We stress at various points that the \*field trial of a planner agent\* is our focus (not the algorithm itself nor the deployment alone).

Instead Reviewer 2 focuses mainly on the planning aspect which, due to lack of space, had to be curtailed to focus on the results of the field trials. We will purchase two additional pages to include details requested (e.g., team value, deadline computation due to the radioactive cloud, lookahead etc…). Our algorithm is anytime (hence scalable). The planning time was set to 5 seconds and used live. We will add extra results and clarifications. In contrast to Muslinger’s work, our responders are heterogeneous, and have to avoid the radiation cloud (which is uncertain) and the planner is deployed live.

While we see the point of Reviewer 3 in terms of multiple foci, we strongly believe a Human-Agent Interaction (HAI) paper cannot have a single focus. There needs to be an AI as well as an HCI contribution for it to be a strong HAI paper (as per AAAI’s keywords). It is a hard balance to strike and we believe this is the best that could be achieved in 6 pages. The results we provide in terms of HCI are completely new to the AI community (and HCI as well). We will add more results to evidence the AI contribution (and link to AtomicOrchid) as requested by reviewers 2&3.

Only flat MCTS can solve our problem given its large state and joint action space. We found it performed no better than the random policy (and greedy and myopic approaches) because of the complexity of the problem. Instead, our algorithm encodes domain knowledge and structures. This will be clarified.

Results show that the commander corrected the planning agent suggestions only 6 times (out of 16 task allocation directives) while it reinforced or complemented the agent’s decision for the rest. Hence, we can generally associate the better team performance to the human+planner agent’s guidance.

The criticism of Reviewer 2 about the number of runs is typical of an AI-mindset requiring large numbers of artificial simulations. Instead, we explicitly highlight the evaluation aims to identify design issues \*in the real world\* rather than validate hypotheses with lots of ‘artificial’ simulations. This is how we believe we can truly elicit design principles for HAI.

We will clarify the different ‘simulations’ (Reviewer1). We agree the ‘expertise’ of the HQ is a key factor – our next field trial will actually involve professional field responders (Reviewer 2). Also, as we showed, ‘simple’ agents perform worse and would probably impose a higher cognitive load on the HQ to correct for their obvious mistakes. The key link between the algorithm and the platform is the ability to deal with radiation uncertainty and planning complexity at the same time, to let the commander focus on user-related challenges (indecision/ability to run)– this will be made clearer. We will also add a summary analysis of our semi-structured interviews to reinforce the qualitative evaluation.