Structural Properties of Optimal Fidelity Selection Policies for Human-in-the-loop Queues ⋆

Piyush Gupta [a], Vaibhav Srivastava [a]

[a] *Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, 48824, USA*

Abstract

We study optimal fidelity selection for a human operator servicing a queue of homogeneous tasks. The agent can service a task with a normal or high fidelity level, where fidelity refers to the degree of exactness and precision while servicing the task. Therefore, high-fidelity servicing results in higher-quality service but leads to larger service times and increased operator tiredness. We treat the human cognitive state as a lumped parameter that captures psychological factors such as workload and fatigue. The operator's service time distribution depends on her cognitive dynamics and the fidelity level selected for servicing the task. Her cognitive dynamics evolve as a Markov chain in which the cognitive state increases with high probability whenever she is busy and decreases while resting. The tasks arrive according to a Poisson process and the operator is penalized at a fixed rate for each task waiting in the queue. We address the trade-off between high-quality service of the task and consequent penalty due to a subsequent increase in queue length using a discrete-time Semi-Markov Decision Process framework. We numerically determine an optimal policy and the corresponding optimal value function. Finally, we establish the structural properties of an optimal fidelity policy and provide conditions under which the optimal policy is a threshold-based policy.

Key words: Fidelity selection; Queueing theory; Human-in-the-loop; Semi-Markov decision process.

1 Introduction

Human-in-the-loop systems are pervasive in areas such as search and rescue, semi-autonomous driving, and robot-assisted surgery. Many safety-critical systems rely on human expertise to ensure safe and efficient operation. Human-robot collaboration allows for integrating human knowledge and perception skills with autonomy. In such systems, it is often of interest to increase the ratio of robots to humans, which leads to a reduction in cost but an increase in human workload. This is detrimental to the system performance as human performance is a function of cognitive factors such as fatigue and workload. Therefore, in environments with constrained human resources, it is critical to facilitate the effective use of limited cognitive resources [2]. In this work, we control the cognitive state of the human operator by optimizing the fidelity level for servicing the tasks, where fidelity refers to the degree of exactness and precision while servicing the task.

We study optimal fidelity selection for a human operator servicing a queue of homogeneous tasks. An example scenario is an airport security system where a human scans the luggage items with different fidelity levels. The term "fidelity" can have different meanings based on the application. For example, in shared-control tasks such as collaborative human-robot search [3], fidelity could refer to the human contribution to the task as compared to autonomy. Similarly, in a dual-task paradigm such as supervising and teleoperating a team of robots [4], servicing single versus both tasks can correspond to different fidelity levels. We incorporate human cognitive dynamics into the fidelity selection problem and study its influence on optimal policy. In particular, we show that servicing the tasks with high fidelity is not always optimal due to larger service times and increased tiredness of the human operator. In fact, we show that the optimal policy depends on the number of tasks awaiting service (queue length) as well as the cognitive state of the human operator. Our results provide insight into the efficient design of human decision support systems. For servicing each task, the human operator receives a reward based on the fidelity level selected for the task.

⋆ A preliminary version of this work [1] was presented at the 2019 American Control Conference, held in Philadelphia. We expand on the work in [1] by establishing structural properties of the optimal value function and consequently the optimal fidelity selection policy. We also introduce additional numerical illustrations. Corresponding author P. Gupta. Tel. +1-517-432-0019

Email addresses: guptapi1@msu.edu (Piyush Gupta), vaibhav@egr.msu.edu (Vaibhav Srivastava).

However, with higher fidelity, the cognitive state quickly rises to higher sub-optimal levels, thereby requiring larger service time for subsequent tasks. Hence, there is a trade-off between the reward obtained by high-fidelity servicing (improved service quality), and the penalty incurred due to the resulting delay in servicing subsequent tasks. We elucidate this trade-off and find an optimal fidelity selection policy. Indeed the optimal policy is problem-specific and depends on the problem parameters. Therefore, without careful system design and parameter tuning such as selecting arrival rates, the optimal policy might behave unexpectedly. This can lead to a bad user experience for the human operator or a lack of trust in the optimal recommendations, for example, in a scenario where the decision-support system recommends frequent switching of the fidelity level. To this end, we establish structural properties [5] of the optimal fidelity selection policy and provide conditions under which, for each cognitive state, there exist thresholds on queue lengths at which optimal policy switches fidelity levels. These structural properties can be used to tune the decision support system parameters such that the optimal policy is well-behaved and the human operator can trust its suggestions. Furthermore, these properties can be leveraged to determine a minimally parameterized policy for specific individuals which can be refined in real-time using a small amount of data.

In our setup, the human operator has a unimodal performance (characterized by its service time) w.r.t. its cognitive state which is inspired by the Yerkes-Dodson law [6]. Intuitively, such unimodal behavior is obtained because excessive stress (high cognitive state) overwhelms the operator and too little stress (low cognitive state) leads to boredom and a reduction in vigilance.

While human-in-the-loop is used as a primary application, this work is applicable to other non-human servers with state-dependent unimodal performance. For example, in the context of traffic flow, the traffic intersection can be interpreted as a server, and traffic flux is a unimodal function of the traffic density [7]. In such a scenario, the control measures may include admitting a vehicle or rerouting it, to maintain the optimal performance of the traffic network.

The major contributions of this work are threefold: (i) we pose the fidelity selection problem in a Semi-Markov Decision Process (SMDP) framework and compute an optimal policy, (ii) we numerically show the influence of cognitive dynamics on the optimal policy, and (iii) we establish structural properties of the optimal fidelity policy and provide sufficient conditions for a threshold-based policy to be optimal.

The rest of the paper is structured in the following way. In Section 2, we discuss some relevant literature. Section 3 presents the problem setup and formulates the fidelity selection problem using an SMDP framework. In Section 4, we numerically illustrate an optimal fidelity selection policy and establish its structural properties in Section 5. Finally, in Section 6, we provide conclusions and discuss the future directions of this work.

2 Related Work

Recent years have seen significant efforts in integrating human knowledge and perception skills with autonomy [8]. A key research theme within this area concerns the systematic allocation of human cognitive resources for efficient performance. Therein, some of the fundamental questions studied include optimal scheduling of the tasks to be serviced by the operator [9], enabling shorter operator reaction times by controlling the task release [2], and determining optimal operator attention allocation [10]. In contrast to the aforementioned works, we consider an SMDP formulation to deal with general (non-memoryless) service time distributions of the human operator. Furthermore, while the above works propose heuristic algorithms, we focus on establishing the structural properties of the optimal policy.

Some interesting recent studies with state-dependent queues are considered in [11, 12]. In these works, authors design scheduling policies that stabilize a queueing system and decrease the utilization rate of a non-preemptive server that measures the proportion of time the server is working. The performance of the server degrades with the increase in server utilization and improves when the server is allowed to rest. In contrast to monotonic server performance with the utilization rate in [11, 12], we model the service time of the human operator as a unimodal function of its cognitive state. Our model for service time is inspired by experimental psychology literature [6] and incorporates the influence of cognitive state and fidelity level on service time.

The optimal control of queueing systems [13] is a classical problem in queueing theory. Of particular interest are the works [14, 15], where authors study the optimal policies for an M/G/1 queue by SMDP formulation and describe its qualitative features. In contrast to a standard control of queues problem, the server in our problem is a human operator with cognitive dynamics that must be incorporated into the problem formulation.

Our mathematical techniques to establish the structural properties of the optimal policy are similar to [5]. In [5], the authors establish structural properties of an optimal transmission policy for transmitting packets over a point-to-point channel in communication networks. The optimal policy of their Markov decision process depends on the queue length, the number of packet arrivals, and the channel fading state. In [16], authors study structural properties of the optimal resource allocation policy for a single-queue system in which a central decision-maker assigns servers to each job. In contrast to [5,16], a major challenge in our problem arises due to SMDP formulation for non-memoryless service time distribution and its unimodal dependence on the cognitive state.

3 Background and Problem Formulation

We now discuss our problem setup, formulate it as an SMDP, and solve it to obtain an optimal policy.

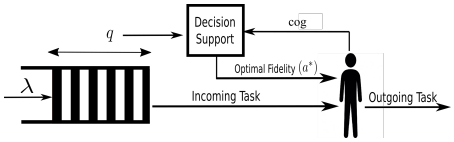


Fig. 1. Overall schematic of the problem setup. The incoming tasks arrive as a Poisson process with a rate λ. The tasks are serviced by the human operator based on the recommended fidelity level by the decision support system.

3.1 Problem Setup

We consider a human supervisory control system in which a human operator is servicing a stream of homogeneous tasks. The human operator may service these tasks with different levels of fidelity. The servicing time of the operator depends on the fidelity level with which she services the task as well as her cognitive state. We assume that the mean service time of the operator increases with the selected fidelity level. For example, when the operator services the task with high fidelity, she may look into deeper details of the task, and consequently take a longer time to service.

In addition to the fidelity level, the human service time may depend on their cognitive state. We treat the cognitive state as a lumped parameter that can capture various physiological measures. It can be a function of stress, workload, arousal rate, operator utilization ratio, etc. Such lumped representation can be obtained by classifying these psychological measurements into different service time distribution parameters. Inspired by the Yerkes-Dodson law, for a fixed level of fidelity, we model the service time as a unimodal function of the human cognitive state. Specifically, the mean service time is minimal corresponding to an intermediate optimal cognitive state (later referred to as the optimal cognitive state cog^*) as shown in Fig. 2c.

We are interested in the optimal fidelity selection policy for the human operator. To this end, we formulate a control of queue problem, where in contrast to a standard queue, the server is a human operator with her cognitive dynamics. The incoming tasks arrive according to a Poisson process at a given rate $\lambda \in \mathbb{R}_{>0}$ and are serviced by the operator based on the fidelity level recommended by a decision support system (Fig. 1). We consider a dynamic queue of homogeneous tasks with a maximum capacity $L \in \mathbb{N}$. The operator is penalized for each task waiting in the queue at a constant rate $c \in \mathbb{R}_{>0}$ per unit delay in its servicing. The set of possible actions available for the operator corresponds to (i) ***Waiting (W)*** when the queue is empty, (ii) ***Resting (R)***, which allows the operator to rest and reach the optimal cognitive state, (iii) ***Skipping (S)***, which allows the operator to skip a task to reduce the queue length and thereby focus on newer tasks, (iv) ***Normal Fidelity (N)*** for servicing the task with normal fidelity, and (v) ***High Fidelity (H)*** for servicing the task more carefully with high precision. The skipping action ensures the stability of the queue by allowing the operator to reduce the queue length by skipping some tasks. Ideally, through appropriate control of the arrival rate, the system designer should ensure that skipping is not an optimal action.

Let $s \in \mathcal{S}$ be the state of the system and $\mathcal{A}_s$ be the set of admissible actions in state s, which we define formally in Section 3.2. The human receives a reward $r : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}_{\geq 0}$ defined by

$$r(s,a) = \begin{cases} r_H, & \text{if } a = H, \\ r_N, & \text{if } a = N, \\ 0, & \text{if } a \in \{W,\, R,\, S\}, \end{cases} \tag{1}$$

where, $r_H, r_N \in \mathbb{R}_{\geq 0}$ and $r_H > r_N$. We intend to design a decision support system that assists the operator by recommending optimal fidelity level to service each task [1]. The recommendation is based on the queue length and the operator's cognitive state which we assume to have real-time access using, e.g., Electroencephalogram (EEG) measurements (see [17] for measures of cognitive load from EEG data) or eye-tracking and pupillometry [18]. We assume that the noisy data from these devices can be clustered into a finite number of bins to estimate the cognitive state. We study the optimal policy under the perfect knowledge of the cognitive state [2].

3.2 Mathematical Modeling

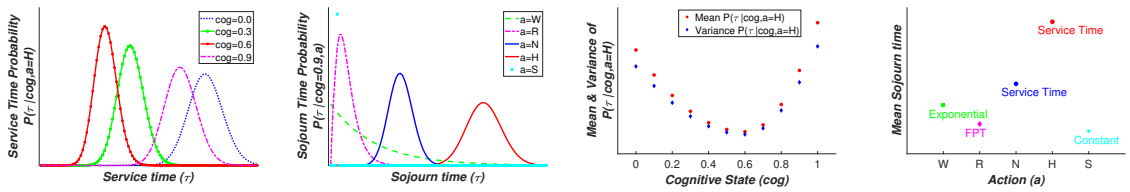
We formulate the control of queue problem as a discrete-time SMDP Γ defined by the following six components:

(i) A finite state space $\mathcal{S} := \{(q, \text{cog}) |\ q \in \{0, 1, ..., L\}, \text{cog} \in \mathcal{C} := \{i/N\}_{i \in \{0, \cdots, N\}}\}$, for some $N \in \mathbb{N}$, where q is the queue length and cog represents the lumped cognitive state, which increases (decreases) when the operator is busy (idle).

(ii) A set of admissible actions $\mathcal{A}_s$ for each state $s \in \mathcal{S}$ which is given by: (i) $\mathcal{A}_s := \{W \mid s \in \mathcal{S},\ q = 0\}$ when queue is empty, (ii) $\mathcal{A}_s := \{\{R,\, S,\, N,\, H\} |\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and $\text{cog} > \text{cog}^*$, where $\text{cog}^* \in \mathcal{C}$ is the optimal cognitive state associated with minimum mean service time, and (iii) $\mathcal{A}_s := \{\{S,\, N,\, H\} |\ s \in \mathcal{S},\ q \neq 0\}$ when queue is non-empty and $\text{cog} \leq \text{cog}^*$.

[1] We assume compliance of the operator with the recommendations. To account for non-compliance, we can introduce p as the probability of compliance and $1-p$ as the probability that the operator will deviate and follow a different behavioral policy. This deviation can be incorporated by using a mixed service time distribution with probabilities p and $1-p$ for the recommended and behavioral actions respectively.

[2] If the cognitive state is not perfectly known, then our policy can be used within algorithms such as Q_{MDP} [19], to derive approximate solutions to the associated partially observable Markov decision process [20].



(a) (b) (c) (d)

Fig. 2. Service time distribution of the human operator with (a) varying cognitive state and high fidelity, (b) varying action and fixed cognitive state, cog = 0.9. (c) Mean and variance of the service time distribution are unimodal functions of the cognitive state. (d) The mean sojourn time distribution takes on different forms based on the selected action.

(iii) A state transition distribution $\mathbb{P}(s'|\,\tau, s, a)$ from state s to s' for each action $a \in A_s$ conditioned on the discrete sojourn time $\tau \in \mathbb{R}_{>0}$ (time spent in state s before transitioning into next state s'). The state transition from $s = (q,\ \text{cog}) \to s' = (q',\ \text{cog}')$ consists of two independent transition processes which are given by (i) a Poisson process for transition from $q \to q'$ (ii) human cognitive dynamics for the transition from cog $\to$ cog$'$. We model the cognitive dynamics of the human operator as a Markov chain in which, while servicing the task, the probability of an increase in cognitive state in small time $\delta t \in \mathbb{R}_{>0}$ is greater than the probability of a decrease in cognitive state. Furthermore, the probability of the increase in the cognitive state increases with the level of fidelity selected for servicing the task. Similarly, while waiting or resting, the probability of a decrease in cognitive state in small time δt is higher than the probability of an increase in cognitive state. Sample parameters of the model used in our numerical simulations are shown in Table 1. This model of cognitive state dynamics is a stochastic equivalent of deterministic models of the utilization ratio considered in [2]. It is assumed that the cognitive state remains unchanged when the human operator chooses to skip the task.

Table 1
Cognitive Dynamics modeled as Markov chain

Action	*Forward Probability*[a] ($\lambda_f \delta t$)	*Backward Probability*[b] ($\lambda_b \delta t$)	*Stay Probability*[c] (1-$\lambda_f \delta t - \lambda_b \delta t$)
W	$\lambda_f = 0.02$ (Noise)	$\lambda_b = 0.5$	$1 - 0.52\delta t$
R	$\lambda_f = 0.02$ (Noise)	$\lambda_b = 0.5$	$1 - 0.52\delta t$
N	$\lambda_f = 0.6$	$\lambda_b = 0.02$ (Noise)	$1 - 0.62\delta t$
H	$\lambda_f = 1.1$	$\lambda_b = 0.02$ (Noise)	$1 - 1.12\delta t$
S	$\lambda_f = 0$	$\lambda_b = 0$	1

[a]Forward Probability does not exist for cog = 1 (reflective boundary)
[b]Backward Probability does not exist for cog = 0 (reflective boundary)
[c]Stay Probability is $1 - \lambda_f \delta t$ for cog = 0 and $1 - \lambda_b \delta t$ for cog = 1

(iv) Sojourn time distribution $\mathbb{P}(\tau|\ s, a)$ of (discrete) time $\tau \in \mathbb{R}_{>0}$ spent in state s until the next action is chosen takes on different forms depending on the selected action (Fig. 2d). The sojourn time is the service time while servicing the task (normal/ high fidelity), resting time while resting, constant time of skipping $t_s \in \mathbb{R}_{>0}$ while skipping, and time until the next task arrival while waiting in case of an empty queue. We model the rest time as the time required to reach from the current cognitive state to the optimal cognitive state cog*. In our numerical illustrations, we model the service time distribution while servicing the task using a hypergeometric distribution (Fig. 2a and 2b), where the parameters of the distribution are chosen such that the mean service time has the desired characteristics, i.e., it increases with the fidelity level (Fig. 2d) and is a unimodal function of the cognitive state (Fig. 2c). While resting, sojourn time distribution is the first passage time (FPT) distribution for transitioning from the current cognitive state cog to cog*. We determine this distribution using matrix methods [21] applied to the Markov chain used to model the cognitive dynamics. Finally, to ensure the stability of the queue, we assume that the constant time of skip is less than $\frac{1}{\lambda}$, i.e., queue length decreases on average while skipping tasks.

(v) For selecting action a at state s, the human receives a bounded reward $r(s, a)$ defined in (1). Additionally, the human incurs a penalty at a constant cost rate of c due to each task waiting in the queue, and consequently, the cumulative expected cost for choosing action a at state $s = (q, \text{cog})$ is given by:

$$\sum_{\tau} \mathbb{P}(\tau|s,a) c\tau \left(\mathbb{E}\left[\frac{q+q'}{2} \middle|\ \tau, s, a \right] \right) = \sum_{\tau} \mathbb{P}(\tau|s,a) c\tau \left(\frac{2q + \lambda\tau}{2} \right),$$

which is obtained by using $\mathbb{E}[q|\tau, s, a] = q$ and $\mathbb{E}[q'|\tau, s, a] = q + \lambda\tau$. The expected net immediate reward received by the operator for selecting an action a in state s is given by:

$$\begin{aligned} R(s,a) &= r(s,a) - \sum_{\tau} \mathbb{P}(\tau|s,a) c \left(\frac{2q + \lambda\tau}{2} \right) \tau \\ &= r(s,a) - c\,\mathbb{E}\left[\tau|s,a\right] q - \frac{c\lambda}{2}\,\mathbb{E}\left[\tau^2|s,a\right], \end{aligned} \tag{2}$$

where $\mathbb{E}[\tau|\ s,a]$ and $\mathbb{E}\left[\tau^2|s,a\right]$ represent the first and the second conditional moment of the sojourn time distribution, respectively.

(vi) A discount factor $\gamma \in [0,1)$, which we choose as 0.96 for our numerical illustration.

Remark 1. *Although we assume a finite skip time, an alternative approach is to incorporate a penalty for the skip action. Note that, unlike a fixed penalty, a finite skip time results in a penalty that increases with queue length (see (2)). Consequently, the current approach is less inclined to skip tasks as the queue length increases compared to a model with a constant penalty.*

Remark 2. *The reward $R(s,a)$ formulation can be interpreted as an unconstrained SMDP corresponding to a constrained SMDP that maximizes $r(s,a)$ subject to a constraint on the average queue length for the stability of the queue. Therefore, the penalty rate c acts as the Lagrange multiplier for the unconstrained problem, and hence, can be obtained by primal-dual methods that use dual ascent for finding the Lagrange multiplier [5].*

3.3 Solving SMDP for Optimal Policy

For SMDP Γ, the optimal value function $V^* : \mathcal{S} \to \mathbb{R}$ satisfies the following Bellman equation [22]:

$$V^*(s) = \max_{a \in A_s} \left[R(s,a) \ + \sum_{s',\tau} \gamma^{\tau} \mathbb{P}\left(s', \tau | s, a\right) V^*\left(s'\right) \right], \tag{3}$$

where $\mathbb{P}\left(s', \tau | s, a\right)$, which is the joint probability that a transition from state s to state s' occurs after time τ when action a is selected can be rewritten as:

$$\mathbb{P}\left(s', \tau | s, a\right) = \mathbb{P}\left(s' | \tau, s, a\right) \mathbb{P}\left(\tau | s, a\right), \tag{4}$$

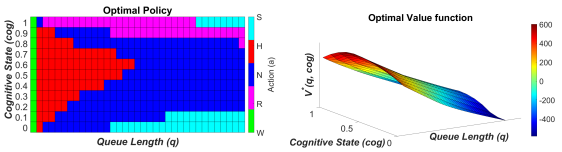
where $\mathbb{P}\left(s' | \tau, s, a\right)$ and $\mathbb{P}\left(\tau | s, a\right)$ are given by the state transition probability distribution and the sojourn time probability distribution, respectively. An optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}_s$ at each state s selects an action that achieves the maximum in (3). We utilize the value iteration algorithm [23] to compute an optimal policy.

4 Numerical Illustrations

We now numerically illustrate the optimal value function and an optimal policy for SMDP Γ.

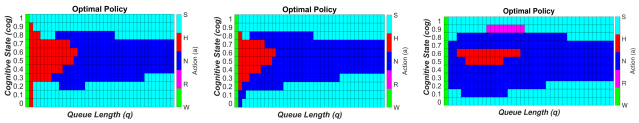
Fig. 3a and 3b show an optimal policy π^*, and the optimal value function V^*, respectively, for the case in which the skip time is not too small compared to the mean service time. If the skip time is too small, the action S is the optimal action almost everywhere to reduce the queue length. For a sufficiently high arrival rate λ such that there is always a task in the queue after servicing the current task, we observe that for any given cog, V^* is monotonically decreasing with q.

Additionally, we observe that for a given q, V^* is an unimodal function of cog, with its maximum value corresponding to the optimal cognitive state ($\text{cog}^* = 0.6$ for numerical illustrations). We observe that π^* selects the high fidelity level around the cog^* for low queue length, and thereafter transitions to a normal fidelity level for higher queue lengths. We also observe that in low cognitive states, the optimal policy is to keep skipping the tasks until the queue length becomes small, and then start servicing the tasks. In higher cognitive states, we observe that resting is the optimal action at smaller queue lengths while skipping tasks is the optimal choice at larger queue lengths. Additionally, we observe the effect of cog on π^*. In particular, we observe that π^* switches from H to N, N to R, and R to S at certain thresholds on q, and these thresholds appear to be a unimodal function of cog. This behavior can be attributed to the mean service time being unimodal w.r.t cog.



(a) (b)

Fig. 3. (a) Optimal Policy π^* and (b) Optimal Value Function V^* for SMDP Γ where the time required to skip the tasks is not too small compared to the mean service time.



(a) (b) (c)

Fig. 4. Optimal policy π^* for $(a)\lambda = 0.5$, (b) $\lambda = 1$, and (c) $\lambda = 4$. In cases (a) and (b), action S in the optimal policy does not have a unique threshold for some cognitive states. Similarly, in case (c), actions S and N in the optimal policy do not have unique thresholds.

Fig. 4 shows some examples of π^* for certain parameters. We observe that for some cognitive states, π^* does not have a unique threshold and the same action reappears after switching to another action. For example, in Fig. 4a, action S is observed between actions H and N, as well as after action N. In the following section, we provide sufficient conditions under which π^* has unique transition thresholds at which actions switch, and the previous action does not re-appear for the same cog.

5 Structural Properties of the Optimal Policy

We establish the structural properties of the optimal infinite-horizon value function by considering the finite horizon case and then extending the results to the infinite horizon by taking the infinite step limit.

Let $V_n^*(s_0), n \geq 0$, be the discounted n-step optimal expected reward when the initial state is s_0, where

$V_0^*(q, \text{cog}) = -Cq$ is the terminal cost for the finite-horizon case for a non-negative constant C. Each step size $k \in \{0, \ldots, n-1\}$ is based on the sojourn time τ_k, spent in a state s_k when action a_k is selected. Let $J_{n,\pi}(s_0)$ denote the discounted n-step expected reward with initial state s_0 under a given policy π. Henceforth, for the brevity of notation, we denote the conditional expectation $\mathbb{E}[\cdot|s_0, \pi]$ by $\mathbb{E}_\pi[\cdot]$. $J_{n,\pi}(s_0)$ is given by:

$$J_{n,\pi}(s_0) = \mathbb{E}_\pi \left[\sum_{i=0}^{n-1} \gamma^{\zeta_i} R(s_i, a_i) - \gamma^{\zeta_n} C q_n \right], \quad (5)$$

where $\zeta_i := \sum_{j=0}^{i-1} \tau_j$ for $i > 0$ and $\zeta_0 := 0$. The discounted n-step optimal expected reward $V_n^*(s)$ is given by:

$$V_n^*(s_0) = J_{n,\pi^*}(s_0), \quad (6)$$

where π^* is the optimal policy that maximizes $J_{n,\pi}(s_0)$ at each s_0.

Let $\mu^1 : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}_{>0}$ and $\mu^2 : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}_{>0}$ be function defined by $\mu^1(s,a) = \mathbb{E}[\tau|s,a]$ and $\mu^2(s,a) = \mathbb{E}\left[\tau^2|s,a\right]$, where τ is the sojourn time. We study the structural properties of the optimal policy for a large queue capacity, i.e. in the limit $L \to +\infty$, and under the following assumptions:

(A1) The task arrival rate λ is sufficiently high so that the queue is never empty with high probability [3].

(A2) For any state $s = (q, \text{cog})$ [4]:

$$\begin{aligned} &\mu^1(s,S) < \mu^1(s,R) < \mu^1(s,N) < \mu^1(s,H), \text{ and} \\ &\mu^2(s,S) < \mu^2(s,R) < \mu^2(s,N) < \mu^2(s,H). \end{aligned} \quad (7)$$

(A3) We assume that $\mathbb{E}_\pi[\gamma^\tau] \le f(\mathbb{E}_\pi[\tau], \text{Var}_\pi(\tau)) < 1$, where $\text{Var}_\pi(\tau) = \text{Var}(\tau|s_0, a = \pi(s_0))$ is the variance of τ in any initial state s_0 under a given policy π, and f is a monotonic function such that $f(\cdot, \text{Var}_\pi(\tau))$ is monotonically decreasing and $f(\mathbb{E}_\pi[\tau], \cdot)$ is monotonically increasing.

We make the assumption (A1) for convenience. Indeed, if the queue is allowed to be empty, then we will need to deal with an extra "waiting" action. Also, high arrival rates are the most interesting setting to study optimal fidelity selection. Assumption (A2) is true for a broad range of interesting parameters that define sojourn time distribution(s). Assumption (A3) holds for a class of light-tail distributions with non-negative support for τ, for example, when the moment generating function (MGF) of τ is upper bounded by the MGF of Gamma distribution, i.e.,

$$\mathbb{E}_\pi[e^{t\tau}] \le \left(1 - \frac{\text{Var}_\pi(\tau) t}{\mathbb{E}_\pi[\tau]}\right)^{\frac{-\mathbb{E}_\pi[\tau]^2}{\text{Var}_\pi(\tau)}}, \text{for all } t < \frac{\mathbb{E}_\pi[\tau]}{\text{Var}_\pi(\tau)}.$$

In this scenario, substituting $t = \ln(\gamma) < 0 < \frac{\mathbb{E}_\pi[\tau]}{\text{Var}_\pi(\tau)}$, we get

$$\begin{aligned} \mathbb{E}_\pi[\gamma^\tau] &\le \left(1 - \frac{\text{Var}_\pi(\tau)\ln(\gamma)}{\mathbb{E}_\pi[\tau]}\right)^{\frac{-\mathbb{E}_\pi[\tau]^2}{\text{Var}_\pi(\tau)}} \\ &=: f(\mathbb{E}_\pi[\tau], \text{Var}_\pi(\tau)). \end{aligned}$$

Let $\rho := \max_{\text{cog},a} f(\mathbb{E}[\tau|\text{cog}, a], \text{Var}(\tau|\text{cog}, a))$. Therefore, $\mathbb{E}_\pi[\gamma^\tau] \le \rho$. For the class of distributions of τ satisfying assumption (A3), and any initial state s_0 and policy π, we have

$$\mathbb{E}_\pi[\gamma^{\zeta_k}] \overset{(1)^*}{=} \prod_{i=0}^{k-1} \mathbb{E}_\pi[\gamma^{\tau_i}] \le \prod_{i=0}^{k-1} f(\mathbb{E}_\pi[\tau_i], \text{Var}_\pi(\tau_i)) \le \rho^k,$$

where $(1)^*$ follows from the independence of τ_i and τ_j, for $i \neq j$. Therefore, we have

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \mathbb{E}_\pi[\gamma^{\zeta_k}] \le \lim_{n\to\infty} \sum_{k=0}^{n-1} \rho^k = \frac{1}{1-\rho}.$$

We will now establish that the optimal policy for SMDP Γ is a threshold-based policy if the following condition holds for each cognitive state cog:

$$\begin{aligned} \min\{&\mathbb{E}[\tau|\text{cog}, H] - \mathbb{E}[\tau|\text{cog}, N],\ \mathbb{E}[\tau|\text{cog}, N] - \\ &\mathbb{E}[\tau|\text{cog}, R],\ \mathbb{E}[\tau|\text{cog}, R] - t_s\} + \\ &\frac{t_s \gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \ge \frac{t_{\max}}{1-\rho} \max_{a\in\mathcal{A}_s} \mathbb{E}[\gamma^\tau|\text{cog}, a], \end{aligned} \quad (8)$$

where $t_{\max} = \mathbb{E}[\tau|\text{cog} = 1, a = H]$ is the maximum expected sojourn time (assuming the largest mean service time in the highest cognitive state), and t_s is the constant time for the skip.

Remark 3. *For tasks with large differences in expected sojourn times, i.e.,* $0 \ll t_s \ll \mathbb{E}[\tau|\text{cog}, R] \ll \mathbb{E}[\tau|\text{cog}, N] \ll \mathbb{E}[\tau|\text{cog}, H]$, $\max_{a\in\mathcal{A}_s} \mathbb{E}[\gamma^\tau|\text{cog}, a] \to 0$, *and* (8) *always holds.*

We introduce the following notation. Let $q_j^* : \mathcal{C} \to \mathbb{Z}_{\ge 0} \cup \{+\infty\}$, for $j \in \{1, 2, 3\}$ be some functions of cog.

Theorem 1 (*Structure of optimal policy*). *For SMDP* Γ *under assumptions (A1-A3) and an associated optimal policy* π^**, if the difference in the expected sojourn times is sufficiently large such that* (8) *holds for any cognitive state* cog*, then the following statements hold:*

[3] Given the service time distributions and the Poisson arrival rate, we can precisely determine the distribution of the number of arrivals that occur between each state transition. Therefore, Chernoff bounds [24] can be utilized to characterize the high probability that the number of arrivals between state transitions while servicing a task exceeds one.

[4] The action R is only available for states with $\text{cog} > \text{cog}^*$.

(i) there exists unique threshold functions $q_1^*(\text{cog})$, $q_2^*(\text{cog})$, *and* $q_3^*(\text{cog})$ *such that for each* $\text{cog} > \text{cog}^*$*:*

$$\pi^*(s=(q,\text{cog})) = \begin{cases} H, & q \le q_1^*(\text{cog}), \\ N, & q_1^*(\text{cog}) < q \le q_2^*(\text{cog}), \\ R, & q_2^*(\text{cog}) < q \le q_3^*(\text{cog}), \\ S, & q > q_3^*(\text{cog}); \end{cases}$$

(ii) there exists unique threshold functions $q_1^*(\text{cog})$ *and* $q_2^*(\text{cog})$ *such that for any* $\text{cog} \le \text{cog}^*$*:*

$$\pi^*(s=(q,\text{cog})) = \begin{cases} H, & q \le q_1^*(\text{cog}), \\ N, & q_1^*(\text{cog}) < q \le q_2^*(\text{cog}), \\ S, & q > q_2^*(\text{cog}). \end{cases}$$

We prove Theorem 1 using the following lemmas.

Lemma 1. ***(Immediate Reward)****: For SMDP* Γ*, the immediate expected reward* $R(s,a)$*, for each* $a \in A_s$

(i) is linearly decreasing with queue length q *for any fixed cognitive state* cog*;*
(ii) is a unimodal function[5] *of the cognitive state* cog *for any fixed queue length* q *with its maximum value achieved at the optimal cognitive state* cog^**.*

Proof. The proof follows by noting that (2) is linearly decreasing in q and the coefficients $\mathbb{E}[\tau|\ s,a]$ and $\mathbb{E}\left[\tau^2|s,a\right]$ are unimodal w.r.t cog. Interested readers can refer to [25] for detailed proof. □

We now provide important mathematical results in Lemma 2 which we use to establish Lemma 3.

Lemma 2. *For the SMDP* Γ*, the following equations hold for any initial state* s_0 *and policy* π*:*

(i) $\mathbb{E}_\pi\left[\gamma^{\zeta_k}\,\mathbb{E}[\tau_k^2|\text{cog}_k,a_k]\right] = \mathbb{E}_\pi\left[\gamma^{\zeta_k}\tau_k^2\right]$*;*
(ii) $\mathbb{E}_\pi\left[\gamma^{\zeta_k}\,\mathbb{E}[\tau_k|\text{cog}_k,a_k]q_k\right] = \mathbb{E}_\pi\left[\gamma^{\zeta_k}\tau_k\,\mathbb{E}_\pi\left[q_k|s_0,\zeta_k\right]\right]$

Proof. The proof utilizes the properties of the expectation operator, and independence of the transition processes for q_k and cog_k. Interested readers can refer to [25] for detailed proof. □

Lemma 3. ***(Value function bounds)****: For SMDP* Γ *under assumptions (A1-A3), for any* $\tilde{q}_0 \ge q_0$*,* $0 \le \frac{ct_s\Delta q}{1-\gamma^{t_{\max}}} \le V^*(q_0,\text{cog}_0) - V^*(\tilde{q}_0,\text{cog}_0) \le \frac{ct_{\max}\Delta q}{1-\rho}$*, where* $\Delta q = \tilde{q}_0 - q_0$*,* ρ *is an upper bound on* $\mathbb{E}_\pi[\gamma^\tau]$*,* $t_{\max} = \mathbb{E}[\tau|\text{cog}=1,a=H]$ *is the maximum expected sojourn time, and* t_s *is the constant time for skip.*

Proof. See Appendix A for the proof. □

Remark 4. *It follows from Lemma 3, that for SMDP* Γ *under assumptions (A1-A3), the optimal value function* $V^*(q,\cdot)$ *is monotonically decreasing with queue length* q*.*

Lemma 4. ***(Thresholds for low cognitive states)****: For the SMDP* Γ *under assumptions (A1-A3), and an associated optimal policy* π^**, the following statements hold for each* $\text{cog} \le \text{cog}^*$*:*

(i) there exists a threshold function $q_1^*(\text{cog})$*, such that N strictly dominates H, for each* $q > q_1^*(\text{cog})$ *if*

$$\mathbb{E}[\tau|\text{cog},H] - \mathbb{E}[\tau|\text{cog},N] + \frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \ge \frac{t_{\max}}{1-\rho}\,\mathbb{E}[\gamma^\tau|\text{cog},N];$$

(ii) there exists a threshold function $q_2^*(\text{cog})$*, such that for each* $q > q_2^*(\text{cog})$*, action S is optimal if*

$$\mathbb{E}[\tau|\text{cog},N] - t_s + \frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \ge \gamma^{t_s}\frac{t_{\max}}{1-\rho}.$$

Proof. See Appendix B for the proof. □

Lemma 5. ***(Thresholds for high cognitive states)****: For the SMDP* Γ *under assumptions (A1-A3), and an associated optimal policy* π^**, the following statements hold for each* $\text{cog} > \text{cog}^*$*:*

(i) there exists a threshold function $q_1^*(\text{cog})$*, such that N strictly dominates H, for each* $q > q_1^*(\text{cog})$ *if*

$$\mathbb{E}[\tau|\text{cog},H] - \mathbb{E}[\tau|\text{cog},N] + \frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \ge \frac{t_{\max}}{1-\rho}\,\mathbb{E}[\gamma^\tau|\text{cog},N];$$

(ii) there exists a threshold function $q_2^*(\text{cog})$*, such that R strictly dominates H & N, for each* $q > q_2^*(\text{cog})$ *if*

$$\mathbb{E}[\tau|\text{cog},N] - \mathbb{E}[\tau|\text{cog},R] + \frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \ge \frac{t_{\max}}{1-\rho}\mathbb{E}[\gamma^\tau|\text{cog},R].$$

(iii) there exists a threshold function $q_3^*(\text{cog})$*, such that for each* $q > q_3^*(\text{cog})$*, action S is optimal if*

$$\mathbb{E}[\tau|\text{cog},R] - t_s + \frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}} \ge \gamma^{t_s}\frac{t_{\max}}{1-\rho}.$$

Proof. Recall that $\mathcal{A}_s := \{\{R,\ S,\ N,\ H\ \}|\ s \in \mathcal{S},\ q \ne 0\}$ when queue is non-empty and $\text{cog} > \text{cog}^*$. The proof of Lemma 5 follows analogously to the proof of Lemma 4.□

Proof of Theorem 1: The proof follows by finding the intersection of the sufficient conditions from Lemmas 4 and 5 to obtain condition (8) for a threshold-based π^*.

6 Conclusions and Future Directions

We studied optimal fidelity selection for a human operator servicing a stream of homogeneous tasks using an SMDP framework. In particular, we studied the influence of human cognitive dynamics on an optimal fidelity

[5] The expected immediate reward under action S is a constant, which we treat as a unimodal function.

selection policy. We presented numerical illustrations of the optimal policy and established its structural properties. These structural properties can be leveraged to tune the design parameters, deal with the model uncertainty, or determine a minimally parameterized policy for specific individuals and tasks.

There are several possible avenues for future research. An interesting direction is to conduct experiments with human subjects, measure EEG signals to assess their cognitive state and test the benefits of recommending optimal fidelity levels. It is of interest to extend this work to a team of human operators servicing a stream of heterogeneous tasks. A preliminary setup is considered in [26, 27], where authors study a game-theoretic approach to incentivize collaboration in a team of heterogeneous agents. In such a setting, finding the optimal routing and scheduling strategies for these heterogeneous tasks is also of interest.

Acknowledgements

This work has been supported by NSF Award IIS-1734272 and ECCS-2024649.

References

[1] P. Gupta and V. Srivastava, "Optimal fidelity selection for human-in-the-loop queues using semi-Markov decision processes," in *2019 American Control Conference (ACC)*, pp. 5266–5271, IEEE, 2019.

[2] K. Savla and E. Frazzoli, "A dynamical queue approach to intelligent task management for human operators," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 672–686, 2012.

[3] I. R. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human-robot teaming for search and rescue," *IEEE Pervasive Computing*, vol. 4, no. 1, pp. 72–78, 2005.

[4] M. Yuan Zhang and X. Jessie Yang, "Evaluating effects of workload on trust in automation, attention allocation and dual-task performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, pp. 1799–1803, SAGE Publications Sage CA: Los Angeles, CA, 2017.

[5] M. Agarwal, V. S. Borkar, and A. Karandikar, "Structural properties of optimal transmission policies over a randomly varying channel," *IEEE Transactions on Automatic Control*, vol. 53, no. 6, pp. 1476–1491, 2008.

[6] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908.

[7] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1393–1403, 2012.

[8] J. Peters, V. Srivastava, G. Taylor, A. Surana, M. P. Eckstein, and F. Bullo, "Human supervisory control of robotic teams: Integrating cognitive modeling with engineering design," *IEEE Control System Magazine*, vol. 35, no. 6, pp. 57–80, 2015.

[9] J. R. Peters, A. Surana, and F. Bullo, "Robust scheduling and routing for collaborative human/unmanned aerial vehicle surveillance missions," *Journal of Aerospace Information Systems*, pp. 1–19, 2018.

[10] V. Srivastava, R. Carli, C. Langbort, and F. Bullo, "Attention allocation for decision making queues," *Automatica*, vol. 50, no. 2, pp. 378–388, 2014.

[11] M. Lin, R. J. La, and N. C. Martins, "Stabilizing a queue subject to activity-dependent server performance," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 4, pp. 1579–1591, 2021.

[12] M. Lin, N. C. Martins, and R. J. La, "Queueing subject to action-dependent server performance: Utilization rate reduction," *arXiv preprint arXiv:2002.08514*, 2020.

[13] P. Gupta and V. Srivastava, "On robust and adaptive fidelity selection for human-in-the-loop queues," in *2021 European Control Conference (ECC)*, pp. 872–877, IEEE, 2021.

[14] S. Stidham Jr and R. R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Operations Research*, vol. 37, no. 4, pp. 611–625, 1989.

[15] L. I. Sennott, "Average cost semi-Markov decision processes and the control of queueing systems," *Probability in the Engineering and Informational Sciences*, vol. 3, no. 2, pp. 247–272, 1989.

[16] R. Yang, S. Bhulai, and R. van der Mei, "Structural properties of the optimal resource allocation policy for single-queue systems," *Annals of Operations Research*, vol. 202, no. 1, pp. 211–233, 2013.

[17] R. P. Rao, *Brain-Computer Interfacing: An Introduction*. Cambridge University Press, 2013.

[18] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 Symposium on Eye-tracking Research & Applications*, pp. 141–144, 2010.

[19] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," in *Machine Learning Proceedings 1995*, pp. 362–370, Elsevier, 1995.

[20] M. T. Spaan, "Partially observable markov decision processes," in *Reinforcement Learning*, pp. 387–414, Springer, 2012.

[21] A. Diederich and J. R. Busemeyer, "Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time," *Journal of Mathematical Psychology*, vol. 47, no. 3, pp. 304–322, 2003.

[22] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dynamic Systems*, vol. 13, no. 1-2, pp. 41–77, 2003.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[24] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[25] P. Gupta and V. Srivastava, "Structural properties of optimal fidelity selection policies for human-in-the-loop queues," *arXiv preprint arXiv:2201.09990*, 2022.

[26] P. Gupta, S. D. Bopardikar, and V. Srivastava, "Achieving efficient collaboration in decentralized heterogeneous teams using common-pool resource games," in *IEEE Conference on Decision and Control*, pp. 6924–6929, 2019.

[27] P. Gupta, S. D. Bopardikar, and V. Srivastava, "Incentivizing collaboration in heterogeneous teams via common-pool

resource games," *IEEE Transactions on Automatic Control*, 2022.

[28] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer Science & Business Media, 2005.

A Proof of Lemma 3

Let w_k be the number of tasks that arrive during stage $k \in \{0, \ldots, n-1\}$ with sojourn time τ_k, in which the state transitions from $s_k = (q_k,\ \mathrm{cog}_k) \to s_{k+1} = (q_{k+1},\ \mathrm{cog}_{k+1})$ and action a_k is selected. Let a_k be an optimal action at state s_k and π be the corresponding optimal policy such that $a_k = \pi(s_k)$. The optimal policy π when applied from an initial state s_0 induces a sequence of states $< s_k >$ and sojourn times $< \tau_k >$ or ζ_{k+1}, where $\zeta_{k+1} = \sum_{j=0}^{k} \tau_j$ and $\zeta_0 = 0$.
Similarly, let $\tilde{s}_0 = (\tilde{q}_0, \mathrm{cog}_0)$ be another initial state with the same initial cognitive state, and $\tilde{q}_0 \geq q_0$. Apply a policy $\tilde{\pi}$ from the initial state $\tilde{s}_0$ such that $\tilde{\pi}(\overline{q}, \overline{\mathrm{cog}}) = \pi(\overline{q} + q_0 - \tilde{q}_0, \overline{\mathrm{cog}})$ for any $(\overline{q}, \overline{\mathrm{cog}})$. Note that $\tilde{a}_0 = a_0$. The optimal policy $\tilde{\pi}$ when applied from an initial state $\tilde{s}_0 = (\tilde{q}_0, \mathrm{cog}_0)$ induces a sequence of realizations $< \tilde{s}_k >$ and $< \tilde{\tau}_k >$. Since cognitive state and sojourn time are independent of the current queue length, for the same action sequence applied from the initial states s_0 and $\tilde{s}_0$, the random process associated with the evolution of cognitive state and sojourn time is almost surely the same except for the offset in the queue length. Hence, the probability of observing a sequence of realizations $< \tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k) >$, $< \tilde{a}_k >$ and $< \tilde{\tau}_k >$ when policy $\tilde{\pi}$ is applied from $\tilde{s}_0$ is equal to the probability of observing a sequence of realizations $< s_k = (q_k, \mathrm{cog}_k) >$, $< a_k >$ and $< \tau_k >$ when policy π is applied from s_0, where $\tilde{q}_k - \tilde{q}_0 = q_k - q_0$, $\tilde{a}_k = a_k$ and $\tilde{\tau}_k = \tau_k$. Therefore, it is easy to show that:

$$\mathbb{E}_{\tilde{\pi}}[\tilde{q}_k | \tilde{s}_0, \zeta_k] - \mathbb{E}_{\pi}[q_k | s_0, \zeta_k] = \tilde{q}_0 - q_0. \tag{A.1}$$

Note that the realization of sequence of actions $< a_k = \pi(s_k) >$, which are optimal for $< s_k = (q_k, \mathrm{cog}_k) >$ might be sub-optimal for $< \tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k) >$. Recall that $\mathbb{E}_{\pi}[\cdot]$ and $\mathbb{E}_{\tilde{\pi}}[\cdot]$ represents $\mathbb{E}[\cdot | s_0, \pi]$ and $\mathbb{E}[\cdot | \tilde{s}_0, \tilde{\pi}]$, respectively. Let $\Delta q := \tilde{q}_0 - q_0$, and $Z := V_n^*(q_0, \mathrm{cog}_0) - V_n^*(\tilde{q}_0, \mathrm{cog}_0)$. We first show the upper bound on Z.

$$\begin{aligned}
Z &\leq V_n^*(q_0, \mathrm{cog}_0) - J_{n,\tilde{\pi}}(\tilde{q}_0, \mathrm{cog}_0) \\
&= \mathbb{E}_{\pi}\left[\sum_{k=0}^{n-1} \gamma^{\zeta_k} R(s_k, a_k) - \gamma^{\zeta_n} C q_n\right] - \\
&\qquad \mathbb{E}_{\tilde{\pi}}\left[\sum_{k=0}^{n-1} \gamma^{\zeta_k} R(\tilde{s}_k, \tilde{a}_k) - \gamma^{\zeta_n} C \tilde{q}_n\right] \\
&= \mathbb{E}_{\pi}\left[\sum_{k=0}^{n-1} \gamma^{\zeta_k} \{r(a_k) - c\,\mathbb{E}[\tau_k | \mathrm{cog}_k, a_k] q_k \right. \\
&\quad \left. - \frac{c\lambda}{2}\,\mathbb{E}[\tau_k^2 | \mathrm{cog}_k, a_k]\} - \gamma^{\zeta_n} C q_n\right] \\
&\quad - \mathbb{E}_{\tilde{\pi}}\left[\sum_{k=0}^{n-1} \gamma^{\zeta_k} \{r(\tilde{a}_k) - c\,\mathbb{E}[\tau_k | \mathrm{cog}_k, \tilde{a}_k] \tilde{q}_k \right. \\
&\quad \left. - \frac{c\lambda}{2}\,\mathbb{E}[\tau_k^2 | \mathrm{cog}_k, \tilde{a}_k]\} - \gamma^{\zeta_n} C \tilde{q}_n\right].
\end{aligned} \tag{A.2}$$

Using statements of Lemma 2, RHS of (A.2) is given by:

$$\begin{aligned}
&\sum_{k=0}^{n-1} \mathbb{E}_{\pi}[\gamma^{\zeta_k} r(a_k)] - c \sum_{k=0}^{n-1} \mathbb{E}_{\pi}\left[\gamma^{\zeta_k} \tau_k\, \mathbb{E}_{\pi}\left[q_k | s_0, \zeta_k\right]\right] \\
&- \frac{c\lambda}{2} \sum_{k=0}^{n-1} \mathbb{E}_{\pi}\left[\gamma^{\zeta_k} \tau_k^2\right] - C\, \mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\, \mathbb{E}_{\pi}\left[q_n | s_0, \zeta_n\right]\right] \\
&- \sum_{k=0}^{n-1} \mathbb{E}_{\tilde{\pi}}[\gamma^{\zeta_k} r(\tilde{a}_k)] + c \sum_{k=0}^{n-1} \mathbb{E}_{\tilde{\pi}}\left[\gamma^{\zeta_k} \tau_k\, \mathbb{E}_{\tilde{\pi}}\left[\tilde{q}_k | \tilde{s}_0, \zeta_k\right]\right] \\
&+ \frac{c\lambda}{2} \sum_{k=0}^{n-1} \mathbb{E}_{\tilde{\pi}}\left[\gamma^{\zeta_k} \tau_k^2\right] + C\, \mathbb{E}_{\tilde{\pi}}\left[\gamma^{\zeta_n}\, \mathbb{E}_{\tilde{\pi}}\left[\tilde{q}_n | \tilde{s}_0, \zeta_n\right]\right] \\
&\overset{(4)^*}{=} c \sum_{k=0}^{n-1} \mathbb{E}_{\pi}\left[\gamma^{\zeta_k} \tau_k \{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_k\ | \tilde{s}_0, \zeta_k] - \mathbb{E}_{\pi}[q_k\ | s_0, \zeta_k]\}\right] \\
&+ C\, \mathbb{E}_{\pi}\left[\gamma^{\zeta_n} \{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_n\ | \tilde{s}_0, \zeta_k] - \mathbb{E}_{\pi}[q_n\ | s_0, \zeta_k]\}\right],
\end{aligned} \tag{A.3}$$

where $(4)^*$ follows by recalling that the probability of observing a sequence of realizations $< \tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k) >$, $< \tilde{a}_k >$ and $< \tilde{\tau}_k >$ when policy $\tilde{\pi}$ is applied from $\tilde{s}_0$ is equal to the probability of observing a sequence of realizations $< s_k = (q_k, \mathrm{cog}_k) >$, $< a_k >$ and $< \tau_k >$ when policy π is applied from s_0, where $\tilde{q}_k - \tilde{q}_0 = q_k - q_0$, $\tilde{a}_k = a_k$ and $\tilde{\tau}_k = \tau_k$. Substituting (A.1) in (A.3), we get,

$$\begin{aligned}
Z &\leq \left\{c \sum_{k=0}^{n-1} \mathbb{E}_{\pi}\left[\gamma^{\zeta_k} \tau_k\right] + C\, \mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\right]\right\} \Delta q \\
&\overset{(5)^*}{=} \left\{c \sum_{k=0}^{n-1} \mathbb{E}_{\pi}\left[\gamma^{\zeta_k}\right] \mathbb{E}_{\pi}\left[\tau_k\right] + C\, \mathbb{E}_{\pi}\left[\gamma^{\zeta_n}\right]\right\} \Delta q \\
&\leq \left\{c t_{\max} \sum_{k=0}^{n-1} \rho^k + C \rho^n\right\} \Delta q,
\end{aligned} \tag{A.4}$$

where $(5)^*$ follows due to independence of $\zeta_k = \sum_{i=0}^{k-1} \tau_k$ and τ_k. Taking the infinite time limit in (A.4), we get, $V^*(q_0, \mathrm{cog}_0) - V^*(\tilde{q}_0, \mathrm{cog}_0) \leq \frac{c t_{\max} \Delta q}{1-\rho},\ \tilde{q}_0 \geq q_0$.
We now show the lower bound on Z. Let $\tilde{a}_k$ be optimal for $\tilde{s}_k = (\tilde{q}_k, \mathrm{cog}_k)$, and choose $< a_k >=< \tilde{a}_k >$ for the sequence $< s_k = (q_k, \mathrm{cog}_k) >$, where $\tilde{s}_0 = (\tilde{q}_0, \mathrm{cog}_0)$ and $s_0 = (q_0, \mathrm{cog}_0)$. Note that (A.1) still holds. Analogous to (A.3), Z is lower-bounded by:

$$Z \geq J_{n,\pi}(q_0, \mathrm{cog}_0) - V_n^*(\tilde{q}_0, \mathrm{cog}_0)$$

$$= c\sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_k}\tilde{\tau}_k\left\{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_k|\tilde{s}_0,\tilde{\zeta}_k]-\mathbb{E}_{\pi}[q_k|s_0,\tilde{\zeta}_k]\right\}\right] + C\,\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_n}\left\{\mathbb{E}_{\tilde{\pi}}[\tilde{q}_n|\tilde{s}_0,\tilde{\zeta}_k]-\mathbb{E}_{\pi}[q_n|s_0,\tilde{\zeta}_k]\right\}\right]. \tag{A.5}$$

Substituting (A.1) in (A.5), we get,

$$\begin{aligned} Z &\geq \left\{c\sum_{k=0}^{n-1}\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_k}\right]\mathbb{E}_{\tilde{\pi}}\left[\tilde{\tau}_k\right]+C\,\mathbb{E}_{\tilde{\pi}}\left[\gamma^{\tilde{\zeta}_n}\right]\right\}\Delta q \\ &\overset{(6)^*}{\geq}\left\{ct_s\sum_{k=0}^{n-1}\gamma^{\mathbb{E}_{\tilde{\pi}}[\tilde{\zeta}_k]}+C\gamma^{\mathbb{E}_{\tilde{\pi}}[\tilde{\zeta}_n]}\right\}\Delta q \\ &\geq\left\{\frac{(1-\gamma^{nt_{\max}})ct_s}{1-\gamma^{t_{\max}}}+\gamma^{nt_{\max}}C\right\}\Delta q, \end{aligned} \tag{A.6}$$

where $(6)^*$ follows by applying Jensens inequality [28] ($\mathbb{E}[g(x)]\geq g(\mathbb{E}[x])$) on the convex function $g(x)=\gamma^x$. Taking the infinite time limit in (A.6), we get, $0\leq\frac{ct_s\Delta q}{1-\gamma^{t\max}}\leq V^*(q_0,\text{cog}_0)-V^*(\tilde{q}_0,\text{cog}_0),\ \tilde{q}_0\geq q_0$. □

B Proof of Lemma 4

Proof. We start by proving the first statement. In the following, we find conditions under which if action N is the optimal choice at queue length q_1 for a given cognitive state $\text{cog}\leq\text{cog}^*$, then for all $q_2>q_1$, N dominates H. Let N be the optimal action in state $s_1=\{q_1,\text{cog}\}$. Let $F(s,a)$ denote the expected future rewards received in state s for taking action a (the second term in the Bellman equation (3)). Then, we have

$$\begin{aligned} &R(s_1,N)-R(s_1,H)+F(s_1,N)-F(s_1,H)>0,\\ \implies\ &M+\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)V^*(\text{cog}',q_1+\overline{q}-1)\times\\ &\left(\mathbb{P}(\text{cog}',\tau|\text{cog},N)-\mathbb{P}(\text{cog}',\tau|\text{cog},H)\right)>0, \end{aligned} \tag{B.1}$$

where $M:=c(\mathbb{E}[\tau|\text{cog},H]-\mathbb{E}[\tau|\text{cog},N])q_1+r_N-r_H+\frac{c\lambda}{2}(\mathbb{E}[\tau^2|\text{cog},H]-\mathbb{E}[\tau^2|\text{cog},N])$, and $\mathbb{P}(q_1+\overline{q}-1|q_1,\tau)$ is replaced by $\texttt{Pois}(\overline{q}|\tau)$, which is the Poisson probability of $\overline{q}$ arrivals during service time τ.
Now for the state $s_2=\{q_2,\text{cog}\}$, with $q_2>q_1$ and identical cog, under the assumption (A1) we show that:

$$R(s_2,N)-R(s_2,H)+F(s_2,N)-F(s_2,H)>0. \tag{B.2}$$

The left-hand side of (B.2) is given by:

$$\begin{aligned} &X+M+\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|,\tau)V^*(\text{cog}',q_2+\overline{q}-1)\times\\ &\left(\mathbb{P}(\text{cog}',\tau|\text{cog},N)-\mathbb{P}(\text{cog}',\tau|\text{cog},H)\right), \end{aligned} \tag{B.3}$$

where $X:=c(\mathbb{E}[\tau|\text{cog},H]-\mathbb{E}[\tau|\text{cog},N])(q_2-q_1)$. To show (B.2), we prove that the difference between LHS of (B.2) and (B.1) is positive. Subtracting LHS of (B.1) from (B.3), we get:

$$\begin{aligned} &X-\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)V_D\times\\ &\left(\mathbb{P}(\text{cog}',\tau|\text{cog},N)-\mathbb{P}(\text{cog}',\tau|\text{cog},H)\right), \end{aligned} \tag{B.4}$$

where $V_D:=\Big[V^*(\text{cog}',q_1+\overline{q}-1)-V^*(\text{cog}',q_2+\overline{q}-1)\Big]$. From Lemma 3, we know that

$$0\leq\beta:=\frac{ct_s(q_2-q_1)}{1-\gamma^{t_{\max}}}\leq V_D\leq\frac{ct_{\max}(q_2-q_1)}{1-\rho}=:\alpha.$$

Therefore, (B.4) is lower bounded by

$$\begin{aligned} &X+\beta\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)\mathbb{P}(\text{cog}',\tau|\text{cog},H)\\ &-\alpha\sum_{\tau}\sum_{\text{cog}'}\sum_{\overline{q}}\gamma^{\tau}\texttt{Pois}(\overline{q}|\tau)\mathbb{P}(\text{cog}',\tau|\text{cog},N)\\ &\geq X+\beta\gamma^{\mathbb{E}[\tau|\text{cog},H]}-\alpha\,\mathbb{E}[\gamma^{\tau}|\text{cog},N], \end{aligned} \tag{B.5}$$

where we utilized Jensen's inequality on convex function γ^{τ} to obtain $\mathbb{E}[\gamma^{\tau}|\text{cog},H]\geq\gamma^{\mathbb{E}[\tau|\text{cog},H]}$. (B.5) is nonnegative when the condition in the first statement holds.
Now we prove the second statement. Using a similar analysis it can be shown that if action S is the optimal choice at queue length q_1 for a given cognitive state $\text{cog}\leq\text{cog}^*$, then for every $q_2>q_1$, S dominates H and N, respectively, under the following conditions:

$$\mathbb{E}[\tau|\text{cog},H]-t_s+\frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},H]}}{1-\gamma^{t_{\max}}}\geq\gamma^{t_s}\frac{t_{\max}}{1-\rho}, \tag{B.6}$$

$$\mathbb{E}[\tau|\text{cog},N]-t_s+\frac{t_s\gamma^{\mathbb{E}[\tau|\text{cog},N]}}{1-\gamma^{t_{\max}}}\geq\gamma^{t_s}\frac{t_{\max}}{1-\rho}, \tag{B.7}$$

respectively, where we have used $\mathbb{E}[\tau|\text{cog},S]=t_s$ and $\mathbb{E}[\gamma^{\tau}|\text{cog},S]=\gamma^{t_s}$ due to constant time of skip. Since $\mathbb{E}[\tau|\text{cog},H]>\mathbb{E}[\tau|\text{cog},N]$, (B.6)-(B.7) can be combined to obtain the condition in the second statement under which action S dominates both H and N. □

Piyush Gupta is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at Michigan State University. He earned his B.Tech. degree in Mechanical Engineering from the Indian Institute of Technology, Delhi, India, in 2015. During the years 2015 to 2017, he served as an R&D Engineer at Honda R&D Co. Ltd., Japan. Subsequently, in 2020, he completed his M.S. degree in Electrical and Computer Engineering at Michigan State University. His research

interests encompass a variety of areas, including human-in-the-loop systems, motion planning and prediction for autonomous vehicles, and machine learning algorithms.

Vaibhav Srivastava received the B.Tech. degree (2007) in mechanical engineering from the Indian Institute of Technology Bombay, Mumbai, India; the M.S. degree in mechanical engineering (2011), the M.A. degree in statistics (2012), and the Ph.D. degree in mechanical engineering (2012) from the University of California at Santa Barbara, Santa Barbara, CA.

Dr. Srivastava is currently an Associate Professor of Electrical and Computer Engineering at Michigan State University. He is also affiliated with Mechanical Engineering, Cognitive Science Program, and Connected and Autonomous Networked Vehicles for Active Safety (CANVAS). He served as a Lecturer and Associate Research Scholar with the Mechanical and Aerospace Engineering Department at Princeton University, Princeton, NJ from 2013-2016. His research focuses on Cyber Physical Human Systems with an emphasis on mixed human-robot systems and networked multi-agent systems.