A Unified Decision-Theoretic Model for Information Gathering and Communication Planning

Jennifer Renoux[1], Tiago Veiga[2,3], Pedro U. Lima[3] and Matthijs Spaan[4]

***Abstract*— We consider in this paper the problem of Communication Planning for Human-Machine Cooperation in stochastic and partially observable environments. POMDPs with Information Reward (POMDPs-IR) form a powerful framework for information gathering tasks in such environments. This paper suggests an extension of POMDP-IR, called the Communicating POMDP-IR (com-POMDP-IR), that allows an agent to proactively plan its communication actions by using an approximation of the human's beliefs. We demonstrate experimentally the capability of our com-POMDP-IR agent to limit its communication to relevant information and its robustness to lost messages.**

I. Introduction

As artificial agents enter human-inhabited environments, we expect them to be capable of communicating relevant information about their knowledge of environment to us, meaning that they should be capable to proactively select relevant information to report to a teammate. We refer to this process as *Communication Planning* and many applications require such communications. For instance, in assisted surveillance domains as the one described in [1], a human operator must monitor many parameters simultaneously (e.g., observe several surveillance cameras for uncommon events) and is at risk of being overwhelmed by the amount of information to process. In such systems, artificial agents can select and communicate about the relevant information to alleviate the operator's workload and improve the efficiency of the surveillance process. Other examples of applications might involve transparency [2] or explainable agency [3] in which the agent should report about its behavior and actions when they might not align with what the user is expecting. Generally speaking, this problem relates to the problem of Active Situation Reporting [4].

Partially Observable Markov Decision Processes (POMDPs) are suited for this type of problem as they are a well studied mathematical framework to perform sequential decision making in uncertain environments. POMDP with Information Rewards [5] is an extension of POMDPs to tackle specifically information-gathering tasks while remaining in the POMDP framework, thus allowing the use of existing POMDP solvers.

The main contribution of the current paper is a decision-theoretic framework, called Communicating POMDP-IR (com-POMDP-IR), integrating Information-Gathering tasks and Communication Planning with more classic goal-oriented tasks. This framework only assumes that the agent receiving the communication is using a Bayesian belief update and does not require any other information about its policy or internal model. For this reason, the com-POMDP-IR is well suited for human-machine collaboration.

Throughout this paper, we will consider the illustrative example presented in Figure 1 and inspired from [5]. One exploring agent is located in an environment with an alarm and must perform three parallel tasks: patrol the environment, observe the current state of the alarm and warn a human operator when the alarm is red. This is an example of a challenging problem where the patrolling agent must reason about its local actions and, simultaneously, decide about the communication to the human operator.

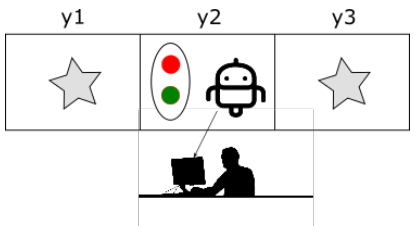


Fig. 1: The surveillance problem. An agent must patrol the environment by traveling between two goals (marked with the stars) while looking at the alarm color and communicate its color to the operator.

The remainder of this paper is organized as follows. Section II presents different studies and models similar to our problem. Section III reviews the key aspects of the POMDP-IR on which our work is based and Section IV presents our contribution: the com-POMDP-IR. Section V evaluates this model on the surveillance problem. Finally, Section VI summarizes our contributions and suggests leads for future work.

II. Related Work

Our work focuses on integrating information-gathering tasks, communication planning and "classic" goal-oriented

[1]Jennifer Renoux is with the Center of Applied Autonomous Sensor Systems, Orebro University, Sweden jennifer.renoux@oru.se

[2]Tiago Veiga is with the Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway tiago.veiga@ntnu.no

[3]Tiago Veiga and Pedro U. Lima are with the Institute for Systems and Robotics, Instituto Superior Tecnico, University of Lisbon, Portugal

[4]Matthijs Spaan is with the Algorithmics Group, Delft University of Technology, The Netherlands

tasks for human-machine cooperation. Information gathering tasks in decision-theoretic settings have received a lot of attention in the last decade, especially with the development of the ρ-POMDP [6] and the POMDP-IR [5], which reward agents based on belief states in addition to environmental states. Renoux [7] and Lauri et al.[8] considered the information gathering in multi-agent systems, using respectively POMDPs and Dec-POMDPs.

The principle of optimizing communication in decision-theoretic multiagent settings has been previously considered, but usually expect each agent to be modeled within the same approach (usually a Dec-POMDP or MTDP) [9], [10], [11]. However, in the case of a human-machine team, the human's actions cannot be controlled and such modeling is impossible. Recently, Wang et al. [12] consider this specific setup but expect the human's policy and observation model to be known. Their work, similarly to ours, introduces some elements of an Artificial Theory of Mind [13] by allowing their agent to reason about the human's beliefs and be rewarded on these beliefs. The Interactive POMDP (I-POMDP) [14], its communicating extension the CIPOMDP [15], and the MAPING model [7] are other attempts at modeling an Artificial Theory of Mind in decision-theoretic models. The IPOMDP and CIPOMDP are very expressive as they allow to model not only the other agent's beliefs but also their whole internal model (transition and observation functions, policy, etc.), and therefore considers that the other agents are modeled as POMDPs. This is not a reasonable hypothesis as soon as the other agent is human. Finally and similarly to our approach, the MAPING model also only considers the belief of the other agents in the planning agent's model. However, the MAPING model considers only information-gathering actions and does not allow the agent to be rewarded on goal-oriented tasks. On the contrary, our approach remains in the standard POMDP framework and can integrate the three types of tasks while benefiting from existing solvers.

POMDPs have already been used successfully in the topic of human-machine interaction. For instance, [16] models HRI-related variables within the POMDP (e.g. intention, satisfaction...) to assist the user more intuitively. More recently, [17] model the behavior of a human user into a POMDP-IR learn latent states of the user. To the best of our knowledge, the problem of explicit communication between humans and artificial agents in Decision-Theoretic settings is still understudied.

III. Background on POMDP-IR

Our work is based on the POMDP with Information Reward (POMDP-IR), described in [5]. In this section, we review the key aspects of the POMDP-IR as well as the notations relevant to the rest of the paper.

A POMDP-IR is represented as a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, R \rangle$ where $\mathcal{X}$ is a set of state factors, $\mathcal{A}$ is a set of action factors, $\mathcal{O}$ is a set of observation factors, $\mathcal{T} : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the transition function, $\Omega : \mathcal{X} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ is the observation function and $R : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the reward function. In a POMDP-IR, the set of state factors $\mathcal{X}$ contains l factors which are called Factors of Interest (FoI) and are the factors that the agent needs to explore. The POMDP-IR introduces the notion of Information Reward (IR) actions. There are as many IR actions as there are FoIs, and their values are either $commit$ or $null$. At each time step, the agent selects simultaneously a domain action and l IR actions. In addition to its domain reward, the agent is also rewarded for each IR action. The IR reward is based on two values: $r_{correct}$ and $r_{incorrect}$. Intuitively, the agent receives $r_{correct}$ when it commits to a correct value for the factor X_i, and $r_{incorrect}$ otherwise. Therefore, the agent should $commit$ on a factor X_i when its belief over X_i's value is high enough. The values of $r_{correct}$ and $r_{incorrect}$ are decided depending on the belief threshold β the system designer wishes to enforce before the agent commit. The relation between $r_{correct}$, $r_{incorrect}$ and β is given by $r_{correct} = \frac{1-\beta}{\beta} r_{incorrect}$

IV. Communicating POMDP-IR

In this section, we present the main contribution of this paper: a decision-theoretic framework rewarding agents for efficient communication planning. This framework is based on three main aspects:

1) an extended state factor space, which includes not only the state factors for the communicating agent but also duplicated state factors which represent what the communicating agent believes the recipient knows about certain state factors of interest.
2) communication actions that can be chosen simultaneously to other domain-dependent actions
3) a reward function that rewards the agent for maintaining synchronized beliefs over its own Factors of Interest and what it believes the human knows about these Factors of Interest.

Formally, we consider one artificial agent, noted ϕ and a human operator, noted ψ. We consider a set $X^\phi = \{X_1^\phi, \dots, X_n^\phi\}$ of state variables, the first l^ϕ of them being POMDP-IR Factors of Interest (FoIs). Within these l FoI, we consider that the first k FoIs are also of interest for the human, and that the agent must communicate to the human about them. We call these k FoIs *shared FoIs*.

Our approach is an extension of the POMDP-IR model which integrates communication actions. Figure 2 presents the Dynamic Bayesian Network representation of our model for the surveillance problem.

A. Extended State Space and Observation Factors

To be able to plan for optimal communication, Agent ϕ needs to model the beliefs of the human ψ in its own belief state, hence leading to nested beliefs. In the com-POMDP-IR, we consider only one level of nested beliefs: we only represent what Agent ϕ believes about the human ψ's beliefs. To do so, we extend the belief state of the POMDP by duplicating each of the k shared FoIs. These duplicated factors represent what Agent ϕ believes the human knows about state factors X_i. For better readability, we will use the notation X_i^ϕ for the classic state factors for Agent ϕ, $X_i^{\psi/\phi}$ for the duplicated state factors, and X_i for any state factor.

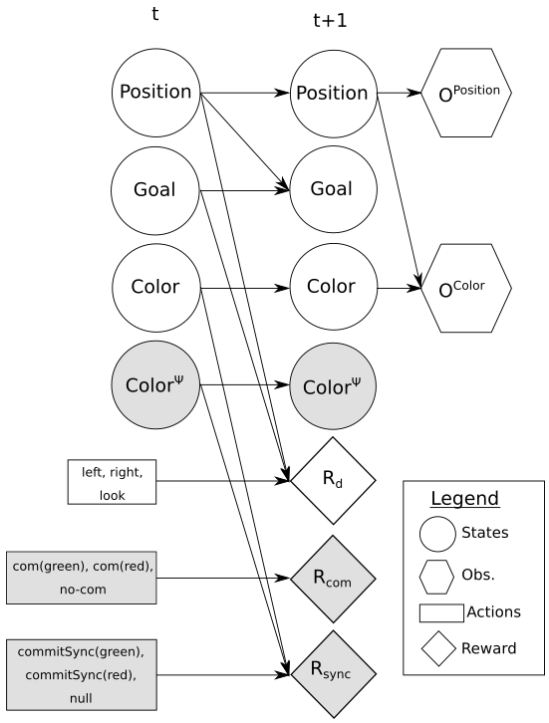


Fig. 2: Dynamic Bayesian Network of the Surveillance Problem model. Greyed nodes are specific to the com-POMDP-IR.

Definition 1 (State Factor Space): The set of state factors $\mathcal{X}$ of a com-POMDP-IR is defined by:

$$\mathcal{X} = \mathcal{X}^{\phi} \cup \mathcal{X}^{\psi/\phi}$$
$$= \{X_1^{\phi}, \dots, X_k^{\phi}, \dots, X_l^{\phi}, \dots, X_n^{\phi}\} \cup \{X_1^{\psi/\phi}, \dots, X_k^{\psi/\phi}\}$$

where $X_1, \dots, X_k$ are the shared Factors of Interest and $X_{k+1}^{\phi}, \dots, X_l^{\phi}$ are the Factors of Interest specific to Agent ϕ. We have $|\mathcal{X}| = 2k + (l-k) + (n-l)$, where n is the number of state factors, $l < n$ the number of FoIs and $. < l$ the number of shared FoIs.

Example 1 (Surveillance Problem - State Factors): In the case of the surveillance problem, the state factors are the following:

$$\mathcal{X} = \{Color^{\phi}, Position^{\phi}, Goal^{\phi}\} \cup \{Color^{\psi/\phi}\}$$

with $Color$ being the color of the alarm (red or $green$), $Position$ being the current position of the robot ($y1$, $y2$ or $y3$) and $Goal$ being the current goal of the robot ($y1$ or $y3$). In this case, only $Color$ is a shared factor of interest.

B. Communication Actions

Agent ϕ should be capable of communicating any possible value for each of the shared FoIs. To do so, we create a $communicate$ action factor, whose possible values are the combination of all the shared FoIs and their respective possible values, plus a $noCom$ action which does not communicate anything. At each time step, the agent must choose a domain-related action and a communication action. Formally, this is described as $\mathcal{A} = A_d \times A_{com}$, where A_d is the set of domain-dependent action factors and A_{com} the communication action, with $DOM(A_{com}) = \bigcup_{i \leq k} DOM(X_i)$. We have $|A_{com}| = 1 + \sum_{i=1}^{k} |X_i|$ We denote $com(X_i, x_i)$ the action of communicating the value x_i for the state factor X_i.

Depending on the domain, it is also possible to create one communication action factor per FoI. In this case, the domain of each communication action factor corresponds to the domain of the FoI, plus the $noCom$ action. The agent would have to choose one domain action and one communication action per FoI at each time step. This would allow the agent to communicate several pieces of information at the same time, with the cost of increasing the number of possible actions and therefore the complexity of the model. We don't consider this option for the remaining of this paper for the sake of simplicity, but all equations and algorithms can be easily adapted to this setup.

As mentioned previously, the state factors in $\mathcal{X}^{\psi}$ represent what Agent ϕ believes the human knows. At this point, it is important to note that this might be an approximation of what the human actually knows. Indeed, in some systems, the human will only get information about the shared FoIs through Agent ϕ, but in others it might get some level of information through another channel, for instance by monitoring him or herself. In this case, it is obvious that $B^{\phi}(\mathcal{X}^{\psi/\phi}) \neq B^{\psi}(\mathcal{X}^{\psi})$. In addition, if the communication is not perfect, the information might not be received by Agent ψ. All these aspects should be captured in the transition function, as presented in Definition 2.

Definition 2 (Transition Function): The transition function of the com-POMDP-IR related to the communication actions is defined as:

$$T(X_{i,t}^{\psi/\phi}, X_{i,t+1}^{\psi/\phi}, com(X_i, x_i)) = \begin{cases} \theta_1 * \theta_2 \text{ if } X_{i,t+1}^{\psi/\phi} = x_i \\ \frac{1-\theta_1 * \theta_2}{|X_i|-1} \text{ otherwise} \end{cases}$$
$$T(X_{i,t}^{\psi/\phi}, X_{i,t+1}^{\psi/\phi}, noCom) = \begin{cases} \theta_2 \text{ if } X_{i,t}^{\psi/\phi} = X_{i,t+1}^{\psi/\phi} \\ \frac{1-\theta_2}{|X_i|-1} \text{ otherwise} \end{cases} \quad (1)$$

where θ_1 represents the probability of the communication to be transmitted successfully and θ_2 represents the probability that the human's beliefs remain the same in the absence of communication.

If the communication is perfect and the human only receives information about from Agent ϕ, then $\theta_1 = \theta_2 = 1$. If the communication is imperfect, $\theta_1 < 1$. If the human ψ receives information from other sources than Agent ϕ, $\theta_2 < 1$. Capturing the different aspects of the system within θ_1 and θ_2 depends on the domain and is left as the responsibility of the system designer.

C. Rewarding Relevant Communication

In the com-POMDP-IR, Agent ϕ should be rewarded for communicating relevant information to the human ψ, which means keeping a belief over $X_i^{\psi/\phi}$ close to the belief over X_i^{ϕ} for all $i \leq k$. To do so, we introduce $commitSync$ actions, similar to the $commit$ actions of the POMDP-IR [5]. There is one $commitSync$ action for each factor $X_i, i \leq k$

and one $commit$ action for each factor $X_i, k < i \leq l$. We must then extend the set of actions described in Section IV-B to obtain the complete action space of the com-POMDP-IR, as presented in Definition 3.

Definition 3 (Action Space): The set of action factors of the com-POMDP-IR is defined as follows:

$$\mathcal{A} = A_d \times A_{com} \times A_1 \times \cdots \times A_k \times \cdots \times A_l \quad (2)$$

with A_d being the set of domain actions, A_{com} the set of communication actions, $A_1, \ldots, A_k$ the set of $commitSync$ actions and $A_{k+1}, \ldots, A_l$ the set of Information Reward actions.

We have for each $X_i, i \leq k$

$$A_i = \{commitSync(x_j), \forall x_j \in DOM(X_i)\} \cup \{null\}$$

and $|A_i| = 1 + \sum_{i=1}^{k} |X_i|$

At each time step, the agent will choose simultaneously a domain action, a communication action, a $commitSync$ action for each shared FoI and a $commit$ action for each non-shared FoI. The $commitSync$ actions, as for the $commit$ actions, have no effect on the agent's belief state but are used for rewarding the agent when it communicates. As for the $commit$ actions, it is used to avoid belief-dependent rewards. Choosing a $commitSync$ action means that the agent commits to a given value for X_i and to a synchronized belief over X_i^{ϕ} and $X_i^{\psi/\phi}$.

Example 2 (Surveillance Problem - Action Space): In the surveillance problem, we have:

$$A_d = \{left, right, look\}$$

$$A_{com} = \{com(color, red), com(color, green), noCom\}$$

$$A_{color} = \{commitSync(red), commitSync(green), null\}$$

Using the com-POMDP-IR action space, the agent receives a positive reward when it commits to a correct synchronized belief, as presented in Definition 4.

Definition 4: The com-POMDP-IR reward function is defined as follows:

$$R(\mathcal{X}, \mathcal{A}) = R_d(X, A_d) + \sum_{i=1}^{k} R_{sync}(X_i, A_i) + \sum_{i=k+1}^{l} R_{commit}(X_i, A_i) \quad (3)$$

where R_d is the domain reward, R_{sync} the reward associated to the $commitSync$ actions, and R_{commit} the Information Reward as defined in [5].

For each $X_i, i \leq k$, R_{sync} is defined as:

$$\begin{aligned} &R_{sync}(X_i, null) = 0 \\ &R_{sync}(X_i, commitSync(x_j)) = \\ &\begin{cases} r_{sync} \text{ if } X_i^{\phi} = x_j \wedge X_i^{\psi/\phi} = x_j \\ -r_{notsync} \text{ otherwise} \end{cases} \end{aligned} \quad (4)$$

with $r_{sync}, r_{notSync} > 0$.

The values of r_{sync} and $r_{notSync}$ have to be chosen carefully to ensure that the agent only commits when its beliefs over X_i^{ϕ} and X_i^{ψ} are certain enough. It is possible to choose different values of r_{sync} and $r_{notsync}$ for different FoIs and even different values of a single FoI. For instance in the surveillance problem, being certain that the alarm is red might be considered more important than being certain it is green.

D. Choosing the parameters

The com-POMDP-IR reward function depends on 2 additional parameters compared to the POMDP-IR: r_{sync} and $r_{notSync}$. From Equation 4, we can compute the expected reward for $commitSync$ actions as follows:

$$\begin{aligned} &R(b^{\phi}, X_i, commitSync(x_j)) \\ &\quad = b^{\phi}(X_i^{\phi} = x_j) \cdot b^{\phi}(X_i^{\psi/\phi} = x_j) \cdot r_{sync} \\ &\quad - (1 - b^{\phi}(X_i^{\phi} = x_j) \cdot b^{\phi}(X_i^{\psi/\phi} = x_j)) \cdot r_{notsync} \end{aligned} \quad (5)$$

We wish the agent to select the $commitSync$ action when it is certain "enough". This translates mathematically to

$$\begin{aligned} &R(b^{\phi}, X_i, commitSync(x_j)) > 0 \\ &\text{iff } b^{\phi}(X_i^{\phi} = x_j) > \beta \text{ and } b^{\phi}(X_i^{\psi/\phi} = x_j) > \beta \end{aligned} \quad (6)$$

β being chosen by the system designer. Using this, we can derive the relation between r_{sync} and $r_{notsync}$:

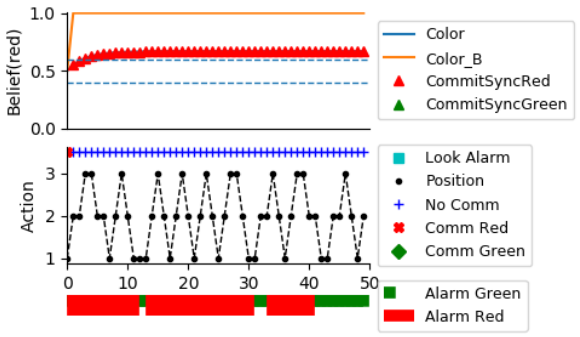
$$\beta^2 r_{sync} - (1 - \beta^2) r_{notsync} = 0 \quad (7)$$

$$\Leftrightarrow r_{sync} = \frac{1 - \beta^2}{\beta^2} r_{notsync} \quad (8)$$

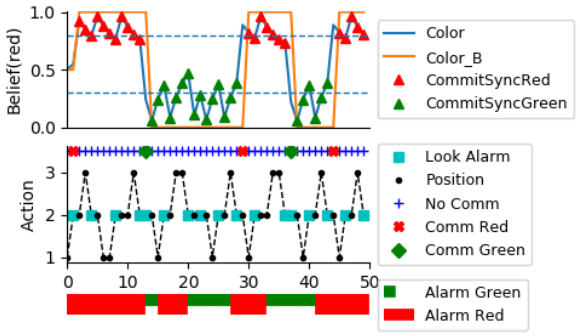
V. Experiments

We evaluated our approach in the case of the Surveillance problem described in Section I. Agent ϕ is patrolling the corridor. When performing a movement action, it has a probability of 0.8 to end up in the intended space. When it reaches one goal at one end of the corridor, the goal switches to the other possible one. The alarm at the center of the corridor starts green and will turn red with a probability of 0.8. Once red, it will turn back to green with a probability of 0.1. The reward for reaching a goal is 15. Unless said otherwise, the cost for a communication is 1. The policy has been calculated with the Symbolic Perseus Solver, modified for Information-Reward actions[5], with a random sampling of 500 belief points and over 100 iterations. Each experiments has been run over 500 episodes. During the experiments, we used $r_{sync} = 10$ and calculated $r_{notsync}$ for each β according to Equation 7.

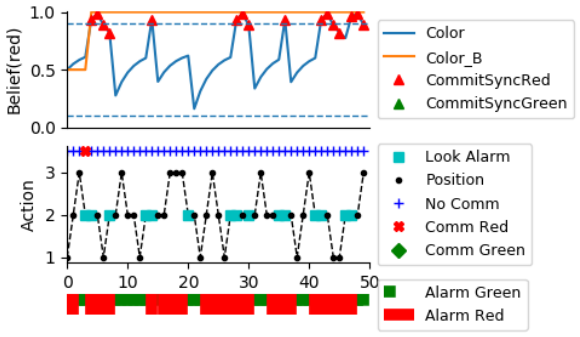
We first evaluated the behavior of the com-POMDP-IR agent in the case $\theta_1 = \theta_2 = 1$. (Section V-A). This allows us to validate the model by ensuring that the agent is exploring and planning its communication appropriately and analyze the influence of the threshold β over the behavior of the agent. We then studied the case where some communication can be lost ($\theta_1 < 1$) (Section V-B) and finally the case where the human might receive information from other sources than Agent ϕ ($\theta_2 < 1$) (Section V-C)



(a) $\beta = 0.6$



(b) $\beta_{red} = 0.8, \beta_{green} = 0.7$



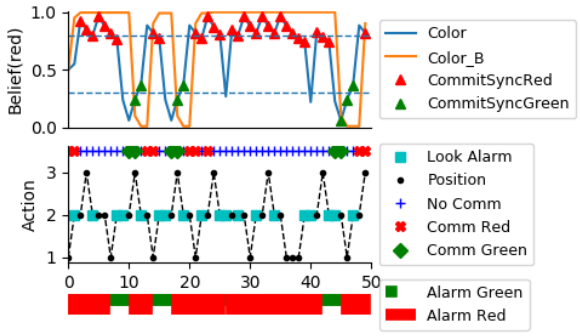
(c) $\beta = 0.9$

Fig. 3: Surveillance problem results with $\theta_1 = \theta_2 = 1$. Each figure shows the belief evolution over $Color^{\phi}$ (called Color) and $Color^{\psi/\phi}$ (called $Color_B$) (top row), the communication action and the robot position (middle row), and the actual color of the alarm (bottom row). The dotted lines on the top row indicates the values for β_{red} and $\beta_{green} = 1 - \beta_{red}$.

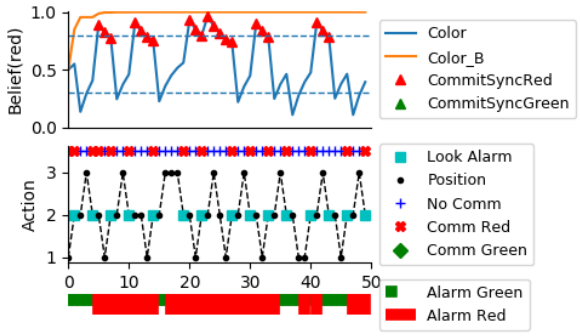
A. Perfect Communication

The threshold β for which the com-POMDP-IR agent should choose to commit depends on the problem at hand and must be carefully chosen by the designer. Figure 3 shows some of the possible thresholds and their effect on the agent's behavior. We see that a too low β (Fig. 3a) causes bad communication patterns. Indeed, in the Surveillance problem, the alarm is more likely to turn red and stay red than green. Therefore, the agent can commit to a synchronized belief state without ever looking at the alarm and only communicating red once. A too high β (Fig. 3c) also causes bad communication patterns as the agent is not capable of reaching such a threshold for one of the values. As the model makes it possible to tailor β for each of the possible values of the factor of interest, we can tune the system for optimal communication (Fig. 3b).

B. Non-perfect communication



(a) $\theta_1 = 0.9, \beta_{red} = 0.8, \beta_{green} = 0.7$



(b) $\theta_1 = 0.7, \beta_{red} = 0.8, \beta_{green} = 0.7,$

Fig. 4: Non-perfect communication

Figure 4 shows the behavior of the com-POMDP-IR agent when 10% and 30% of the messages are lost. The system is relatively robust to lost messages, provided that the β are carefully chosen (Fig. 4a). As expected, when the risk of lost messages is too high, the agent does not communicate anymore about the less probable value of the alarm (green) as it cannot reach the expected belief threshold, even if it is low (Fig. 4b). However, the agent is still capable to communicate about the more probable value.

The model was also evaluated when communication is not perfect. To do so, we modeled the human operator as a purely reactive agent which performs an action *raise-alarm* when it receives a communication that the alarm is red. The system receives a positive reward when the alarm is raised appropriately and a negative reward otherwise. This experiment allows us to ensure that the communication from Agent ϕ is enough to ensure good performance of the system, without any double-check from the human. In a more realistic scenario, the human would use another way to confirm the value communicated before acting. We ran this experiment

for different values of θ_1. To ensure that a system with perfect communication ($\theta_1 = 1$) is performing optimally, we also computed the reward gathered by a centralized POMDP-IR, controlling the agent performing the patrolling and raising the alarm, and solved optimally. Since the values of θ_1 and β are linked, the values of β for this experiment have been chosen in order to ensure the best result for each value of θ_1 and shown in Table I.

	$\theta = 1$	$\theta = 0.99$	$\theta = 0.9$	$\theta = 0.8$	$\theta = 0.7$
β_{red}	0.8	0.8	0.8	0.8	0.8
β_{green}	0.8	0.8	0.7	0.7	0.6

TABLE I: Values of β_{red} and β_{green} for each θ_1

Figure 5 shows the box plots of the reward obtained at the end of the simulation for each configuration.

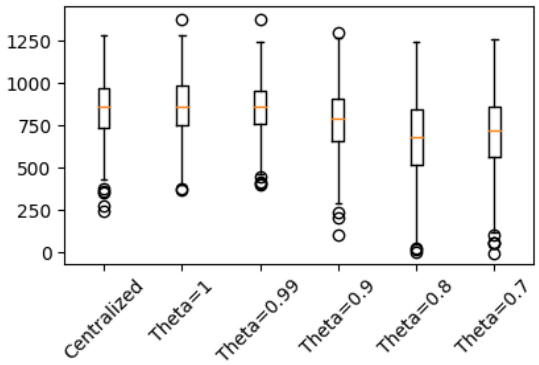


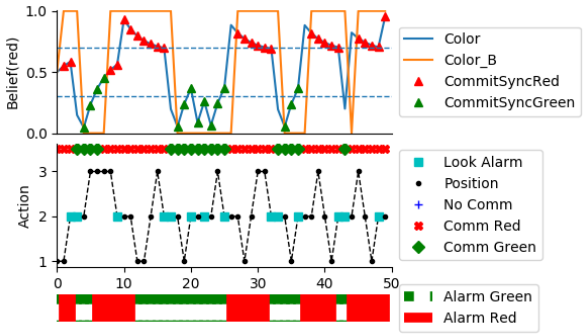
Fig. 5: Accumulated reward for different values of θ_1

The com-POMDP-IR agent performs as well as the centralized POMDP-IR agent when $\theta_1 = \theta_2 = 1$. We also note that a loss of 1% of the communications ($\theta_1 = 0.99$) does not affect significantly the performance of the system and that a loss of 10% of the message still gives in average good results, even though more variability is observed. For these configurations where the communication is highly unreliable, the need for a confirmation of the value by the human operator is obvious.

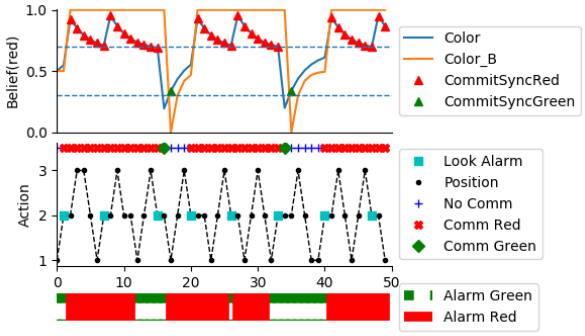
C. Varying recipient's beliefs

The parameter θ_2 allows us to model how the beliefs of Agent ψ evolve without communications from Agent ϕ. In this section, we consider a perfect communication ($\theta_1 = 1$) and various values for θ_2. Figure 6 shows the results of the surveillance problem with varying values of θ_2.

Figure 6a show that when θ_2 is low, the agent tends to communicate more to maintain a high level of beliefs over $\mathcal{X}^{\psi/\phi}$. However, this can be mitigated by increasing the communication cost (Figure 6a). One could also want to impose a certain number of steps between two successive communications by introducing a bookkeeping variable in the model for instance.



(a) $\theta_2 = 0.7$ and $cost = 1$



(b) $\theta_2 = 0.7$, $\beta = 0.7$ and $cost = 3$

Fig. 6: Surveillance problem results with $\theta_1 = 1$, $\beta = 0.7$ and various θ_2.

VI. Conclusion

We presented the Communicating POMDP-IR, a model that integrates information management in an environment with active and humans (as passive agents). In particular, we focus on scenarios in which active agents must perform information-oriented tasks while, simultaneously, they must decide when information is needed to the human operator. The com-POMDP-IR extends the POMDP-IR with nested beliefs and communication actions, enabling the agent to plan its communication depending on what it believes the human believes. We tested our approach over a surveillance problem and showed that our system is robust to imperfect communication and human belief's evolution, provided that its parameters are tuned properly.

As future work, we intend to deal with scalability issues. One aspect of com-POMDP-IR is that the number of actions grows exponentially with the number of agents (including humans). Reducing the communication action to a choice between to communicate or not and decide on the fly which information to send and to whom might ease the complexity of the model. In addition, the agent currently receives no feedback about whether the human received the communication or not. In cases where the communication is imperfect, such feedback could allow the agent to avoid re-sending the information when it has been received.

REFERENCES

[1] S. Witwicki, J. C. Castillo, J. Messias, J. Capitan, F. S. Melo, P. U. Lima, and M. Veloso, "Autonomous surveillance robots: A decision-making framework for networked muiltiagent systems," *IEEE Robotics & Automation Magazine*, vol. 24, no. 3, pp. 52–64, 2017.

[2] J. B. Lyons, "Being transparent about transparency: A model for human-robot interaction," in *2013 AAAI Spring Symposium Series*, 2013.

[3] P. Langley, "Explainable agency in human-robot interaction," in *AAAI Fall Symposium Series*, 2016.

[4] J. Renoux, "Active situation reporting: Definition and analysis," in *Proceedings of the European Conference on Multi-Agent Systems* (F. Belardinelli and E. Argente, eds.), vol. 10767 of *Lecture Notes in Computer Science (LNCS)*, pp. 70–78, Springer International Publishing, 2017.

[5] M. T. J. Spaan, T. S. Veiga, and P. U. Lima, "Decision-theoretic planning under uncertainty with information rewards for active cooperative perception," *Autonomous Agents and Multi-Agent Systems*, vol. 29, pp. 1157–1185, nov 2015.

[6] M. Araya, O. Buffet, V. Thomas, and F. Charpillet, "A POMDP extension with belief-dependent rewards," in *Advances in Neural Information Processing Systems 23* (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds.), (Vancouver, Canada), pp. 64–72, Curran Associates, Inc., 2010.

[7] J. Renoux, A. I. Mouaddib, and S. L. Gloannec, "A decision-theoretic planning approach for multi-robot exploration and event search," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5287–5293, Sept 2015.

[8] M. Lauri, J. Pajarinen, and J. Peters, "Information gathering in decentralized pomdps by policy graph improvement," in *Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

[9] C. V. Goldman and S. Zilberstein, "Optimizing information exchange in cooperative multi-agent systems," in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pp. 137–144, ACM, 2003.

[10] M. T. J. Spaan, "Cooperative active perception using POMDPs," in *AAAI 2008 workshop on advancements in POMDP solvers*, 2008.

[11] F. S. Melo, M. T. J. Spaan, and S. J. Witwicki, "Querypomdp: Pomdp-based communication in multiagent systems," in *Multi-Agent Systems. EUMAS 2011. Lecture Notes in Computer Science* (M. Cossentino, M. Kaisers, K. Tuyls, and G. Weiss, eds.), vol. 7541, (Berlin, Heidelberg), pp. 189–204, Springer Berlin Heidelberg, 2011.

[12] A. Wang, R. Chitnis, M. Li, L. P. Kaelbling, and T. Lozano-Pérez, "A unifying framework for social motivation in human-robot interaction," in *AAAI Workshop on Plan, Activity, and Intent Recognition (PAIR)*, 2020.

[13] D. Strouse, M. Kleiman-Weiner, J. Tenenbaum, M. Botvinick, and D. J. Schwab, "Learning to share and hide intentions using information regularization," in *Advances in Neural Information Processing Systems*, pp. 10270–10281, 2018.

[14] P. J. Gmytrasiewicz and P. Doshi, "A framework for sequential planning in multi-agent settings," *Journal of Artificial Intelligence Research*, vol. 24, pp. 49–79, 2005.

[15] P. Gmytrasiewicz and S. Adhikari, "Optimal sequential planning for communicative actions: A bayesian approach," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1985–1987, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[16] T. Taha, J. V. Miró, and G. Dissanayake, "A pomdp framework for modelling human interaction with assistive robots," in *2011 IEEE International Conference on Robotics and Automation*, pp. 544–549, IEEE, 2011.

[17] J. A. Garcia and P. U. Lima, "Improving human behavior using pomdps with gestures and speech recognition," in *Cognitive Architectures*, pp. 145–163, Springer, 2019.