We would like to thank all reviewers for their time and helpful comments.

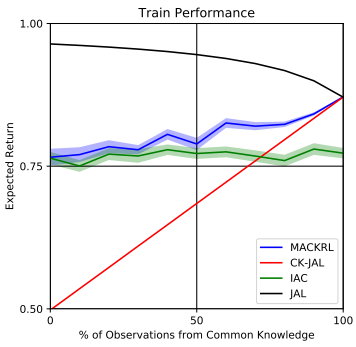
Revised matrix game (@R1,2) We are sorry that the matrix game caused confusion and will use this feedback to make the exposition more intuitive. Briefly: all policies use tabular representations, for reward matrices see Fig 3, for policy inputs and game details see Fig 7 in Supplement. We will add details on the training procedure.

JAL: Indeed, a confusing name. Since we focus on decentralized execution, we used a JAL that only conditions on CK, which we will rename 'CK-JAL'. We have added a true JAL which receives the joint observation. As expected this upper-bounds performance.

IAC: Independent learners converge to a local optimum (e.g. rows 2 and 3 for agent A) when the other agent is not reliably coordinating for the global maximum (as will be the case during training). So even when the state is known to both agents, the coordination problem leads to suboptimal IAC performance. All $P(CK)$ are trained separately and there is no meta/transfer learning. MACKRL gets around the IAC limitation by allowing agents to explore and learn in the joint action space using CK; it does not suffice for either agent to know independently which game is played. In contrast to IAC, MACKRL and JAL will specialize on actions 1 and 5 (as suggested by R1).

Presentation of results: R1 makes a great suggestion, we have now changed the game such that the unconditional probability (i.e. $p(\texttt{independent knowledge})$) of an agent observing the game is fixed at 75%. On the x-axis, we change which % of these observations is due to CK and which is due to independent observations. Note that the JAL benefits from having a low % of CK observations, since it's less likely that neither of the agents observes the state. @R2 - we are investigating the wagging which likely stems from sub-optimal tuning of the gradient-based optimiser.

Practical applications and scaling (@R1,3) We use deterministic policies in decentralised execution so there is no need for any further coordination. If we wish to use a stochastic policy, it is easy to share a random seed either from the centralised training phase, or using, e.g., synchronised clocks or a highly limited communication channel. Deploying agents at different times is a different setting (ad hoc teamwork) that we don't address in this work.



While the formal definition of CK is strict, in practice, humans easily relax these requirements to perform commonsense reasoning. If Alice tells Bob, "Meet me at King's Cross" there's no point for Alice to actually go there without *assuming* the CK that Bob heard and understood what Alice said, and the CK concerning social conventions. We find that MACKRL, perhaps surprisingly, is quite robust to a naive relaxation to a type of probabilistic CK. Note that in a normal sensor, noisy observations will be closely correlated with the true observation. In our game, by contrast, 10% bit flip noise is quite extreme, indicating a completely different matrix. We will elaborate in the paper on softer forms of CK (see e.g. Halpern & Moses 2000) and our use of correlated sampling. We are also excited about future work which would extend MACKRL even more robustly to these settings.

Any MARL approach faces challenges in scaling with the number of agents. What matters is how they address it: MACKRL does so with a sampling approach that we show is quite effective (Fig 5), more so than independent learning (which scales well, but at the cost of highly limited capacity for coordination).

Baselines and ablations (@R3) HRL methods (including FuN) target temporal abstraction, and are not relevant or comparable to MACKRL. We will clarify that 'policy hierarchy' in no way refers to HRL. In Appendix C, we perform one possible introspective study, showing the pair controllers prefer not to delegate when coordination is difficult and CK is large (i.e. when many enemies are present in the CK).

Further clarifications We will update the paper to reflect the following and address all other minor issues raised.

(R1-3) Nothing disallows C from using its knowledge about A and B's CK, if the pair controller delegates to C. However, it is not commonly known by A and B that C knows about their CK.
(R1-4) An "independent override" heuristic is an interesting avenue for future research which we believe lies outside the scope of this paper. We predict the current form of MACKRL would simply learn to delegate to the individual agents in order to mitigate the risk of the tiger.
(R1-5) If agents incentives are not fully aligned, then establishing conventions is more difficult because agents may have incentives to violate the conventions. Nonetheless, coordination devices play an important role in game theory and CK could help agents to coordinate in these settings. However, this is well beyond the scope of this paper.
(R1-8) Should read "a unique middle ground" rather than "uniquely..."
(R1-9) In our setting the action space can depend on the unit-type of the agent and hence the agent ID, while the observation space is common across all agents.
(R1-11) $\tau_t^{\mathcal{G}}$ is a set and line 101 should use $\supseteq$ rather than $\geq$. We will define $\tau_t^{\mathcal{G}}$ more precisely.
(R1-12) Pairwise selection based on proximity is an interesting avenue, but note that the selection itself needs to be based on CK (which doesn't in general include global proximity information).
(R2) S3.1 should read that 'pair controllers' have shared parameters.