factors, collectively known as the *curses of dimensionality*, which prevent the computation of an optimal policy in large-scale problems ([17], pp. 5-6): size of the state space S, size of the decision space $\mathcal{X}$, and size of the outcome space $\mathcal{W}$. When any of these become too large, it becomes too computationally expensive to determine an optimal policy.

The multi-domain evacuation problem suffers from all of these curses. For example, the scenario description in Section 2.0 states there are 1,900 individuals in the white triage category at S_0. Given the transition model between triage categories, after the first loading decision is made, any number of these individuals may

transition to any lower category, including the black category. This set of transitions alone creates a very large state space – if zero individuals transition to the black category, then there are ~1.1 billion ways in which these 1900 individuals may transition to the four categories captured in the state variable; if one individual transitions to the black category then there are an additional ~1.1 billion potential transitions; and so on. Although the number of possible transitions reduces as the scenario proceeds, the size of the state space means that it is not feasible to solve for an optimal policy via Dynamic Programming in a reasonable amount of time.

ADP aims to overcome the curses of dimensionality, albeit at the cost of producing a near-optimal policy rather than one that is optimal. To do so, it employs the concept of the Post-Decision State Variable (PDSV), labelled as S_k^x. It also approximates the value function $V(S_{k+1})$ by $\overline{V}(S_k^x)$, here labelled as the Approximate Value Function (AVF), which is a function of the PDSV rather than the state variable. The PDSV describes the state of the system after a decision x_k has been applied but before the stochastic processes occur and the associated exogenous information W_{k+1} has arrived. As depicted in Figure 3, this allows the transition function $S^M(S_k, x_k, W_{k+1})$ to be broken into two steps,

$$S_k^x = S^{M,x}(S_k, x_k) \tag{9}$$

$$S_{k+1} = S^{M,w}(S_k^x, W_{k+1}) \tag{10}$$

where Equation (9) is deterministic and Equation (10) is stochastic. The result is that the expectation operator moves outside of the max operator in Equation (7) and the decision problem in Equation (8) becomes a deterministic optimization problem, in this case given as

$$X^{\pi}(S_k) = \underset{x_k \in \mathcal{X}_k(S_k)}{\operatorname{argmax}} \left(C(x_k) + \overline{V}(S_k^x)\right) \tag{11}$$

In our recent study that explored the use of ADP to seek a near-optimal helicopter loading decision policy in a mass evacuation scenario, we compared policies generated using three different PDSVs [16]: the number of individuals in the red category at the evacuation site; a two-dimensional vector that represented the number of individuals that needed a stretcher and did not need a stretcher; and a four-dimensional vector that represented the number of individuals in each triage category. Our results indicated that a policy based on the four-dimensional PDSV performed the best, outperforming a benchmark Policy Function Approximation (PFA) by 42 ± 3%. Based on these results, in this study we use the four-dimensional PDSV, specifically $S_k^x = [\rho_{k,w}, \rho_{k,g}, \rho_{k,y}, \rho_{k,r}]$.

Regarding the AVF $\overline{V}(S_k^x)$, there are a variety of options from which to choose. These options may be grouped into three strategies [17]: a lookup table that returns a discrete value for each PDSV; parametric representation, which is an analytical function that involves a vector of tuneable parameters θ and a set of basic functions ($\varphi_f(S_t^x)$ and $f \in \mathcal{F}$) where $\mathcal{F}$ is a set of features based on information from the PDSV; and nonparametric representation, which builds local approximations based on observations, such as with neural networks, kernel regression, k-nearest neighbour, etc. Similar to our previous study [16], in this study we use a lookup table strategy; however, rather than a one-to-one lookup table, we employ state aggregation, which is a "simple form of generalizing function approximation in which states are grouped together, with one estimated value (one component of the weight vector w) for each group" ([44], p. 203). Our approach divides the state space into multiple overlapping encodings (see Figure 4 for examples of 2d encodings). This is accomplished by dividing each of the four axes of the state space into n bins, with the i^{th} axis containing n_i bins. A set of four bins $\{n_w, n_g, n_y, n_r\}$ defines an encoding, and a set of encodings defines the state aggregation scheme. The number of bins in the j^{th} encoding, N_j, is given by Equation (13), and the total number of bins in a state aggregation scheme, $|\mathcal{G}|$, where $\mathcal{G}$ is the set of all bins, is given by Equation (14).

$$N_j = \prod_i n_i \tag{13}$$

$$|\mathcal{G}| = \sum_j N_j \tag{14}$$

Given this approach, the AVF is defined as

$$\overline{V}(S_k^x) = \sum_{g \in \mathcal{G}} w_g y_g \tag{15}$$

where y_g is an indicator variable: 1 if the PDSV resides within g^{th} bin of the scheme and 0 otherwise.

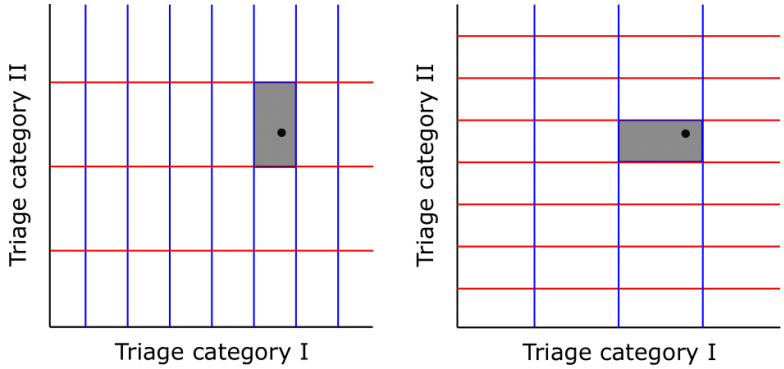


Figure 4: Example state aggregation scheme over a 2D state space. In this example, two encodings are defined using asymmetric bins along each of the state space axes. The black dot represents a PDSV in the state space, the filled in bin for each encoding represents the y_m, which would be unity, and empty bins have a y_m of zero.

4.0 COMPUTATIONAL RESULTS

This section describes how a near-optimal policy for the multi-domain evacuation problem described by the MDP outlined in Section 2.0 was generated using the state aggregation scheme outline in Section 3.2 and the Approximate Value Iteration (AVI) learning algorithm listed in Algorithm 1. The ADP-generated policy was then compared to four benchmark policies to determine which of the five performed best. A policy's performance is defined as the average number of individuals that have been successfully evacuated by the end of the MDP when the policy is followed. To estimate this quantity, the MDP was simulated many times, recording the number of individuals evacuated after each simulation. To determine the statistical significance of each policy's performance, the mean and variance of the distributions were computed and used to compare the performance of policies. The learning and policy evaluation stages were performed on a 2.5 GHz i5-10300H quad core processor with 16 GB of memory running Ubuntu 20.04. The source code was written in C++11 and compiled using the g++ 9.3.0v compiler with optimization level O3.

During both the learning and policy evaluation stages, the MDP parameters describing the evacuation remained the same. The initial population distribution and their mean transition times correspond to the values listed in Table 1. The helicopter and ship used to evacuate the individuals are described by: total capacity h^H/h^S, initial arrival time i, return time r, and a vector of capacity requirements Δ^H/Δ^S. The values for these parameters are summarized in Table 4.

Algorithm 1 Approximate Value Iteration (AVI) with state aggregation.

$n \leftarrow 1$; define ϵ; $w_g^0 \leftarrow 0, \forall g \in \mathcal{G}$
$k \leftarrow 0$.
Initialize S_k^n.
Initialize step size α_0.
for $n \leq N$ **do**
  $k \leftarrow 0$
  while $S_k \neq S^{end}$ **do**
    if $r < \epsilon$ **then**
      Randomly select $x_k \in \mathcal{X}_k$
    else
      $x_k \leftarrow \arg\max_{x_k \in \mathcal{X}_k} \left(C(S_k, x_k) + \sum_{g \in \mathcal{G}} w_g^{n-1} y_g\right)$
    end if
    Generate $\hat{v}_k^n$ associated with x_k.
    if $k > 0$ **then**
      for $\{g | y_g = 1, S_{k-1}^{x,n}\}$ **do**
        $w_g^n \leftarrow (1 - \alpha_{n-1}) w_g^{n-1} + \alpha_{n-1} \hat{v}_k^n$
      end for
    end if
    Generate the post-decision state variable: $S_k^{x,n} \leftarrow S^{M,x}(S_k, x_k)$
    Get exogenous information W_{k+1}
    Generate next pre-decision state variable $S_{k+1} \leftarrow S^M(S_k, x_k, W_{k+1})$
    $k \leftarrow k + 1$
  end while
  $n \leftarrow n + 1$
  Update α_n
end for

Table 4: Helicopter and ship parameters.

Type	h^H/h^S	i	r	Δ^H/Δ^S
Helo	10	48	3	[1,1,3,3]
Ship	50	4	16	[1,1,3,3]

There are three parameters for the learning algorithm (see Algorithm 1) which need to be set: the learning rate α, the exploration rate ε, and the number of learning iterations, *N*. A dynamic learning rate α(n) was used and updated at the end of each iteration n. The update rule for α is shown in Equation (16), which is called the Generalized Harmonic Step-size function [17]. The parameter a was chosen to be $5.62e^5$ so that the learning rate would be 0.01 at the end of the learning runs (as is suggested in Ref. [17], p. 430). The remaining two parameters were defined as follows. The exploration rate was held constant at 0.25, and the number of learning iterations was 10^7.

$$\alpha(n) = \frac{a}{a + n - 1} \tag{16}$$

Before a near-optimal policy could be learned, an appropriate state aggregation scheme had to be defined. The state aggregation scheme selected after extensive testing is listed in Table 5. Using this configuration, the total run time to determine the ADP-generated policy was 23.3 hours.

Table 5: Encodings used to define the state aggregation scheme.

Encoding 1	Encoding 2	Encoding 3	Encoding 4
{50,50,50,100}	{50,100,50,50}	{50,50,100,50}	{100,50,50,50}

4.1 Policy Evaluation

The ADP-generated policy was compared to a set of four benchmark policies. The set of benchmark policies contained two policies based on Policy Function Approximation (PFA), which are essentially rule-based policies. The first PFA policy, referred to as critical-first, evacuates individuals in the red triage category (critically ill) first and then moves on to the less severe medical states (yellow, green, and finally white). This policy represents what might be perceived as the most humane, or compassionate policy. The second PFA policy, named green-first, evacuates individuals in the green triage category first, then those in the white, red and, finally, yellow triage categories. This policy represents one that is focused on extracting those that do not need a stretcher before those that do, and in addition puts greater emphasis on those in a worse triage category in each case (no stretcher: green, then white; stretcher: red, then yellow). The next benchmark policy was a random policy, which simply chooses a decision at random from the set of allowable decisions. The final benchmark policy was a myopic policy, which similar to the PFA-based policies only used the presently available information to form its decision. However, in this case, decisions are made via a knapsack optimization model that aims to maximize the number of individuals evacuated at each epoch. If a tie between decisions was encountered, then the decision was selected randomly from among the tied decisions.

The results of the policy evaluation are illustrated in Figure 5, and the mean and variance of each distribution are listed in Table 6, which were used to perform a statistical comparison between the ADP-generated policy and each of the benchmark policies. The results of this comparison are listed in second last column of the table. As can be seen in Figure 5, the random and critical-first policies performed substantially worse than the other benchmarks and ADP-generated policy. The ADP-generated policy was able to improve upon these benchmark policies by 42 ± 16% and 31 ± 7%, respectively. The myopic policy, though a considerable improvement over the random and critical-first, still performed below the ADP-generated policy by 12 ± 4%. Finally, the green-first policy performed the best, and was statistically quite similar to the ADP-generated policy but outperformed it by 5 ± 3%.

The statistically similar results between the ADP-generated policy and the best performing benchmark policy, green-first, warrants further investigation. To gain a qualitative understanding of the strategies used by the ADP-generated policy we can examine the average cumulative number of individuals evacuated from each triage category as a function of event number. A series of these graphs are shown in Figure 6 at various points throughout the learning process. The graphs in the left column illustrate the changing distribution of individuals evacuated by helicopter, and the second column contains the graphs for the ship. The dashed lines show the average cumulative number of individuals evacuated from each triage category by the green-first policy, whereas the solid lines depict the ADP-generated policy.

As can be seen in Figure 6, after 1 million learning runs, the ADP-generated policy for both the helicopter and the ship selects fairly evenly between the white and green triage categories, almost entirely ignoring the more critical triage categories. As the learning process progresses, the number of individuals selected from the green category grows at the expense of all other categories. From this it can be seen that the ADP-generated policy is approaching the green-first benchmark policy. Although the end results are statistically similar, the amount of time required to complete the evacuation is considerably less. The green-first policy required roughly three quarters the amount of time as compared to the ADP-generated policy. This gives confidence to the idea that the green-first policy is a near-optimal policy.

Given that the ADP-generated policy does not outperform the green-first policy and that its prioritization of individuals in the green triage category tends towards this benchmark, these results indicate that for this scenario, incorporating the future value of a decision within the policy provides no value to the decision maker. This result tends to agree with guidance provided by Powell: “it should not be surprising to find out that if it is possible to move from any state to any other state (instantly and with no cost), then a myopic policy will be optimal” ([17], p. 594), where myopic may be interpreted as either a cost function approximation or a PFA. This is the situation considered in this scenario; that is, the contribution is received immediately upon loading

individuals onto either the helicopter or ship. However, this guidance does not indicate the optimal structure of this policy. In particular, Figure 5 demonstrates the wide range of results achieved for different policies that ignore a decision's downstream impact, ranging from random policy that performs rather poorly to the green-first policy that performs nearly 1.75 times better in terms of lives saved.

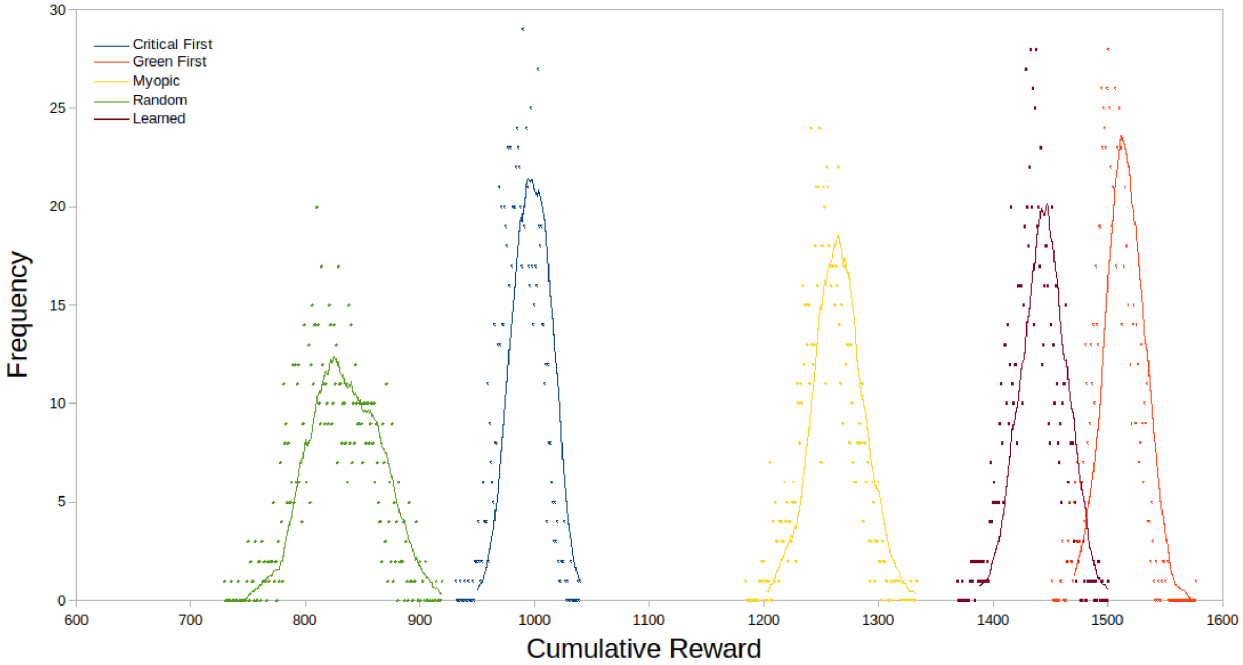
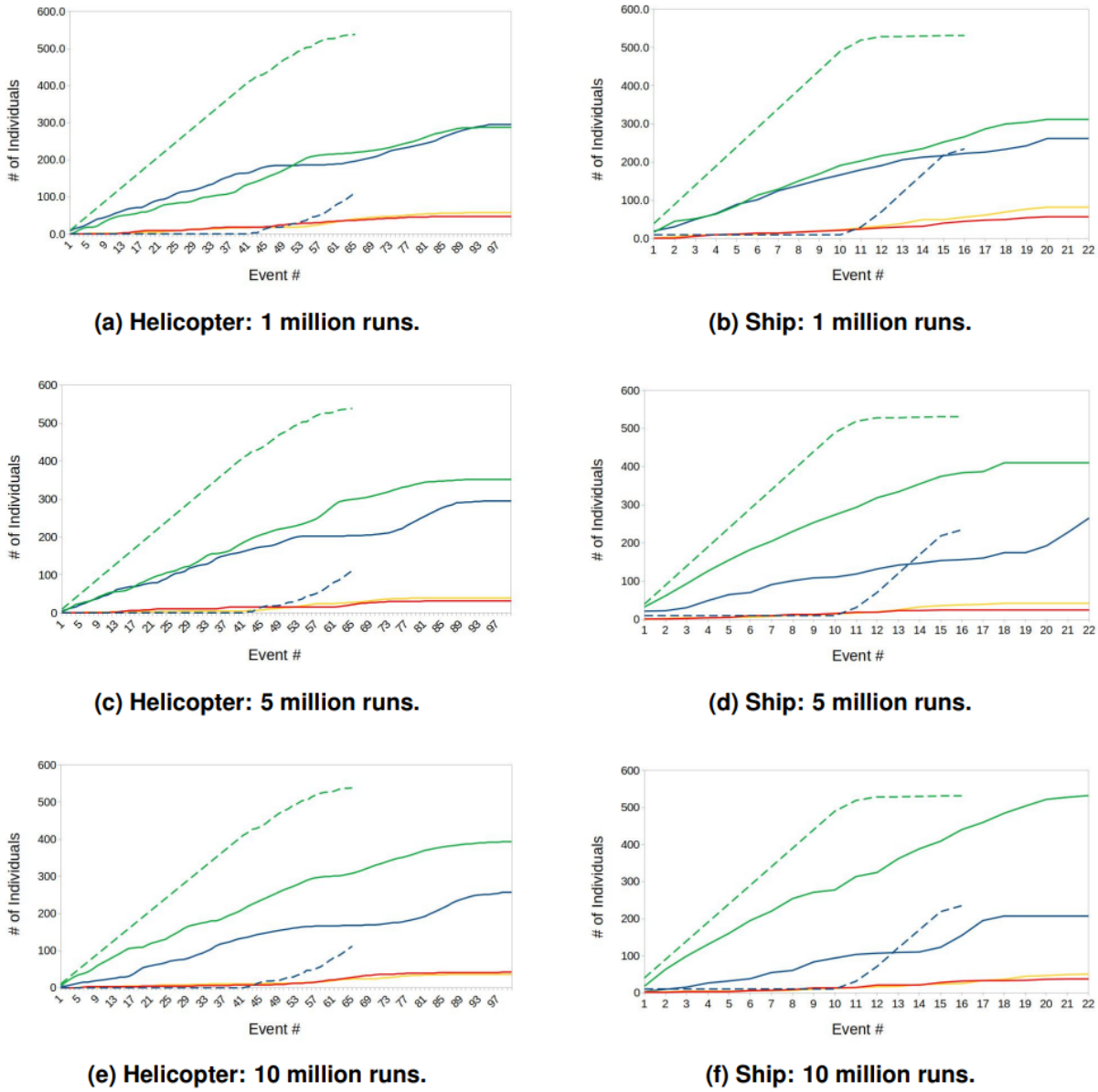


Figure 5: Distributions of the expected number of individuals evacuated when following each of the five policies.

Table 6: Summary of policy performance. Statistical comparison between the ADP-generated policy and the four benchmark policies are shown in the second last column. The values show the 95% confidence intervals (CI)s of the expected difference between the performance of the ADP-generated and benchmark policies.

Policy	Mean	Variance	Comparison (95% CI)	Rank
Green-First	1504	17.0	-5 ± 3%	1
ADP-generated	1434	21.2	-	2
Myopic	1255	23.0	12 ± 6%	3
Critical-First	987	16.5	31 ± 17%	4
Random	823	31.3	42 ± 16%	5



(a) Helicopter: 1 million runs.

(b) Ship: 1 million runs.

(c) Helicopter: 5 million runs.

(d) Ship: 5 million runs.

(e) Helicopter: 10 million runs.

(f) Ship: 10 million runs.

Figure 6: The solid lines show the average cumulative number of individuals evacuated from each triage category as a function of event number when following the ADP-generated policy. The dashed lines show the average cumulative number of individuals evacuated from each triage category when following the green-first policy. Key: Blue: white triage category, Green: green triage category, Yellow: yellow triage category, Red: red triage category.

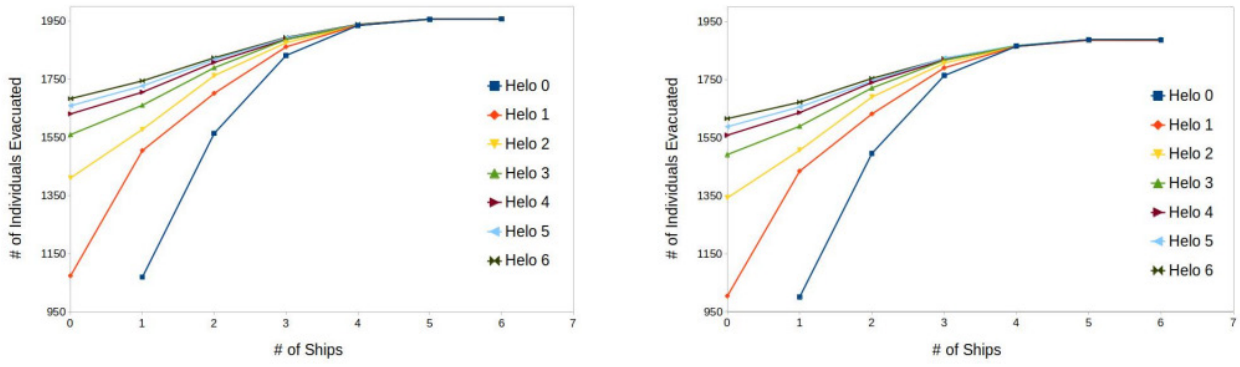
4.2 Test Scenarios

Next, we apply the ADP-generated policy within a range of test scenarios and compare the outcome to that achieved using the green-first policy. The test scenarios selected are similar to the representative planning scenario, with the exception of the number of helicopters and ships employed. Specifically, we evaluate each policy in 48 scenarios, where the number of helicopters and ships ranged between zero and six (with at least one of either being present). The MDP was simulated 35 times for each scenario, and policy combination and the expected number of evacuated individuals, depicted in the scenario, was computed. The initial arrival time of each ship and helicopter is listed in Table 7.

Table 7: Initial arrival times of helicopters and ships in test scenarios.

	Initial arrival Time [h]	
	Helicopter	Ship
1	48	4
2	49	5
3	50	6
4	51	7
5	52	8
6	53	9

The results indicate that the green-first policy, in which the helicopter and ship policies are non-coordinated, outperforms the ADP-generated policy regardless of the number of ships or helicopters (Figure 7). In addition, for both policies, the results indicate that when four ships are available, the presence of helicopters provides little benefit within the context of the scenario studied. However, it should be noted that these results may be sensitive to the scenario's assumptions, including the helicopter and ship revisit times, their capacities, the number of individuals to be evacuated, and so on. Regardless, although the ADP-generated policy did not outperform the green-first policy in either the representative or test scenarios, modelling the problem as an MDP and searching for a near-optimal policy demonstrates how these methodologies may be used to provide decision support to multi-domain operations.



(a) Green-first policy. (b) ADP-generated policy.

Figure 7: Test scenario results.

5.0 CONCLUSION

This article examined a MAJMAR scenario in which a large number of individuals, whose health stochastically deteriorates over time, are stranded at a remote location, and must be evacuated. Within this context, a multi-domain evacuation operation was examined, where individuals are evacuated either by air or sea, with the aim of the operation being to maximize the number of survivors. This problem was modelled as an MDP, and due to the curses of dimensionality, an ADP approach was employed to search for a near-optimal policy. Our search focused on a decision policy based on a VFA that used a lookup table representation based on state aggregation. The ADP-generated policy, which included two sub-policies – one for the helicopter and one for

the ship – was evaluated using a baseline scenario and compared to four benchmark policies: a random policy, a myopic policy, a PFA that prioritizes time-critical individuals first, and a PFA that prioritizes healthy individuals first.

The results of the policy evaluation demonstrated that in the context of the scenario and the MDO being studied, prioritizing the evacuation of the healthy individuals first maximizes the number of survivors. This benchmark policy outperformed the ADP-generated policy by 5 ± 3%; however, the ADP-generated policy was shown to prioritize individuals in the green triage category in a similar way to the green-first policy, indicating that both may be near-optimal policies. This result generally agrees with Ref. ([21], p. 813) which suggested that "[starting] from the most urgent jobs and [moving] onto those that are less urgent as resources become available" is not the best policy to save the most lives and that "who gets the top priority [to be evacuated] should ideally depend on the number of patients as well their injury characteristics" (Ref. [21], p. 827). Simply stated, to maximize the number of survivors, it is beneficial to evacuate those who have less critical care needs than those with a lower chance of survival.

In addition, our results indicated that the coordination of loading policies between domains may not be required. This is for two reasons. First, the ADP-generated policy, which incorporated the downstream impact of a decision made here and now, did not outperform the green-first benchmark policy that does not consider the future value of a decision. Second, the ADP-generated policy followed a similar prioritization scheme to the green-first benchmark in both domains. Given that considering downstream impact may provide no value and that maximum number of survivors occurs when both domains employ the same policy, it is concluded that within the scenario studied each domain's loading policy may operate independently.

Although our results indicate that the coordination of loading decision policies between the air and maritime domain may not be required, the scenario considered does not account for an individual's health while en route to an FOL, while at an FOL, or during transport to a southern location. These dynamics are accounted for to a degree in Ref. [15], however, the authors do so in a deterministic manner within a single optimization model rather than stochastically. Consideration of these dynamics, along with incorporation of a wider range of uncertainties (e.g., weather, breakdowns, transit time, etc.), and an evaluation of the sensitivity of the candidate decision policies to variations in the mean transition times between triage categories, is required to gain a more fulsome understanding of how the policies perform. In addition, incorporating the abovementioned dynamics will result in a wider range of decision policies that must be considered, including which individuals are loaded onto fixed-wing transport at an FOL, refuelling decisions at the FOL and southern location, and – as the number of transports are increased – decisions regarding their coordination due to limited space at an FOL may be required. In addition, incorporating these components will result in a reward being received when individuals arrive at the southern location, rather immediately upon loading at the evacuation site, which may alter the structure of the loading policy. To seek near-optimal policies within the context of such a model, it is anticipated that RL/ADP will play a critical role towards providing effective decision support.

6.0 REFERENCES

[1] J.E. Walsh, F. Fetterer, J.S. Stewart, and W.L. Chapman, A database for depicting Arctic sea ice variations back to 1850. Geographical Review, vol. 107, no. 1, pp. 89-107, 2017.

[2] L.R. Mudryk, J. Dawson, S.E.L. Howell, C. Derksen, T.A. Zagon, and M. Brady, Impact of 1, 2 and 4 ◦C of global warming on ship navigation in the Canadian Arctic. Nature Climate Change, vol. 11, no. 8, pp. 673-679, 2021.

[3] Government of Canada, Arctic and Northern policy framework: Safety, security, and defence chapter, Sep. 2019.

[4] D. Dalaklis and M. Drewniak, Search and Rescue capabilities in the Arctic: Is the high north prepared at an adequate level? in Crisis and Emergency Management in the Arctic: Navigating Complex Environments (N. Andreassen and O. Borch, eds.), Routledge, 2020.

[5] A.B. Farré, S.R. Stephenson, L. Chen, M. Czub, Y. Dai, D. Demchev et al., Commercial Arctic shipping through the Northeast Passage: Routes, resources, governance, technology, and infrastructure, Polar Geography, vol. 37, no. 4, pp. 298-324, 2014.

[6] International Maritime Organization, International code for ships operating in polar waters (Polar Code). https://www.imo.org/en/OurWork/Safety/Pages/polar-code.aspx. Accessed: 13 May 2022.

[7] K. Solberg, O. Gudmestad, and B. Bjarte, SARex Spitzbergen: Search and rescue exercise conducted off North Spitzbergen: Exercise report. Tech. Rep. Report 58, University Stravanger, 2016. https://uis.brage.unit.no/uis-xmlui/bitstream/handle/11250/2414815/Rapport_58.pdf?sequence=3&isAllowed=y.

[8] National Defence, Department of National Defence and Canadian Armed Forces 2021 – 22 Departmental Plan. 2021.

[9] T. Browne, B. Veitch, R. Taylor, J. Smith, D. Smith, and F. Khan, Consequence modelling for Arctic ship evacuations using expert knowledge. Marine Policy, vol. 130, p. 104582, 2021.

[10] D. Clark, J. Ford, L. Berrang-Ford, T. Pearce, S. Kowal, and W. Gough, The role of environmental factors in search and rescue incidents in Nunavut, Canada. Public Health, vol. 137, pp. 44-49, 2016.

[11] B.I. Kruke and A.C. Auestad, Emergency preparedness and rescue in Arctic waters. Safety Science, vol. 136, p. 105163, 2021.

[12] C.N. Mills and G.H. Mills, Mass casualty incident response and aeromedical evacuation in Antarctica. The Western Journal of Emergency Medicine, vol. 12, pp. 37-42, Feb. 2011.

[13] R. Sheehan, D. Dalaklis, A. Christodoulou, M. Drewniak, P. Raneri, and A. Dalaklis, The Northwest Passage in the Arctic: A brief assessment of the relevant marine transportation system and current availability of search and rescue services. Logistics, vol. 5, no. 2, p. 23, 2021.

[14] M.C. Camur, T.C. Sharkey, C. Dorsey, M.R. Grabowski, and W.A. Wallace, Optimizing the response for Arctic mass rescue events. Transportation Research Part E: Logistics and Transportation Review, vol. 152, p. 102368, 2021.

[15] D.G. Hunter, J. Chan, and M. Rempel, Assessing the operational impact of infrastructure on Arctic operations. Tech. Rep. DRDC-RDDC-2021-R024, Defence Research and Development Canada, Feb. 2021.

[16] M. Rempel, N. Shiell, and K. Tessier, An approximate dynamic programming approach to tackling mass evacuation operations. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 01-08, 2021.

[17] W. Powell, Approximate Dynamic Programming: Solving the Curses of Dimensionality. Hoboken, New Jersey: Wiley, second ed., 2011.

[18] NATO Allied Command Transformation, Multi-Domain Operations: Enabling NATO to Out-pace and Out-Think its Adversaries. https://www.act.nato.int/articles/multi-domain-operations-out-pacing-and-out-thinking-nato-adversaries (Accessed 22 December 2022).

[19] W. Sacco, D. Navin, K. Fiedler, R. Waddell II, W. Long, and R. Buckman Jr., Precise formulation and evidence-based application of resource-constrained triage. Academic Emergency Medicine, vol. 12, no. 8, pp. 759-770, 2005.

[20] M. Argon, S. Ziya, and R. Righter, Scheduling impatient jobs in a clearing system with insights on patient triage in mass casualty incidents. Probability in the Engineering and Informational Sciences, vol. 22, no. 3, pp. 301-332, 2008.

[21] E. Jacobson, N. Argon, and S. Ziya, Priority assignment in emergency response. Operations Research, vol. 60, no. 4, pp. 813-832, 2012.

[22] A. Mills, M. Argon, and S. Ziya, Resource-based patient prioritization in mass-casualty incidents. Manufacturing & Service Operations Management, vol. 15, no. 3, pp. 361-377, 2013.

[23] M. Dean and S. Nair, Mass-casualty triage: Distribution of victims to multiple hospitals using the SAVE model. European Journal of Operational Research, vol. 238, no. 1, pp. 363-373, 2014.

[24] A. Mills, A simple yet effective decision support policy for mass-casualty triage. European Journal of Operational Research, vol. 253, no. 3, pp. 734-745, 2016.

[25] K. Shin and T. Lee, Emergency medical service resource allocation in a mass casualty incident by integrating patient prioritization and hospital selection problems. IISE Transactions, vol. 52, no. 10, pp. 1141-1155, 2020.

[26] A.K. Childers, G. Visagamurthy, and K. Taaffe, Prioritizing patients for evacuation from a health-care facility. Transportation Research Record, vol. 2137, no. 1, pp. 38-45, 2009.

[27] F. Duchateau, L. Verner, O. Cha, and B. Corder, Decision criteria of immediate aeromedical evacuation. Journal of Travel Medicine, vol. 16, pp. 391-394, Sep. 2009.

[28] K. Klein and N. Nagel, Mass medical evacuation: Hurricane Katrina and nursing experiences at the New Orleans Airport. Disaster Management and Response, vol. 5, no. 2, pp. 56-61, 2007.

[29] N. Caglayan and S.I. Satoglu, Multi-objective two-stage stochastic programming model for a proposed casualty transportation system in large-scale disasters: A case study. Mathematics, vol. 9, no. 4, 2021.

[30] S. Keneally, M. Robbins and J. Lunday, A Markov decision process model for the optimal dispatch of military medical evacuation assets. Health Care Management Science, vol. 19, pp. 111-129, 2016.

[31] P. Jenkins, B. Lunday and M. Robbins, Robust, multi-objective optimization for the military medical evacuation location-allocation problem. Omega, vol. 97, 102088, 2020.

[32] M. Robbins, P. Jenkins, N. Bastian and B. Lunday, Approximate dynamic programming for the aeromedical evacuation dispatching problem: Value function approximation using multiple level aggregation. Omega, vol. 91, 102020, 2020.

[33] D. Rawat, Secure and trustworthy machine learning/artificial intelligence for multi-domain operations. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III (T. Pham and L. Solomon, eds.), vol. 11746 of Proceedings of SPIE, SPIE, 2021. Conference on Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III, Electr Network, Apr. 12 – 16, 2021.

[34] J.Z. Hare, B.C. Rinderspacher, S. Kase, S. Su and C.P. Hung, Battlespace: Using AI to understand friendly vs. hostile decision dynamics in MDO. In Artificial Intelligence and Machine Learning for Multidomain Operations Applications III (T. Pham and L. Solomon, eds.), vol. 11746 of Proceedings of SPIE, SPIE, 2021. Conference on Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III, Electr Network, Apr. 12 – 16, 2021.

[35] D.C. Verma, S.S. Gartner, D.H. Felmlee, D. Braines, and R. Yarlagadda, Using ai/ml to predict perpetrators for terrorist incidents. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II (T. Pham, L. Solomon, and K. Rainey, eds.), vol. 11413 of Proceedings of SPIE, SPIE, 2020. Conference on Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II Part of SPIE Defense + Commercial Sensing Conference, Electr Network, Apr 27 – May 08, 2020.

[36] D. Verma, G. White, and G. de Mel, Federated AI for the enterprise: A web services based implementation. In 2019 IEEE International Conference on Web Services (ICWS), pp. 20-27, 2019.

[37] S. Kriegler, Artificial Intelligence Guided Battle Management: Enabling Convergence in Multi-Domain Operations. Master's thesis, School of Advanced Military Studies, US Army Command and General Staff College, Fort Levenworth, Kansas, May 2020.

[38] D.E. Asher, A. Basak, R. Fernandez, P.K. Sharma, E.G. Zaroukian, C.D. Hsu, M.R. Dorothy, T. Mahre, G. Galindo, L. Frerichs, J. Rogers, and J. Fossaceca, Strategic maneuver and disruption with reinforcement learning approaches for multi-agent coordination. 2022.

[39] M. Rempel and J. Cai, A review of approximate dynamic programming applications within military operations research. Operations Research Perspectives, vol. 8, p. 100204, 2021.

[40] Bloomberg Businessweek, Apocalypse tourism? Cruising the melting Arctic Ocean. https://www.bloomberg.com/features/2016-crystal-serenity-northwest-passage-cruise/

[41] International Maritime Organization, SOLAS (Consolidated Edition, 2020). London, United Kingdom, 2020.

[42] M. Puterman, Markov Decision Processes: Discrete stochastic dynamic programming. Hoboken, New Jersey: Wiley, 2005.

[43] R. Bellman, Dynamic Programming. Princeton University Press, 1957.

[44] R. Sutton and A. Barto, Reinforcement Learning: An Introduction. London, England: MIT Press, second ed., 2020.