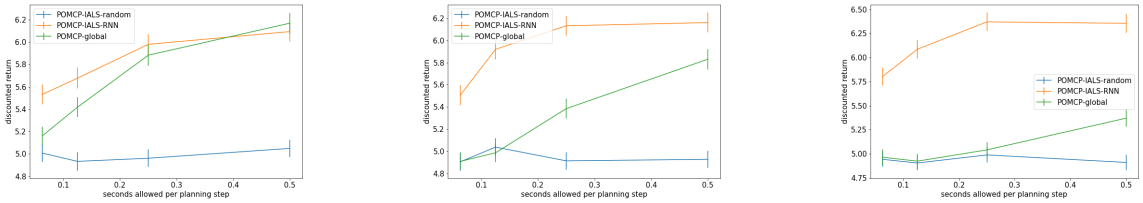

Figure 1: Time-constrained evaluation in the grab a chair domain with 33 (left), 65 (middle) and 129 (right) agents.

We thank all reviewers for the encouraging and constructive feedback and the relevant pointers.

Reproducibility (@**R2**, **R3**): Our codebase is available at `https://www.dropbox.com/sh/sg2qyfpdwfmkfqd/AACq5R5BQS-jpSS6OMFS3FrLa?dl=0` with source code, pretrained models and instructions to reproduce every figure in the submission. We will make the codebase available on Github after the review phase.

Comparison to exogenous variables (@**R4**): Thanks for the pointers to the related work which indeed are similar in spirit. We will add the discussion of these works to our related work section. Nevertheless, we would like to point out that exogenous variables are fundamentally different from the non-local variables: by definition, exogenous variables refer to those variables that are **beyond the control of the agent** in the sense that **the values of which are not affected, directly or indirectly, by the agent's actions** (Boutilier et al., 1999; Zhang et al., 2017; Chitnis et al., 2019). However, the non-local variables in the influence-based abstraction (IBA) (Oliehoek et al., 2012) refer to those variables that **do not directly affect the agent's observation and reward**. Therefore, the exogenous variables and non-local variables are in general two different sets of variables that can be exploited to reduce the state space size. For instance, in the traffic problem of Figure 4a, there are *no* exogenous variables as our action can directly or indirectly effect the transitions at other intersections (by taking or sending vehicles from/to them). In other words, **our approach allows us to reduce the state space of this problem *beyond* the exogenous variables**.

Clarifications on data collection and influence learning (@**R3**, **R4**): In Section 3.1, we formalize influence prediction as a classification problem where inputs are trajectories of local states and actions, and outputs are influence source variables. In this work, the training trajectories are sampled **from the global simulator** (cf. line 162) with an exploratory policy and therefore no human annotation is required.

In all our experiments, we collected 1000 episodes (line 199) and trained the influence predictor to convergence, which did not take more than an hour. While such offline computation could be costly (as noted by **R4**), we believe that in many settings the real-time constraints during the online phase are the bottleneck for application and these can be significantly brought down with our approach.

Clarifications on experimental design and results (@**R1**, **R2**, **R3**, **R4**): As **R1** and **R3** noted, we only investigate the performance of our approach in those domains for which it was designed: problems with clear factorized structure. For these problems, we think our benchmarking domains are general enough so that the results can be easily translated. We certainly would not want to claim anything about a broader class of problems. We do stress, however, that there is a large literature on factored (multiagent) decision making that has been motivated by many different applications, such as smart grids, warehouse commissioning, etc, and that Bayesian networks (the tool we use to represent the factored structure) are one of the most applied AI techniques in history (even if perhaps overtaken by NNs by now).

When sufficient time is available for online planning, the benefits in the traffic domain are indeed moderate (**R1**). However, in many real-world decision problems, the online decision time is limited, and precisely for those cases our approach significantly outperforms the baselines (e.g., when <10s per action allowed in Fig. 4d). To further test this, Figure 1 shows real-time planning results in the grab a chair domain, as suggested by **R3**. These further demonstrate that the advantage of our approach in time-constrained settings, as the global model of the problem gets more complex.

Further clarifications: **R3**: learning models that matter with value, goal and task awareness is indeed relevant to the idea of capturing the impact from the rest of the environment by predicting only the influence source variables. However, these approaches do not perform explicit abstraction as ours does and therefore are complimentary. **R2**: we chose to use POMCP in unmodified form in our experiments to make it easier to interpret the benefits of our approach, but indeed our approach can be combined generally with other planners and more advanced particle filters, and therefore all known improvements from SIR can be applied. **R4**: it is not possible to define a baseline local simulator with "no influence" because by definition the influence destination variables have dependency on the source variables. We agree that attention mechanisms are a promising avenue for learning influence representations (not the focus of this paper) and thus are a promising direction of future work. As for the grab a chair domain (**R4**), we performed online planning for a finite horizon of 10 (line 198-199), which explains the choice for undiscounted reward $\gamma = 1$. At every time step, each agent can only decide to grab the chair on their left or right and therefore it is not possible for an agent to obtain more than one chair at a time step (line 33, 388-389).