MIT Libraries | DSpace@MIT

MIT Open Access Articles

Formal contracts mitigate social dilemmas in multi-agent reinforcement learning

Massachusetts Institute of Technology

Check for
updates

Formal contracts mitigate social dilemmas in multi-agent reinforcement learning

Andreas Haupt[1] · Phillip Christoffersen[1] · Mehul Damani[1] · Dylan Hadfield-Menell[1]

Abstract

Multi-agent Reinforcement Learning (MARL) is a powerful tool for training autonomous agents acting independently in a common environment. However, it can lead to sub-optimal behavior when individual incentives and group incentives diverge. Humans are remarkably capable at solving these social dilemmas. It is an open problem in MARL to replicate such cooperative behaviors in selfish agents. In this work, we draw upon the idea of formal contracting from economics to overcome diverging incentives between agents in MARL. We propose an augmentation to a Markov game where agents voluntarily agree to binding transfers of reward, under pre-specified conditions. Our contributions are theoretical and empirical. First, we show that this augmentation makes all subgame-perfect equilibria of all Fully Observable Markov Games exhibit socially optimal behavior, given a sufficiently rich space of contracts. Next, we show that for general contract spaces, and even under partial observability, richer contract spaces lead to higher welfare. Hence, contract space design solves an exploration-exploitation tradeoff, sidestepping incentive issues. We complement our theoretical analysis with experiments. Issues of exploration in the contracting augmentation are mitigated using a training methodology inspired by multi-objective reinforcement learning: Multi-Objective Contract Augmentation Learning. We test our methodology in static, single-move games, as well as dynamic domains that simulate traffic, pollution management, and common pool resource management.

Keywords Social dilemma · Decentralized training · Formal contracts

---

✉ Andreas Haupt
  haupt@mit.edu

✉ Phillip Christoffersen
  philljkc@mit.edu

  Mehul Damani
  mehul42@mit.edu

  Dylan Hadfield-Menell
  dhm@csail.mit.edu

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge 02139, MA, USA


 Springer

1 Introduction

We study the problem of how to get selfishly motivated agents to act pro-socially through the lens of Multi-Agent Reinforcement Learning (MARL). To solve this problem, we wish to give agents the option to propose and accept contracts with the aim of encouraging prosocial behavior.

As an example, consider the Cleanup domain, depicted in Fig. 1. Agents get reward for picking apples that only grow if a nearby river is unpolluted. In a pro-social solution to Cleanup, agents need to work together: one cleans while the other eats apples. However, self-interested agents cannot sustain this solution. Cleaning has no direct benefit, so selfish agents focus exclusively on picking apples. A social dilemma ensues.

This article studies contracts as voluntary commitments by all agents to zero-sum modifications of the environment reward, so-called *social contracts*, as a detail-free method of mitigating social dilemmas. More specifically, contracts transfer rewards between agents depending on *contractible observations*. Contracts are proposed by agents, and can be vetoed by any single agent—enacting the contract is voluntary. Even upon acceptance, agents may still choose any action, only their rewards for these actions are changed by the contract. We show, both game-theoretically and empirically, that providing agents with the option to contract allows for optimal system performance, in the sense of maximizing sum of agents' rewards, regardless of the initial reward signal.

Consider again the problem of teaching agents to harvest in a polluted orchard where agents are only rewarded for picking apples. In a contract, an agent could propose that an agent receives a small reward bonus r_{clean} for each polluted river square that is cleaned. All other agents are "charged" a reward penalty of $-\frac{r_{\text{clean}}}{n-1}$ to "fund" this bonus. Harvesting agents prefer this to an outcome with no apples as long as the bonus from picking apples offsets the cost. Non-harvesting agents prefer this contract as long as the bonus is small enough to be offset by the benefits of having apples grow.

(a) Cleanup Domain   (b) Illustration of Contracts
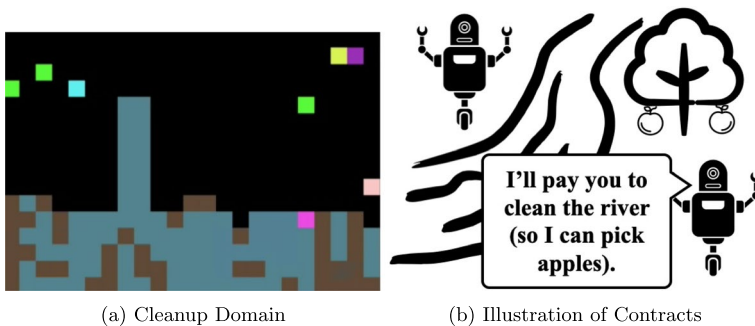
Fig. 1 We evaluate our method in the Cleanup domain [14]. **a** A representation of the environment. The different agents correspond to the pink, yellow, and purple tiles. Agents are rewarded for picking apples (green), but apples will only grow if the river (blue) is clean of pollution (brown). Agents can clean up pollution, but aren't directly rewarded for cleaning. This creates a *social dilemma* where no agents clean because they don't expect to benefit from cleaning directly. **b** An illustration of the solution that our contracting augmentation facilitates. In the Cleanup domain, one agent commits to "pay" the other to clean the river. As a result, the agents are able to coordinate on policies that maximize the total reward across both agents

To understand the workings of contracting quantitatively, consider the classic Prisoner's Dilemma [39]. The tables in Table 1 show the payoffs for the unmodified game (Table 1a) and the modified incentives (Table 1b) under the following contract:

Any agent who defects is fined 1.5 *units of reward by the other agent.*

If both defect, both pay, and the payments cancel. In this modified game, cooperation is dominant, and hence (*C*, *C*) is the only Nash equilibrium.

Would the agents agree to this contract if proposed? If it is rejected, they subsequently play the game in Table 1a which has a unique Nash equilibrium of (*D*, *D*), yielding a reward of −2 for both agents. However, if it is accepted, they subsequently play the game in Table 1b which has a Nash equilibrium of (*C*, *C*), yielding a reward of −1 for both agents. Thus, agents want to accept the contract and subsequently play the socially optimal outcome. Hence, merely the option to commit to a state-action dependent reward transfer, or enter into a formal contract, mitigates a social dilemma, even among selfish agents.

In a real-world multi-agent system, contractible observations may be thought of as *indices* of system behavior, e.g., a tally of desirable actions taken by person. Accepted contracts are held in a ledger of contracts, and impact rewards of agents directly.

In this article, we advance the understanding of contracting for multi-agent systems with the following four contributions:

1. We formalize *Formal Contracting* as a generic augmentation of Markov games (i.e., non-cooperative MARL);
2. We prove that this augmentation makes socially optimal behavior a *subgame-perfect equilibrium* (SPE), and that every SPE of the augmented game is socially optimal, if sufficiently complex contracts are allowed and the game is sufficiently observable in a sense we specify;
3. We prove in more general settings that more expressive contracts allow for higher system performance in terms of the sum of rewards;
4. We implement the contracting augmentation using state-of-the-art deep reinforcement learning. Contracting improves system performance in several complex game environments. We propose an algorithm to overcome exploration challenges, which performs

Table 1 (a) Prisoner's Dilemma (b) Prisoner's Dilemma after signing a contract in which a defector transfers 1.5 reward to a cooperator

(a) Without Contract		
	C	*D*
C	−1,−1	−3,0
D	0,−3	−2,−2
(b) With Contract		
	C	*D*
C	−1,−1	−1.5,−1.5
D	−1.5,−1.5	−2,−2

With this contract in force, cooperating becomes a dominant action for both players

close to, or better than, a joint controller after a fixed number of time periods in complex dynamic domains such as Cleanup and Harvest, see Hughes et al. [14].

The rest of this article is structured as follows. We provide preliminaries and define our augmentation in Sect. 2. In Sect. 3, we show that formal contracting mitigates social dilemmas as long as contracts can detect deviations from a socially optimal policy profile. In Sect. 4 we formalize the notion of contractible features, and prove that the system performance improves with richer contracting features. We describe our evaluation methodology and introduce a learning algorithm exploiting the notion of subgame perfection, multi-objective contract augmentation learning, in Sect. 5. Experimental results are presented in Sect. 6. We review related work in Sect. 7. In Sect. 8, we discuss the real-world application and enforcement of contracts, fairness concerns, and avenues for future work. Appendices contain proofs, additional statements and experiments, a formal definition of a more general contracting augmentation, and hyper-parameter settings for our experiments.

2 Formal contracting

2.1 Background

2.1.1 Fully observable markov games

We define an N-agent Fully Observable Markov Game (FOMG) as a 6-tuple, $M = \langle S, s_0, \mathbf{A}, T, \mathbf{R}, \gamma \rangle$, where

- S is a state space;
- $s_0 \in S$ is the initial state;
- $\mathbf{A} = A_1 \times A_2 \times \cdots \times A_n$ is the space of action profiles $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ for N agents;
- $T : S \times \mathbf{A} \to \Delta(S)$ is a transition function;
- $\mathbf{R} : S \times \mathbf{A} \to [-R_{\max}, R_{\max}]^n$ is a (bounded) reward function mapping state-action profiles to reward vectors for the n agents; and
- $\gamma \in [0, 1)$ is a discount factor.

We will use bold symbols to denote sets of tuples. Agents choose policies $\pi_i : S \to \Delta(A_i)$, $i = 1, 2, \ldots, N$. We write $\boldsymbol{\pi} := (\pi_1, \pi_2, \ldots, \pi_N)$ to denote a *policy profile*. For a policy profile $\boldsymbol{\pi}$, we denote by

$$V_i^{\boldsymbol{\pi}}(s_0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, \mathbf{a}_t)\right] \tag{1}$$

the *value* to agent $i \in [N] := \{1, 2, \ldots, N\}$. The expectation in (1) is on the process $s_t \sim T(s_{t-1}, \mathbf{a}_{t-1})$ and $a_{t,i} \sim \pi_i(s_t)$, $t = 1, 2, \ldots$ A subscript $()_{-i}$ denotes a partial profile of policies or actions or policies excluding agent i, e.g., $\boldsymbol{\pi}_{-i} := (\pi_1, \pi_2, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_N)$.

2.1.2 Partially Observable Markov Games

An N-agent Partially Observable Markov Game (POMG) is an 8-tuple

$$P = \langle S, s_0, \Omega, \mathbf{A}, T, \mathbf{O}, \mathbf{R}, \gamma \rangle.$$

S, s_0, A and T are defined as in Fully Observable Markov Games. Ω is an observation space, and $\mathbf{O} : S \to \Delta(\Omega)^n$ an observation model, mapping world states to a tuple of observations. In POMGs, agents' policies condition on sequences of observations and actions, $\pi_i : H \to \Delta(A_i)$, where H denotes the set of observation-action histories. Denote

$$V_i^{\pi}(h) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, \mathbf{a}_t) \middle| h_t \right]$$

where the expectation is given with respect to the process $h_{t+1} = (h_t : o_t, a_t)$, : denoting concatenation, $s_t \sim T(s_{t-1}), o_t \sim \mathbf{O}(s_t)$ and $a_{t,i} \sim \pi_i(o_t)$. Write $V_i^{\pi}(h_0) := V_i^{\pi}(\emptyset)$ for the empty history $\emptyset$. For the choice $\Omega = S$, $O_i(s, \mathbf{a}) = (s, \mathbf{a})$ for all $s \in S$, the class of POMGs recovers FOMGs.

We will consider an additional observation model which we call the *contractible observations*. Let P be a POMG, and $O_0 : S \times \mathbf{A} \to \Delta(\Omega_0)$ (where Ω_0 is a distinct observation space). We denote the observations depending on a particular strategy profile π by a superscript, i.e. o_t^{π} is the observation at timepoint t when agents follow the strategy profile π.

Definition 1 Observation model O_0 is *sufficient to detect deviators from π* in POMG P if, for all $i = 1, 2, \ldots, n$ and $\pi_i' \neq \pi_i$, we have that

$$\mathbb{P}\left[\exists t \in \mathbb{N} : o_{0t}^{(\pi_i', \pi_{-i})} \neq o_{0t}^{\pi} \right] = 1.$$

To unpack this equation, an observation model is sufficient to detect deviation if some future observations are able to distinguish π from any policy profile in which a single agent deviates, (π_i, π_{-i}). For example, an observation model that observes which agents defect in the prisoners dilemma is sufficient to detect deviations from the profile π where all agents cooperate.

2.1.3 System performance and optimality

This article measures system performance through *welfare*,

$$W^{\pi}(s_0) := \sum_{i=1}^{N} V_i^{\pi}(s_0).$$

We refer to a policy profile that maximizes welfare as *jointly optimal*: $\pi^* \in \arg\max_{\pi} W^{\pi}(s_0)$. A policy profile π is *Pareto-optimal* if there is no policy profile π' such that $V_i^{\pi}(s_0) \leq V_i^{\pi'}(s_0)$ for all $i = 1, 2, \ldots, N$, with a strict inequality for at least one agent. Intuitively, in such profiles, there are no "win-wins": no agent can attain higher reward without at least one other agent losing reward.

2.1.4 Stable policy profiles and equilibria

In social dilemmas, social and individual incentives diverge. In our game-theoretic analysis, we use an equilibrium notion to capture outcomes of selfish incentives. One potential solution concept is *Nash equilibrium*. A policy profile is Nash equilibrium if unilateral

deviation is suboptimal for all agents. Formally, a policy profile $\boldsymbol{\pi}$ is a Nash equilibrium if for any agent $i \in [N]$ and any policy $\pi'_i : S \rightarrow \Delta(A_i)$ (or $\pi'_i : H \rightarrow \Delta(A_i)$ in POMGs), $V^{\boldsymbol{\pi}}_i(s_0) \geq V^{(\pi'_i, \boldsymbol{\pi}_{-i})}_i(s_0)$.

While this solution concept is common, it has its drawbacks, because it is too permissive. For example, in Cleanup, a policy profile in which agents never clean (likely a non-cooperative outcome of the Cleanup game from Sect. 1) under some contract even if it is in their best interest, and another agent not proposing this contract, would be a Nash equilibrium: no agent would benefit unilaterally from changing their behavior. The "threat" of one of the agents to not clean, however, is non-credible, as, if the contract were active, they would rather clean. Compare [29, Section 5.5] on non-credible threats and subgame perfection.

To avoid non-credible threats, we model selfish incentives in MARL with *subgame-perfect equilibria* (SPE). Subgame perfection requires that for any state s or history h, there cannot be a profitable deviation to another policy, for any agent. This is stronger than a Nash equilibrium, which only requires this to hold at the initial state s_0.

Definition 2 (Subgame-Perfect Equilibrium) A policy profile $\boldsymbol{\pi}$ is a *subgame-perfect Nash equilibrium* or *subgame-perfect* for a POMG if for all histories $h \in H$, agents $i = 1, 2, \ldots, N$ and policies $\pi'_i : H \rightarrow \Delta(A_i)$, we have[1]

$$V^{\boldsymbol{\pi}}_i(h) \geq V^{(\pi'_i, \boldsymbol{\pi}_{-i})}_i(h).$$

2.2 The contracting augmentation

Contractible observation-dependent reward transfers are the main object of study in this article. Here, we formalize these augmentations within Markov games (Fig. 2).

Definition 3 (Contract) A contract is a function $\theta : \Omega_0 \rightarrow \mathbb{R}^N$ whose range consists of zero-sum vectors, i.e.

$$\sum_{i=1}^{N} \theta_i = 0$$

for any $(\theta_1, \theta_2, \ldots, \theta_N) \in \text{range}(\theta)$. We denote a generic set of contracts by $\boldsymbol{\Theta}$.

Contracts, once accepted, act on the reward that agents receive by transferring reward from one agent to another. In this sense, no *new reward* is created through contracting. Moreover, the information used by the contracting logic is modeled by a corresponding contracting observation.

Definition 4 (Contracting Model) A contracting model is a 2-tuple $\mathcal{C} = \langle O_0, \boldsymbol{\Theta} \rangle$, where $O_0 : S \times \mathbf{A} \rightarrow \Delta(\Omega_0)$ is a contractible observation model and $\boldsymbol{\Theta}$ is a contract space.

---

[1] Some histories might never be reached. We will assume for this article that such "off-path" values can be chosen arbitrarily. The research program on *equilibrium refinements* studies how to discipline off-path beliefs, compare [9, ch. 4].
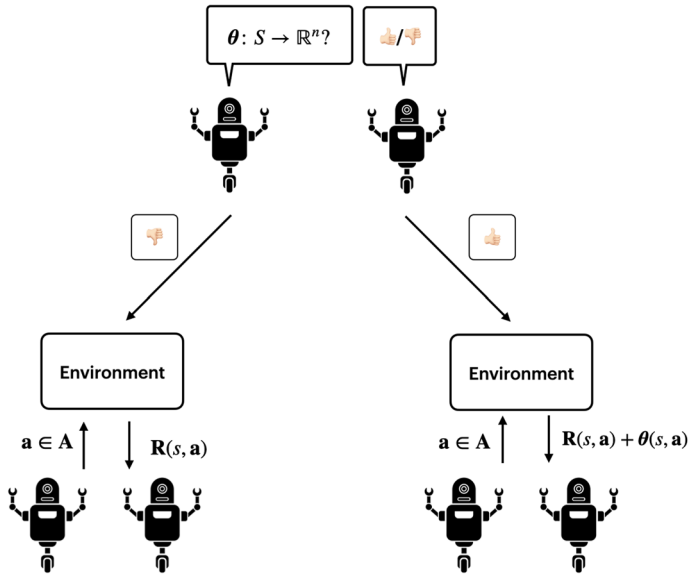
Fig. 2 The Contracting Augmentation. Top: Agents can propose *contracts*: state-dependent, zero-sum, additive augmentations to their reward functions. Other agents can accept or decline proposed contracts. Left: In case of declination, the interaction between agents happens as before. Right: In case of acceptance of the contract, the reward of the agents is altered according to the rules of the contract

This allows us to define the contract augmentation.

Definition 5 ($\mathcal{C}$-Augmented Game) Let $M = \langle S, s_0, \mathbf{A}, T, \mathbf{R}, \gamma \rangle$ be an FOMG and $\mathcal{C} = (O_0, \boldsymbol{\Theta})$ be a contracting model. The i-proposing, $\mathcal{C}$-*augmented game* is $M^{\mathcal{C}} = \langle S', (i, \mathbf{0}), \mathbf{A}', T', \mathbf{R}', \gamma \rangle$, with the components listed below. If the contractible observations are clear from context (e.g. when the full state S is provided to the contract), we write $M^{\boldsymbol{\Theta}}$ instead of $M^{\mathcal{C}}$.

We describe the notation first for FOMGs.

2.2.1 States

The augmented state space is

$$S' = ([N] \cup S) \times (\{\mathbf{0}\} \cup \boldsymbol{\Theta}).$$

States have the following meanings:

- $(i, \mathbf{0})$: Agent i has the opportunity to propose a contract $\theta \in \boldsymbol{\Theta}$;
- (i, θ): $\theta \in \boldsymbol{\Theta}$ awaits acceptance or rejection by all agents;
- $(s, \mathbf{0})$: The game is in state $s \in S$ with a null contract, $\mathbf{0}(s, a) = 0$, for all $s \in S, \mathbf{a} \in \mathbf{A}$, in force;
- (s, θ): The system is in state s with contract $\theta \in \boldsymbol{\Theta}$ in force.

We will implicitly assume that contractible observation models always perfectly observe the proposed contract, and whether contracts were accepted.

2.2.2 Actions

The action spaces for the agents are

$$A_i' = A_i \cup \Theta \cup \{\text{acc}\}$$

which corresponds to actions in the game (A_i), proposal actions (Θ) and the acceptance action ($\{\text{acc}\}$).

2.2.3 Transitions

There are deterministic transitions, given by

$$T'((i, \mathbf{0}), (\theta, \mathbf{a}_{-i})) = \begin{cases} (i, \theta), & \text{for any } (\theta, \mathbf{a}_{-i}) \\ (i, \mathbf{0}) & \text{else.} \end{cases}$$

$$T'((i, \theta), \mathbf{a}) = \begin{cases} (s_0, \theta) & \text{if } \mathbf{a} = \mathbf{acc} \\ (s_0, \mathbf{0}) & \text{otherwise.} \end{cases}$$

for any contract $\theta \in \Theta$ and any action profile $\mathbf{a} \in \mathbf{A}$. Here, we denoted $\mathbf{acc} := (\text{acc}, \text{acc}, \dots, \text{acc})$ the profile of unanimous acceptance of a contract.

Transitions in states $(s, \mathbf{0})$ and (s, θ) are as in the underlying game M,

$$T'((s, \theta), \mathbf{a}) = T(s, \mathbf{a})$$

for any $s \in S$, $\theta \in \Theta$ and $a \in A$.

2.2.4 Rewards

$$\mathbf{R}'((s, \theta), \mathbf{a}) = \mathbf{R}(s, \mathbf{a}) + \theta(s, a),$$
$$\mathbf{R}'((i, \theta), \mathbf{acc}) = \theta(\text{acc}).$$

for $\theta \in \Theta$ and $s \in S$. In states (s, θ), the actions acc and θ yield a high negative reward. The first line means that depending on a state-action profile pair, reward is transferred between the agents. The second line refers to reward being transferred on signing a contract.

2.2.5 Partially observable markov games

In order to generalize the above construction to POMGs, we consider a set of contracts $\Theta = \{\Omega_0 \to \mathbb{R}^N\}$ which depend on the contractible observation model $O_0 : S \times \mathbf{A} \to \Delta(\Omega_0)$, in general distinct from the observation models for each agent. We recover contracting in a FOMG for contractible observations $O_0(s, \mathbf{a}) = (s, \mathbf{a})$. The environment dynamics are the same as in the augmentation for FOMGs. It remains, hence, to generalize the observation and reward model to POMGs.

The observation space for the agents then becomes $\Omega' = ([N] \cup \Omega) \times (\{\mathbf{0}\} \cup \Theta)$. Agents observe the proposer, the contract proposed or in force, and, if they are in a game state, they also recieve their individual observations in the base POMG, as below

$$O'_i((s, \theta), \mathbf{a}) = \begin{cases} (O_i(s, \mathbf{a}), \theta) & s \in S \\ (s, \theta) & s \in [n] \end{cases}$$

To unpack this notation, observe that $(s, \theta) \in S'$ is a state in the augmented game, and $\mathbf{a}$ is an action profile.

In the contracting augmentation, once enforced, the expected rewards of agents are directly changed, according to

$$R'_i(s, \mathbf{a}) = R_i(s, \mathbf{a}) + \mathbb{E}_{o \sim O_0(s, \mathbf{a})}[\theta_i(o)].$$

That is, the rewards are sampled according the contractible observation model. As in FOMGs, $R((i, \theta), \mathrm{acc}) = \theta(\mathrm{acc})$. Note that agents maximize their reward as modified under the contract, so there is no concept of "breaking" a contract. The incentives that align agents' behavior with pro-social goals are encoded in the reward function.

2.2.6 Expressiveness

We define a notion of comparison for a subclass of contracting models. A contracting model $\mathcal{C} = (O_0, \Theta)$ has greater *expected transfer expressiveness* than $\mathcal{C}' = (O'_0, \Theta')$, written $\mathcal{C}' \preceq_{\mathrm{ETE}} \mathcal{C}$ if, for all $\theta' \in \Theta'$, there exists $\theta \in \Theta$ such that

$$\mathbb{E}_{o \sim O'_0(s, \mathbf{a})}[\theta'(o)] = \mathbb{E}_{o \sim O_0(s, \mathbf{a})}[\theta(o)]. \tag{2}$$

Note that the inequality in ETE comes from the quantifier order—if one contract space has less ETE than the other, any contract in the first can be replicated in the second in expectation.

3 Formal contracting mitigates social dilemmas

In this section, we show that formal contracting, given a sufficiently rich set of contracts, mitigates social dilemmas in many fully and partially observable settings. Find a summary of our results in the first column of Fig. 3.

Theorem 1 (Optimality of Contracting) *Let $M = \langle S, s_0, \mathbf{A}, T, \mathbf{R}, \gamma \rangle$ be a Fully Observable Markov game. If all observations are contractible, $O_0(s, \mathbf{a}) = (s, \mathbf{a})$, then the contracting space*

$$\Theta_{\mathrm{full}} = \{(S \times \mathbf{A}) \cup \{\mathrm{acc}\} \to \mathbb{R}\},$$

1. *There exists an SPE of M^C;*
2. *For any SPE $\boldsymbol{\pi}$ of M^C, there is a jointly optimal policy profile $\boldsymbol{\pi}^*$ of M such that $\boldsymbol{\pi}((s, \theta), \mathbf{a}) = \boldsymbol{\pi}^*(s, \mathbf{a})$ for the contract θ that agent i chooses in $\boldsymbol{\pi}$.*

The theorem shows that, under the assumption of richness, contracting mitigates social dilemmas in game-theoretic equilibrium.

	Optimal When $\Theta \supset \Theta_{full}$	Monotone in Θ	Monotone in Ψ
Full Information Games + arbitrary unconditional transfers	✅	✅	✅
POMGs with Detectable Deviation + arbitrary unconditional transfers	✅	✅	✅
POMGs with Detectable Deviation	✅	❌	❌
POMGs + arbitrary unconditional transfers	❌	✅	✅
POMGs	❌	❌	❌

Fig. 3 A table summarizing the theoretical results proven in Sects. 3 and 4. The columns along the top list theorem types, which are (1) optimality when you have a contracting space with sufficient richness, (2) monotonicity in the size of the contract space (3) monotonicity in the space of contractible features. The rows list the various problems setups considered in these sections

Proof Sketch For any socially optimal policy profile π^*, we can construct a "forcing contract" θ^*, which sets a high penalty for not playing π^*. A signing bonus for this contract transfers all utility gained or lost by playing θ^* for all agents but the proposer to the proposer. The main part of the proof is to show that this is the (only) optimal choice for the proposer from Θ_{full}, subject to agents accepting the contract. □

The full proof can be found in Appendix A. The conclusions of this theorem, along with those of Theorem 4, are illustrated in Fig. 4.
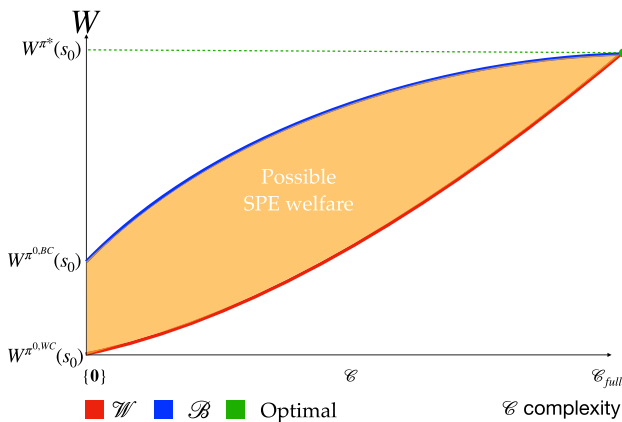
Fig. 4 Under detectable deviations, possible welfare levels at equilibrium improve to optimality as the complexity of Θ grows, if spaces allow for arbitrary unconditional transfers (see discussion in Sect. 4)

The same result continues to hold for POMGs if deviations are detectable, even if the whole state is not necessarily observed. In this case, it might not be immediately possible to punish deviations. However, the contracting logic can simply wait a finite amount of time until a deviation occurs, and the precise agent that committed it is determined, which must occur by the detectable deviation assumption. This is more precisely articulated below.

Theorem 2 (Optimality of Contracting with Detectable Deviations) *Let* $P = \langle S, s_0, \Omega, \mathbf{A}, T, \mathbf{O}, \mathbf{R}, \gamma \rangle$ *be a POMG, and* O_0 *a contractible observation model that has detectable deviations from some socially optimal policy profile* π^*. *Then, the same conclusions as in Theorem* 1 *hold, if we require* $\mathbf{\Theta} = \mathbf{\Theta}_{\text{full}}$ *and* $\mathcal{C} \succeq_{\text{ETE}} \mathcal{C}_{\text{full}} = (\mathbf{\Theta}_{\text{full}}, O_0)$.

This result, however, fails to hold if the joint observations are unable to detect deviations from π^*. If this were the case, the mechanism behind the forcing contract no longer goes through, because contracts can't necessarily sanction agents appropriately for behaviour deviating from π^*, and so in general can't compel all agents to act in a way that exactly maximizes welfare.

For an extreme example of this, suppose contractible observations are perfectly uninformative, $\Omega_0 = \{0\}$. In this case, reward transfers cannot condition on actions played in the course of the game at all, and hence are irrelevant for any agent in computing best responses. The augmented game then has the same subgame-perfect equilibria as the original game.

4 More possible contracts improve welfare

Even without full observability or detectable deviations, contracts can still be useful. We show that in both FOMGs and POMGs, system performance as measured by social welfare improves for more expected-transfer-expressive ($\preceq_{\text{ETE}}$) contracting models.

Restrictions on the capability to contract correspond to some contracts being impermissible in the environment. Reasons for such a restriction in practice may be desirable to incorporate domain knowledge into the contract design, or simply to improve exploration landscape for reinforcement learning.

In defining our order on system performance, we use $M + \theta$ as the game following the acceptance of a contract θ. Note that the system performance definition is complicated by the existence of potentially multiple equilibria, compare [7]. We can show, though, that the performance of the worst *and* the best stable outcome, that is, a variant of welfare of the worst and best subgame-perfect equilibria, improve. As it might be possible that the proposing agent has an optimal contract they would like to propose, e.g. if there is a limit point of utility for them that is not achievable, we define a notion of best and worst equilibrium that considers cases in which the proposer takes an arbitrarily close to optimal contract. For finite contract spaces, this issue disappears, and the reader can view the following results as being about the worst and best subgame-perfect equilibria of $M^{\mathcal{C}}$.

Definition 6 (WCSPW and BCSPW) Let M be a Markov game, and $\mathcal{C} = \langle O_0, \mathbf{\Theta} \rangle$ be a contracting model. For $\theta \in \mathbf{\Theta}$, we define the *worst-case* resp. *best-case subgame-perfect welfare under* θ as

$$\mathrm{WCSPW}_{\mathcal{C}}(\boldsymbol{\theta}) := \inf_{\boldsymbol{\pi} \in \mathrm{SPE}(M+\theta)} W^{M+\theta,\boldsymbol{\pi}}(s_0)$$

$$\mathrm{BCSPW}_{\mathcal{C}}(\boldsymbol{\theta}) := \sup_{\boldsymbol{\pi} \in \mathrm{SPE}(M+\theta)} W^{M+\theta,\boldsymbol{\pi}}(s_0). \tag{3}$$

We also define the supremum over the contract choices.

$$\mathcal{W}(\mathcal{C}) = \sup_{\boldsymbol{\theta} \in \Theta} \mathrm{WCSPW}_{\mathcal{C}}(\boldsymbol{\theta})$$

$$\mathcal{B}(\mathcal{C}) = \sup_{\boldsymbol{\theta} \in \Theta} \mathrm{BCSPW}_{\mathcal{C}}(\boldsymbol{\theta}).$$

These are bounds on the welfare under subgame perfection, which is attained if the proposer's choice of a contract has an optimal choice. Note that otherwise a subgame-perfect equilibrium fails to exist, as Proposition 3 shows.

Before stating the main result of this section, we formalize the intuition that the proposer may specify an arbitrary transfer at the moment a contract is signed.

Definition 7 We say a contracting space Θ admits arbitrary unconditional transfers if and only if, for any $\boldsymbol{\theta} \in \Theta$, and any $v \in \mathbb{R}^N$, we have that $\boldsymbol{\theta}' \in \Theta$, where $\boldsymbol{\theta}'$ is identical to $\boldsymbol{\theta}$, except for an additional transfer of v at contract acceptance.

The above means that the contracting space is rich enough to allow the proposing agent to redistribute a certain amount of reward unconditionally, at the start of the episode. With this defined, we have:

Proposition 3 ($\mathcal{W}, \mathcal{B}$ bound equilibrium welfare) *Let Θ admit arbitrary unconditional transfers, augmenting a Markov game M. Then, we have that, for any $\boldsymbol{\pi} \in \mathrm{SPE}(M^{\mathcal{C}})$*

$$\mathcal{W}(\mathcal{C}) \le W^{\boldsymbol{\pi}}(s_0) \le \mathcal{B}(\mathcal{C}).$$

If $\mathcal{W}(\mathcal{C}) = \sup_{\boldsymbol{\theta} \in \Theta} \mathrm{WCSPW}_{\mathcal{C}}(\boldsymbol{\theta})$ resp. $\mathcal{B}(\mathcal{C}) = \sup_{\boldsymbol{\theta} \in \Theta} \mathrm{BCSPW}_{\mathcal{C}}(\boldsymbol{\theta})$ are attained by some $\boldsymbol{\theta}^*$, and moreover for such $\boldsymbol{\theta}^*$ we can take $\boldsymbol{\pi}^*$ attaining $\mathrm{WCSPW}_{\mathcal{C}}(\boldsymbol{\theta}^*)$ resp. $\mathrm{BCSPW}_{\mathcal{C}}(\boldsymbol{\theta}^*)$, then there is $\boldsymbol{\pi} \in \mathrm{SPE}(M^{\mathcal{C}})$ such that $W^{\boldsymbol{\pi}}(s_0) = \mathcal{W}(\mathcal{C})$ resp. $W^{\boldsymbol{\pi}}(s_0) = \mathcal{B}(\mathcal{C})$.

Proof Sketch We sketch this proof for first inequality. The second is analogous.

We first show the second assertion assuming that the supremum in (3) and (4) each are attained. Let $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta} \in \Theta} \mathrm{WCSPW}_{\mathcal{C}}(\boldsymbol{\theta})$ and denote by $\boldsymbol{\pi}^*$ be an SPE attaining (3) and $\boldsymbol{\pi}$ any subgame-perfect equilibrium in M. We construct an SPE for $M^{\mathcal{C}}$, by first defining $\boldsymbol{\theta}_{\mathrm{tr}}^*$ (tr referring to "transfer") as

$$\boldsymbol{\theta}_{\mathrm{tr}}^*(o) = \begin{cases} \mathbf{v} & o = \mathrm{acc} \\ \boldsymbol{\theta}^*(o) & o \ne \mathrm{acc} \end{cases}$$

where the vector $\mathbf{v}$

$$\mathbf{v} = \begin{pmatrix} \sum_{i=2}^n V_i^{M+\theta^*,\pi^*}(s_0) - V_i^{M,\pi}(s_0) \\ V_2^{M,\pi^*}(s_0) - V_2^{M+\theta^*,\pi}(s_0) \\ \vdots \\ V_n^{M,\pi^*}(s_0) - V_n^{M+\theta^*,\pi}(s_0) \end{pmatrix}$$

We have $\theta_{tr}^* \in \Theta$ as Θ admits arbitrary unconditional transfers. It is an SPE as no agent at no point has an incentive to deviate. All agents except for the proposing agent are compensated exactly to the point of indifference; the proposing agent maximizes its reward, under the constraint of agents accepting the contract. Now, taking an arbitrary SPE, we have that *full surplus extraction*, i.e. unconditionally transferring all excess reward from the non-proposers to the proposer, must occur at every SPE (otherwise, the proposing agent could unilaterally change by simply performing this extraction, and changing nothing else). But this directly ties the proposer's reward at all SPE to the welfare attained across all agents: hence, in order for the propose to get better payoff than $\mathcal{W}(\mathcal{C})$ at equilibrium, the welfare attained at equilibrium must be higher. Therefore, $\mathcal{W}(\mathcal{C})$ acts as a lower bound on the possible welfare attained across all SPE of the contracting game.

By choosing θ^* to be close to $\sup_{\theta \in \Theta} \mathrm{WCSPW}_{\mathcal{C}}(\theta)$ we can generalize to the case where the exact value is not attained. □

Arbitrary unconditional transfers are necessary for Proposition 3 to hold. We provide a counterexample to Proposition 3 without arbitrary unconditional transfers in Sect. 4.5. In this, the proposer cannot internalize social welfare due to limits of transferring reward, and hence has an incentive to propose a contract decreasing equilibrium welfare in order to improve its individual reward.

The main theorem of this section is the following monotonicity result.

Theorem 4 (Monotonicity in Contract Spaces) *We have that the bounds $\mathcal{W}(\mathcal{C})$ and $\mathcal{B}(\mathcal{C})$ are both monotonically increasing quantities in $\mathcal{C}$: that is, when $\mathcal{C}_1 \preceq_{\mathrm{ETE}} \mathcal{C}_2$, we have $\mathcal{W}(\mathcal{C}_1) \leq \mathcal{W}(\mathcal{C}_2)$ and $\mathcal{B}(\mathcal{C}_1) \leq \mathcal{B}(\mathcal{C}_2)$*

Proof Let $\mathcal{C}_1 \preceq_{\mathrm{ETE}} \mathcal{C}_2$. We show the proof only for $\mathcal{W}$: the proof for $\mathcal{B}$ is identical. In order to establish the desired inequality, given that the expression

$$\mathcal{W}(\mathcal{C}) = \sup_{\theta \in \Theta} \mathrm{WCSPW}_{\mathcal{C}}(\theta)$$

is defined by a supremum, it is enough to show that, for any $\theta_1 \in \Theta_1$, there exists a $\theta_2 \in \Theta_2$ such that $\mathrm{WCSPW}_{\mathcal{C}_1}(\theta_1) \leq \mathrm{WCSPW}_{\mathcal{C}_2}(\theta_2)$. But by definition of $\preceq_{\mathrm{ETE}}$, for any $\theta_1 \in \Theta_1$, there exists a $\theta_2 \in \Theta_2$ which matches expected transfers at all state-action transitions—hence all payoffs in the POMG augmented by $\mathcal{C}_1$ are matched exactly by such $\theta_2 \in \Theta_2$ at all states and observation histories. Hence, all subgame-perfect equilibria attainable under $\mathcal{C}_1$ are attainable in $\mathcal{C}_2$, as needed. □

The implications of this result are illustrated in Figs. 4 and 5—specifically, whenever contracting models are nested in expressiveness, welfare in subgame-perfect equilibrium increases, whether or not optimal performance is attained in the limit of complexity. We apply Theorem 4 in special cases.

4.1 Deterministic contractible features

Here, we formalize the relationship between contracting models and contracting on a set of features ϕ of the state $s \in S$. Let $\psi : S \to \Psi$ be a generic (deterministic) feature map. For instance, in the harvest problem from Sect. 1, one might contract based on the number of apples near to an agent, and indicator variables for whether or not those
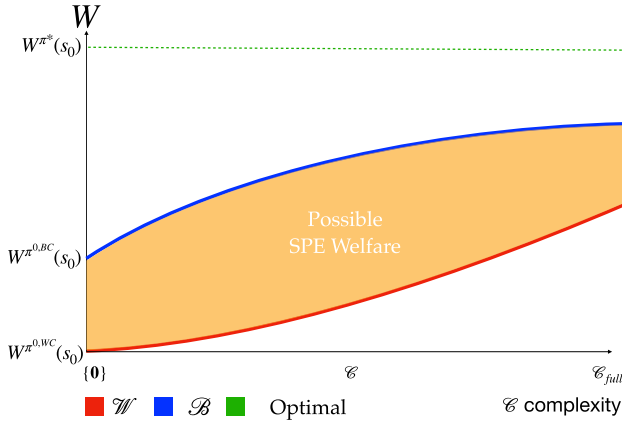
Fig. 5 Under general partial observability, possible welfare levels at equilibrium improve, without necessarily reaching optimal social welfare, assuming arbitrary unconditional transfers

apples have a sufficiently high density of neighboring apples (recall that apples only grow in high-density regions of apples). We define the feature-based contracting model on feature mapping ψ through the observation model $O_0(s, \mathbf{a}) = \psi(s)$, when paired with the space of all ψ-dependent contracts Θ_{full}, and denote it by $C_\psi = (\Theta_\psi, \psi)$.

Corollary 5 (Feature Monotonicity) *Let M be a Markov game and ψ_1, ψ_2 two feature models such that $\kappa : \Psi_2 \to \Psi_1$ and $\kappa \circ \psi_2 = \psi_1$. Then,*

$$\mathcal{W}(C_{\psi_1}) \leq \mathcal{W}(C_{\psi_2})$$
$$\mathcal{B}(C_{\psi_1}) \leq \mathcal{B}(C_{\psi_2}).$$

Proof By Theorem 4, it suffices to show that $C_{\psi_1} \preceq_{\text{ETE}} C_{\psi_2}$. Let $\theta_1 \in \Theta_{\psi_1}$. We have that

$$\theta_1 \circ \psi_1(s, \mathbf{a}) = \theta_1 \circ (\kappa \circ \psi_2)(s, \mathbf{a})$$
$$= (\theta_1 \circ \kappa) \circ \psi_2(s, \mathbf{a}).$$

Hence, in the definition of $\preceq_{\text{ETE}}$, we can choose $\theta_2 = \theta_1 \circ \kappa$ as a contract in Θ_{ψ_2}. As it is pointwise the same contract, they match in expectation, as in Equation (2). □

This result gives that, by Proposition 3, adding features to your contracting space shifts the region of possible equilibrium welfares upwards. This motivates the monotonicity in Figs. 4 and 5.

4.2 Random contractible features in a fully observable markov game

Noisy features still give rise to monotonicity. Concretely, let $\psi : S \times Z \to \Psi$ be a mapping for a random variable Z. This gives rise to a contractible observation model $O_0(s, \mathbf{a}) = \psi(s, z)$, which may be random.

Proposition 6 (Monotonicity of Random Features) *Let M be an FOMG, and suppose we have that $s \to \psi_2 \to \psi_1$ form a Markov chain. Then, $\mathcal{W}(\mathcal{C}_{\psi_1}) \leq \mathcal{W}(\mathcal{C}_{\psi_2})$ and $\mathcal{B}(\mathcal{C}_{\psi_1}) \leq \mathcal{B}(\mathcal{C}_{\psi_2})$.*

Note that this result subsumes the deterministic case, as the second arrow in the above Markov chain takes the role of κ in that definition.

Proof We proceed, analogously to the deterministic feature case, by showing that any random contract $\theta_1 \in \Theta_{\psi_1}$ has expected transfers matched by some contract $\theta_2 \in \Theta_{\psi_2}$, hence verifying $\mathcal{C}_{\psi_1} \preceq_{\text{ETE}} \mathcal{C}_{\psi_2}$, and then applying Theorem 4 for the desired result.

Let $\theta_1 \in \Theta_{\psi_1}$. Then, since $S \to \psi_2 \to \psi_1$ is a Markov chain, we know there exists a Markov kernel κ s.t. $\psi_1 = \kappa \circ \psi_2$. From this, we define the contract $\theta_2 \in \Theta_{\psi_2}$ by the following expectation

$$\theta_2 := \mathbb{E}[\theta_1 | \psi_2],$$

which is a function only of ϕ_2. This contract choice is as desired since, by the tower property of conditional expectation:

$$\mathbb{E}_{\phi_2 \sim \psi_2(s,\mathbf{a})}[\theta_2(\phi_2)] = \mathbb{E}_{\phi_1 \sim \psi_1(s,\mathbf{a})}[\mathbb{E}_{\phi_2 \sim \kappa(\phi_1)}[\theta(\phi_2)|\phi_1]] = \mathbb{E}_{\phi_1 \sim \psi_1(s,\mathbf{a})}[\theta_1(\phi_1)]$$

Hence, $\mathcal{C}_{\psi_1} \preceq_{\text{ETE}} \mathcal{C}_{\psi_2}$, and we can apply Theorem 4, as needed. □

4.3 Random contractible features in a partially observable markov game

Note that, in no step for the above sections do we rely on full observability for indiviual agents. Moreover, the case of random features in a deterministic game is subsumed by different choices of contractible observations O_0, and so no further extension is required on that front. Therefore, we have the fully general result:

Theorem 7 *Let $P = \langle S, s_0, O, \mathbf{A}, T, \mathbf{O}, \mathbf{R}, \gamma \rangle$ be a POMG, and let $\mathcal{C} = \langle O_0, \Theta \rangle$ be a contracting model. Then, the conclusions in Proposition 3, Corollary 5, Proposition 6 all continue to hold for $P^\mathcal{C}$, when $\text{WCSPW}_\mathcal{C}$ and $\text{BCSPW}_\mathcal{C}$ minimize and maximize over the space of all equilibria of recurrent policy profiles $\Pi : H_i \to \Delta A_i$, and with contracts depending on contracting observation O_0.*

4.4 Contracting and decentralized markov processes

Above, we have argued that sufficiently expressive contracting removes strategic pressures on individual agents, and ensures that agents are incentivized to enact a socially optimal policy profile. One way to interpret this is that sufficiently expressive contracting in fully-observed Markov games is a reduction from solving social dilemmas to solving decentralized Markov decision processes (DEC-MDPs) at equilibrium [2]. DEC-MDPs are Markov games where agents all share a common reward, and our theory says that at all SPE of a (sufficiently expressive) contract-augmented game, agents act as if they are attempting to coordinate on optimizing a common scalar reward of $\sum R_i$.

Analogously, in decentralized *partially observed* Markov decision processes (DEC-POMDPs, Bernstein et al. [2]), all agents are given an identical scalar reward to maximize, but potentially different observation models. A natural follow-up question is whether DEC-POMDPS admit a similar reduction. Unfortunately, the answer is negative. Partially-observed contracting with even $\mathcal{C}_{\text{full}}$ poses an additional restrictions beyond DEC-POMDPs. Agent policies depend on observation-action histories, and the contracts depend on the state through a different observation model. This means that it may be impossible for contracts which are functions of the true state to perfectly track all agent's current observation-action histories—thus, it may be impossible to enforce socially optimal policy profiles as a function of these histories. (This restriction can be relaxed if observation-action histories are deterministic function of the true state, which we could call *history-transparent* POMDPs, for which such a reduction holds.)

4.5 Necessity of arbitrary unconditional transfers

As a last result in this section, we show that our restriction to contract spaces that admit arbitrary unconditional transfers was necessary. Welfare does not necessarily increase when moving to $\preceq_{\text{ETE}}$-greater contract spaces when the spaces involved do not admit arbitrary unconditional transfers.

Example 1 (Bribing the Contract Proposer) Recall the Prisoner's Dilemma of Table 1a, with the row player as the contract proposer. We can consider $\Theta = \{\mathbf{0}, \theta\}$ (recalling the null contract $\mathbf{0}$ having no effect on the original game), where θ imposes the choice (C, C) for both agents with high penalty (say, -10) of defecting, with an additional unconditional transfer of 0.5 from the proposer to the column player, and $\Theta' = \{\mathbf{0}, \theta, \theta'\}$, where θ' imposes a penalty of -10 on only the column player for defecting, and a transfer of 1.25 for cooperating. Then, for Θ, we get the proposer has an incentive to select θ over null contract $\mathbf{0}$, attaining reward -1.5 at equilibrium, but in the second space, the proposer has a stronger incentive to select θ', where it gets a reward of -1.25 at equilibrium (in all cases, these contracts give reward greater than -2 for both proposing and accepting agents, and hence are incentive-compatible in proposal, acceptance, and in selection at the given equilibrium). Thus, even though Θ' has more possible contracts to choose from, worse welfare are attained in the limit due to poor incentives for the proposing agent.

5 Evaluation

We now evaluate the performance of the contracting augmentation. First, we introduce the baseline methods used to evaluate our approach. Then, we introduce our experimental domains. Finally, we provide details on MOCA, our training procedure for contracting. Replication code can be found at https://github.com/Algorithmic-Alignment-Lab/contracts.

5.1 Evaluation

We evaluate MOCA by comparing to the following baselines.

- *Joint Training* a centralized algorithm with joint action space $\mathbf{A} = \times_{i=1}^{N} A_i$ chooses actions to maximize welfare (e.g. with reward $\sum R_i$);
- *Separate Training* Agents selfishly maximize their reward, with no modification;
- *Gifting* Agents can "gift" [23] another agent at every timestep by directly transferring some of their reward;
- *Vanilla Contracting* Run off-the-shelf deep RL on the contract-augmented versions of the respective domains.

We train all domains with 2, 4, and 8 agents, using Proximal Policy Optimization (PPO) [35] with ray rllib's [22] implementation (we report our hyperparameter settings in Appendix C). In each domain, we train gifting agents with a lower bound of 0 and an upper bound equivalent to the transfer value in the given constract space. This allows the same magnitude of transfers in gifting and contracts, for fair comparison. In one of the games, Emergency Merge, we reduced the gifting values to 10 per timestep, as this improved gifting's performance. On the Prisoner's Dilemma and the Public Goods game, we trained agents for 1 M environment steps, and the complex dynamic games are trained for 10 M environment steps. In the study on monotonicity (Sect. 6.3), we run the environment for 15 M environment steps.

5.2 Games

We test on several classes of games. We use Prisoner's Dilemma and a Public Goods game as static, simultaneous-move games, and Harvest, Cleanup, and Emergency Merge as dynamic domains.

5.2.1 Prisoner's dilemma

The N-agent Prisoner's Dilemma has two actions per agent, cooperate and defect. If all agents cooperate, they each get reward N, and if all defect, they all get reward 1. However, if some defect and some cooperate, the ones that cooperate get reward 0 and the ones that defect get reward $N + 1$. Again, the socially optimal outcome is the one where all agents cooperate, but only Nash equilibrium is where all agents defect. We run an additional timestep after the matrix game is played for gifting actions to take place.

5.2.2 Public goods

We study the following public goods game [17]. Agents choose an *investment* $a_i \in [0, 1]$, and get reward $R_i(\mathbf{a}) = \frac{1.2}{N} \sum_{j=1}^{N} a_j - a_i$, i.e. they are given their share of the public returns, the investment returning 20%, minus their own investment level. At social optimum, all agents choose $a_i = 1$ to get optimal joint reward. However, selfish agents are not incentivized to invest at this high level, as they would like to free-ride on the other agents' efforts.

5.2.3 Harvest

In Harvest, from Hughes et al. [14], agents move along a square grid to consume apples, gaining a unit of reward. Apples grow faster if more apples are close by, which leads to incentives to overconsume now, leading to an intertemporal dilemma. We choose

engineered features to limit the amount of computational resources needed. In particular, agents receive their position and orientation, the coordinates and orientation of the closest other agent, the position of the nearest apple, the number of apples close to the agent, the total number of apples, and the number of apples eaten by each agent in the last timestep. We don't allow agents to use a punishment beam following Lupu and Precup [23]. The environment runs for 1,000 timesteps per episode.

5.2.4 Cleanup

In Cleanup, also from Hughes et al. [14], agents similarly move along a square grid to consume apples and gain one unit of reward. Apples only spawn if a nearby river contains a number of waste objects lower than a threshold. Removing a waste object is a costless, but also rewardless, task. Apple-picking agents can free-ride on other agents, which leads to degraded performance. The observation space used for agents is simplified to limit computational requirements, and agents are passed their position and orientation, the position and orientation of the closest agent, the positions of the closest apple and waste object, and the number of current apples and waste objects. The environment runs for 1,000 timesteps per episode.
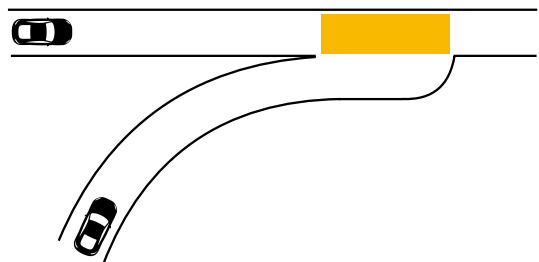
5.2.5 Emergency merge

A set of $N - 1$ cars approaches a merge, an ambulance behind them, compare Fig. 6. The ambulance incurs a penalty of 100 per timestep that it has not reached the end of a road segment past the merge. The cars in front also want to get to the end of the road segment, but incur a penalty of only 1 per timestep. They are limited to one-fourth of the velocity that the ambulance can go. We assume access to controllers preventing cars from colliding (stopping cars short of crashing into another car) and managing merging, and so the actions $a_i \in [-0.1, 0.1]$ only control the forward acceleration of each vehicle. A dilemma arises as cars prefer to drive to the merge fast, not internalizing the strong negative effect this has on the ambulance. The environment resets after 200 rounds. Note that here, due to the asymmetry of agent capabilities and rewards, attaining optimal social welfare cannot be done via Pareto improvement within the original game.

5.3 Contract spaces

We consider low-dimensional contract spaces for different domains.
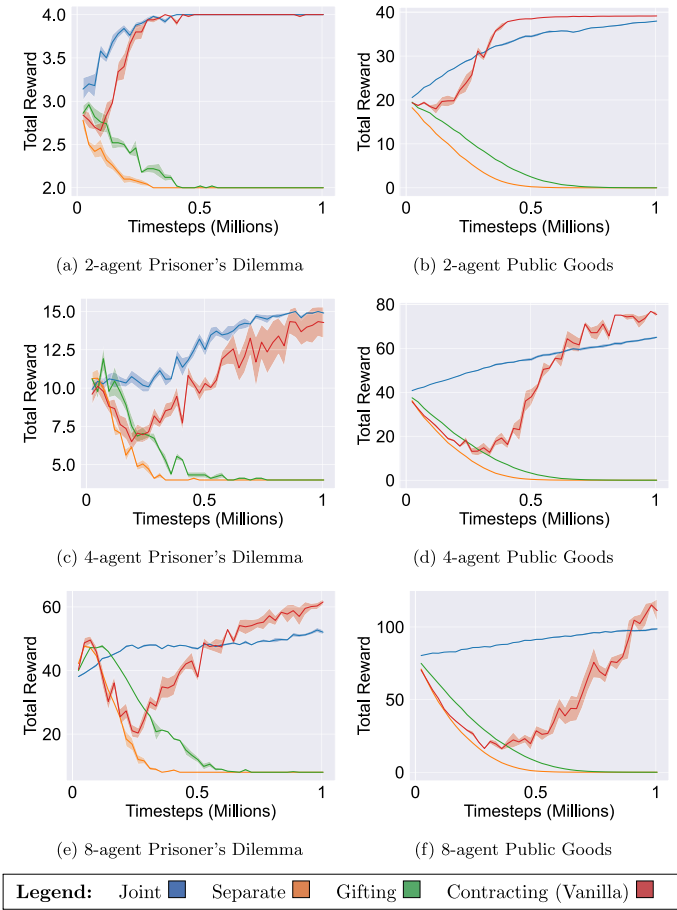

Fig. 6 A depiction of the emergency merge domain

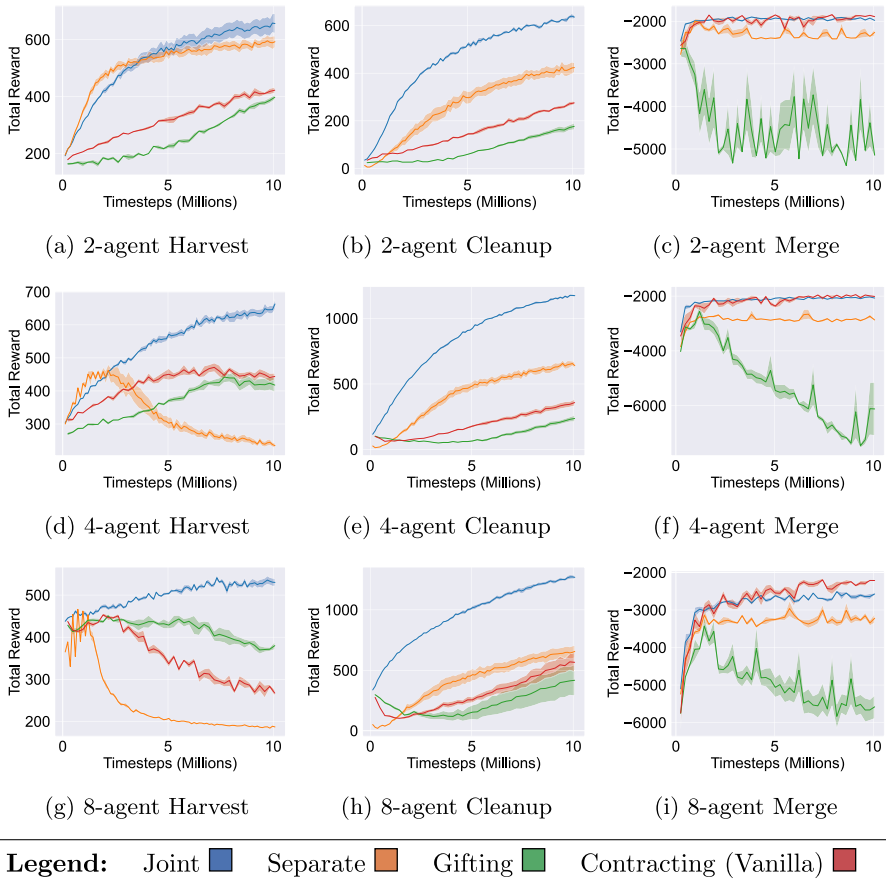|	
Prisoner's dilemma	Contracts are parameterized by a transfer $\theta \in [0, N]$ for defecting, which is distributed to the other agents in equal proportions.
Public goods	Contracts are parameterized by a transfer $\theta \in [0, 1.2]$, agents transfer $\theta(1 - a_i)$, which is distributed to the other agents in equal proportions.
Harvest	Contracts are parameterized by $\theta \in [0, 10]$. When an agent takes a consumption action of an apple in a low-density region, defined as an apple having less than 4 neighboring apples within a radius of 5, they transfer θ to the other agents, which is equally distributed to the other agents.
Cleanup	Contracts are parameterized by $\theta \in [0, 0.2]$, which correspond to a payment per garbage piece cleaned, paid for evenly by the other agents.
Emergency car	The ambulance can propose a per-unit subsidy of $\theta \in [0, 100]$ to the cars at the time of ambulance crossing. Each car is transferred θ times its distance behind the ambulance at time of merge by the ambulance. If a car is ahead of the ambulance at time of reward, it pays the ambulance θ times its distance ahead of the ambulance.

5.4 Training

The contracting augmentation yields a Markov game, for which one could directly train agents with deep reinforcement learning (we will call this *vanilla contracting*). However, as can be observed from Figs. 7, 8, 9, and 10 this implementation of contracting does not outperform joint training in problems with more complex dynamics, or higher-dimensional state and action spaces. To fix this, we propose an algorithm inspired by multi-objective reinforcement learning, compare [1], which we call *Multi-Objective Contract Augmentation Learning* (MOCA). We present it in Algorithm 1. MOCA consists of two phases: first, the algorithm draws random contracts (which, in the language of multi-objective reinforcement learning, can be viewed as different "objectives"). This can be used to estimate $V_i^\pi(s_0, \theta)$, $i = 1, 2, \ldots, N$ for the initial state s_0 and any contract θ, i.e. the values for agents when contract $\theta \in \Theta$ is in force. This allows it to learn estimates of the utility agents will get under a particular contract. Due to random sampling, these estimates are not biased by contract exploration, which may be an issue when using deep reinforcement learning directly.

In a second phase, we freeze play following (s_0, θ) for any contract θ and the policy at states $(i, \mathbf{0})$ and (i, θ). We do so by choosing a contract repeatedly from the policy $\pi_i(i, \mathbf{0})$, and use as a proxy for acceptance the expected probability of acceptance, $\prod_{j=1}^{N} \pi_j(i, \theta)$. In order to help exploration of the contract space in this stage, we sample ν agents from the space of non-proposing agents, and only use these agent's accept-reject probabilities in determining contract acceptance. Here, the introduced ν becomes a tunable hyper-parameter, for which $\nu = 2$ obtained strong performance across all domains, which we report in Sect. 6. We update the weights for the actions of all agents at $\pi_i(i, \mathbf{0})$ and $\pi_j(i, \theta)$, for $j = 1, 2, \ldots, n$. Finally, the algorithm returns the so-obtained policy profile.

(a) 2-agent Prisoner's Dilemma

(b) 2-agent Public Goods

(c) 4-agent Prisoner's Dilemma

(d) 4-agent Public Goods

(e) 8-agent Prisoner's Dilemma

(f) 8-agent Public Goods

Legend: Joint ■   Separate ■   Gifting ■   Contracting (Vanilla) ■

Fig. 7 Welfare throughout training the benchmark algorithms for the simple tabular games (including a *vanilla* implementation of contracting using off-the-shelf deep reinforcement learning). We have joint training (in blue), separate training (in orange), gifting (in green), and vanilla contracting (in red), presented. In simple static domains, contracting achieves welfare that attains or exceeds the returns from jointly training agents. For each figure, the *x*-axis plots number of environment steps (e.g. all agents taking an action is one step), and error is one standard deviation over 5 independent runs

(a) 2-agent Harvest   (b) 2-agent Cleanup   (c) 2-agent Merge

(d) 4-agent Harvest   (e) 4-agent Cleanup   (f) 4-agent Merge

(g) 8-agent Harvest   (h) 8-agent Cleanup   (i) 8-agent Merge

Legend: Joint ■   Separate ■   Gifting ■   Contracting (Vanilla) ■

Fig. 8 Welfare throughout training the benchmark algorithms for the more complicated dynamic games (including a *vanilla* implementation of contracting using off-the-shelf deep reinforcement learning algorithms). We have joint training (in blue), separate training (in orange), gifting (in green), and vanilla contracting (in red), presented. In the simpler Merge domain, contracting achieves welfare that is close to joint training, but more complex domains (i.e. Harvest and Cleanup), biased policy exploration due to the difficulty of learning best responses has a greater effect. Therefore, vanilla contracting suffers in performance. For each figure, the *x*-axis plots number of environment steps (e.g. all agents taking an action is one step), and error is one standard deviation over 5 independent runs
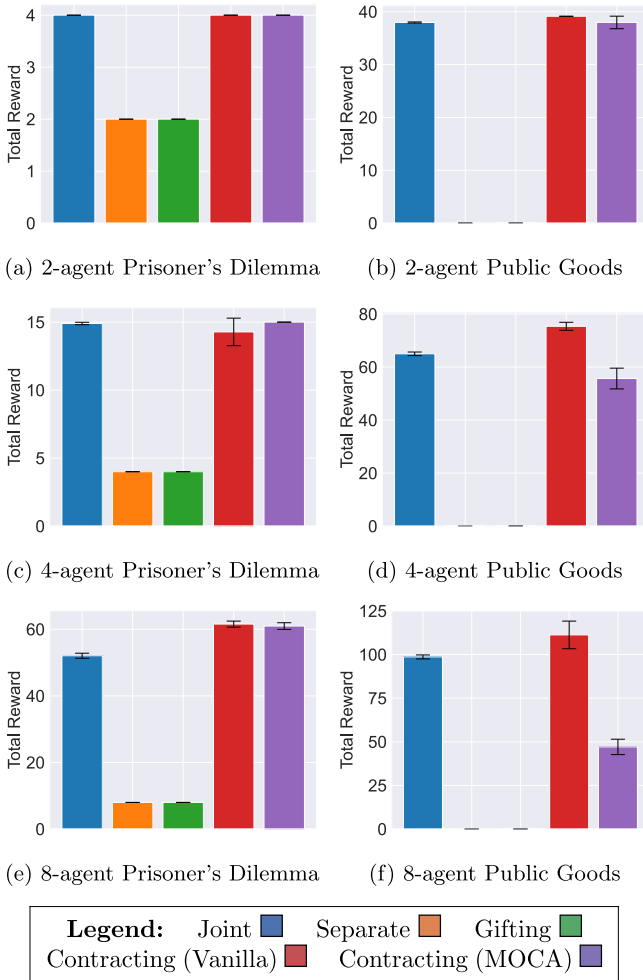
Fig. 9 Experimental results including MOCA, for the simple tabular domains. From left to right, we have joint training (in blue), separate training (in orange), gifting (in green), vanilla contracting (in red), and MOCA (in purple), presented. Every plot is the mean social reward per episode at the end of training (1 M plays of the static games) for each of the 5 algorithmic setups tested. Cells vary across number of agents present (2, 4, 8), environments (Prisoner's Dilemma, Public Goods) with each cell comparing different algorithms. Error bars denote one standard deviation over five independent runs. For Prisoner's Dilemma, MOCA attains higher social reward than both gifting and separate training. However, it fails to match joint training in Public Goods, since this is a domain with simple environment dynamics where learning to respond to contracts is difficult, so early best responses are more likely to be informative, and biasing towards these early on is likely to help performance. In 8-agent prisoner's dilemma, contracting outperforms joint training, which is a result of the large action space of the joint problem
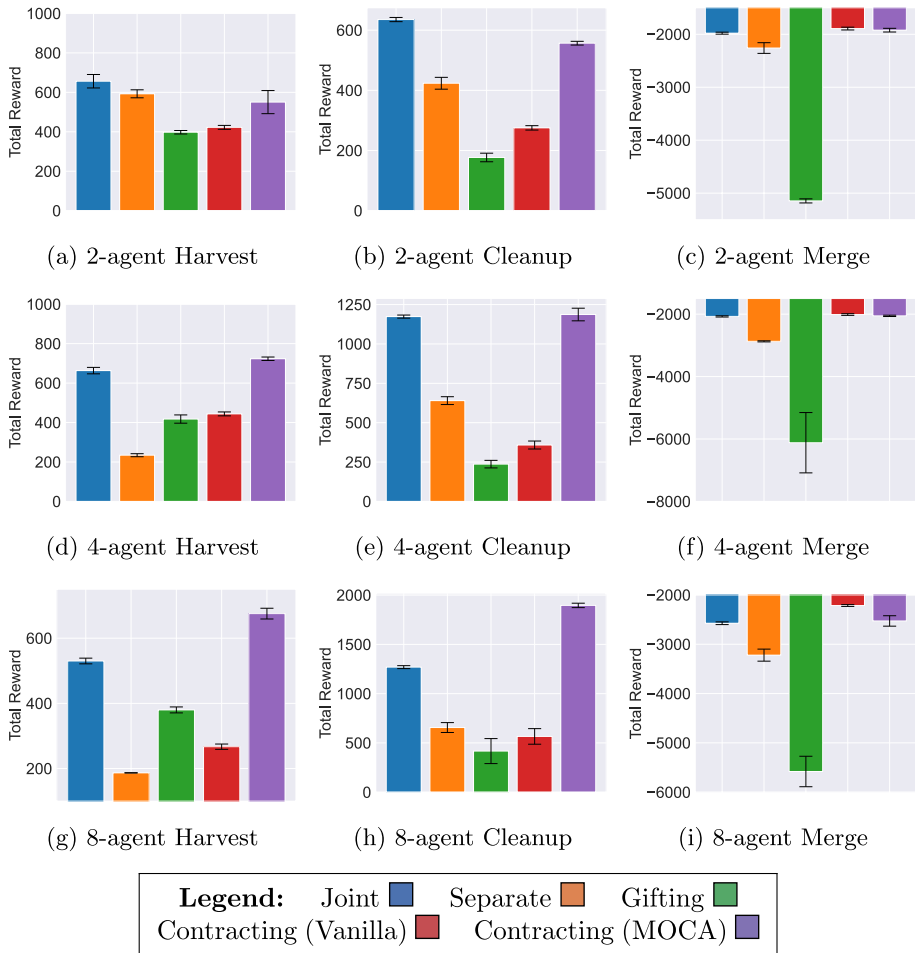
Fig. 10 Experimental results including MOCA, for the more complex dynamic games. From left to right, we have joint training (in blue), separate training (in orange), gifting (in green), vanilla contracting (in red), and MOCA (in purple), presented. Every plot is the mean social reward per episode at the end of training (10 M environment steps) for each of the 5 algorithmic setups tested. Cells vary across number of agents present (2, 4, 8), environments (Harvest, Cleanup, Emergency Merge) with each cell comparing different algorithms (joint, gifting, separate, vanilla contracting, MOCA). Error bars denote one standard deviation over five independent runs. For these more complex dynamic games, contracting leads to higher social reward than gifting and separate training, and always at least matches that of joint training, except for 2-agent Harvest and Cleanup, where sufficient resources are available to make these very mild social dilemmas, as the higher welfare resulting from separate training compared to joint training shows. In Emergency Merge, both vanilla contracting and MOCA contracting significantly outperforms separate training. In several domains, contracting outperforms joint training, which is a result of the large action space of the joint problem

Algorithm 1 Multi-objective contract-augmentation learning (MOCA)

> **Data:** Contract Space Θ including the null contract $\mathbf{0}$, Markov Game M,
>   probability distribution $P(\Theta)$
> **Result:** Policy Profile π
> $\pi \leftarrow$ initialize_policies();
> **for** $t = 1$ **to** $\frac{9}{10}$ num_episodes **do**
> | $\theta \sim P(\Theta)$;
> | train_subgame_episode($\pi(s_0, \theta)$)
> Freeze $\pi|_{S \times \Theta}$;
> **for** $i = 1$ **to** $\frac{1}{10}$ num_episodes **do**
> | $\theta \sim \pi_i(i, \mathbf{0})$;
> | **if** rand() $< \prod_{j=1}^{n} \pi_j(i, \theta)$ **then contract** $\leftarrow \theta$;
> | **else contract** $\leftarrow \mathbf{0}$;
> | $\mathbf{R} \leftarrow$ sample_episode_reward($\pi, \mathbf{contract}$);
> | train_with_rewards($\pi, \mathbf{R}$);
> **return** π;

We evaluate the performance of the final trained algorithm on rollouts. The choice of length of the two periods (e.g. the $\frac{9}{10}$th for the first phase) is arbitrary—what matters is that both phases of training get adequate samples for learning to (1) converge in the subgame conditioned on a choice of contract in all cases, and (2) converge in the subgame of contract proposal, before playing the base Markov game. For the study in Sect. 6.3, we only run the second stage for 1% of the environment timesteps of the first, which was sufficient for close to optimal performance. Failure to do so in the first case leads to bad best responses to a given choice of contract, which may negatively affect social welfare as agents no longer respond (subgame) rationally to the proposed contract. Failure in the second case means the proposer chooses a suboptimal contract, leading to potentially suboptimal welfare.
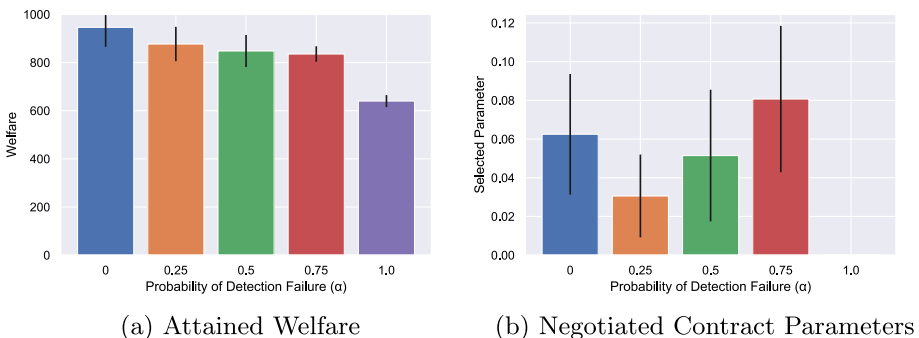
5.5 Monotonicity

Finally, in addition to assessing the performance of contracting (vanilla and MOCA) on these domains, we also wish to empirically validate the monotonicity analysis for policies shown in Sect. 4. To study this, we considered Cleanup with 4 agents, and explored whether or not Proposition 6 held in a complex, dynamic game. Recall that our choice of contract for this domain depends on agents being compensated for cleaning a river. To study feature monotonicity, we take the same underlying contracting logic, but introducing a probability $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ that a given cleaning action is unobserved by the contract. This is equivalent to keeping Θ constant, but altering O_0, in the contracting model we used earlier for the Cleanup domain. Notice that $\alpha = 0$ represents is equivalent to the contracting experimented with earlier, and $\alpha = 1$ is equivalent to the separate training baseline. Our theory would suggest that, because the 1-0 random features for detection of contracts form a Markov chain with decreasing α, we ought to have decreasing welfare as α increases.

6 Results

We first present a sample of our experiments with our baselines, which motivate the need for MOCA (Algorithm 1), in Figs. 7 and 8. Then, we discuss overall trends from experiments including MOCA, Figs. 9 and 10. Finally, we discuss results analysing the performance of contracting under randomized features, as compared to predictions from theory, as seen in Fig. 11.

6.1 Contracting using deep reinforcement learning in the contracting augmentation

Consider first Figs. 7 and 8. We observe that, in Prisoner's Dilemma and Public Goods, the baseline implementation of contracting is sufficient to achieve optimal or near-optimal performance, as can be seen by contracting either matching or surpassing the social welfare of training all agents jointly, and vastly surpassing the welfare of both gifting and separate training (both of which converge to socially suboptimal Nash equilibrium welfare). However, in more complex domains, such as Cleanup, this ceases to be the case. One potential reason for this is that, in these domains, learning the best responses to contracts becomes much more challenging, and so estimates of value for given contracts are less reliable early in training. Therefore, the proposing agent may benefit from additional exploration of the space of contracts, the main feature of MOCA. As seen in Figs. 9 and 10, MOCA again attains higher social welfare than joint training, separate training, and gifting. However, since intermediate levels of reward are not directly comparable with the baselines (since contracts are randomly sampled in the first stage of training, and are not run for the same number of timesteps in the second stage), MOCA is omitted from Figs. 7 and 8. For this,

(a) Attained Welfare   (b) Negotiated Contract Parameters

Fig. 11 Experimental validation of contract feature monotonicity. To evalutate the theory of Sect. 4, we study the performance of MOCA contracting after introducing a probability α of failing to detect contractible behaviour. This frames the detection of contractible behaviour as a *random feature*, and our contracts can be understood as contracting on such random features. Viewed in this way, we can predict from Proposition 6 that welfare should decrease with increasing α, which is exactly what we observe in Fig. 11a. The decay, however, is far from linear, as we see a plateau for middle values of α, and a sharp drop from $\alpha = 0.75$ to $\alpha = 1.0$. Moreover, our theory does not predict anything about which parameter is selected at equilibrium, and none in particular emerges from this experiment, as seen in Fig. 11b. This experiment was run on Cleanup with 4 agents, with the contractible behaviour being the agent cleaning the river, and was averaged over 5 seeds, the error plotted being one standard deviation. Only 1% of environment steps used in the first stage are used in the second stage here – see Appendix C for details

results are presented in Figs. 9 and 10 with bar plots summarizing welfare at the end of training, for all evaluated methods.

6.2 Multi-objective contracting augmentation

Now, we take a closer look at the full results in Figs. 9 and 10. In the simpler domains (Fig. 9), MOCA, like vanilla contracting, attains higher social welfare than for separately trained agents and agents trained with gifting. In Prisoner's Dilemma, contracting reaches joint optimality for 2, 4, and 8 agents. A smaller action space (and hence easier exploration) is a potential reason for why contracting can perform *even better* than joint training, since the action space for joint training grows exponentially in the number of agents. In Public Goods, especially for higher number of agents, joint training interestingly outperforms MOCA, but not vanilla contracting. One possible reason for this is that, uniquely in our suite of environments, learning best responses to each contract is challenging, while the socially optimal policy is itself trivial to execute. Therefore, early in training, it is likely that socially optimal play is learned as a response to some of the contracts, particularly for those θ which are near-optimal. Therefore, biasing contract exploration early on is good for performance. In complex environments, since the socially optimal contracts are harder to execute, early biasing of contract exploration is unlikely to be well-informed, and so converging onto a poor contract proposal algorithm is likely in vanilla contracting at scale.

In the more complex domains (Fig. 10), MOCA in almost all cases at least matches the level of social welfare of joint training, and often far exceeds it. The exceptions to this trend are 2-agent Harvest and 2-agent Cleanup—in both cases, joint training and separate training exhibit strong performance (the former performing comparably to MOCA in both cases, the latter only for Harvest). The reason for this is that the two-agent settings for this problem are not strong social dilemmas, since the agents do not interact enough to create the social dilemmas, either by overconsuming the same patch of land in Harvest, or by consuming each other's apples in Cleanup. Notably, this trend even applies in cases where the vanilla contracting fails (particularly in Harvest and Cleanup), motivating MOCA. In the Emergency Merge domain, MOCA, contracting with a standard PPO training, and joint training, all perform similarly, given the fact that this is a substantially lower-dimensional, and simpler in terms of best-response policies, than Harvest or Cleanup.

6.3 Experiments on monotonicity

From our theory, we would predict that the average welfare attained by contracting, if agents have learned to enact subgame perfect equilibria in the contracting game, should decrease as we inject noise into the contracting process. Indeed, we find that this happens, see Fig. 11a. Specifically, we find that, in Cleanup with 4 agents, as we increase the probability that agents cleaning the river go unnoticed in the contract, the lower that attained social welfare becomes when learned with MOCA. However, we note that the monotonicity we see in this dynamic setup is not uniform: we see a significant drop when adding any noise at all, then a relative plateau for middle values of α, and a sharp drop when going from $\alpha = 0.75$ to $\alpha = 1.0$. However, our theory doesn't say anything about which contracts are actually selected in each case, and as we see in Fig. 11b, no strong pattern emerges between α and contract parameter. Moreover, there seems to be a strong amount of variance regarding the precise selected contract at all levels.

7 Related work

We review related work in Computer Science and Economics.

Our work aims to mitigate social dilemmas in games. In addition to classical static social dilemmas such as Prisoner's Dilemma (compare [39]) and a public goods game (compare [17]), we also consider more complex social dilemmas such as the Harvest and Cleanup domains of Hughes et al. [14]. Cooperation and prosociality in complex domains are of keen interest to MARL researchers, with related challenges including dilemmas like Gathering and Wolfpack [21], the StarCraft challenge [33], or more recently with MARL results in Diplomacy [20].

We relate to the study of augmentations of Markov games to enhance pro-social behavior. Gifting [23, 42] is an augmentation expanding agents' action spaces to allow for reward transfers to other agents. Wang et al. [42] prove that gifting is unable to change the set of Nash equilibria of the underlying game. Our approach differs from gifting in that contracting forces commitment to a given modification of reward *before* taking an action in the original game, and that this commitment is binding for the length of time the contract is in force. This improves total welfare by incentivizing play that was not a Nash equilibrium in the unaugmented game. Relatedly, contracts in the sense of Hughes et al. [15], which we will call *binding contracts*, are contracts that force certain actions and do not allow for reward transfers. They will only ever be enacted when all agents are made better off when a contract is accepted, greatly limiting the effectiveness of the system design. For example, no agent will ever consent to a binding contract cleaning trash in Cleanup. Further, since binding contracts are action-level contracts, the system designer would have to manually encode a "clean trash" policy, instead of relying on a reward signal to *incentivise* them to clean. Other work augments Markov games with auxiliary rewards to incentivize socially beneficial outcomes, without using $\sum R_i$ directly, which can be unstable. For instance, Radke et al. [31] uses multi-task learning to examine the effect of trading off individual, small team, and overall social objectives in designing proxy rewards for agents in a multiagent system. They find that considering auxiliary rewards comprised of combinations of these can be more effective surrogate objectives than pure social welfare in promoting prosocial outcomes—we achieve a similar effect by designing contract spaces which make credit assignment easier for learning prosocial behaviour in MARL.

A large class of prior work has considered different forms of agent commitments. Smith [36] proposes the contract net protocol, which allocates tasks among agents and commits them to complete a particular task. Hughes et al. [15] is a recent study in this line of work. The paper considers multi-agent zero-sum games in which agents may give other agents the option to commit to taking a particular action. Hughes et al. [15]'s contracts— as opposed to our concept of *formal contracts*—might be insufficient to induce the desired joint behavior, as agents will only commit to actions which improve their own individual welfare over the basic game, leading to Pareto optima, which may not be jointly optimal. Han et al. [11] considers the idea of commitments, without transfers, in evolutionary game theory. McAleer et al. [24] lets an auxiliary agent propose a Pareto-optimal equilibrium in a game. Sandholm and Lesser [34] suggests endowing agents with the ability to decommit from a task and paying a side payment. Our approach can be seen as "soft commitment" in which agents always only incur a cost in terms of reward when taking different actions, but are not forced to take a particular action. Sodomka et al. [37] considers the proposal of commitment and side payments, and reaches social cooperation in 2-player games. We provide a general approach that only considers reward transfer without the need to commit

to actions. The idea of negotiations between rounds to arrive to a commitment to an action, was considered in De Jonge and Zhang [4]. Formal contracting does not have the dynamic structure of a negotiation and lets a proposing agent make a take-it-or-leave-it offer. Other related publications study the emergence and learning of social norms [19, 41] as well as conventions [18]. In contrast to the present paper, these do not require explicit consent by agents.

Another related paradigm to ours is Stackelberg Learning. In such models, typically, a special agent, the principal, optimizes incentives for other agents in a bi-level optimization problem. Stackelberg learning has received a lot of attention in strategic Machine Learning [12, 25, 44] and has been used to learn large scale mechanisms such as auctions [3, 6]. Also, Yang et al. [43]'s approach of learning to incentivize other agents may be seen as an instance of Stackelberg Learning. Monderer and Tennenholtz [26] considers a mediator's problem in Normal-Form Games. In contrast to Stackelberg Learning, the focus of formal contracts is that no additional agents—i.e. Stackelberg leaders—are introduced into an environment, but instead contract-proposing agents are embodied in the environment.

Formal contracts have been considered as an alternative to relational contracts in the fields of organizational Economics, see [10] and [8, 5.2.3]. The setting of an agent proposing state-dependent reward transfers has received considerable attention in contract economics, compare the literature following [27], and mechanism design, compare [16, 28, 40]. In our proof of Theorem 1, we use a *forcing contract*, compare [13].

8 Discussion

We expand on the assumption that a single agent proposes a contract, remark that is crucial for our results, and examine its resulting fairness implications, in Sect. 8.1. We discuss approaches to scaling formal contracting to more complex domains in Sect. 8.2. We discuss fairness concerns for contracting in Sect. 8.3.

8.1 Limitations for formal contracting

One crucial assumption in our analysis is that a single agent proposes contracts. Game-theoretic analysis, given in Appendix B.1, shows that if two or more agents are allowed to propose in a game, SPEs may be socially suboptimal. The intuition is that an agent i may choose a contract to affect the state distribution in a way that gives the agent a rejection reward when j proposes a contract, hence increasing their reward when contracting. Our game-theoretic analysis also showed that unfair outcomes might result from contracts. Insofar as proposal by different agents and joint optimal behavior are incompatible, system designers who would like to ensure fairness need to design contracts in a way that limit the number of transfers that can be made, potentially at the expense of welfare.

8.2 Scaling formal contracting

The clearest avenue for future work is in scaling contracts to more realistic domains. Here, we outline three ways to do that.

First, contracts in this work were hand-engineered with relevant internal logic, in order to make the transfers a useful signal. For this approach to scale, a complexity tradeoff must be managed: Contracts need to be flexible enough to extract enough relevant information

to incentivize welfare-optimal play, while being simple enough for MARL agents to allow fast learning of which contracts to choose and accept. Domain-independent techniques for designing contract spaces would greatly improve the scalability of the method.

Manually managing this complexity tradeoff is undesirable. In particular, there may be domains whose social dilemmas are not transparent, or where it might be hard to design concise yet powerful contractible observation models. Therefore, learning which aspects of a state are useful for contracting will allow us to scale the approach to more realistic scenarios while keeping contract space sizes low.

Even with a fixed contract space, sample efficiency may be improved. MOCA took a first step into improving contract learning, by decreasing the bias in estimated $V_i^{\pi}(s_0, \boldsymbol{\theta})$ values. MOCA outperformed benchmarks in all, even complex, dynamic, environments that exhibited a social dilemma. Increasing sample efficiency will allow the use of formal contracting to mitigate social dilemmas in even more complex domains. One potential way to increase the sample efficiency of the first phase of MOCA is to leverage more of the literature on multi-task reinforcement learning methods [5, 30, 32, 38].

8.3 Fairness

One striking observation in the proof of Proposition 3 is that agents $j \in [N] \setminus \{i\}$ are indifferent between accepting the contract θ^* and not accepting it. Hence, the contract may lead to an improvement in welfare relative to $\boldsymbol{\pi}$, but no agents except agent $i = 1$ gets any benefit from this improvement. In many decentralized learning tasks, this is not of concern. For example, if a robot fleet needs to coordinate which locations to service, low reward for some of the agents is not of concern, as long as system performance (that is, welfare) is high. In others, this behavior may be unfair. Trading off fairness and welfare is an avenue for future work.

9 Conclusion and future work

To close, in this work, we propose a novel augmentation to POMGs, termed formal contracting, which helps mitigate the occurrence of social dilemmas that may arise as a result of conflicting agent incentives. This augmentation can be applied in a modular way, to any POMG, requiring access only to the reward function, which we modify only with zerosum transfers (hence leaving welfare unaffected in all states). Theoretically, we show that this augmentation carries several desirable theoretical properties: (1) contracting in settings when deviations from prosocial policy profiles are detectable have only welfare-maximizing SPEs, and (2) whenever a contracting space admits arbitrary unconditional transfers, contracting can only improve SPE welfare compared to the original POMG, and improves as contracting spaces get richer (in the sense of an order $\prec_{ETE}$ we define). We discuss the implications of these results in several settings and discuss the relationship between contracting and decentralized Markov processes.

We also demonstrate the power of contracts in tabular games (Prisoner's Dilemma and Public Goods), and more complex sequential-move games (Harvest, Cleanup, and a novel self-driving car domain Merge). We propose a specialized algorithm for contracting, Multi-Agent Contract Augmentation Learning, and matches or exceeds the welfare of training all agents with one controller. We also verify that our theoretical predictions hold. Contracts

with more observability into agent deviation attains higher welfare. In a discussion, we discuss scaling contracting, fairness concerns, and generalizability of our theory.

There are several avenues for future work, of which we highlight three: one is locality of contracting, which refers to contracts that are only binding for subsets of the observation space. A second direction is to allow contract spaces to change over time and be adjusted by agents. Finally, contracting in the presence of collusion by agents is an avenue for future work—both our theoretical guarantees and empirical experiments are in a setting where agents have selfish incentives but do not collude to take advantage of other agents.

Omitted proofs

We provide the proof of Theorem 2, from which Theorem 1 is a direct consequence.

Proof of Theorem 2 Consider any $\Theta \supseteq \Theta_{\text{full}}$. Fix an arbitrary socially optimal policy profile $\boldsymbol{\pi}^*$, and consider a forcing contract θ^* which punishes any deviation from $\boldsymbol{\pi}^*$ by $-\frac{R_{\max}}{\gamma^T(1-\gamma)}$, where T is the time that a deviation is detected. Let the signing transfer for θ^* be such that all payoff gain for a non-proposer due to θ^* over no contract is transferred to the proposer. Moreover, let $\boldsymbol{\pi}_0$ be an SPE of the unaugmented Markov game M. Note that the punishment is a larger negative reward than any reward that can be achieved in the game from the outset. Hence, after acceptance of the contract, $\boldsymbol{\pi}^*$ constitutes an SPE. By construction, after rejection of the contract, it also constitutes an SPE. By definition, all agents but the proposer are indifferent between accepting and rejecting the contract. In particular, accepting it is an SPE. It remains to show that the proposer cannot get more payoff than under θ^*. All other agents need to get at least $V_i^{M,\boldsymbol{\pi}_0}(s_0)$ to accept a contract, it is impossible for the proposer, in any contract following a contract that is accepted, to get more welfare than

$$W^{\boldsymbol{\pi}^*}(s_0) - \sum_{i=2}^{N} V_i^{M,\boldsymbol{\pi}_0}(s_0).$$

This is exactly like the payoff that θ^* yields.

It remains to show that all SPE must lead to a jointly optimal policy profile in $M + \theta$ for the contract proposed by θ. First observe that it cannot be optimal for the proposer to propose a contract that wouldn't be accepted (proposing θ^* would yield higher payoff). Hence, only accepted contracts are proposed at SPE. Note that if a jointly sub-optimal policy profile $\boldsymbol{\pi}'$ would be played following the proposed contract θ, the proposer payoff would be bounded by

$$W^{\boldsymbol{\pi}'}(s_0) - \sum_{i=2}^{N} V_i^{M,\boldsymbol{\pi}_0}(s_0) < W^{\boldsymbol{\pi}^*}(s_0) - \sum_{i=2}^{N} V_i^{M,\boldsymbol{\pi}_0}(s_0),$$

which is lower than what the proposer would get under θ^*. $\square$

Proof of Proposition 3 As in the proof sketch, we only consider the first inequality in depth here only: the upper inequality follows by exchanging WCSPW for BCSPW, "worst case subgame-perfect equilibrium" with "best case subgame-perfect equilibrium", changing arg min into arg max, and exchanging inequalities in the second part of the proof and the section headings.

We first show the second claim – that is, assuming θ^* and π^* are attained, we wish to construct an SPE attaining $\mathcal{W}(\mathcal{C})$. Fix a Markov game M, a contracting space Θ that admits arbitrary unconditional transfers to augment it and an arbitrary observation model O_0, giving contracting model $\mathcal{C}$. As the supremum in (3) is attained, there is an SPE attaining θ^* in $M^{\mathcal{C}}$. Let $\theta^* \in \arg\max_{\theta \in \Theta} \text{WCSPW}(\theta)$, and

$$\pi^* \in \arg\min_{\pi \in \text{SPE}(M+\theta^*)} W^{\pi}(s_0)$$

Then, take

$$\pi \in \arg\min_{\pi \in \text{SPE}(M)} W^{\pi}(s_0)$$

Moreover, define the following vector

$$\mathbf{v} = \begin{pmatrix} \sum_{i=2}^{N} V_i^{M+\theta^*,\pi^*}(s_0) - V_i^{M,\pi}(s_0) \\ V_2^{M,\pi}(s_0) - V_2^{M+\theta^*,\pi^*}(s_0) \\ \vdots \\ V_N^{M,\pi}(s_0) - V_N^{M+\theta^*,\pi^*}(s_0) \end{pmatrix}$$

and the following contract

$$\theta_{\text{tr}}^*(o) = \begin{cases} \mathbf{v} & o = \text{acc} \\ \theta^*(o) & o \neq \text{acc} \end{cases}$$

Since θ admits arbitrary unconditional transfers, we have $\theta_{\text{tr}}^* \in \Theta$. Then, the following is an SPE: first, the proposing agent proposes θ_{tr}^*, all agents accept this contract: for other contract choices, agents accept or reject based on whether the expected value under worst-case SPE in $M + \theta$ is higher than $V_i^{M,\pi}(s_0)$ (and accepting in the case of a tie). Once contracts θ are selected, in all subgames $M + \theta$, agents play the worst-case SPE in $M + \theta$.

To argue this is an SPE, we proceed by backwards induction: once contracting is accepted or rejected, each agent plays a worst-case SPE, and moreover agents decide to accept or reject contracts appropriately according to the expected value under both. Then, it simply suffices to argue that θ_{tr}^* gives the proposer maximal value amongst all choices of contract, given the responses of other agents subsequently. Upon playing θ_{tr}^*, agents have their value under the contract extracted exactly to the point of indifference ($V_i^{M,\pi}(s_0)$), and hence accept the contract. They then play π^*, given that $M + \theta_{\text{tr}}^*$ has exactly the same SPE as $M + \theta^*$, as all comparisons of value are unaffected by a constant shift of reward in all episodes. The proposing agent therefore gets payoff:

$$V_1^{M+\theta^*,\pi^*}(s_0) + \left(\sum_{i=2}^{N} V_i^{M+\theta^*,\pi^*}(s_0) - V_i^{M,\pi}(s_0) \right) = W^{\pi^*}(s_0) - \sum_{i=2}^{N} V_i^{M,\pi}(s_0)$$

This is maximal: suppose for contradiction a contract giving strictly higher payoff for the proposer was available in θ', which leads to equilibrium profile π'. If the proposer proposes a contract rejected in the SPE, then no contract is implemented and agents play π in M, giving proposer value

$$V_1^{M,\pi}(s_0) = W^{\pi}(s_0) - \sum_{i=2}^{n} V_i^{M,\pi}(s) \leq W^{\pi^*}(s) - \sum_{i=2}^{n} V_i^{M,\pi}(s_0).$$

Therefore, for a contract θ' which is accepted, and with worst-case equilibrium π' played in $M + \theta'$, transfers from non-proposing agent $i > 1$ to the proposing agent cannot be greater than $V_i^{M+\theta',\pi'}(s_0) - V_i^{M,\pi}(s_0)$. Further, the proposing agent has a reward of (1) its own value under the game, plus (2) the time-discounted transfers through the game $M + \theta'$. From this:

$$V^{\mathcal{C}}((1,\theta')) = V_1^{M+\theta',\pi'}(s_0) + \mathbb{E}_{\pi'}\left[\sum_{t=0}^{\infty} \gamma^t \theta'_1(o_{0,t})\right] + \theta'_1(\text{acc})$$

$$\leq V_1^{M+\theta',\pi'}(s_0) + \left(\sum_{i=2}^{N} V_i^{M+\theta',\pi'}(s_0) - V_i^{M,\pi}(s_0)\right)$$

$$= W^{\pi'}(s_0) - \sum_{i=2}^{N} V_i^{M,\pi}(s_0)$$

$$\leq W^{\pi^*}(s_0) - \sum_{i=2}^{N} V_i^{M,\pi}(s_0).$$

This establishes that the policy profile here is indeed an SPE for the extended game $M^{\mathcal{C}}$.

Now on to the first claim. Let π be any SPE for $M^{\mathcal{C}}$, π_0 the equilibrium of M that is played absent a contract in force. Denote π_θ the profile of sub-policies following the proposal of contract θ. We show that π must achieve at least welfare $\mathcal{W}(\mathcal{C})$.

Let $\epsilon > 0$. As the contract space admits arbitrary unconditional transfers, for any choice of contract $\theta \in \Theta$, the proposer can attain, by selecting some $\theta' \in \Theta$ and unconditionally transferring welfare at SPE to itself,

$$V_1^{\mathcal{C}}((1,\theta')) \geq \text{WCSPW}_{\mathcal{C}}(\theta) - \sum_{i=2}^{N} V_i^{M,\pi_0}(s_0) - (N-1)\epsilon \tag{4}$$

where each agent gets $\epsilon > 0$ beyond their non-contract value as an incentive to accept the contract. Under such θ',

$$V_i^{\mathcal{C}}((1,\theta')) = V_i^{M,\pi_0}(s_0) + \epsilon$$

for $i = 2, 3, \ldots, N$. Hence, as $\epsilon > 0$ can be arbitrarily small, substituting in and rearranging (4), we get

$$W^{M+\theta',\pi_{\theta'}}(s_0) \geq \text{WCSPW}_{\mathcal{C}}(\theta). \tag{5}$$

Now, for all θ_{SPE} in the support of SPE π, these must be a best response for the proposer among all choices of contract. In particular, they must be better than any choice of θ' – hence, substituting (5), supremizing over Θ, and letting $\epsilon \to 0$ gives

$$W^{M+\theta_{SPE},\pi_{\theta_{SPE}}}(s_0) \geq \sup_{\theta \in \Theta} \text{WCSPW}_{\mathcal{C}}(\theta) = \mathcal{W}(\mathcal{C}),$$

which proves the claim.   □

Generalizations of contracting

Repeated proposals

Next, we prove that the statements of Theorem 1 continue to hold if we have an agent i that repeatedly proposes contracts. For a Markov game M, we say that a state $s \in S$ is unreachable from state s_0 is there is no sequence of action profiles $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{t-1} \in \mathbf{A}$ such that $s_t = s$.

Proposition 8 *Consider a general formulation in the sense of Appendix B.2 such that there is an agent i such that $(\mathbf{0}, j)$, $j \neq i$ is unreachable from $(\mathbf{0}, i)$, and that $(s, \mathbf{0})$ is unreachable from $(s', \boldsymbol{\theta})$ for any $s, s' \in S, \theta \in \boldsymbol{\theta}$. Then, the conclusions of Theorem 1 continue to hold.*

Proof First, observe that the strategies in which agent i repeatedly proposes a contract θ^* as in the proof of Theorem 1 is an SPE: At each proposal and acceptance state, the agent can assume that the value is as if the agent proposes an efficient contract in the next proposal state. As in previous proofs more informally, we invoke the single-deviation principle for SPE (compare [29]) stating that to check an SPE, it is sufficient to only consider deviations by one agent at a time. First, observe that there is no reason for the proposer to propose another contract to the other agents, as the upper bound on value demonstrated in Theorem 1 continues to hold. The forcing contracts still exist and incentivize all agents to follow $\boldsymbol{\pi}^*$. Finally, rejection of all contracts is again not possible as the proposing agent's best response would not be defined. Hence, the statement continues to hold for a repeatedly proposing agent.

The uniqueness of such an equilibrium follows equivalently, as an upper bound on the proposal agent is exactly attained at each proposal point. □

Finally, we show by means of an example that this statement breaks if multiple agents may propose a contract.

Example 2 Consider a repeated Prisoner's Dilemma in which agent 1 may propose as long as (C, D) has been played. Then, agent 2 proposes thereafter. Consider Prisoner's Dilemma, Table 1a. An SPE of this game involves proposal of a forcing contract on (C, D) with values after a signing bonus that leads to values $(-2 + \varepsilon, -1 - \varepsilon)$. All contracts that give agent 1 more value are rejected. It is a best response for agent 2 to accept, as long as $(-1 - \varepsilon)/(1 - \gamma)$, the value of accepting this contract, is larger than $-2 + 0 \times \gamma/(1 - \gamma) = -2$. The best such contract is the one with the largest ε for which this inequality holds.

Conditional proposals

In this section, we present the augmentation for several proposing agents. Let $M = \langle S, s_0, \mathbf{A}, T, \mathbf{R}, \gamma \rangle$ be an N-agent Markov game, $\boldsymbol{\theta}$ be a contract space as before. To generalize, we require one additional component: define the *contracting initiation dynamics* to be a function

$$P : \{(\boldsymbol{\theta} \cup \{\mathbf{0}\}) \times S \times \mathbf{A}\} \cup \{\text{init}\} \rightarrow \Delta([N] \cup \{\mathbf{0}\} \cup \{\bullet\})$$

which determine whether or not a new contract phase begins or ends at a given state of M, and if one begins, which agent is proposing the contract. More specifically, sampling from $P(\text{init})$ at the start of an episode, or from $P(\theta, s, a)$ at state-action $(s, \mathbf{a})$ under contract θ, either:

1. Agent i is given the option to propose a new contract, and the game is frozen ($i \in [N]$);
2. None of the agents proposes a contract, but the current contract stays in force, and the game continues ($\mathbf{0}$), or;
3. The current contract becomes void, and no agent proposes a new contract, and again the game continues ($\bullet$)

We again assume that the contracting space θ contains the null contract $\mathbf{0}(s, \mathbf{a}) \equiv 0$. Define the contract-augmented Markov game $M^\theta = \langle S', S_0, \mathbf{A}', T', \mathbf{R}', \gamma \rangle$ as

$$S' = (S \cup (S \times \{\mathbf{0}\})) \times (\theta \cup [N])$$

$$A' = A \cup \theta \cup \{\text{acc}\}$$

$$T'((s, \theta), a) = \begin{cases} T(s, \mathbf{a}) & P(\theta, s, \mathbf{a}) = \mathbf{0} \\ (s, \mathbf{0}) & P(\theta, s, \mathbf{a}) = \bullet \\ (\mathbf{0}, i) & P(\theta, s, \mathbf{a}) = i \end{cases}$$

$$T'(((s, \mathbf{0}), i), (\theta, \mathbf{a}_{-i})) = ((s, \mathbf{0}), \theta)$$

$$T'(((s, \mathbf{0}), \theta), \mathbf{a}) = \begin{cases} (s, \theta) & \mathbf{a} = \text{acc} \\ (s, \mathbf{0}) & \text{else.} \end{cases}$$

$$R((s, \theta), \mathbf{a}) = R(s, \mathbf{a}) + \theta(s, \mathbf{a}).$$

with S_0 being state $((s_0, \mathbf{0}), i)$ (e.g. agent i proposing a contract at the start of an episode) with probability $\mathbb{P}[P(\text{init}) = i]$ and $(s_0, \mathbf{0})$ otherwise. Note that, in contrast to the definition of S' in Sect. 2.2, we use tuples $(s, \mathbf{0})$ more generally to denote cases where state s is frozen in the game while agents are negotiating a contract (which can now happen during the game, and not just at the start). For example, to capture the contracting initiation dynamics of Sect. 2.2 in this general model, we have $P(\text{init}) = 1$ almost surely, and $P(\cdot) = \mathbf{0}$ almost surely for any other state (i.e., contracting is initiated and proposal rights given to agent 1 with certainty at the start of an episode, and that contract is held in force for the remainder of the episode).

More complex augmentations can be constructed in a similar way. Examples of such augmentations are multiple contracts in force at the same time, majority of acceptance as opposed to unanimity, and contract overriding only if new contracts are accepted. Exploring such augmentations is an avenue for future work.

Hyperparameter settings

In this section, we describe the hyperparameters we experimented with in the process of developing the study of Sect. 6.

Optimization details

In all methods, we used a learning rate of $\alpha = 10^{-4}$. Experimentation with $\alpha = 10^{-2}, 10^{-6}, 10^{-8}, 10^{-10}$, as well as a linearly decaying learning schedule from $\alpha = 10^{-2}$ to 10^{-10}, none of which improved performance on any method, the smaller learning rates significantly degrading the sample efficiency of the methods involved.

Training happened using stochastic gradient descent, with 30 SGD updates per training iteration and a momentum of 0.99. For all methods second phase of MOCA, training batches consisted of 12000 sampled timesteps for the static dilemmas, and 120000 sampled timesteps for the dynamic games, all of them with minibatches of size 4092. Other minibatch values attempted for these methods were 128, 256, 1024 and 40000, all of which were found to degrade the performance, the first two very significantly.

However, in training $\pi(s, \theta)$, we restricted ourselves to 10% of timesteps used in training $\pi(0, 0)$ (and the baseline algorithms), and therefore found that sampling 128 episodes per training iteration, and using a minibatch size of 128 (in all domains) was sufficient for strong performance. For the monotonicity study, we used 100,000 timesteps, 1% of total environment steps.

Model architectures

On all algorithms, we experimented with 32×32, 64×64, 256×256, $64 \times 64 \times 64$, $256 \times 256 \times 256$, and 1024×1024 MLP networks, and took the models of minimal complexity attaining maximal performance. For separate training, gifting, and contracting, we found 64×64 MLP networks to be this value, while for joint training, due to the added complexity of the joint policy, we found that expanding the network to a 256×256 MLP architecture yielded the highest performance.

Environment horizon and discount factors

All domains have discount factor $\gamma = 0.99$. Moreover, the maximum horizon for the matrix domains was 2, for the Public Goods Dilemma was 100, for Emergency Merge was 200 (although episodes can terminate earlier by cars reaching the end early), and for both Harvest and Cleanup was 1000.

PPO-specific parameters

We used a standard KL-coefficient of 0.2, KL-target of 0.01, clip parameter of 0.3, value function coefficient of 1.0, entropy coefficient of 0.0, for all experiments in all domains.

Novel hyperparameters

As mentioned in the main text, we introduced a novel hyperparameter ν, governing the number of agents sampled in the accept-reject decision for contracts. In all experiments, we used $\nu = 2$.

Other unmentioned hyperparameters are as found in the default settings in the RLLib implementation of PPO [22].

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Andrychowicz, M., Wolski, F., Ray, A., et al. (2017). Hindsight experience replay. Advances in neural information processing systems. 30
2. Bernstein, D. S., Givan, R., Immerman, N., et al. (2002). The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research, 27*(4), 819–840.
3. Curry, M., Sandholm, T., & Dickerson, J. (2022). Differentiable economics for randomized affine maximizer auctions. arXiv preprint arXiv:2202.02872
4. De Jonge, D., & Zhang, D. (2020). Strategic negotiations for extensive-form games. *Autonomous Agents and Multi-Agent Systems, 34*(1), 1–41.
5. Devin, C., Gupta, A., Darrell, T., et al. (2017). Learning modular neural network policies for multi-task and multi-robot transfer. In: *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, pp 2169–2176.
6. Dütting, P., Feng, Z., Narasimhan, H., et al. (2023). Optimal auctions through deep learning: Advances in differentiable economics. *Journal of the ACM*. https://doi.org/10.1145/3630749
7. Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica, 54*(3), 533–554.
8. Gibbons, R. (2005). Four formal (izable) theories of the firm? *Journal of Economic Behavior & Organization, 58*(2), 200–245.
9. Gibbons, R. S. (1992). *Game theory for applied economists*. Princeton University Press.
10. Gil, R., & Zanarone, G. (2017). Formal and informal contracting: Theory and evidence. *Annual Review of Law and Social Science, 13*, 141–159.
11. Han, T. A., Pereira, L. M., & Lenaerts, T. (2017). Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems, 31*(3), 561–583. https://doi.org/10.1007/s10458-016-9338-4
12. Hardt, M., Megiddo, N., Papadimitriou, C., et al. (2016). Strategic classification. In: *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp 111–122.
13. Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics, 13*, 74–91.

14. Hughes, E., Leibo, J.Z., Phillips, M.G., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. arXiv preprint arXiv:1803.08884.

15. Hughes, E., Anthony, T.W., Eccles, T., et al. (2020). Learning to resolve alliance dilemmas in many-player zero-sum games. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp 538–547.

16. Hurwicz, L. (1973). The design of mechanisms for resource allocation. *The American Economic Review, 63*(2), 1–30.

17. Janssen, M., & Ahn, T. (2003). Adaptation versus anticipation in public-good games. In: *Annual meeting of the American Political Science Association*, Philadelphia, PA.

18. Köster, R., McKee, K.R., Everett, R., et al. (2020). Model-free conventions in multi-agent reinforcement learning with heterogeneous preferences. arXiv preprint arXiv:2010.09054.

19. Köster, R., Hadfield-Menell, D., Everett, R., et al. (2022). Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. In *Proceedings of the National Academy of Sciences* 119(3)

20. Kramár, J., Eccles, T., Gemp, I., et al. (2022). Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nature Communications, 13*(1), 7214.

21. Leibo, J.Z., Zambaldi, V., Lanctot, M., et al. (2017). Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv:1702.03037.

22. Liang, E., Liaw, R., Nishihara, R., et al. (2018). Rllib: Abstractions for distributed reinforcement learning. In: *International Conference on Machine Learning*, PMLR, pp 3053–3062.

23. Lupu, A., & Precup, D. (2020). Gifting in multi-agent reinforcement learning. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp 789–797

24. McAleer, S., Lanier, J., Dennis, M., et al. (2021). Improving social welfare while preserving autonomy via a pareto mediator. arXiv preprint arXiv:2106.03927.

25. Milli, S., Miller, J., Dragan, A.D., et al. (2019). The social cost of strategic classification. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp 230–239

26. Monderer, D., & Tennenholtz, M. (2004). K-implementation. *Journal of Artificial Intelligence Research, 21*(1), 37–62.

27. Mussa, M., & Rosen, S. (1978). Monopoly and product quality. *Journal of Economic theory, 18*(2), 301–317.

28. Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research, 6*(1), 58–73.

29. Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT press.

30. Parisotto, E., Ba, J.L., & Salakhutdinov, R. (2015). Actor-mimic: Deep multitask and transfer reinforcement learning. arXiv preprint arXiv:1511.06342.

31. Radke, D., Larson, K., & Brecht, T. (2022). The importance of credo in multiagent learning. arXiv preprint arXiv:2204.07471.

32. Rusu, A.A., Colmenarejo, S.G., & Gulcehre, C., et al. (2015). Policy distillation. arXiv preprint arXiv:1511.06295.

33. Samvelyan, M., Rashid, T., De Witt, C.S., et al. (2019). The starcraft multi-agent challenge. arXiv preprint arXiv:1902.04043

34. Sandholm, T.W., & Lesser, V.R. (1996). Advantages of a leveled commitment contracting protocol. In: *AAAI/IAAI*, Vol. 1, Citeseer, pp 126–133

35. Schulman, J., Wolski, F., Dhariwal, P., et al. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347

36. Smith, R. G. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers, 29*(12), 1104–1113.

37. Sodomka, E., Hilliard, E., Littman, M., et al. (2013). Coco-q: Learning in stochastic games with side payments. In: *International Conference on Machine Learning*, PMLR, pp 1471–1479.

38. Teh, Y., Bapst, V., Czarnecki, W. M., et al. (2017). Distral: Robust multitask reinforcement learning. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1707.04175

39. Tucker, A. W., & Straffin, P. D., Jr. (1983). The mathematics of tucker: A sampler. *The Two-Year College Mathematics Journal, 14*(3), 228–232.

40. Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance, 16*(1), 8–37.

41. Vinitsky, E., Köster, R., Agapiou, J.P., et al. (2021). A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. arXiv preprint arXiv:2106.09012.

42. Wang, W.Z., Beliaev, M., Bıyık, E., et al. (2021). Emergent prosociality in multi-agent games through gifting. arXiv preprint arXiv:2105.06593.

43. Yang, J., Li, A., Farajtabar, M., et al. (2020). Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems, 33*, 15208–15219.

44. Zrnic, T., Mazumdar, E., Sastry, S., et al. (2021). Who leads and who follows in strategic classifica-tion? *Advances in Neural Information Processing Systems, 34*, 15257–15269.