Stochastic Policies in Morally Constrained (C-)SSPs

Charles Evans
The Australian National University
Canberra, Australia
u6942700@anu.edu.au

Claire Benn
The Australian National University
Canberra, Australia
Claire.Benn@anu.edu.au

Ignacio Ojea Quintana
The Australian National University
Canberra, Australia
Ignacio.Ojea@anu.edu.au

Pamela Robinson
The Australian National University
Canberra, Australia
Pamela.Robinson@anu.edu.au

Sylvie Thiébaux
The Australian National University
Canberra, Australia
Sylvie.Thiebaux@anu.edu.au

ABSTRACT

Stochastic policies often outperform deterministic ones. This is especially true for Constrained Stochastic Shortest Path (C-SSP) problems, a popular approach to planning under uncertainty with multiple objectives. Nevertheless, there are moral concerns about stochastic policies that should deter us from selecting them. In this paper, we identify some of these moral concerns and offer 'acceptability constraints' that allow only certain stochastic policies to be selected. We propose a novel C-SSP solver able to integrate our moral acceptability constraints, we evaluate its performance in a relevant test problem, and we show that our approach can successfully produce acceptable policies in morally significant domains.

CCS CONCEPTS

• **Computing methodologies → Planning under uncertainty**; **Philosophical/theoretical foundations of artificial intelligence**.

KEYWORDS

ethical decision making; automated planning; uncertainty; constrained Stochastic Shortest Path problems; stochastic policies; moral constraints; risk of harm

ACM Reference Format:
Charles Evans, Claire Benn, Ignacio Ojea Quintana, Pamela Robinson, and Sylvie Thiébaux. 2022. Stochastic Policies in Morally Constrained (C-)SSPs. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3514094.3534193

1 INTRODUCTION

Many planning problems involve morally-loaded objectives, such as *protecting personal safety* or *improving health outcomes*. Often these are accompanied by non-moral objectives, such as time or monetary constraints, and significant uncertainty. A popular approach is to model such problems as Constrained Stochastic Shortest Path (C-SSP) problems, where the various objectives are captured by different cost functions, one of which is optimised, while the others are constrained. The optimal policy in a C-SSP is usually a stochastic policy (as are the vast majority of near-optimal policies) [1, 7].

Being optimal with respect to expected cost is a moral good when that expected cost captures morally significant objectives. Nevertheless, stochastic policies can have moral downsides. Choosing a stochastic policy over a deterministic one can introduce morally relevant risks and disparate outcomes, raising moral concerns about harm, risk, inequality and unfairness. Despite the benefits of adopting stochastic policies, we might sometimes have most reason not to choose them.

Our paper addresses these moral concerns directly. We identify the various moral concerns that might be had about stochastic policies and propose several additional constraints to distinguish *acceptable* policies, which address our concerns, from *unacceptable* ones, which do not. We also propose a method for weighing the gains with respect to expected cost against our acceptability measures, relative to a baseline (often deterministic) policy. This allows us to select stochastic policies that are not only *feasible* (that satisfy our objectives) but are also *acceptable* (the moral reasons to avoid some stochastic policies don't apply to them). We offer an approximate anytime algorithm that can do just that, selecting plans that are both feasible and acceptable in the context of morally significant C-SSPs. Our work, therefore, offers concrete ways to choose stochastic policies while being sensitive to moral concerns. This, in turn, enables us to use stochastic policies more effectively in a range of morally significant domains.[1]

The paper is structured as follows. Section 2 identifies two areas of related work and explains how our paper extends the existing literature. Section 3 provides the required technical background on C-SSPs. Section 4 contains the bulk of our analysis and philosophical contribution. There, we tackle the problem of how to further constrain a C-SSP to avoid the various moral concerns we may have about stochastic policies. Section 5 introduces our novel solver, which allows us to enforce the constraints developed in Section 4. Section 6 demonstrates our solver's effectiveness in an example domain. Section 7 summarises our findings and provides suggestions for future work.

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3534193

[1] Due to their greater robustness, stochastic policies are widely used beyond C-SSPs. In particular, they are also frequently sought in the context of ordinary SSPs (which feature a single objective and no constraint), even though there always exists a deterministic optimal policy for such a problem. Our moral concerns and accompanying constraints apply equally to stochastic policies for these simpler problems.

2 RELATED WORK

Ethically-Conscious Planning

There has lately been increased interest in (i) finding ethically acceptable plans under uncertainty in morally relevant domains, and (ii) evaluating existing policies for their ethical content [3, 6, 14, 16, 20]. In research on (i), the state-of-the-art approach is an 'Ethically Compliant Autonomous System' (ECAS) [16, 20]. In an ECAS, a Markov Decision Process (MDP) – a close relative to the plain SSP – is defined for some task, and the solver must find the optimal policy for this MDP that satisfies a 'moral principle', which is a mapping from policies to booleans that purportedly evaluates the moral acceptability of the policy. Multiple ECAS formulations are proposed to represent particular moral theories, such as the divine command theory or virtue ethics [20]. The main benefit of these approaches is that only the objective of task completion is minimised, while ethical considerations are introduced as constraints. This avoids convoluting various conflicting incentives into the objective function, which can lead to unpredictable and undesirable behaviours in policies. Similarly, in [6], autonomous agents minimise the cumulative 'concern' associated with the principles they break in completing a task.

In the research on (ii), evaluating the permissibility of existing plans, there is recent work such as [14], which again attempts to define the formal properties of plans that are (im)permissible from the perspective of moral viewpoints (e.g., utilitarianism, deontology, etc.). It then explores the computational complexity of evaluating whether a given plan adheres to the properties of some such viewpoint. (This is done for planning in certain, rather than uncertain, environments.) A theme in this body of literature is the method of focusing on optimising the completion of a non-moral task so long as it has no morally bad side effects. Consider, for example, an autonomous car tasked with reaching a destination without injuring or endangering pedestrians.

We address an important and under-examined set of problems within this broader context. The combination of optimising for a *morally* significant objective, while remaining responsive to additional non-moral constraints, is a natural problem-framing in many areas. Consider, for example, an agent tasked with coordinating a disaster response, or with treating a patient at a hospital [12, 13, 23]. The primary objective in such scenarios is to ensure the safety and wellbeing of the affected people, though it would also be irresponsible and impractical to disregard competing objectives, like the conservation of time and money. Our paper, therefore, makes a significant contribution to this existing body of work by extending consideration to multi-objective contexts as well as by focusing on primary tasks which are explicitly morally significant.

Risk and Variance Constraints

There is existing research on limiting the risk and variance induced by policies in MDPs. For example, [9] proposes a method for penalising the expected reward of a particular policy for a MDP if the expectation is subject to excessive variance. [15] proposes a move towards constrained mean-variance optimisation, in which a policy is sought for which the expected cumulative reward obeys some lower bound and the variance over cumulative reward obeys an upper bound. Other proxies for risk besides variance have also been employed. [4, 5]

However, none of this work focuses explicitly on risk in a moral setting. Further, insofar as it has upshots for this kind of risk, it only aims to mitigate risk and variance arising from the stochastic effects of actions. Our work contributes to this existing literature by focusing specifically on the other potential source of risk and variance: stochastic policies. Stochastic action effects are an instance of environmental stochasticity. While one policy might be riskier than another in such an environment, all risk is ultimately attributable to the environment as opposed to the policy itself. Stochastic policies *introduce* risk by using probabilistic devices to determine the exact strategy to be taken on a particular execution of the policy. Given that C-SSP solvers often settle on stochastic policies, if we want to solve morally-significant problems with C-SSPs, we must consider whether the risk introduced by these policies is morally relevant. We'll first offer some background on C-SSPs.

3 BACKGROUND: (CONSTRAINED) STOCHASTIC SHORTEST PATH PROBLEMS

An (unconstrained) **Stochastic Shortest Path Problem (SSP)** is a planning problem in an uncertain environment with a well-defined start point and end points in the state space [2]. Formally, such a problem is defined as the tuple $\langle S, s_0, G, A, P, C\rangle$, where:

- S is the set of states, and:
  - $s_0 \in S$ is the 'initial' or 'start' state
  - $G \subset S$ is the set of valid end or 'goal' states
- A is the set of actions
  - $A(s)$ denotes the actions that are applicable from a state $s \in S$
- P is the transition probability function, where $P(s'|s,a)$ is the probability of transitioning from state s to state s' when applying action $a \in A(s)$
- $C(s,a) : S \times A \rightarrow \mathbb{R}^+$ defines a strictly positive real cost associated with taking action a from state s

In a SSP, the problem is to find a *policy* enabling an agent to reach the goal G starting from the initial state s_0, while incurring the minimum cumulative expected cost. A policy may be either *deterministic* or *stochastic*, where:

- A *deterministic* policy is a function $\pi : S \rightarrow A$ which, for any given state, returns a particular action.
- A *stochastic* policy can be interpreted in two ways, which are effectively equivalent in expectation but are both useful interpretations in certain scenarios [10]:
  - A function $\pi : S \times A \rightarrow [0,1]$, which returns the probability with which a particular action should be taken from a particular state. In this paper, this will be referred to as a **distributed stochastic policy**. This is the canonical definition of a 'stochastic policy'; however, as we will make use of both this and the next interpretation in this paper, it is worth introducing this distinguishing terminology.
  - A function $\pi : \{\pi_D | \pi_D \in \Pi_D\} \rightarrow [0,1]$, or a function which defines a distribution over the SSP's set of deterministic policies Π_D, and returns a probability with which a particular deterministic policy π_D should be followed. In this paper, this will be referred to as a **concentrated stochastic policy**.

Note that a concentrated policy $\pi_{\text{conc.}}$ may easily be converted to a distributed stochastic policy $\pi_{\text{dist.}}$ that is equivalent in expectation, by defining:

$$\pi_{\text{dist.}}(s,a) = \sum_{\pi_D \in \Pi_D} \pi_{\text{conc.}}(\pi_D) \cdot \pi_D(s,a)$$

where, for a deterministic policy π_D, $\pi_D(s,a) = 1$ if $\pi_D(s) = a$ and 0 otherwise.

When following a certain policy π in an SSP, we write $V^\pi(s)$ for the 'value of s under π', i.e., the expected cumulative cost (according to the cost function C) incurred if π is adhered to from s until a goal state is reached. Similarly, $Q^\pi(s,a)$ represents the 'Q-value of s, a under π', i.e., the expected cumulative cost if action a is taken from state s, and policy π is adhered to thereafter. We can therefore denote the expected cost or value of a policy π in totality as $V^\pi(s_0)$. An optimal policy is one that minimises $V^\pi(s_0)$. It is well known that, for any SSP, at least one *optimal* policy is deterministic.

A **Constrained Stochastic Shortest Path Problem (C-SSP)** is an extension of the SSP model (described above) to capture situations in which multiple competing cost functions need to be considered. For a search and rescue drone, for example, one might consider two cost functions: the number of victims located, and the energy levels in the drone's battery. The problem is to find a policy optimising one of these cost functions (say, the number of victims located) while keeping the others (here, just the energy) under given bounds. Much of the groundwork described above for (unconstrained) SSPs holds, but there are some key differences.

Formally, a C-SSP introduces an additional k 'secondary' cost functions to coexist with what we now call our 'primary' cost, C. For clarity, C is instead referred to as C_0 in the C-SSP, while our other k functions are referred to as $C_1, C_2, ..., C_k$; this defines a cost function vector $\overrightarrow{C} = [C_0, C_1, ..., C_k]$. In the C-SSP context, states and state-action pairs are valued with respect to the i^{th} cost under policy π with the extended notations $V_i^\pi(s)$ and $Q_i^\pi(s,a)$ for some state s and some action a. That is, the subscript on these functions denotes the cost with respect to which that evaluation is being made.

The objective is to find a policy π that minimises the expected cumulative primary cost $V_0^\pi(s_0)$, and keeps each expected cumulative secondary costs $V_i^\pi(s_0)$ below a given bound $\hat{c}_i$, for $1 \leq i \leq k$. Writing $\overrightarrow{\hat{c}} = [\hat{c}_1, \hat{c}_2, ..., \hat{c}_k]$ for the vector of these secondary cost bounds allows us to define a C-SSP as a tuple $\langle S, s_0, G, A, P, \overrightarrow{C}, \overrightarrow{\hat{c}} \rangle$. We will also introduce to our vocabulary the notion of *feasibility*: a policy π for some C-SSP is feasible if and only if it satisfies the secondary cost constraints.

One crucially important property that we wish to draw the reader's attention to is that, unlike in the unconstrained SSP case, the optimal feasible policy for a C-SSP is no longer guaranteed to be deterministic; in the general case, it will be a stochastic policy [7, 8]. Therefore, when executing a policy for a C-SSP, two sources of stochasticity typically arise. The first is the stochasticity of the policy: it captures a probabilistic *choice* of actions. The second is the stochasticity of the state transition function: it captures the probabilistic *effects* of actions. A planner generating policies will have control over the former but not the latter. This paper is concerned with the ethical implications of stochastic *policies*. For this reason, and to avoid any confusion between these two sources of uncertainty, the running example we use in Section 4 when discussing ethical considerations deliberately features deterministic action effects only. However, our proposed constraints and computational approach also apply when the policies' actions have stochastic effects. The more complex example we use in Section 6 to evaluate our approach does feature such stochastic effects.

4 MITIGATING ETHICAL CONCERNS ABOUT STOCHASTICITY

As we have said, stochastic policies can often do the best, in expectation, at realising our primary objectives within the constraints given by our secondary objectives. Nevertheless, there may be ethical reasons not to adopt them. In this section, we consider three sets of ethical concerns about stochastic policies, offering standards that policies must meet if they are to address them. Let's begin with an example.

4.1 The Autonomous Medic

4.1.1 Problem. A medic is operating in a hospital. The medic's stated purpose is to administer painkillers in order to reduce each patient's level of self-reported pain, measured on an integer scale of zero to ten, inclusive.[2]

At each timestep during treatment, the actions available to the medic are to (a) administer a particular painkiller from its available set, or (b) discharge the patient, which will end the treatment episode. Each painkiller in the available set will reduce the patient's pain by x with probability p. If x is greater than the patient's current pain level, the patient will be left a pain level of 0 (i.e., the patient's pain level can only be reduced to zero and no further). We assume that each timestep leaves sufficient time for any painkiller to take complete effect during the transition to the next state, and that any administered painkillers' effect does not degrade during an episode. Applying the 'discharge' action transitions us to a goal state of the C-SSP.

In the initial state s_0, no painkillers have been administered. When determining whether to administer a painkiller or discharge, and which painkiller to administer, there are two relevant costs:

- C_0 is the morally-loaded cost function and the primary cost of the C-SSP. It's constituted by the pain level of the discharged patient. If the medic applies the discharge action, then, for some state s, $C_0(s, \text{'discharge'})$ will be equal to the current pain level of the patient at s; for any other action a, $C_0(s,a)$ will be a negligible positive constant.
- C_1 is a secondary cost in the C-SSP which represents the monetary cost of an administered painkiller. Each painkiller a will have an associated monetary cost a_{cost}, and so $C_1(s,a) = a_{\text{cost}}$ for any state s. Discharging has no monetary cost; i.e., $C_1(s, \text{'discharge'}) = 0$ for any s.

The objective of the C-SSP is to minimise C_0 (minimise the patient's pain upon discharge), subject to an upper bound $\hat{c}_1$ on the expected cost of the policy from the initial state according to C_1.

[2] We make the simplifying assumption that the patients' self-reported pain levels are accurate.

4.1.2 Example Instance. In this instance, which we will call T, we assume for simplicity (i) that there are no stochastic effects, and hence that all stochasticity is introduced by the medic; and (ii) that in the initial state s_0, the patient's self-reported pain is 10.

The medic in T has three available painkillers:

- A, which reduces pain by 9 and costs \$1200
- B, which reduces pain by 7 and costs \$1000
- C, which reduces pain by 4 and costs \$200

The objective of the C-SSP is to minimise the patient's pain upon discharge, subject to an upper bound $\hat{c}_1$ on the expected cost. Let us assume that $\hat{c}_1 = 1000$, i.e., there is an upper bound of \$1000 on the expected combined cost of the painkillers administered to the patient. Let us also assume that the same painkiller cannot be given more than once.

So what is the best deterministic policy? Let π_A be the deterministic policy that has the medic administer painkiller A at time 0, then discharge the patient at time 1. π_B and π_C are defined identically with respect to painkillers B and C. Each policy can be evaluated as follows:

- $V_0^{\pi_A}(s_0) = 1$, and $V_1^{\pi_A}(s_0) = 1200$
- $V_0^{\pi_B}(s_0) = 3$, and $V_1^{\pi_B}(s_0) = 1000$
- $V_0^{\pi_C}(s_0) = 6$, and $V_1^{\pi_C}(s_0) = 200$

Policy π_A has too high a secondary cost to be feasible: it violates the constraint to cost less than \$1000. While both policies π_B and π_C are feasible (they cost \$1000 or less), π_B is best because it results in less patient pain upon discharge (and thus has a lower primary cost than π_B).

However, if we allow ourselves to consider stochastic policies as well as deterministic ones, it looks like we can do better with respect to the primary objective while satisfying the secondary cost constraints. Consider the (concentrated) stochastic policy π_S, defined as $\pi_S(\pi_A) = 0.8$, $\pi_S(\pi_C) = 0.2$, and $\pi_S(\pi_X) = 0$ for any other deterministic policy π_X. With this policy, the agent would follow π_A 80% of the time and π_C 20% of the time. With this policy,

- $V_0^{\pi_S}(s_0) = (0.8)(1) + (0.2)(6) = 2$
- $V_1^{\pi_S}(s_0) = (0.8)(1200) + (0.2)(200) = 1000$

Policy π_S is therefore feasible: it costs \$1000 per patient *on average*. It also results in a better outcome for patients on average than the best deterministic policy, as π_S results in patients being discharged with a reported pain level of 2 on average. This is achieved because a cheap but bad deterministic policy is taken with probability 0.2, saving enough money in expectation to feasibly perform (with a probability of 0.8) a deterministic policy with better patient outcomes and which would otherwise be too expensive.

We therefore have some reasons to prefer the stochastic policy in this instance. However, there may still be other reasons to prefer the deterministic policy. We next discuss three sets of ethical concerns which, if they can't be addressed, give us reason to prefer deterministic policies. We propose several different constraints on policy choice to address these concerns.

4.2 Caring about the worst off

4.2.1 The worst off individual outcome. In the Autonomous Medic Problem instance T, deterministic policies ensure that the pain levels of all patients are reduced by the same amount. Stochastic policies reduce some patients' pain levels by a lot and some by a little. There may, therefore, be possible policy executions that are so bad for some patients that these policies should be rejected. In general, stochastic policies tend to allow for a range of possible outcomes for those affected by them. This raises concerns for those negatively affected. We might wonder how bad the stochastic policy could be for its unluckiest moral subject (i.e., the one subject to the worst execution of the policy), and whether this potential low level of wellbeing is acceptable. Let's suppose that there's a threshold below which levels of wellbeing are unacceptable. Even if a person's actual level of wellbeing remains above this threshold as a result of an application of the policy, the person may still be wronged in virtue of being exposed to the *risk* that their wellbeing could fall below the threshold. Many ethicists hold that being subjected to the *risk* of harm is a harm, even if you are not in fact harmed [11, 17].

To ensure that no policy is chosen that allows anyone's wellbeing to fall below some predefined threshold, we can constrain policy choice by placing an upper bound h on the primary cost of the worst-case outcome of any stochastic decision that is a feasible policy. Let's consider what this looks like for both a concentrated stochastic policy and for a distributed stochastic policy.

- For a concentrated stochastic policy π for a C-SSP C, the worst-case primary cost will be the maximum expected primary cost of any deterministic policy for C that π gives us a nonzero chance of sampling; call this value w. We can calculate w as follows:
  - Let Π_D denote the set of deterministic policies for C. Throughout the paper, we will use π_D as a variable for the deterministic policies in Π_D.
  - Then $w = \max_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0)$, where s_0 is the initial state of C.
  - As discussed above, we want to define an upper bound on w, here h, where w exceeding h would indicate a level of unacceptable harm or risk of harm to our subject(s). So, the desired constraint will take the following form:
$$\max_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0) \leq h \tag{C1.1}$$
- For a distributed stochastic policy π for a C-SSP C, and a non-goal state s in the state space of C, the worst-case primary cost will be the maximum primary Q-value associated with any action that π gives us a nonzero chance of sampling in s; let us call this value w_s. We can calculate w_s as follows:
  - $w_s = \max_{\{a\in A(s)\,|\,\pi(s,a)>0\}} Q_0^{\pi}(s,a)$
  - As in the concentrated case, we want to introduce an upper bound on w_s, here h_s, where w_s exceeding h_s would indicate an unacceptable level of harm or risk of harm to our subject(s). So the desired constraint will take the following form:[3]
$$\max_{\{a\in A(s)\,|\,\pi(s,a)>0\}} Q_0^{\pi}(s,a) \leq h_s \tag{C1.2}$$

4.2.2 A more robust notion of 'the worst off'. The constraints just proposed (C1.1 and C1.2) are sensitive to the plight of the *absolute* worst-off moral subject under a particular stochastic policy. They

[3]Note that (C3.2) applies just to an individual state s, but could (should) be applied for all non-goal states in the morally significant C-SSP.

require that the expected harm (or risk of harm) faced by the worst-off does not exceed some upper bound h. But these constraints don't track how likely (or unlikely) it is that the worst cases manifest, and the likelihood that they do is also morally relevant.

Compare, for example, the *distributions* of possible outcomes in the Example Instance T. There, stochastic policy π_S followed action C with probability 0.2 (pain level 6), and action A with probability 0.8 (pain level 1). It might be wrong to release patients with pain level 6 *simpliciter*, but a more sophisticated approach would also factor in the *likelihood* that a patient will actually be released with that pain level [5, 19]. To capture this, we can use the notion of 'value at risk'. When we consider the value at risk, we look at the worst case scenarios that jointly carry a fixed cumulative probability. More precisely, the value at risk of a random variable X with 'confidence' $\alpha \in [0, 1]$ is the minimum value in X's distribution with cumulative probability greater than α, i.e., $VaR_\alpha(X) = \min\{x | P(X \leq x) \geq \alpha\}$. Hence, we can have confidence that with probability α costs will not exceed $VaR_\alpha(X)$. We can then introduce the 'conditional value at risk' (CVaR) of X with confidence α as $CVaR_\alpha(X) = \mathbb{E}[X | X \geq VaR_\alpha(X)]$ [5, 19], i.e., the expected value of the tail of the distribution that exceeds the value at risk or the expected value of the $(1-\alpha)$ (worst-off) quantile.

Observe that the whole C-SSP can be summarized by looking at a random variable X_π, where sampling from X_π allows us to compute how much expected cost the planner would incur if it were to execute policy π in the process. Since we have two ways of doing the computation (we can use concentrated or distributed policies), we use two random variables, X_π and $Y_{\pi,s}$, which take the values $V_0^{\pi_D}(s_0)$ and $Q_0^\pi(s,a)$ with probabilities $\pi(\pi_D)$ and $\pi(s,a)$ respectively.

We can then generalize constraints (C1.1 and C1.2) to consider not only the absolute worst case scenarios, but the expected cost of all of the worst case scenarios with joint probability $(1-\alpha)$:

$$CVaR_\alpha(X_\pi) \leq h \tag{C2.1}$$

$$CVaR_\alpha(Y_{\pi,s}) \leq h_s \tag{C2.2}$$

4.3 Caring about distribution

In the above discussion, we were concerned with the worst off. However, some reasons to be concerned about the worst off involve only their *relative* levels of wellbeing – their levels of well-being relative to others affected by the policy – as opposed to their *absolute* levels of wellbeing. If there are moral reasons to promote equality and fairness, then there are moral reasons to consider how the wellbeing levels of different people affected by a policy compare to one another.

We show how to take into consideration three different measures of inequality in constraining our choice of policies. Each allows us to set upper bounds on how much of each type of disparity in outcomes we allow from a policy.

4.3.1 The worst off and the average. The first measure involves the disparity between the *worst off* and *average* subjects. We can define acceptable policies to be those on which the worst case scenario is not too far from the policy's expected cost. We capture this as follows:

$$\max_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0) - \sum_{\pi_D \in \Pi_D} \pi(\pi_D) \cdot V_0^{\pi_D}(s_0) \leq m \tag{C3.1}$$

$$\max_{\{a \in A(s) | \pi(s,a)>0\}} Q_0^\pi(s,a) - \sum_{a \in A(s)} \pi(s,a) \cdot Q_0^\pi(s,a) \leq m_s \tag{C3.2}$$

4.3.2 The worst off and the best off. A second measure involves the disparity between the *worst off* and the *best off* under a policy. Two policies might have the same distance between the worst off and the average subject. However, if we care about inequality understood as the distance between the best off and worst off, we might object to using a policy that has an especially large *spread*.

Importantly, this concern is independent of the mean of that distribution. In a scenario in which the worst outcome is near the mean, but the best outcomes are exceedingly better, the previous constraint will not have a significant effect. If we want to capture this, we can require the distance between the worst and the best case scenarios to be bounded:

$$\max_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0) - \min_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0) \leq d \tag{C4.1}$$

$$\max_{\{a \in A(s) | \pi(s,a)>0\}} Q_0^\pi(s,a) - \min_{\{a \in A(s) | \pi(s,a)>0\}} Q_0^\pi(s,a) \leq d_s \tag{C4.2}$$

4.3.3 Variance. The best known measure of dispersion of a random variable, such as X_π, is its variance $Var(X_\pi) = \mathbb{E}[(X_\pi - \mu_{X_\pi})^2]$, which captures its expected normalized squared distance to its mean. Consider a scenario in which we have two policies, π_1 and π_2. Under both π_1 and π_2, patients will leave the hospital with the same level of pain in expectation, say 3. Assume also that, under both policies, all the worst off patients have the same pain level upon discharge, as do all the best off patients – say 4 and 2 respectively. These assumptions guarantee that constraints (C3) and (C4) are easily met. But suppose that π_1 has greater variance than π_2, because π_1 is bimodal – most people are in the extremes – and π_2 has most people near the mean. We might have reason to object to π_1 on grounds of inequality, and this can be captured straightforwardly by requiring policies to have a bounded variance:

$$Var(X_\pi) = \mathbb{E}[(X_\pi - \mu_{X_\pi})^2] \leq v \tag{C5.1}$$

$$Var(Y_{\pi,s}) = \mathbb{E}[(Y_{\pi,s} - \mu_{Y_{\pi,s}})^2] \leq v_s \tag{C5.2}$$

4.4 Caring about trade-offs between policies

The constraints outlined above help decide whether a policy, considered on its own, is morally acceptable. This enables basic comparisons *between* policies in terms of acceptability (one might be acceptable; another unacceptable). However, we might also want to make more sophisticated comparisons between policies. We might want to know whether the benefits of a stochastic policy with respect to the expected primary cost come at too high a price given moral concerns we've been discussing, relative to some other available policy. This is significant because the best stochastic policy is often better in expectation than the best deterministic one – this was borne out in our Example Instance T – but often with the consequence that variance, risk, and other morally concerning elements are introduced into the policy.

For this analysis, let us begin with a *baseline* policy β: a known feasible solution to the C-SSP.

As we have seen, not everything that is morally relevant is captured by the expected cost; this is why we have offered our earlier constraints. Our constraints all have a general form. Given some acceptability measurement on a policy $\phi(\pi)$, and an upper bound b on the value of that measurement:

$$\phi(\pi) \leq b \tag{AM}$$

For example, consider constraint (C4.1). There, $\phi(\pi)$ is $\max_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0) - \min_{\{\pi(\pi_D)>0\}} V_0^{\pi_D}(s_0)$ – that is, we take as our acceptability measurement the difference between the worst- and the best-off outcome of the stochastic decision. *Variance*, similarly, would be our acceptability measurement in constraint (C5.1), giving us an overall measure of distribution spread; and so forth with the other proposed constraints.

This commonality raises the possibility of expressing a comparative constraint in a general form. Consider a new 'candidate' policy π which we *may* want to choose over β on account of it having a lower expected primary cost. In order to decide whether to prefer π to β, we can constrain our choice by saying π can be preferred only if its relative decrease in expected primary cost justifies the increase of $\phi(\pi)$ relative to $\phi(\beta)$ for some acceptability measurement ϕ. Grounding this in the Example Instance T, we may not prefer the stochastic policy π_S over the best deterministic policy π_B as, for the decrease in expected cost that it offers, it worsens the outcomes of the worst-off by too much, or introduces too much variance.

Let us get a bit more concrete. Assume that β is (i) *the best* (ii) *feasible* (iii) *deterministic* policy.[4] Let π be a feasible stochastic policy with a lower expected primary cost; hence, $V_0^{\beta}(s) > V_0^{\pi}(s)$. In comparative terms, this leads straightforwardly to a quantification of the 'upside' of π over β in terms of the Expected Primary Cost Reduction (EPCR):

$$V_0^{\beta}(s) - V_0^{\pi}(s) \tag{EPCR}$$

Meanwhile, taking any of our acceptability measurements ϕ, we can consider another point of difference: how much of an Acceptability Measurement Increase (AMI) do we get by choosing π over β?

$$\phi(\pi) - \phi(\beta) \tag{AMI}$$

One way to codify this idea of trading off between the two policies, then, is with the following constraint:

$$V_0^{\beta}(s) - V_0^{\pi}(s) \geq \theta(\phi(\pi) - \phi(\beta)) \tag{C6}$$

(C6) requires that the EPCR outweighs the AMI (subject to a weighting parameter θ applied to the AMI). In practice, this means that, with β in hand, (C6) forces the planner to find only policies whose EPCR justifies any AMI they introduce.[5]

[4]Importantly, such a policy does not always exist in C-SSPs; in some cases there might not be any feasible deterministic, or even stochastic policy to work with. But it does exist in our example T (π_B is such a policy).

[5]The present example starts from the assumption that the test policy π outperforms the baseline policy β with respect to expected cost while it under-performs with respect to the acceptability measurement; but there are other interesting scenarios. Consider the opposite case, namely when the baseline policy β is better in expectation but worse with respect to the acceptability measurements, relative to π. If that is the case, then the appropriate constraint might require reversing the inequality:

$$V_0^{\pi}(s) - V_0^{\beta}(s) \leq \theta(\phi(X_\beta) - \phi(X_\pi)) \tag{C7}$$

We leave this constraint untreated, because it is not significantly different from (C6).

Here θ plays an important conceptual role: it is the parameter that allows us to compare the moral value captured by the two different measures represented in the constraint. Choosing the right θ requires consideration of (i) the scale of the acceptability measurement being used, and (ii) the actual importance we want the acceptability measurement to hold relative to expected cost. That is, do we want the expected outcome of, e.g., the worst-off person to hold equal weight with the average person, or less weight, or more? We expect the answer to this question to depend on the decision context.

Opting for a stochastic policy over a deterministic one can have both moral upsides and moral downsides. In this section, we have provided additional acceptability constraints to capture the moral downsides of stochastic policies not captured by the expected cost. We've also offered a way to compare the moral upsides and downsides of stochastic policies with any feasible baseline policy.

5 STAN-MS: A STOCHASTIC ANYTIME ALGORITHM FOR MORALLY SIGNIFICANT C-SSPS

In this section, we introduce a novel C-SSP solution method designed to integrate the constraints offered in the previous section. We call this a **St**ochastic **An**ytime algorithm for **M**orally **S**ignificant C-SSPs (**StAn-MS**). This method iteratively improves an initial distribution over the deterministic policy space (i.e., a concentrated stochastic policy) by randomly sampling batches of deterministic policies and introducing them to the distribution if they allow a closer-to-optimal (but still feasible and acceptable) concentrated policy to be defined. While, in general, the algorithm does not guarantee optimality, we show that it will converge to the optimal solution under certain conditions – conditions which will plausibly be met by many practical problems.

5.1 Algorithm Overview

The basic structure of our algorithm is as follows. We first need an initial concentrated policy: a distribution over the deterministic policy space Π_D that is (i) a feasible C-SSP policy, (ii) reasonably performant with regards to the primary objective, and (iii) acceptable with regards to any of the constraints from Section 4 that we may want to apply in some problem instance. Let us call the subset of the deterministic policies which are active under this distribution $\hat{\Pi}$. Following this, we require:

- a process for sampling new deterministic policies from $\Pi_D \backslash \hat{\Pi}$
- a solver for re-optimising the distribution p over the union of the current $\hat{\Pi}$ and the sampled new policies, such that the distribution remains feasible and acceptable

On a high level, the reader should be able to see that this structure enables a process of converging upon an optimal or at least 'good' and feasible solution to the raw C-SSP that *also* obeys our moral acceptability constraints, and therefore can balance these competing considerations. We now describe our implementation of the three components mentioned above. The pseudo code for our algorithm is given in Appendix A.1.

5.1.1 Initial policy. For the first requirement noted above – the initial concentrated policy – we leverage the optimal feasible *deterministic* policy to the C-SSP.[6] The attraction of this option is that the policy will, in general, be relatively performant with respect to the primary objective – although not optimal, in general, for the C-SSP. Moreover, it will trivially satisfy all acceptability constraints raised in Section 4 by virtue of introducing no policy stochasticity. So we initialise $\hat{\Pi} = \{\pi_D^*\}$ for the optimal deterministic policy π_D^*. Such a policy can be computed by solving a Mixed-Integer Linear Program (MILP) in the space of occupation measures [7]. This is the method we use in our implementation. A slight modification of the heuristic search algorithm i-dual [22] would also be possible.

5.1.2 Stochastic policy re-optimisation. We now address the third requirement above (we'll examine the second next). Given a subset of the deterministic policy space, a distribution over which will become our new current-best concentrated stochastic policy, how do we find the optimal distribution which is (i) feasible and (ii) acceptable with respect to our constraints? To do this, we extend the 'Reduced Master Problem' for column-generation-based C-SSP solving methods, a Linear Program (LP) which offers the ability to find the optimal feasible distribution [10].

This LP is structured as follows. If we consider a probability p_{π_D} associated with each $\pi_D \in \hat{\Pi}$, we optimise for the values of p that minimise the overall expected primary cost of the distribution while ensuring that the secondary costs do not exceed their bounds in expectation:

$$\min_{p} \sum_{\pi_D \in \hat{\Pi}} p_{\pi_D} V_0^{\pi_D}(s_0) \text{ s.t. (CX.1) - (CX.3)} \tag{LP}$$

$$p_{\pi_D} \geq 0 \qquad \forall \pi_D \in \hat{\Pi} \tag{CX.1}$$

$$\sum_{\pi_D \in \hat{\Pi}} p_{\pi_D} = 1 \tag{CX.2}$$

$$\sum_{\pi_D \in \hat{\Pi}} p_{\pi_D} V_j^{\pi_D}(s_0) \leq \hat{c}_j \qquad \forall j \in \{1, \ldots, k\} \tag{CX.3}$$

While this is a helpful start, so far it doesn't enforce our moral acceptability constraints. However, because our acceptability constraints are all framed as non-strict inequalities, they can be integrated into this program relatively easily. For example, if we especially care about the worst off, we might wish to enforce (C1.1), an upper threshold on the worst-case expected cost. To do this, we would just need to make a trivial rewrite of $\pi(\pi_D)$ for the equivalent variable p_{π_D} to have the following constraint which could be introduced to the above program:

$$\max_{\{\pi_D \in \Pi_D | p_{\pi_D} > 0\}} V_0^{\pi_D}(s_0) \leq h \tag{CX.4}$$

If we were also concerned with distribution, we might want to enforce (C3.1) – a bound on the disparity between the worst-off and average expected cost. With the same substitution, this constraint is ready for integration into the program.

$$\max_{\{\pi_D \in \Pi_D | p_{\pi_D} > 0\}} V_0^{\pi_D}(s_0) - \sum_{\pi_D \in \Pi_D} p_{\pi_D} \cdot V_0^{\pi_D}(s_0) \leq m \tag{CX.5}$$

[6] We acknowledge that there will not always be such a policy for a given C-SSP, but we take it as a simplifying assumption here; indeed there are other reasonable options for this initial distribution where such a policy does not exist.

There is a computational price to pay for these constraints: they involve taking the maximum value over a set, which leads to the introduction of binary variables that turn the original LP into a harder-to-solve MILP.

We also give particular attention to constraint (C6), which introduced the notion of a comparative constraint relative to some baseline (as opposed to appealing to an upper threshold on some acceptability measure, as the preceding constraints did). Integrating the constraint form of (C6) is a particular strength of our algorithm's structure. At all times during execution, we already possess a 'current-best' feasible and acceptable policy due to its anytime nature. This effectively gives us a good baseline concentrated policy β for free at each timestep – we simply need to record the current best policy's expected primary cost as well as the relevant acceptability measures (i.e., its worst-case expected cost, or its conditional value at risk, etc.).

Using this baseline β, we can then integrate (C6) into the program, with $V_0^{\pi}(s_0)$ replaced with its equivalent representation in terms of p and $\hat{\Pi}$ on the left hand side:

$$V_0^{\beta}(s) - \sum_{\pi_D \in \hat{\Pi}} p_{\pi_D} V_0^{\pi_D}(s_0) \geq \theta \cdot (\phi(\pi) - \phi(\beta)) \tag{CX.6}$$

For brevity, we omit auxiliary constraints that may need to be added to the problem depending on which acceptability measure is integrated into this constraint. For example, if conditional value at risk is used, several auxiliary constraints will be needed – we follow the formula for calculating it over discrete distributions presented by [18]. One thing to note is that computing this particular acceptability measure does require the introduction of bilinear auxiliary constraints to the program, rendering it a non-convex Mixed-Integer Quadratically Constrained Program (MIQCP). However, as the number of decision variables in the program scales linearly with respect to $|\hat{\Pi}|$, which we generally do not expect to be an extraordinarily large set, this is not too computationally prohibitive a result.

To summarise, the constraints that we formalised in Section 4 for concentrated stochastic policies can be integrated with nearly no overheads in terms of design. (C1.1)-(C6) can all be introduced or removed independently from the program – with the caveat that enforcing them also requires introducing the auxiliary variables and constraints for calculating the acceptability measures themselves.

5.1.3 Sampling new deterministic policies. Finally, we discuss the second requirement for our algorithm, which is a method for exploring new deterministic policies which could be added to the current-best concentrated stochastic policy's active set. While similar approaches for solving plain C-SSPs – like column generation – are able to use guided approaches which identify deterministic policies in Π_D which are guaranteed to facilitate improving upon the current best solution, these approaches do not extend well to our situation, where the distribution optimisation problem solved at each timestep is more complex and no longer linear.

We get around this by simply sampling batches of n deterministic policies from Π_D at each timestep for some preconfigured $n \in \mathbb{N}$, and adding them to $\hat{\Pi}$ before invoking our redistribtion program. After re-optimising, we redefine $\hat{\Pi}$ as $\{\pi_D | \pi_D \in \hat{\Pi} \wedge p_{\pi_D} > 0\}$,

discarding unused deterministic policies to avoid $\hat{\Pi}$ growing needlessly. Sampling new deterministic policies in batches rather than one at a time is a mitigation strategy for the fact that we are no longer adding policies which provably improve our solution. A larger sample size opens the possibility of finding good distributions which assign nonzero probabilities to a larger number of deterministic policies. In our approach, we sample from a uniform distribution over Π_D.

5.2 Advantages and Limitations

StAn-MS produces concentrated stochastic policies. However, as explained in Section 3, these can easily be converted to an in-expectation equivalent distributed stochastic policy.[7]

As an algorithm, StAn-MS is not without limitations. It is an approximate approach, and inexact relative to the column generation approach for regular C-SSPs. In general, it cannot guarantee convergence to the optimal policy. However, we can trivially observe that if our batch sample size n is larger than the number of deterministic policies that take a nonzero probability in the optimal (morally acceptable) concentrated stochastic policy π^*, and we are sampling from the full set Π_D, then in unlimited iterations we will provably find π^*. So, in certain problem domains where the concentrated representations of optimal stochastic policies tend to utilise a smaller number of deterministic policies than what we can comfortably sample from Π_D at each iteration, we may find that we converge to the optimal policy for most problem instances.

Additionally, more sophisticated sampling strategies may be needed to apply StAn-MS in real-world, large-scale problems. Similarly, devising a method for finding an initial solution which is easier to compute than π^*_D but still feasible and acceptable would benefit the scalability of this approach.

Nonetheless, our move away from existing leading C-SSP solvers is, we argue, well motivated. The dual linear-program-based solvers such as i-dual and i^2-dual introduced by [21, 22] are problematised when attempting to integrate our constraints on two fronts. First, these solvers produce distributed stochastic policies. Enforcing any of our constraints that invoke states' Q-values requires a significant modification to the dual linear program underlying these approaches that renders that program a mixed-integer quadratically constrained program – which is far more computationally intensive. Second, they do not offer a well-justified method for defining and using a 'baseline' policy of the kind needed for constraint (C6). This is largely because it introduces the need for each transient state to have its own notion of an acceptable baseline policy, which ultimately leads to unworkable inconsistencies between these baselines. Meanwhile, the column-generation-based approach introduced by [10] could not be straightforwardly applied to our use case either. Since our acceptability constraints are universally non-linear, the linearity of the underlying reduced master problem of that approach could not be maintained, which hamstrings attempts to use a column-generation approach.

[7] However, it should be noted that after such a conversion from an acceptable concentrated policy, it does not necessarily follow that the distributed equivalent satisfies the corresponding distributed constraints we presented in Section 4 from all states.

6 RESULTS

In this section, we evaluate the performance and effectiveness of the StAn-MS algorithm in a test instance of the autonomous medic C-SSP environment. This instance, which is detailed in Appendix A.2, includes stochastic action effects (i.e. pain reduction) and requires the administration of many painkillers in sequence. We present visualisations of the evolution of the current best solution through StAn-MS's improvement iterations. This allows us to compare the change over time of the policy's expected primary cost with the corresponding change in our acceptability measures. (The worst-case primary cost, expected- vs. worst-case primary cost difference, and conditional value at risk of the stochastic policy.) We do this with and without our acceptability constraints to demonstrate that they have the desired effect on the solving process. We also report and comment on the relative differences in solve time and primary expected cost when we do enforce the acceptability constraints, and when we do not.

6.1 Visualising Constraint Enforcement

We visualise the evolution of policies in the StAn-MS algorithm, from the initial deterministic policy to the final policy returned by the algorithm after t improvement iterations. We visualise this change over time with respect to two metric categories at each timestep:

- The expected primary cost of the current best feasible and acceptable policy, i.e., the expected remaining patient pain after discharge.
- The value of a couple of the various acceptability measures that we have introduced throughout this paper. To keep things simple we avoid plotting all of these measures, but instead choose just two:
  - The 'Conditional Value at Risk' of the policy, which gives us an example of a measure thats' useful when 'caring about the worst off'. For these results, we use an α of 0.9 – meaning we can interpret this measure as the expected primary cost of the 90th percentile of the distribution over deterministic policies (i.e., the 'most costly' top 10%).
  - The 'Expected-Worst Primary Cost Difference' of the policy, which gives us an example of a measure thats' useful when 'caring about distribution' – this is the difference between the expected primary cost of our current best solution and its worst-case primary cost.

All of the following results were produced with the StAn-MS algorithm using 100 improvement iterations and a random sample size of 20 deterministic policies at each iteration. Each problem was solved 20 times, so the following lineplots represent the mean values achieved over these repetitions with the shaded regions representing the confidence interval.

We begin by presenting the policy evolution when we solve StAn-MS without enforcing any of the moral acceptability constraints that we have formulated. This means that the algorithm will treat the automated medic instance as an ordinary (morally insignificant) C-SSP. Figure 1 visualises this convergence. We can see that within 100 iterations, the expected primary cost of the current best feasible and acceptable policy converges from the optimal deterministic policy's (approximately 0.84) to within a negligible distance of the

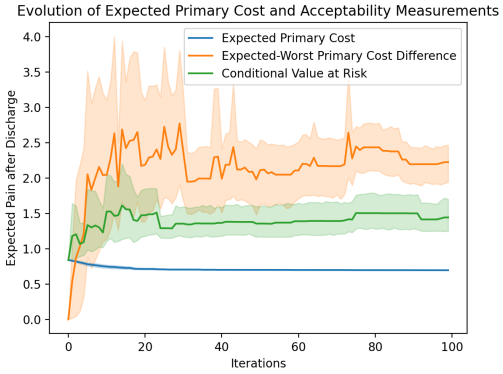


Figure 1: Change in the current best feasible policy over 100 improvement iterations for StAn-MS with no moral acceptability constraints. The decrease in expected primary cost of the policy is plotted against our example acceptability measures.

optimal stochastic policy's (approx. 0.69). However, we also see on the right of this figure that our acceptability measures increase significantly from the initial to the eventual solution, with the expected-worst difference settling at about 2.2 and the conditional value at risk settling at about 1.5 – and these are alarmingly high in context.

With this baseline behaviour established, we can evaluate the effectiveness of our various constraints in inducing morally acceptable stochastic policies.

6.1.1 Enforcing a Conditional Value at Risk Threshold. Figure 2(a) visualises the policy evolution over time if we enforce our upper threshold on conditional value at risk (C2.1); we enforce an upper threshold of 1.2 in this case. We can see that StAn-MS is able to find a best solution that is competitive with the optimal stochastic policy in terms of expected primary cost, but also obeys the upper bound upon the conditional value at risk at all iterations.

6.1.2 Enforcing an Expected-Worst Primary Cost Difference Threshold. Figure 2(b) visualises the policy evolution over time if we enforce our upper threshold on expected-worst primary cost difference (C3.1); we enforce an upper threshold of 0.5. Again, this bound is obeyed while the solver still manages to significantly improve upon the initial solution by introducing acceptable policy stochasticity.

6.1.3 Trading off Expected Primary Cost Against Conditional Value at Risk. We additionally give an example in Figure 2(c) of operationalising the constraint form (C6), leveraging the conditional value at risk acceptability measure as our ϕ in this specific case. Recall the intuition of this constraint form. At each iteration i, the decrease in expected primary cost must be more than the θ-weighted increase in the acceptability measure (conditional value at risk), relative to iteration $i-1$. This is shown with θ simply configured to 1.0. The conditional value at risk is visibly compliant, and increases less than was observed in our initial unconstrained solution.

6.2 Discussion

Overall, our results demonstrate that StAn-MS is capable of producing concentrated stochastic policies which avoid introducing policy stochasticity in morally unacceptable ways. All of our proposed constraints induce their intended effects in terms of how the algorithm develops and improves upon its best feasible and acceptable solution over time. Moreover, as the raw expected primary costs of the found solutions indicate, in many cases we can induce these acceptable planning behaviours with minimal effect on the solution quality in terms of the expected primary cost of the returned policy. Let us not forget that when our primary cost is morally significant, as in this domain, achieving low expected primary cost is a moral good, all else being equal. Specifically, compared to solving the C-SSP without enforcing any acceptability constraints with StAn-MS – which produced solutions with on average an expected primary cost 17.06% lower than the optimal deterministic policy's in our experiments – we saw improvements over the optimal deterministic policy of 16.63% with an upper threshold of 1.2 on conditional value at risk, 16.53% with an upper threshold of 0.5 on expected-worst primary cost difference, and 14.49% with (C6) applied to conditional value at risk with $\theta = 1.0$. Given the already-stated benefits of enforcing our constraints, these are relatively small optimality gaps being given up for considerable gain in moral policy acceptability.

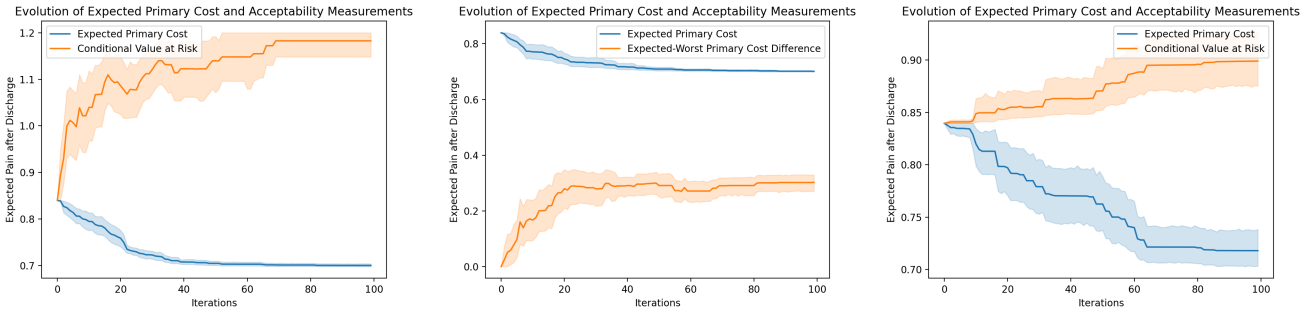
It is also worthwhile to consider the times taken to solve the instance under various constraint combinations. Relative to a solve time of 4.315 seconds to solve the instance with StAn-MS with no acceptability constraints, we saw solve times of 4.813 seconds when bounding expected-worst primary cost difference, 10.607 seconds when bounding conditional value at risk, and 13.196 seconds when "trading off" conditional value at risk with (C6).[8]

The solve times for the expected-worst primary cost difference bound is promisingly quite similar to that of the acceptability-unbounded solve time, given that this constraint makes our modified reduced master problem into a mixed integer program. This has positive implications for all the constraints we have introduced which only raise the complexity of the distribution reoptimisation problem to this mixed integer level. The solve time does increase significantly once constraints are introduced which require computing the conditional value at risk. This is to be expected, since it introduces bilinear (non-convex) constraints into the modified reduced master problem. These results raise interesting design decisions between using robust but complex acceptability measures such as conditional value at risk, versus some other easier-to-compute (but less expressive) acceptability measure.

7 CONCLUSION

In this paper, we have studied the moral significance of stochasticity in C-SSPs. The results are articulated in Section 4, where we provided different arguments for different constraints on what we called acceptability measurements ϕ. We concluded that section by presenting a constraint (C6), that allows trade-offs between the moral upsides and moral downsides of stochasticity. Constraints of the form (C6) are justified on moral grounds, but are new to the literature. For this reason, we designed a novel C-SSP solver, **St**ochastic **An**ytime algorithm for **M**orally **S**ignificant C-SSPs (**StAn-MS**),

[8] Computed on an instance with a 2019 2.6 GHz i7 CPU and 16GB of memory.



(a) CVaR, upper bounded at 1.2

(b) Expected-worst primary cost difference, upper bounded at 0.5

(c) CVaR when (C6) is enforced with θ set to 1.0

Figure 2: Change in expected primary cost and acceptability measure of best feasible, acceptable solution over time, for 3 different acceptability constraints.

capable of handling those more complex constraints. Finally, we presented experimental results.

Section 4 has the bulk of the philosophical argument. There we identified two locations for moral value in our framework: one associated with the expected primary cost; one associated with caring about the worst off and caring about the distribution. We do not claim to have given an exhaustive account of all possible moral concerns about stochastic policies. The main lesson from the section is that stochasticity has both moral upsides and moral downsides, and that we should weigh them against each other.

Sections 5 has the bulk of the computer science contribution. Constraints of the form (C6) are new and morally motivated, and there is no literature that we are aware of that can handle such C-SSP constraints well. In particular, dual-LP-based solvers face difficulties with regards to computational complexity and lack a natural way to introduce a 'baseline' policy for (C6), while column-generation-based solvers do not extend particularly well to nonlinear constraints. While there is still progress to be made in the areas of efficient initial policy computation and in informed sampling strategies for new deterministic policies in our algorithm, we claim it is able to handle the constraints we have raised far more naturally than current leading C-SSP solvers.

The constraints and computational approach proposed in our paper can accommodate stochastic effects of policy actions. However, a deeper philosophical question is whether the very same ethical issues that arise when choosing a stochastic policy also arise when actions have stochastic effects. We suspect that the ethical issues will be very similar (and thus that our proposed acceptability measures are appropriate in those contexts). However, further work must be done to determine if there are significant moral differences, given that policy decisions are under our control while the stochastic effects arising from actions are not.

To conclude, we hope to have made progress by providing a clear way to encode complex moral considerations with formal constraints, and a novel algorithmic approach to deal with them.

ACKNOWLEDGMENTS

This research is part of the *Humanising Machine Intelligence* project at ANU. We thank the project's members, Felipe Trevizan, and the anonymous reviewers for useful discussions and comments.

REFERENCES

[1] Ethan Altman. 1999. *Constrained Markov Decision Processes.* Chapman and Hall.
[2] Dimitri P. Bertsekas and John N. Tsitsiklis. 1991. An Analysis of Stochastic Shortest Path Problems. *Mathematics of Operations Research* 16, 3 (1991), 580–595.
[3] Vicky Charisi, Louise A. Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Sombetzki, Alan F. T. Winfield, and Roman Yampolskiy. 2017. Towards Moral Autonomous Systems. *CoRR* abs/1703.04741 (2017).
[4] Yinlam Chow and Mohammad Ghavamzadeh. 2014. Algorithms for CVaR Optimization in MDPs. In *Proc. 27th Annual Conference Advances on Neural Information Processing Systems (NIPS'14).* 3509–3517.
[5] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. 2015. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In *Proc. 28th Annual Conference Advances on Neural Information Processing Systems (NIPS'15).* 1522–1530.
[6] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14.
[7] Dmitri Dolgov and Edmund Durfee. 2005. Stationary Deterministic Policies for Constrained MDPs with Multiple Rewards, Costs, and Discount Factors. In *Proc. 19th International Joint Conference on Artificial Intelligence (IJCAI'05).* 1326–1331.
[8] Eugene A. Feinberg and Adam Shwartz. 1995. Constrained Markov Decision Models with Weighted Discounted Rewards. *Mathematics of Operations Research* 20, 2 (1995), 302–320.
[9] Jerzy A. Filar, Lodewijk C. M. Kallenberg, and Huey-Miin Lee. 1989. Variance-Penalized Markov Decision Processes. *Mathematics of Operations Research* 14, 1 (1989), 147–161.
[10] Florian Geißer, Guillaume Povéda, Felipe Trevizan, Manon Bondouy, Florent Teichteil-Königsbuch, and Sylvie Thiébaux. 2020. Optimal and Heuristic Approaches for Constrained Flight Planning under Weather Uncertainty. In *Proc. 30th International Conference on Automated Planning and Scheduling (ICAPS'20).* 384–393.
[11] Seth Lazar. 2017. Anton's Game: Deontological Decision Theory for an Iterated Decision Problem. *Utilitas* 29 (2017), 88–109.
[12] Hyun-Rok Lee and Taesik Lee. 2018. Markov decision process model for patient admission decision at an emergency department under a surge demand. *Flexible Services and Manufacturing Journal* 30, 1 (2018), 98–122.
[13] Hyun-Rok Lee and Taesik Lee. 2021. Multi-agent reinforcement learning algorithm to solve a partially-observable multi-agent problem in disaster response. *European Journal of Operational Research* 291, 1 (2021), 296–308.
[14] Felix Lindner, Robert Mattmüller, and Bernhard Nebel. 2020. Evaluation of the moral permissibility of action plans. *Artificial Intelligence* 287 (2020), 103350.
[15] Shie Mannor and John N. Tsitsiklis. 2011. Mean-Variance Optimization in Markov Decision Processes. In *Proc. 28th International Conference on International Conference on Machine Learning (ICML'11).* 177–184.

[16] Samer Nashed, Justin Svegliato, and Shlomo Zilberstein. 2021. Ethically Compliant Planning within Moral Communities. In *Proc. 4th AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*. 188–198.
[17] Adriana Placani. 2017. When the Risk of Harm Harms. *Law and Philosophy* 36, 1 (2017), 77–100.
[18] R.Tyrrell Rockafellar and Stanislav Uryasev. 2002. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* 26, 7 (2002), 1443–1471.
[19] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. 2014. *Lectures on Stochastic Programming - Modeling and Theory, Second Edition*. MOS-SIAM Series on Optimization, Vol. 16. Society for Industrial and Applied Mathematics.
[20] Justin Svegliato, Samer B. Nashed, and Shlomo Zilberstein. 2021. Ethically Compliant Sequential Decision Making. In *Proc. 35th AAAI Conference on Artificial Intelligence (AAAI'21)*. 11657–11665.
[21] Felipe Trevizan, Sylvie Thiébaux, and Patrik Haslum. 2017. Occupation Measure Heuristics for Probabilistic Planning. In *Proc. 27th International Conference on Automated Planning and Scheduling (ICAPS'17)*. 306–315.
[22] Felipe W. Trevizan, Sylvie Thiébaux, Pedro Henrique Santana, and Brian Charles Williams. 2016. Heuristic Search in Dual Space for Constrained Stochastic Shortest Path Problems. In *Proc. 26th International Conference on Automated Planning and Scheduling (ICAPS'16)*. 326–334.
[23] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2023. Reinforcement Learning in Healthcare: A Survey. *Comput. Surveys* 55, 1 (2023), 1–36.

A APPENDIX

A.1 Pseudo Code for StAn-MS

Our algorithm combines the elements described in Section 5 as follows.[9]

Algorithm 1 StAn-MS

Input: MC:: C-SSP, params:: Tuple⟨Float⟩, t:: int, n:: int
Output: $\pi :: (\Pi_D \rightarrow [0, 1])$

1: $\pi_D^* \leftarrow$ optimalDetPolicy(MC)
2: $\hat{\Pi} \leftarrow \{\pi_D^*\}$
3: $p \leftarrow \{\pi_D^* : 1.0\}$
4: $obj \leftarrow V_0^{\pi_D^*}(MC.s_0)$ ▷ Expected Primary Cost of solution
5: $m \leftarrow$ calculateInitialAcceptabilityMeasures($MC, \hat{\Pi}, p$)
6: **while** $t > 0$ **do**
7: $\quad \Pi_{\text{new}} \leftarrow$ sampleDetPolicies(MC, n)
8: $\quad (obj', \hat{\Pi}', p', m') \leftarrow$ reoptimise($\{\hat{\Pi} \cup \Pi_{\text{new}}\}$, params, m)
9: $\quad$ **if** $obj' < obj$ **then** ▷ Only update if a better solution found
10: $\quad\quad obj \leftarrow obj'$
11: $\quad\quad \hat{\Pi} \leftarrow \hat{\Pi}'$
12: $\quad\quad p \leftarrow p'$
13: $\quad\quad m \leftarrow m'$
14: $\quad$ **end if**
15: $\quad t \leftarrow t - 1$
16: **end while**
17: $\pi :: (\Pi_D \rightarrow [0, 1])$
18: **for** $\pi_D \in \hat{\Pi}$ **do**
19: $\quad \pi(\pi_D) \leftarrow p_{\pi_D}$
20: **end for**
21: **return** π

We pass to the algorithm:

- A C-SSP
- A tuple of parameters (each constraint we have introduced requires one, e.g., (C1)-(C5) require constants for the bounds and (C6) requires the balancing constant θ.)
- The number of improvement iterations to perform
- The number of deterministic policies to sample at each iteration

In lines 1-6, we initialise our solution, as well as calculate the relevant acceptability measures for the solution that are needed to define the baseline for our tradeoff constraints (stored in the map m). In lines 6-16, we iteratively sample new solutions, find a new solution, and store the new solution if it is an improvement. Note that we are guaranteed never to find a worse solution at t than at $t - 1$ since $t - 1$'s solution will trivially also be a solution at t; nonetheless, for efficiency we avoid updating our solution if the new solution has an equal objective value to the old one. Finally, in lines 19-22 we convert the solution into a concentrated stochastic policy π (we assume that for any deterministic policy π_D that we don't define a mapping for at this stage, $\pi(\pi_D) = 0$).

A.2 Medic Instance Used in Experiments

For evaluating our algorithm, we introduce a new, more detailed instance of the automated medic domain than the instance T that was referred to in Section 4. In this instance, we have the following defining factors:

- Three medications A, B and C, where:
  - A costs \$1000, and reduces patient pain by 10 with probability 0.5, 6 with probability 0.25, and 5 with probability 0.25.
  - B costs \$600, and reduces patient pain by 6 with probability 0.5, 5 with probability 0.25, and 3 with probability 0.25.
  - C costs \$500, and reduces patient pain by 5 with probability 0.8, and 0 with probability 0.2.
- An upper bound on expected monetary cost of the policy at \$1200.

Recall that a state of the problem is defined in terms of the self-reported pain of a patient at that timestep, and the list of already-administered painkillers.[10] So to provide an example of the transition function of the problem for one state-action pair (we avoid enumerating further for the sake of brevity), we might have:

$$P(\langle 4, [B, C]\rangle \mid \langle 10, [C]\rangle, B) = 0.5$$
$$P(\langle 5, [B, C]\rangle \mid \langle 10, [C]\rangle, B) = 0.25$$
$$P(\langle 7, [B, C]\rangle \mid \langle 10, [C]\rangle, B) = 0.25$$
$$P(s \mid \langle 10, [C]\rangle, B) = 0 \text{ for all other } s \in S$$

Looking at the cartesian product of possible pain levels $\{0, ..., 10\}$ and the powerset of $\{A, B, C\}$, we can see that this instance will have as many as $11 \cdot 2^3 = 88$ transient states, and 88 goal states (with a one to one transition mapping between these two sets when the 'discharge' action is applied in the former). However, in practice the reachable state space of the instance is smaller on account of the fact that not all reported pain levels between 0 and 10 can possibly be reached by a certain combination of the available painkillers.

Recall also that if a painkiller's randomly sampled pain reduction exceeds the pain of the patient at the state where it is being applied, then the patient's pain will be 0 rather than negative at the successor

[9] Our implementation can be accessed at https://github.com/chevans-lab/ethical-stoch-policies.

[10] Technically, we implement this as a list of already-administered actions (so including the discharge action), as this allows us to differentiate transient and goal states more simply.

state. So for example, if B were administered to a patient with a pain level of 5, then their pain level would transition to 0 with probability 0.75 and 2 with probability 0.25.

Finally, recall that only discharge actions incur a non-negligible primary cost (which equals the patient's pain at time of discharge) and only painkiller actions incur a secondary/monetary cost.