Auton Agent Multi-Agent Syst (2018) 32:26–58
https://doi.org/10.1007/s10458-017-9372-x

Severity-sensitive norm-governed multi-agent planning

Luca Gasparini[1] · **Timothy J. Norman**[2] ·
Martin J. Kollingbaum[1]

Published online: 7 July 2017

Abstract In making practical decisions, agents are expected to comply with ideals of behaviour, or norms. In reality, it may not be possible for an individual, or a team of agents, to be fully compliant—actual behaviour often differs from the ideal. The question we address in this paper is how we can design agents that act in such a way that they select collective strategies to avoid more critical failures (norm violations), and mitigate the effects of violations that do occur. We model the normative requirements of a system through contrary-to-duty obligations and violation severity levels, and propose a novel multi-agent planning mechanism based on Decentralised POMDPs that uses a qualitative reward function to capture levels of compliance: N-Dec-POMDPs. We develop mechanisms for solving this type of multi-agent planning problem and show, through empirical analysis, that joint policies generated are equally as good as those produced through existing methods but with significant reductions in execution time.

Keywords Norms · Multi-agent planning · Dec-POMDPs

1 Introduction

With increased automation, the need for systems to act in such a way that they are cognizant of normative expectations is critical. Norms declare ideals of behaviour, but they are inherently violable: the actual behaviour of agents may differ from the ideal. Sub-ideal behaviour may,

✉ Timothy J. Norman
t.j.norman@soton.ac.uk

Luca Gasparini
luca.gasparini@live.it

Martin J. Kollingbaum
m.j.kollingbaum@abdn.ac.uk

[1] Department of Computing Science, University of Aberdeen, Aberdeen, UK

[2] Department of Electronics and Computer Science, University of Southampton, Southampton, UK

however, be inevitable. It may not be possible for an agent (or a group of agents) to be fully compliant, given resource limitations. An agent may decide to violate a norm now in order to avoid a more serious violation in the future. A norm may be violated due to an unexpected outcome of a sequence of actions. In fact, inaction may not be sufficient to avoid the violation of a norm: the world may change into a sub-ideal state unless an agent acts. The challenge we address in this paper is how to develop effective reasoning mechanisms for agents such that they operate robustly, both individually and as a collective, under normative expectations. By robust, we mean that the agents act so that they are as compliant with the ideal as possible, given prevailing circumstances.

Any solution to this general problem must take into account uncertainties due to non-deterministic action outcomes, dependencies between and among agents with respect to actions and resources, and environmental changes that are not under their control. There are two other important considerations we take into account: there may be expectations on agents regarding how they should repair, or recover from, non-ideal states of affairs (so called contrary-to-duty obligations); and the violation of norms may vary in severity. The former of these is widely considered to be an important characteristic of real-world domains. Prakken and Sergot [21] use an example derived from regulations about the appearance of holiday cottages to illustrate this: there must be no fence (the primary obligation); and if there is a fence, it must be white (the contrary-to-duty rule). In the case where there is a fence (the primary obligation is violated), it is the duty of the owner to ensure that it is painted white.

The idea that norms (or, strictly, the violation of norms) vary in severity is also widely recognised, but, we argue, often poorly modelled for the purposes of practical reasoning. In computational models, severity is often modelled through pre-defined sanctions [1]. Further, the vast majority of examples used are fines, or some other loss of utility, implying an underlying additive assumption. The argument we present against this rather simplistic approach is grounded, again, on how violations are classified in real-world domains. Distinguishing different qualitative levels of violation is an important principle in law, often referred to as "fair labelling". According to Ashworth [3], for example, fair labelling is (in part) where "offences are subdivided and labelled so as to represent fairly the nature and magnitude of the law-breaking". This is a principle reflected in various legal systems; e.g. misdemeanour versus felony in the US. Similarly, in security contexts, information is often classified in terms of levels (restricted, secret, etc.), representing the idea that the revelation of any document at a higher security classification is always more severe than revealing any amount of information at a lower classification. Of course, revealing any classified information is undesirable, but severity levels give tipping points of compliance. There is an important pragmatic reason that qualitative levels of violation are specified in this way: sanctions are imposed *after the fact* and *given an assessment of the context* in which the norm was violated. All we know in advance (i.e. at the point where we need to make decisions about how to act) is that violations of some norms are more/less severe than others. Further, specifying sanctions for all norms over a single interval scale, equating to some loss of utility, can lead to additive fallacies [18], where some number of violations at a lower level of severity are taken to be as bad as, or worse than one at a qualitatively higher level. Such fallacies would lead to poor practical decisions.

Contrary-to-duty obligations and violation severity provide complementary means to specify requirements for system robustness. The use of contrary-to-duty obligations enables us to reason about behaviour that *goes some way to repair a failure*. The use of severity levels enables us to reason about behaviour that *avoids critical levels of failure* and that *minimises accumulated failures at some level*.

Our starting point is a deontic logic for the specification of normative systems that may contain contrary-to-duty structures [29], along with a strict partial order over obligations that declares the relative severity of their violation. From this, we compute a preference relation over possible worlds that captures levels of system robustness (Sect. 3), which we prove to be both transitive and acyclic. A transitive and acyclic preference relation is necessary for reliable practical reasoning: with this input, an agent can compare worlds and hence possible courses of action for compliance with the normative specification. Next, we propose a novel model of multi-agent planning under uncertainty that is suitable for reasoning about domains with qualitative reward functions such as those representing levels of system robustness. This multi-agent planning mechanism is grounded on Dec-POMDPs [2]: Normative Decentralised Partially Observable Markov Decision Processes, or N-Dec-POMDPs (Sect. 4). We provide an algorithm for computing joint policies that uses a sequence of linear programs that optimise against levels of robustness, iteratively introducing additional constraints at less critical levels until no additional improvement can be found (Sect. 4.2). The analysis of more/less preferred possible worlds is also exploited in the planner through a Most-Critical-States (MCS) heuristic that is used to identify belief states to optimize an N-Dec-POMDP policy towards a more compliant behaviour in a team of agents (Sect. 4.3). We demonstrate through empirical analysis (Sect. 5) that this approach offers significant reductions in execution times (by 50% in the most challenging problem considered) for the N-Dec-POMDP solver with no loss in solution quality. Before moving on to present the two key contributions of this research (in Sects. 3 and 4), we outline a scenario that both illustrates the normative concepts that are core to the model and gives an intuition of the practical reasoning problem we address. We defer our review of related research in norm-governed and preference-based planning, and discussion of the model and possible avenues for future research to Sect. 6.2.

The core contributions we claim of this research are twofold. First, we propose a mechanism to efficiently compute a preference relation over possible worlds from a normative specification that correctly reflects both contrary-to-duty structures and violation severity. Second, we present a novel multi-agent planning model, N-Dec-POMDPs, and associated heuristic, MCS, that can compute effective joint plans given a qualitative reward function, such as one that represents levels of compliance derived from a normative system specification. We, therefore, contribute both to modelling and practical reasoning in normative multi-agent systems, and to algorithms for decentralised planning under uncertainty.

2 Motivating scenario

Consider a force protection scenario in which various agents are deployed to protect critical infrastructure in a harbour. This may involve, among other things, establishing and maintaining a restricted area off shore, around the harbour, through which only authorised vessels may pass. This restricted area is to be continuously monitored by assets such as UAVs, and patrolled by boats. In specifying the norms for this scenario, we first consider the surveillance task. Let us assume we have UAVs and helicopters available, and suppose that, ideally, surveillance should be done by a UAV.

O_1 A UAV must always monitor the restricted area.
O_2 If no UAV is monitoring the area, a helicopter must monitor the area.

Norms O_1 and O_2 capture a contrary-to-duty structure, with the primary obligation being that a UAV is monitoring. If this is violated, we should at least have a helicopter monitoring the area. Now, we can specify what should be done if an unauthorised vessel (boat) is detected by

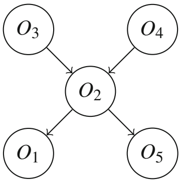


Fig. 1 Partial order over severity of norm violation

an agent conducting surveillance. There are two interventions that we consider: interception and reporting.

O_3 If an unauthorised boat is detected, at least one agent must intercept it.

O_4 If an unauthorised boat is detected and no agent intercepts it, the incursion must be reported to headquarters.

Again, we have a contrary-to-duty structure, this time triggered by detection of an unauthorised boat, where the primary obligation is that at least one agent intercepts it. The 'at least one' part of this obligation clearly introduces a requirement for agents to coordinate. Finally, we have a further obligation on UAVs:

O_5 UAVs should not reveal their location.

Given the focus is on harbour protection, the most severe violation would be not to intervene when an unauthorised boat is detected (violation of O_3 and O_4). The next most severe violation is if the restricted area is not being monitored (violation of O_2). Given this, we can specify an ordering over the severity of norm violation in our scenario (see Fig. 1).

Within the scenario, O_5 (UAV revealing its location) indirectly interacts with decisions regarding which agent intercepts a detected unauthorised vessel: the location of a UAV is revealed if it intercepts a suspicious vessel (a causal constraint). This is, however, one of the least severe violations, and so it would be better for a UAV to intercept an unauthorised boat if it is the only agent available to do so. This is just one example of interactions among normative (violable) constraints on agents' actions, violation severity, dependencies between agents' actions, causal constraints, and stochastic events. In order to bore down on the details of this kind of problem, consider a single UAV and a helicopter in this harbour protection scenario and suppose that an unauthorised boat has been detected in the restricted area. In Fig. 2, we illustrate some of the states that might occur and transitions between them. A transition $\langle\langle\beta,\alpha\rangle, 1.0\rangle$ indicates that there is a joint action $\langle\beta,\alpha\rangle$ where the UAV does β and the helicopter does α, and if they perform this joint action this transition occurs with probability 1.0. A transition $\langle\langle\_,\_\rangle, p\rangle$ indicates that this transition will occur with probability p regardless of the actions of the agents.

Now, suppose that α is the act of intercepting the unauthorised boat, and β is to monitor the restricted area. If, in the initial state, the UAV intercepts the unauthorised boat (does α), then, regardless of what the helicopter does, we reach state **B** ('bad' state) in which the UAV's location is revealed. Subsequent transitions may also mean that the UAV's location is known, or we may return to a fully compliant state (the terminal state in Fig. 2), which we summarise using the transition probability p. What if the UAV chooses to continue to conduct surveillance (action β)? The outcome depends on the actions of the other agent. If the helicopter intercepts the unauthorised boat (joint action $\langle\beta,\alpha\rangle$), then all is well. If not, and the helicopter conducts surveillance, then there is some chance that the system will transition to the state **W** ('worse' state) in which the unauthorised boat is neither intercepted nor reported. Depending on the probabilities p and q in this summarised situation, the likelihood of the system entering state **W** and the proability of multiple violations of O_5 (entering state **B**)

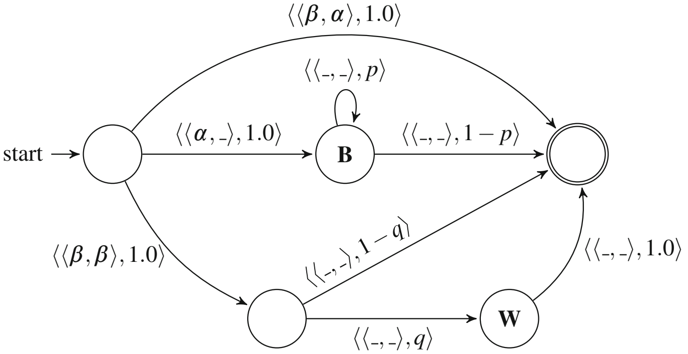


Fig. 2 A simple 2-agent decision problem

may vary. What we want is a mechanism that produces plans for multiple agents such that the qualitative differences between the possible execution paths that the system as a whole may take are taken into account to drive more compliant behaviour. The resulting plans need to provide guidance to agents for situations in which fully compliant behaviour is not possible. In this small example this could occur if, for example, the helicopter is low on fuel, leaving only a choice between a path with 1 or more violations of O_5 versus violation of O_3.

3 Levels of robustness in normative system specifications

Given a normative specification, such as the one described in the previous section, we need to identify how *compliant* each state of affairs is so that we can guide the planning process. In essence, our aim is to compute a preference relation over possible worlds that reflects the level of compliance of those worlds with a set of norms. For example, we want worlds in which O_5 is violated (**B** in Fig. 2) to be preferred to those in which O_3 is violated (due to the severity specification), and worlds in which O_3 is violated to be preferred to those in which O_4 is violated (due to the contrary-to-duty structure linking O_3 and O_4). We then use this preference relation to build a ranking of possible worlds, which allows the space of possible worlds to be partitioned into different severity ranges. We first present the semantics of our model and define the notions of *compliance* of a world with a norm and *coherence* between an obligation and a pair of worlds. With these definitions, we specify a transitive and acyclic preference relation, P_W, and present a method to efficiently compute a ranking of possible worlds from this relation.

3.1 Normative system semantics

Our semantics is inspired by Prohairetic Deontic Logic (PDL) [29], and other preference-based deontic logics, such as that proposed by Prakken and Sergot [22], where dyadic (conditional) obligations are represented through a preference relation between worlds. These logics rely on a normative specification that is free from logical conflict, which is adequate for our purposes because we are interested in effective multi-agent decision making mechanisms in the presence of *functional* conflicts [10]; i.e. conflicts between a consistent set of social norms and the actions of agents in a non-deterministic environment. The key advantage of these logics, however, is that they adequately capture the concept of contrary-to-duty obligations, avoiding well-known paradoxes of normative reasoning [11,14]. In addition to a set of norms, we declare a strict partial order over these norms that represents the relative severity of their violation. We, therefore, define a model $M = \langle W, PV, VA, OS, P_o \rangle$[1] where:

- $W = \{w_1, \ldots, w_i, \ldots, w_n\}$ is a set of n possible worlds.
- PV is a set of propositional variables, and ϕ, ϕ_2 denote individual propositions. The set of well formed formulae of propositional logic, F, is such that $PV \subset F$, and if $p, q \in F$ then $\neg p \in F$, $p \wedge q \in F$, etc.
- $VA : W \rightarrow 2^{PV}$ is a valuation function that assigns, to each world $w \in W$, the set of propositional variables that hold true in w.
- $OS = \{O_1 = \mathbf{O}(p_1 \mid q_1)), \ldots, O_m = \mathbf{O}(p_m \mid q_m)\}$ is a normative specification, where p_i and q_i are two formulae in F. Intuitively, $\mathbf{O}(p_i \mid q_i)$ represents a dyadic obligation to achieve (or maintain) p_i that applies to worlds in which q_i holds: an obligation to achieve p_i that is *conditional* on q_i.
- $P_o \subseteq OS \times OS$ is a strict partial order over obligations that reflects the relative severity of their violation. Given two obligations O_i and O_j, $(O_i, O_j) \in P_o$ (or alternatively $O_i \succ_o O_j$) means that a violation of O_i is considered more severe than one of O_j. P_o is a transitive relation, thus, if we consider a graph G, where each node represents an obligation, and each edge a member of P_o, we say that violating O_a is more severe than violating O_b if and only if the node representing O_b is reachable from O_a through the edges of G.

Propositional logic formulae are evaluated as usual over possible worlds. Given a world $w_i \in W$, we define the logical entailment relation $\models_{w_i}$ as follows:

- $M \models_{w_i} \phi$ iff $\phi \in VA(w_i)$
- $M \models_{w_i} \neg\phi$ iff $\phi \notin VA(w_i)$
- $M \models_{w_i} \phi_1 \wedge \phi_2$ iff $M \models_{w_i} \phi_1$ and $M \models_{w_i} \phi_2$

The other boolean operators are defined as usual. Prohibition is defined in terms of obligation: $\mathbf{F}(p \mid q)$ (p is forbidden whenever q holds) is equivalent to $\mathbf{O}(\neg p \mid q)$ ($\neg p$ is obliged whenever q holds). Contrary-to-duty structures are specified in this logic in the following way: suppose that p is a state of affairs that is prohibited ($\mathbf{F}(p \mid \top)$ or $\mathbf{O}(\neg p \mid \top)$ where $\top$ is a tautology) and that the achievement of the state of affairs q in some way mitigates the violation of this norm, then we state that $\mathbf{O}(q \mid p)$. In this way we capture the intuition that in states of affairs where p holds, and hence norm $\mathbf{F}(p \mid \top)$ is violated, q is obliged. Furthermore, we assume that everything that is not forbidden is permitted.

We now define the *compliance* of a world with a dyadic obligation, and the *coherence* of an ordered pair of worlds with respect to an obligation. These two concepts are used to define the relationship between the normative and severity specifications and the preference relation over worlds.

Definition 1 A world w_i is *compliant* with an obligation $O_j = \mathbf{O}(p \mid q)$ if $M \models_{w_i} \neg q \vee p$; in other words, if the obligation does not apply to w_i ($\neg q$) or the obligation is satisfied (p). We denote this by *compliant*(w_i, O_j).

Definition 2 A preference for world w_i over world w_j is coherent with respect to $O_k \in OS$, written *coherent*(w_i, w_j, O_k), iff *compliant*(w_i, O_k) and $\neg$*compliant*(w_j, O_k).

[1] In van der Torre and Tan [29] and in our prior research [15], a model also includes an accessibility relation $R \subseteq W \times W$ in order to evaluate temporal logic formulae. This is not necessary here because our aim is only to compute a ranking of possible worlds for use within a multi-agent planning mechanism.

Table 1 Norms in the harbour protection scenario

Id	Norm
O_1	$\mathbf{O}(m_u \mid \top)$
O_2	$\mathbf{O}(m_h \mid \neg m_u)$
O_3	$\mathbf{O}(i_u \vee i_b \vee i_h \mid \top)$
O_4	$\mathbf{O}(rep \mid \neg(i_u \vee i_b \vee i_h))$
O_5	$\mathbf{O}(\neg r_u \mid \top)$

Definition 3 A preference for w_i over w_j is incoherent with respect to $O_k \in OS$, written *incoherent*(w_i, w_j, O_k), iff *compliant*(w_j, O_k) and $\neg$*compliant*(w_i, O_k).

This concept of (in)coherence is used in considering whether or not a pair of worlds (w_i, w_j) is part of the preference relation over worlds representing their relative "ideality", or compliance with a normative specification. Informally, the pair (w_i, w_j) is coherent with obligation O_k if and only if, taking into account *only* compliance with O_k, w_i would be preferred to w_j; i.e. if w_i satisfies the obligation but w_j does not. Note that incoherence does not simply mean that w_i is not preferred to w_j, but that w_j is preferred to w_i; i.e. that obligation O_j is incompatible with the preference (w_i, w_j). Therefore, while *incoherent*(w_i, w_j, O_k) implies that *coherent*(w_i, w_j, O_k) does not hold, the fact that *coherent*(w_i, w_j, O_k) is false does not imply incoherence. A pair of worlds can be neither coherent nor incoherent with an obligation; e.g. if both the worlds comply with the obligation. We chose the term incoherence, rather than conflict, in order to avoid confusion with the concept of conflicts among norms.

We can now formalise the norms described in Sect. 2: Table 1. For simplicity of presentation, we assume there is an unauthorised boat in the restricted area (norm O_3 is triggered). We also simplify this illustration by using propositional variables: m_u, the UAV is monitoring the restricted area; m_h, the helicopter is monitoring; i_u, i_h and i_b the UAV, helicopter or boat is intercepting the unauthorised boat; r_u the UAV's position is revealed; and *rep* a report is made to headquarters.

There will be causal constraints on possible worlds in any domain model. In the harbour protection scenario, for example, we have $i_u \rightarrow r_u$ (if the UAV is intercepting the unauthorised boat, its location is revealed) and $\neg m_h \vee \neg i_h$ (the helicopter cannot both monitor and intercept). Possible worlds are then all the joint assignments of values for the propositional variables that satisfy these constraints. Consider, for example, the following two possible worlds: w_3 (where *rep* and m_u are true with all other propositions false) and w_{16} (where m_h, r_u and i_u are true). World w_3 violates obligation O_3 because none of the agents is intercepting. World w_{16} violates O_1 (the UAV is not monitoring) and O_5 (the UAV's location has been revealed). This means that obligation O_5 is violated in world w_{16}, but not in world w_3: *coherent*(w_3, w_{16}, O_5). Similarly, *coherent*(w_{16}, w_3, O_3) holds.[2]

3.2 A preference relation over possible worlds

We define $P_W \subseteq W \times W$ as a preference relation among worlds. We write $(w_i, w_j) \in P_W$, or alternatively $w_i \prec_w w_j$, if w_i is preferred to w_j according to the normative and severity specification. P_W is computed from M according to Eq. 1.

[2] These world IDs (w_3 and w_{16}) are the same as those used in Table 2 and are generated as part of the ranking mechanism that we will present in Sect. 3.3; we simply use the same IDs here for consistency.

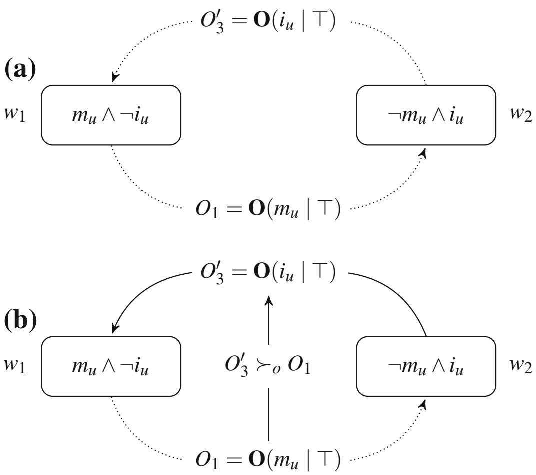

(a)
$O_3' = \mathbf{O}(i_u \mid \top)$
w_1
$m_u \wedge \neg i_u$
$\neg m_u \wedge i_u$
w_2
$O_1 = \mathbf{O}(m_u \mid \top)$
(b)
$O_3' = \mathbf{O}(i_u \mid \top)$
w_1
$m_u \wedge \neg i_u$
$O_3' \succ_o O_1$
$\neg m_u \wedge i_u$
w_2
$O_1 = \mathbf{O}(m_u \mid \top)$

Fig. 3 The effect of a severity specification on the preference relation over worlds. **a** No severity relation. **b** Failing to intercept is more severe than failing to monitor

$$\begin{aligned} w_i \prec_w w_j \leftrightarrow \exists\ O_k \in OS : &\ \mathit{coherent}(w_i, w_j, O_k) \text{ and} \\ &(\forall\ O_l \in OS \text{ s.t. } \mathit{incoherent}(w_i, w_j, O_l) : \\ &\quad \exists\ O_m \in OS : \mathit{coherent}(w_i, w_j, O_m) \text{ and } (O_m \succ_o O_l)) \end{aligned} \tag{1}$$

Informally, we say that w_i is preferable to w_j if all the obligations O_l that are complied with by w_j, and are violated by w_i, are strictly less severe than at least one obligation O_m that is complied with by w_i and is violated by w_j. The requirement that there exists at least one obligation O_k that is complied with by w_i, and that is violated by w_j is introduced in order to make incomparable two worlds that violate exactly the same obligations. If we assume that all obligations are incomparable in terms of their severity, a possible world w_i is preferred to another possible world w_j if and only if w_i violates a strict subset of the obligations violated in w_j. In this case, the second part of the equation (the universal quantification) is used to solve the strong preference problem, making two worlds incomparable if they violate incomparable obligations.

In order to illustrate this concept, and how the introduction of severity preferences affects the resulting preference order over possible worlds, consider the situation depicted in Fig. 3. We consider two obligations: $O_1 = \mathbf{O}(m_u \mid \top)$, the UAV should monitor the restricted area; and $O_3' = \mathbf{O}(i_u \mid \top)$, a simplification of O_3 in Table 1 that requires the UAV to intercept an unauthorised boat. These are enforced over two possible worlds w_1 and w_2 such that $M \models_{w_1} m_u \wedge \neg i_u$ and $M \models_{w_2} \neg m_u \wedge i_u$. Clearly, w_1 complies with O_1 but violates O_3', whereas w_2 complies with O_3' but violates O_1. An arrow (solid or dotted) labelled with an obligation and directed from a world w_i to a world w_j represents the fact that the obligation is coherent with w_i being preferred to w_j. Figure 3a represents the situation where no severity relation is specified, whereas Fig. 3b illustrates the result of introducing a severity relation $O_3' \succ_o O_1$, which reflects the requirement in our scenario that intercepting is more critical than monitoring. In the first case we have that $\mathit{coherent}(w_1, w_2, O_1)$ and $\mathit{coherent}(w_2, w_1, O_3')$ hold. Since the two obligations are incomparable, no preference between the two worlds can be inferred. In the second case, since violations of O_3' are defined to be more severe than those of O_1, and there is no other obligation coherent with (w_1, w_2) being included in the preference relation, we have that w_2 is preferred to w_1. We can think of the arrow labelled with an obligation O_i as overriding the arrows labelled with any O_j such that $O_i \succ_o O_j$.

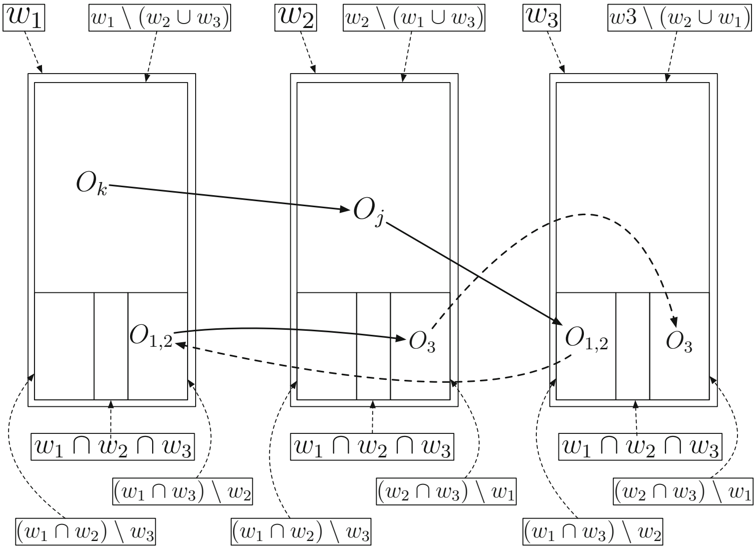

w_1
$w_1 \setminus (w_2 \cup w_3)$
w_2
$w_2 \setminus (w_1 \cup w_3)$
w_3
$w3 \setminus (w_2 \cup w_1)$
O_k
O_j
$O_{1,2}$
O_3
$O_{1,2}$
O_3
$w_1 \cap w_2 \cap w_3$
$w_1 \cap w_2 \cap w_3$
$w_1 \cap w_2 \cap w_3$
$(w_1 \cap w_3) \setminus w_2$
$(w_2 \cap w_3) \setminus w_1$
$(w_2 \cap w_3) \setminus w_1$
$(w_1 \cap w_2) \setminus w_3$
$(w_1 \cap w_2) \setminus w_3$
$(w_1 \cap w_3) \setminus w_2$

Fig. 4 Partition over three generic possible worlds: cases 2.2 and 2.2.1

Equation 1 can be interpreted as saying that w_i is preferred to w_j if for each arrow from w_j to w_i there is one arrow from w_i to w_j that overrides it, and there is at least one such arrow from w_i to w_j.

Our "ideality" preference relation, computed according to Eq. 1, must be guaranteed to be *transitive* and *acyclic*. Transitivity is an intuitive property for a preference relation, including those used in other preference-based deontic logics. Acyclicity is required in order for us to be able to rank the possible worlds from the most to the least compliant, which we do in Sect. 3.3. Moreover, given transitivity, and given the fact that our preference relation is strict, the presence of a cycle would imply that each world in the cycle is less compliant than itself. These properties are, therefore, necessary for this relation to effectively guide practical reasoning.

Lemma 1 *Given a set of possible worlds, W, a set of obligations, OS, and an acyclic severity specification that contains no infinite chain of preferences, P_o, the preference relation over possible worlds computed according to Eq. 1 is transitive.*

Proof Without loss of generality, let a possible world w_i be characterised by the subset of obligations that are violated in w_i. In doing so, we will assume that all obligations that are in OS, but not in the set w_i, are complied with in the world w_i. Consider three possible worlds w_1, w_2 and w_3, such that $w_3 \prec_w w_2$ and $w_2 \prec_w w_1$. These possible worlds can be partitioned such that, for example, world w_1 consists of the subsets of obligations that are violated in that world, and that are, or are not, violated in worlds w_2 and w_3; thus: $w_1 = (w_1 \setminus (w_2 \cup w_3)) \cup ((w_1 \cap w_2) \setminus w_3) \cup ((w_1 \cap w_3) \setminus w_2) \cup (w_1 \cap w_2 \cap w_3)$ (see Figs. 4, 4 and 6).

In order to prove that $w_3 \prec_w w_1$ holds, we first need to show that each obligation O_3 in w_3 is either in w_1, or it is less severe than an obligation that is coherent with the pair (w_3, w_1); that is, an obligation that is in w_1 but not in w_3. We reason by cases, and consider each separately:

1 $O_3 \in (w_1 \cap w_2 \cap w_3)$ or $O_3 \in (w_1 \cap w_3) \setminus w_2$. In these situations, O_3 is also in w_1, and therefore O_3 is not incoherent with (w_3, w_1).

2 $O_3 \in ((w_2 \cap w_3) \setminus w_1)$. In this case, the obligation O_3 is also violated in w_2. Since $w_2 \prec_w w_1$ holds, and since $O_3 \notin w_1$, there must be an obligation $O_{1,2} \in w_1 \setminus w_2$ such that $O_{1,2} \succ_o O_3$. We can distinguish between two sub-cases:

2.1 $O_{1,2} \in w_1 \setminus (w_2 \cup w_3)$. In this case, $O_{1,2}$ is coherent with (w_3, w_1), and is more severe than O_3.

2.2 $O_{1,2} \in ((w_1 \cap w_3) \setminus w_2)$. In this situation, $O_{1,2}$ is not coherent with (w_3, w_1) because it is also violated in w_3. Since $O_{1,2}$ is also in w_3, but not in w_2, it is incoherent with (w_3, w_2). Therefore, there must exist an obligation $O_j \in w_2 \setminus w_3$ that is more severe than $O_{1,2}$. We can distinguish between two further sub-cases:

2.2.1 $O_j \in w_2 \setminus (w_3 \cup w_1)$. This situation is depicted in Fig. 4. Since O_j is incoherent with (w_2, w_1) there must be an obligation $O_k \in w_1 \setminus w_2$ that is more severe than O_j. From the transitivity of P_o we have that $O_k \succ_o O_3$. If O_k is not in w_3, (that is, $O_k \in w_1 \setminus (w_2 \cup w_3)$), then O_k is coherent with the pair (w_3, w_1). If O_k is also in w_3, that is, $O_k \in ((w_1 \cap w_3) \setminus w_2)$ we can apply again Case 2, but taking $O_{1,2} = O_k$. Note that at each recursive application of Case 2, O_k must be different from any previous value of $O_{1,2}$, otherwise P_o would be cyclic. Since P_o does not contain any infinite chain of preferences, it follows that this recursive reasoning must eventually terminate with an $O_k \in w_1 \setminus (w_2 \cup w_3)$ or with case 2.2.2 detailed below.

2.2.2 $O_j \in (w_2 \cap w_1) \setminus w_3$. In this case, O_j is also in w_1, and therefore is coherent with (w_3, w_1). Moreover, for the transitivity of P_o, we have that $O_j \succ_o O_3$.

3 $O_3 \in (w_3 \setminus (w_1 \cup w_2))$. Given $w_3 \prec_w w_2$, and since O_3 is incoherent with (w_3, w_2) there must be an obligation $O_{2,3} \in (w_2 \setminus w_3)$ such that $O_{2,3} \succ_o O_3$. We distinguish two sub-cases:

3.1 $O_{2,3} \in (w_2 \setminus (w_1 \cup w_3))$. Given $w_2 \prec_w w_1$, and since $O_{2,3}$ is incoherent with (w_2, w_1), there must be an obligation $O_{1,2} \in (w_1 \setminus w_2)$ such that $O_{1,2} \succ_o O_{2,3}$. We distinguish between two further sub-cases:

3.1.1 $O_{1,2} \in (w_1 \setminus (w_2 \cup w_3))$. For the transitivity of P_o, we have that $O_{1,2} \succ_o O_3$. Moreover, $O_{1,2}$ is coherent with (w_3, w_1).

3.1.2 $O_{1,2} \in ((w_1 \cap w_3) \setminus w_2)$. This situation is depicted in Fig. 5. In this case, obligation $O_{1,2}$ is neither coherent, nor incoherent with (w_3, w_1). By reasoning in a similar way to Case 2.2.1, it is easy to see that there must be an obligation $O_k \in (w_1 \setminus w_3)$ that is more severe than O_3.

3.2 $O_{2,3} \in (w_2 \cap w_1) \setminus w_3$. Since $O_{2,3}$ is also in w_1, it is coherent with (w_3, w_1).

This case-by-case analysis proves that, for each obligation O_3 that is incoherent with (w_3, w_1), there exists at least one obligation O_1 that is coherent with (w_3, w_1), and that is more severe than O_3.

It remains to be shown that there is always at least one obligation (O_k in Eq. 1) that is coherent with (w_3, w_1). We prove this again through an exhaustive strategy. Since $w_3 \prec_w w_2$, there must be at least one obligation that is coherent with (w_3, w_1), that is, $O_{2,3} \in w_2 \setminus w_3$. There are two possible cases:

4 $O_{2,3} \in ((w_1 \cap w_2) \setminus w_3)$. Since $O_{1,2} \in w_1$, it is also coherent with (w_3, w_1).

5 $O_{2,3} \in (w_2 \setminus (w_1 \cup w_3))$. Since $w_2 \prec_w w_1$, and *incoherent*$(w_2, w_1, O_{2,3})$, there must be an obligation $O_{1,2} \in (w_1 \setminus w_2)$ that is more severe than $O_{2,3}$. There are two sub-cases:

5.1 $O_{1,2} \in (w_1 \setminus (w_2 \cup w_3))$. If so, $O_{1,2}$ is coherent with (w_3, w_1).

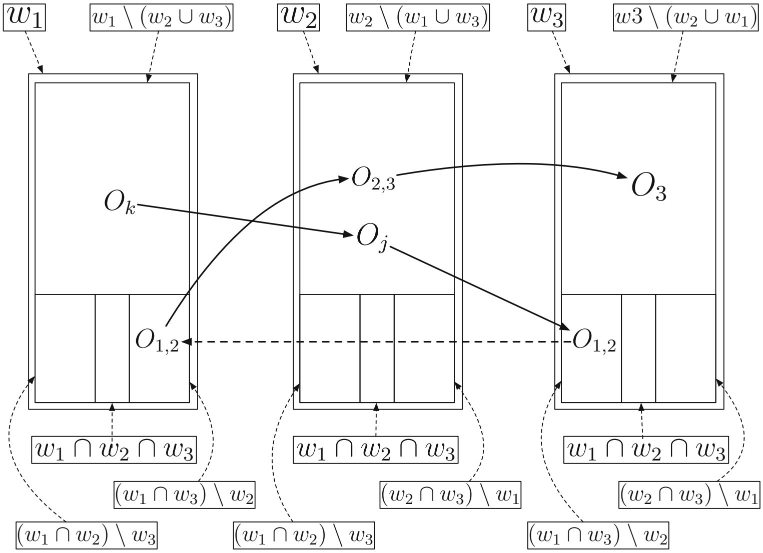

w_1
$w_1 \setminus (w_2 \cup w_3)$
w_2
$w_2 \setminus (w_1 \cup w_3)$
w_3
$w3 \setminus (w_2 \cup w_1)$
O_k
$O_{2,3}$
O_3
O_j
$O_{1,2}$
$O_{1,2}$
$w_1 \cap w_2 \cap w_3$
$w_1 \cap w_2 \cap w_3$
$w_1 \cap w_2 \cap w_3$
$(w_1 \cap w_3) \setminus w_2$
$(w_2 \cap w_3) \setminus w_1$
$(w_2 \cap w_3) \setminus w_1$
$(w_1 \cap w_2) \setminus w_3$
$(w_1 \cap w_2) \setminus w_3$
$(w_1 \cap w_3) \setminus w_2$

Fig. 5 Partition over three generic possible worlds: case 3.1.2

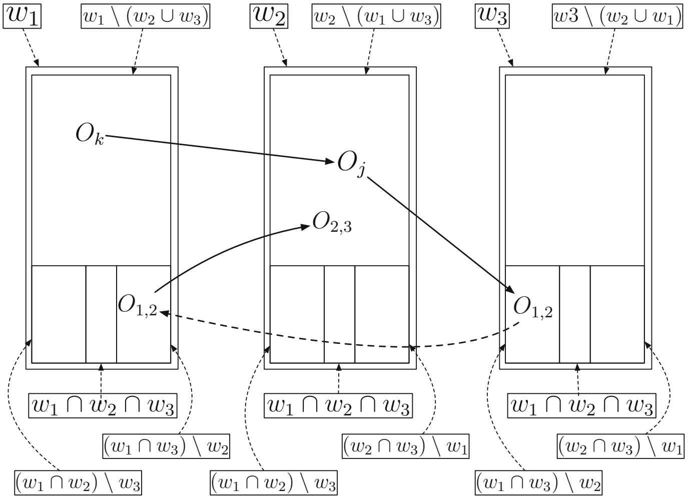

w_1
$w_1 \setminus (w_2 \cup w_3)$
w_2
$w_2 \setminus (w_1 \cup w_3)$
w_3
$w3 \setminus (w_2 \cup w_1)$
O_k
O_j
$O_{2,3}$
$O_{1,2}$
$O_{1,2}$
$w_1 \cap w_2 \cap w_3$
$w_1 \cap w_2 \cap w_3$
$w_1 \cap w_2 \cap w_3$
$(w_1 \cap w_3) \setminus w_2$
$(w_2 \cap w_3) \setminus w_1$
$(w_2 \cap w_3) \setminus w_1$
$(w_1 \cap w_2) \setminus w_3$
$(w_1 \cap w_2) \setminus w_3$
$(w_1 \cap w_3) \setminus w_2$

Fig. 6 Partition over three generic possible worlds: case 5.2

5.2 $O_{1,2} \in ((w_1 \cap w_3) \setminus w_2)$. This situation is depicted in Fig. 6. Since $O_{1,2}$ is also in w_3, and $w_3 \prec_w w_2$, there must be an $O_j \in w_2 \setminus w_3$ that is more severe than $O_{1,2}$. This, in turn, must be different from $O_{2,3}$, otherwise there would be a cycle in P_o. If O_j is in $((w_2 \cap w_1) \setminus w_3)$, then O_j is also coherent with (w_3, w_1). If, on the other hand, $O_j \in (w_2 \setminus (w_1 \cup w_3))$, then, by reasoning in a similar way to Case 2.2.1 we can see that there must be an O_k in $(w_1 \setminus (w_2 \cup w_3))$ that is coherent with (w_3, w_1). □

Lemma 2 *Given a set of possible worlds, W, a set of obligations, OS, and an acyclic severity specification that contains no infinite chain of preferences, P_o, the preference relation among possible worlds, P_W, computed according to Eq. 1 does not contain any finite cycle.*

Proof Assume that there is a cycle in P_W. From the transitivity of P_W, we have that, for all possible worlds w_i in the cycle, $w_i \prec_w w_i$ holds. Consider now a possible world w_i in the cycle. Since there is no obligation that is coherent with (w_i, w_i), and from Eq. 1 it follows that $w_i \prec_w w_i$ does not hold. Therefore, there must be no cycle in P_W. □

3.3 Computing a ranking over possible worlds

We can now use P_W to rank worlds from the most to the least compliant.

Definition 4 Given a set of possible worlds, W, and a preference relation, P_W, we define the ranking of the set as a function $rank_{(P_W)} : W \to \mathbb{N}$ where:

$$rank_{(P_W)}(w_i) = \begin{cases} 1 & \text{if } \nexists (w_j, w_i) \in P_W \\ \max\limits_{(w_j, w_i) \in P_W} \big(rank_{(P_W)}(w_j) + 1\big) & \text{otherwise} \end{cases} \tag{2}$$

Since there are no cycles in P_W, such a ranking can always be computed, but the question remains: how to do this efficiently?

Suppose there is a function $VI : W \times OS \to 2^{OS}$ that, given a set of obligations, associates with each possible world the set of obligations that are violated in that world. If the satisfaction of a formula in a possible world can be computed in constant time, then the complexity of a naïve algorithm for this function will be $O(|OS| \cdot |W|)$.

Consider now two possible worlds w_1 and w_2. Given a set of obligations OS, and a severity relation P_o, we want to verify whether $w_1 \prec_w w_2$. Notice that, in Eq. 1, only obligations that are either coherent or incoherent with (w_1, w_2) are considered. Therefore, all violated obligations in $VI(w_1, OS) \cap VI(w_2, OS)$ can be disregarded. In order to verify whether $w_1 \prec_w w_2$, we need to check that all violated obligations in $V_1 = VI(w_1, OS) \setminus VI(w_2, OS)$ are strictly less severe than at least one violated obligation in $V_2 = VI(w_2, OS) \setminus VI(w_1, OS)$ and that the set V_2 is non-empty. A naïve algorithm would consist of verifying, using Depth-First Search over the graph induced by the severity relation, for each violated obligation $o_1 \in V_1$ and each $o_2 \in V_2$, whether o_1 is reachable from V_2. This algorithm would run in time $O(|OS|^2 \cdot (|OS| + |P_o|)) = O(|OS|^3 + |OS|^2 \cdot |P_o|)$, because the complexity of verifying reachability is $(|OS| + |P_o|)$.

Given two worlds w_1 and w_2, Algorithm 1 verifies whether $w_1 \prec_w w_2$. The algorithm first computes the set of violated obligations V_2 that are coherent with (w_1, w_2) and the set V_1 of those that are incoherent with (w_1, w_2). If V_2 is empty, then we can conclude that $w_1 \prec_w w_2$ does not hold. The algorithm then proceeds to use a Depth First Search from multiple starting points (all the violated obligations in V_2) to compute the set of all obligations that are reachable from at least one violated obligation in V_2; that is, all those violated obligations that are less severe than at least one in V_2. Finally, we just need to verify whether V_1 is included in the set of reachable states. Since we run a single depth first search, we visit every violated obligation, and every member of P_o at most once: the algorithm runs in time $O(|OS| + |P_o|)$.

To compute the preference relation, we need to compare each ordered pair of possible worlds, or each pair of subsets of OS, depending on which one is smaller. The resulting algorithm has complexity $O(\min(|W|^2, 2^{2|OS|}) \cdot (|OS| + |P_o|))$. Some properties of the preference relation can be used in order to decrease the number of comparison that are needed.

Algorithm 1 Computing the preference relation, P_W.

```
Input: $W, OS, P_o, VI, w_1, w_2$
Output: true iff $w_1 \prec_w w_2$
$V_1 = VI(w_1, OS) \setminus VI(w_2, OS)$
$V_2 = VI(w_2, OS) \setminus VI(w_1, OS)$
if $V_2$ is empty then
    return false
else
    $S$ = empty stack
    $reachable = \emptyset$
    for all $o \in V_2$ do
        push $o$ in $S$
        $reachable = reachable \cup o$
    end for
    while $S$ is not empty do
        $o = S.pop$
        for all $o'$ s.t. $o \succ_o o'$ do
            if $o' \notin reachable$ then
                $reachable = reachable \cup o'$
                push $o'$ in S
            end if
        end for
    end while
    if $V_1 \subseteq reachable$ then
        return true
    else
        return false
    end if
end if
```

In particular, from Eq. 1, it is straightforward that, if a world w_i violates a set of obligations $VI(w_i, OS)$, then every world w_k such that $VI(w_k, OS) \subset VI(w_i, OS)$ is preferable to w_i. Since P_W is transitive, given two worlds w_i and w_j, once we have established that $w_i \prec_w w_j$ holds, we can infer that, for all worlds w_k such that $VI(w_k, OS) \subseteq VI(w_i, OS)$, $w_k \prec_w w_j$.

Now we can rank the worlds according to Eq. 2, obtaining a ranking where the more compliant worlds are in a higher position; that is, they are associated with a lower ranking number. To do so, we extend the topological sorting algorithm developed by [19], computing the ranking while sorting the worlds in a linear extension of the partial order. Instead of saving the nodes in an ordered list, we iteratively increase the ranking after eliminating each level of the topologically sorted graph. We denote the set of pairs $(w_i, w_j) \in P_W$, for any w_j as *from*(w_i, P_W), and the set of pairs $(w_j, w_i) \in P_W$, for any w_j as $to(w_i, P_W)$. Given a graph that represents the preference relation P_W, *from*(w_i, P_W) and $to(w_i, P_W)$ denote the outgoing and incoming edges of w_i, respectively. In Algorithm 2 we iteratively find all the nodes that have indegree 0 (those nodes with no incoming edges), and assign to them the current ranking. After doing so, we remove these nodes and their outgoing edges, and increase the current ranking value. We repeat this procedure until all nodes are visited.

This ranking of possible worlds, computed on the basis of a normative system specification that captures both contrary-to-duty obligations and varying severity of violation, can be used to guide agents within a team to make effective decisions about what to do. The challenge now is that these decisions need to take into account *strategies* of action rather than simply considering the compliance of agents with isolated states of affairs. Agents need to take into account future possible compliance with a set of norms in making (collective) action decisions now. Decision mechanisms need to take into account uncertainties in terms of action

Algorithm 2 Computing the ranking over possible worlds.

```
Input: $W, P_W$
Output: $rank_{(P_W)}$
$toVisit = W$
$relation = P_W$
$currentRank = 1$
while $toVisit$ is not empty do
    $noIncoming = \{w_i \in toVisit : to(w_i, relation)$ is empty $\}$
    $toVisit = toVisit \setminus noIncoming$
    for all $w_i \in noIncoming$ do
        $rank_{(P_W)}(w_i) = currentRank$
        $relation = relation \setminus from(w_i, relation)$
    end for
    $currentRank = currentRank + 1$
end while
```

outcomes and exogenous influences on the state of the environment, and enable agents to coordinate their behaviour with others with influence over the environmental state. In order to model decisions in this context, we propose a novel, decentralised planning mechanism that is driven by *qualitative* rewards reflecting this norm-based ranking of possible worlds.

Returning to the harbour protection scenario, our main objective is to preserve the properties $i_u \vee i_h \vee i_b$ (unauthorised boats are intercepted) and, whenever $i_u \vee i_h \vee i_b$ does not hold, to preserve *rep* (incursion into the restricted area is reported). Violations of O_3 or O_4 are more severe than other violations. Moreover, since we want to specify that having someone monitoring the area is more important than not revealing the UAV location, we want to say that violations of O_2 are more severe than violations of O_1 or O_5. This partial order is illustrated in Fig. 1. Looking at, for example, worlds w_3 and w_{16} (Table 2), this ordering means that w_3 is considered worse than w_{16}, even though fewer obligations are violated, because the unauthorized boat is not intercepted in w_3. Similarly, w_{15} should be considered less preferable than w_{16} because it violates O_2, which is more severe than O_5, whereas the two are incomparable with regard to obligation O_1. Part of the ranking over possible worlds in our example computed using Algorithm 2 in this scenario is presented in Table 2. The most and least compliant worlds are w_9 and w_{22} respectively, and we use Λ to refer to the ranking of the least compliant world. World w_3 appears, as expected, at a higher ranking than worlds w_{15} and w_{16} with w_{16} considered more compliant than w_{15} given the relative severity of violation of obligations O_2 and O_5. States that are incomparable with respect to violation severity and CTD structures are ranked equally; e.g. worlds w_1 and w_2 where O_3 is violated in both and O_1 and O_5 are equally severe violations.

Given that we can reliably compute a ranking over possible worlds that takes into account normative constraints, we now turn to the problem of norm-governed planning for a team of agents.

4 Norm-governed multi-agent planning

Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) are an effective means to model collective, distributed decision making where multiple agents, each of them with a particular view of the environment, must coordinate their actions in a decentralized fashion in order to optimize some joint-reward [2]. Existing Dec-POMDP formalisations are founded on a real-valued reward function that specifies the value an agent obtains from

Table 2 Ranking of possible worlds in the harbour protection scenario

R	Id	World							Violations
1	w_9	$\neg i_h$	$\neg rep$	i_b	$\neg m_h$	$\neg r_u$	m_u	$\neg i_u$	
	…								…
3	w_{16}	$\neg i_h$	$\neg rep$	$\neg i_b$	m_h	r_u	$\neg m_u$	i_u	O_1, O_5
4	w_{15}	i_h	$\neg rep$	$\neg i_b$	$\neg m_h$	$\neg r_u$	$\neg m_u$	$\neg i_u$	O_1, O_2
	…								…
6	w_3	$\neg i_h$	rep	$\neg i_b$	$\neg m_h$	$\neg r_u$	m_u	$\neg i_u$	O_3
7	w_1	$\neg i_h$	rep	$\neg i_b$	m_h	$\neg r_u$	$\neg m_u$	$\neg i_u$	O_1, O_3
7	w_2	$\neg i_h$	rep	$\neg i_b$	$\neg m_h$	r_u	m_u	$\neg i_u$	O_3 ,O_5
	…								…
$\Lambda = 15$	w_{22}	$\neg i_h$	$\neg rep$	$\neg i_b$	$\neg m_h$	r_u	$\neg m_u$	$\neg i_u$	O_1, O_2, O_3, O_4, O_5

performing some action in some state of affairs. Our problem is different, however. We require a model of decentralised planning in which agents are rewarded for remaining as compliant with social norms as possible. We have shown that norms are most naturally organised as levels of compliance. We want agents to operate in such a way that they maximise their compliance, and in order to motivate agent behaviour in this way we require a model of rewards that reflects these *qualitative* levels of compliance. To achieve this aim, here we propose a novel model of Dec-POMDPs with qualitative rewards, which we dub N-Dec-POMDPs to reflect our aim of developing a model of severity-sensitive and norm-governed multi-agent planning.

4.1 N-Dec-POMDPs

An N-Dec-POMDP is defined as a tuple, $\langle I, S, b^0, \{A_i\}, P_s, \{E_i\}, P_\mathbf{e}, R\rangle$ where: I is a set of agents, and S is the set of states; b^0 is an initial belief state, i.e. a probability distribution over possible initial states; A_i is a finite set of actions available to agent i and $\mathbf{a} = \langle a_1, \ldots, a_n\rangle$ is a joint-action (one for each agent); $P_s(s_j|s_i, \mathbf{a})$ represents the probability that taking joint-action $\mathbf{a}$ in state s_i will result in a transition to state s_j; E_i is a finite set of observations available to agent i and $\mathbf{E}$ is the set of joint observations $\mathbf{e}$ consisting of one local observation for each agent; $P_\mathbf{e}(\mathbf{e} \mid s_j, \mathbf{a})$ specifies the probability of observing $\mathbf{e}$ when performing a joint-action $\mathbf{a}$ that leads to a state s_j; R is a reward function, the definition of which we provide below.

We focus on finite-horizon N-Dec-POMDPs, and so assume that the execution terminates after H steps. An action-state history as a sequence of joint actions, each followed by a state $(\mathbf{a}^1, s^1, \ldots, \mathbf{a}^t, s^t)$, and an action-observation history is a sequence of local actions each of them followed by a local observation $(a^1, e^1, \ldots, a^t, e^t)$ up to an instant of time t. Agents decide how to act only according to their local observations, and so a solution for a N-Dec-POMDP is a joint-policy $\mathbf{q} \in \mathbf{Q}$, consisting of a local policy $q_i \in Q_i$ for each agent i; i.e. $\mathbf{Q} = Q_1 \times \cdots \times Q_n$. Each local policy maps action-observation histories to stochastic sub-policy choices.

To get an intuition of the strategies that are developed for agents during this planning process, consider again the harbour protection scenario introduced in Sect. 2. A good policy for the UAV may be to continue surveillance, even if it has observed an unauthorised boat in the restricted area, but this depends on the context. If it is operating in a team with some other agent (e.g. a helicopter), it may keep monitoring with a high probability if it observes only one boat in the area (assuming the other agent will intercept it), and with a low probability it will intercept the boat. In situations where the UAV intercepts more than one boat in the area, the UAV may decide to intercept (or report) one of the boats with a higher probability. These policies are, of course, stochastic, and observations (e.g. detection of an incursion) lead to a choice between different sub-policies for each agent in the team. The objective we have in this planning process is to find, given the initial belief state b^0, a joint-policy for the agents in the team that maximizes the total expected value of the joint-reward over the horizon H. Given a t-steps-to-go joint-policy $\mathbf{q}^t$, q_i^t is the local policy for agent i, $\mathbf{a}_{\mathbf{q}^t}$ is the joint-action prescribed by the policy, and $\pi_i^t : E \times Q_i^{t-1} \rightarrow [0, 1]$ is the stochastic mappings that return, for each agent i and observation e, the probability of selecting local sub-policy $q_i^{t-1} \in Q_i^{t-1}$ after observing e.

The expected value of executing a policy $\mathbf{q}^t$ from a state s_i can be computed recursively:

$$V(\mathbf{q}^0, s_i) = R(s_i, \mathbf{a}_{\mathbf{q}^0}) \tag{3}$$

$$V(\mathbf{q}^t, s_i) = R(s_i, \mathbf{a}_{\mathbf{q}^t}) + \left(\sum\nolimits_{s_j, \mathbf{e}} P_s(s_j | s_i, \mathbf{a}_{\mathbf{q}^t}) \times P_{\mathbf{e}}(\mathbf{e} | s_j, \mathbf{a}_{\mathbf{q}^t}) \times \sum\nolimits_{\mathbf{q}^{t-1}} \left(V(\mathbf{q}^{t-1}, s_j) \times \prod\nolimits_i \pi_i^t(\mathbf{e}, q_i^{t-1}) \right) \right) \tag{4}$$

We compute the immediate reward obtained from executing action $\mathbf{a}_{\mathbf{q}^0}$ from state s_i (Eq. 3). Then we consider each possible outcome state s_j, each possible joint-observation, and all possible resulting joint sub-policy choices with their probabilities and recursively evaluate these sub-policies from s_j (Eq. 4). This equation can be generalized for a generic belief state b as shown in Eq. 5.

$$V(\mathbf{q}^t, b) = \sum_{s_i} b(s_i) \cdot V(\mathbf{q}^t, s_i) \tag{5}$$

Most of this characterisation of an N-Dec-POMDP mirrors that of a standard Dec-POMDP; the differences occur in the reward function and the way in which we optimise agents' policies.

One way to define the reward function in terms of norm violations is:

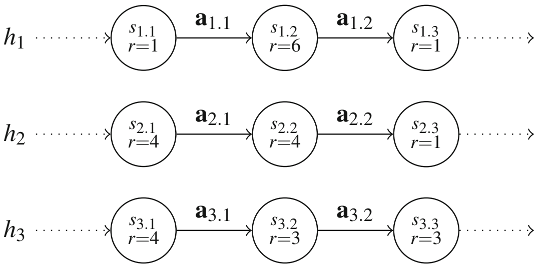
$$\forall s_i \in S, \mathbf{a}_j \in \mathbf{A} : \ R(s_i, \mathbf{a}_j) = -rank_{(\eta)}((s_i)) \tag{6}$$

Essentially, we assign a higher penalty (a negative reward) to states at a lower level of compliance (higher ranking). This approach, however, does not avoid the fallacies in reasoning that we highlight in the introduction. Recall that, in solving a Dec-POMDP, we want to find a joint-policy that maximises the expected value of the *sum* of the rewards accumulated during execution (i.e. minimises the penalties for norm violation). A reward function, then, entails a preference relation over possible histories, where histories that have a higher total reward are preferred.

Consider the three execution histories h_1, h_2 and h_3 depicted in Fig. 7. History h_1 visits three states $s_{1.1}$, $s_{1.2}$ and $s_{1.3}$ with rankings $r = 1$, $r = 6$ and $r = 1$, respectively. History h_2 visits states with rankings 4, 4 and 1, and history h_3 visits states with rankings 4, 3 and 3. Using the reward function of Eq. 6, these three histories would be associated with rewards of -8 (h_1), -9 (h_2) and -10 (h_3), and hence history h_1 will be preferred to h_2, and h_2 will be preferred to h_3. This is inconsistent with our view of norm compliance as a series of partially ordered levels. Our goal is to find policies that minimise the likelihood of reaching states at lower compliance levels, and so we require a reward function such that:

1. Histories that include states with lower compliance levels are less preferred; and

Fig. 7 Three example histories: h_1, h_2 and h_3



2. If two histories are incomparable with respect to violations for all compliance levels lower than level i, the history with fewer violations at level i is preferred.

In order to capture these requirements, we exploit the qualitative theory of MDPs proposed by Bonet and Pearl [6]. This is based on an order of magnitude approximation for utility and probability values, and was developed to model problems where only imprecise information about quantitative parameters of an MDP is available. They define polynomials and an infinite series of elements of the set $\mathcal{Q}$ of *extended reals*. Given a variable ε which represents a small unknown quantity, an extended real is a rational function p/q, with p and q being polynomials in ε. As an example, the quantity ε^{-1} can be used to represent an unknown high utility, while ε^5 can be used to represent a very small utility. Bonet and Pearl define operations over the set Ψ of infinite series $\psi = \sum_k c_k \varepsilon^k$ and generalize the value iteration algorithm [23] for Qualitative MDPs and POMDPs. Let ψ^1 and ψ^2 denote two members of Ψ such that $\psi^i = \sum_k c^i_k \varepsilon^k$, $\alpha \in \mathbb{R}$ and $o(\psi^i)$ be the *order* of ψ^i, defined as $o(\psi^i) := min\{k : c^i_k \neq 0\}$. Equations 7–10 define the sum and comparison between two extended reals and the product of an extended real with a real. Note that the comparison of extended reals is reduced to a lexicographic comparison of their coefficients.

$$(\psi^1 + \psi^2) := \sum\nolimits_k (c^1_k + c^2_k)\varepsilon^k \tag{7}$$

$$\psi^1 \prec 0 \ \text{ iff } c^1_{o(\psi^1)} < 0 \tag{8}$$

$$\psi^1 \prec \psi^2 \text{ iff } \psi^1 - \psi^2 \prec 0 \tag{9}$$

$$(\alpha * \psi^1)_k := \sum\nolimits_k (\alpha \ * c^1_k)\varepsilon^k \tag{10}$$

Given a real parameter, ρ, the magnitude of an extended real is defined as:

$$\|\psi^i\|_\rho := \sum\nolimits_k |c_k|\rho^{-k} \tag{11}$$

Bonet and Pearl [6] then show that, for positive coefficients and large enough ρ, we have $\psi^1 \prec \psi^2$ if and only if $\|\psi^1\|_\rho < \|\psi^2\|_\rho$.

Given our requirements, therefore, we can characterise the reward function for an N-Dec-POMDP: $R : S \times \mathbf{A} \rightarrow \mathcal{Q}$ is a reward function, and $R(s_i, \mathbf{a})$ specifies the reward obtained by performing $\mathbf{a}$ in s_i. We can then associate with each history, h, an extended real utility $\psi^h = \sum_{0 \leq i < \Lambda} -c^h_i \varepsilon^i$ where $\Lambda$is the maximum ranking level of the given N-Dec-POMDP, and each component c^h_i corresponds to the number of states in the history that have rank $\Lambda - i$. This is equivalent to assigning a state-based reward of $-\varepsilon^{\Lambda - i}$ to each state with rank i. Such a reward function can be interpreted as the agents incurring in a higher cost for visiting states with higher rank. Equation 12 captures this formally.

$$\forall s_i \in S, \mathbf{a}_j \in \mathbf{A} : R(s_i, \mathbf{a}_j) = -\varepsilon^{\Lambda - rank_{(P_W)}(s_i)} \tag{12}$$

Consider again the example of Fig. 7. Assuming a maximum rank, $\Lambda = 6$, history h_1 is now associated with a value of $-\varepsilon^0 - 2\varepsilon^5$, history h_2 with $-2\varepsilon^2 - \varepsilon^5$, and history h_3 with $-\varepsilon^2 - 2\varepsilon^3$. From Eq. 9 we have that $(-\varepsilon^0 - 2\varepsilon^5) \prec (-2\varepsilon^2 - \varepsilon^5) \prec (-\varepsilon^2 - 2\varepsilon^3)$, and therefore, as expected, h_3 is preferred to h_2 and h_2 is preferred to h_1. Maximizing the total expected reward, therefore, implies minimizing the probability of reaching higher ranking levels (lower levels of compliance).

4.2 Policy optimisation in N-Dec-POMDPs

Given that finding a γ-approximation of an optimal policy for a Dec-POMDP is NEXP-complete, finding such a policy for an N-Dec-POMDP will be at least as hard given that the introduction of qualitative levels of reward, representing norm compliance, does not simplify the underlying decision problem. Due to this complexity, our approach is to develop an algorithm that can efficiently find solutions without providing any guarantees on the solution quality with respect to the optimum. One of the most successful existing algorithms in solving large instances of Dec-POMDPs is Point-Based Policy Generation (PBPG) [32]. As discussed in Sect. 6.1, PBPG starts from the last time step and moves backwards using the t-steps-to-go policies as possible sub-policies for the $(t + 1)$-steps-to-go policies. At each step, a heuristic is used to select a set of reachable belief states. A set of candidate policies is then generated and evaluated from those belief states and only the best *maxTrees* policies are retained. In the policy generation phase, one candidate for each possible joint-action is created, and a linear program is used to find sub-optimal stochastic mappings for the given belief state and joint-action. The mappings for each agent are iteratively improved while the other agents' policies are fixed. We adapt PBPG in order to approximately solve N-Dec-POMDPs, and, in Sect. 4.3, propose a novel heuristic for qualitative reward domains to restrict the selected belief states.

The reward function of an N-Dec-POMDP has its co-domain in Ψ, and so we need to define a procedure for policy optimisation that accepts reward values and returns expected values in Ψ. Note that P_s, $P_\mathbf{e}$, and π_i^t are defined as functions with real co-domain. To do this we can use a combination of Eqs. 7 and 10 to evaluate joint policies.

Since the linear program used in PBPG to improve the stochastic mappings of policy candidates optimises a real valued expected total reward, it cannot be directly applied to improve the policy of an N-Dec-POMDP. One alternative is to simply substitute each extended real with its magnitude, and use the linear program to maximise the magnitude of the expected reward. The correctness of this approach relies on the fact that, for each pair $\psi^1, \psi^2 \in \Psi$ we can find a large enough ρ such that $\psi^1 \prec \psi^2$ if and only if $\|\psi^1\|_\rho < \|\psi^2\|_\rho$. Figure 8 gives the linear program for this method. The value of the current policy is $V^{t+1}(\boldsymbol{\pi}, b)$, the variables $\pi_i'(q_i^t|e_i)$ represent the new values for the stochastic mappings of agent i, while $\pi_{-i}(q_{-i}^t|e_{-i})$ are the fixed mappings for agents other than i, and δ is the improvement that needs to be maximized. This procedure is repeated until no further improvement is possible. The solution corresponds to an equilibrium, where no agent can unilaterally improve its own policy. We use the inverse of the magnitude in order to account for the fact that our extended reals rewards have only negative coefficients. Tijs [28] shows that it is always possible to find a ρ large enough such that lexicographic optimization reduces to linear programming. Tijs' proof does not show how to find a good value of ρ, however. As a result, this approach can only approximate the solution of a lexicographic optimisation.

An alternative approach is to consider a series of Λ linear programs, each of them maximizing one coefficient of the value function. In Fig. 9 we present the linear program that maximises the value of the jth coefficient $V_j^t(\boldsymbol{\pi}, b)$ of the value function. The improvement

Variables: $\delta, \pi'_i(q^t_i|e_i)$

Objective: maximize δ

Subject to:

$$\begin{array}{ll}
 & \delta - \|V^{t+1}(\pi,b)\|_\rho \leq \sum\limits_{s,\mathbf{e}} \left(P(\mathbf{e},s|\mathbf{a},b) \cdot \sum\limits_{\mathbf{q}} \left(\pi'_i\left(q^t_i|e_i\right) \cdot \pi_{-i}\left(q^t_{-i}|e_{-i}\right) \cdot \left(-\|V^t(\mathbf{q}^t,s)\|_\rho\right)\right)\right) \\
\forall e_i \in E_i & \sum\limits_{q^t_i \in Q^t_i} \pi'_i\left(q^t_i|e_i\right) = 1 \\
\forall e_i \in E_i, q^t_i \in Q^t_i & \pi'_i\left(q^t_i|e_i\right) \geq 0
\end{array}$$

Fig. 8 Linear program for joint-policy optimisation through reward magnitude

Variables: $\delta_j, \pi'_i(q^t_i|e_i)$

Objective: maximize δ

Subject to:

$$\begin{array}{ll}
 & V^{t+1}_j(\pi,b) + \delta_j \leq \sum\limits_{s_i,\mathbf{e}} \left(P(\mathbf{e},s_i|\mathbf{a},b) \cdot \sum\limits_{\mathbf{q}} \left(\pi'_i\left(q^t_i|e_i\right) \cdot \pi_{-i}(q^t_{-i}e_{-i}) \cdot V^t_j(\mathbf{q}^t,s_i)\right)\right) \\
\forall\, 0 \leq k < j & V^{t+1}_k(\pi,b) + \delta_k \leq \sum\limits_{s_i,\mathbf{e}} \left(P(\mathbf{e},s_i|\mathbf{a},b) \cdot \sum\limits_{\mathbf{q}} \left(\pi'_i\left(q^t_i|e_i\right) \cdot \pi_{-i}\left(q^t_{-i}|e_{-i}\right) \cdot V^t_k(\mathbf{q}^t,s_i)\right)\right) \\
\forall e_i \in E_i & \sum\limits_{q^t_i \in Q^t_i} \pi'_i(q^t_i|e_i) = 1 \\
\forall e_i \in E_i, q^t_i \in Q^t_i & \pi'_i(q^t_i|e_i) \geq 0
\end{array}$$

Fig. 9 Linear program to optimise the jth coefficient of the Ψ value function

of each jth component is represented by δ_j. We start by maximizing the value of the coefficient of ε^0. Then for each subsequent LP, we improve only the jth components of the value functions of the policies. However, for each kth component with $k < j$ we introduce an additional constraint to ensure that we do not decrease the kth component.

Note that improving the jth component might result in situations where, for some lth component, with $l > j$, only a negative improvement is possible. These improvements represent a trade-off where we accept a decrease in one component of $V^t(\boldsymbol{\pi}, b)$ in order to improve one that is associated with a higher ranking level (lower compliance). We say that an improvement sequence $\delta_0, \ldots, \delta_\Lambda$ is *acceptable* if and only if, for each negative improvement $\delta_j < 0$, there exists a $\delta_k > 0$ such that $k < j$. Informally, an improvement is not acceptable if it leads to a policy that has lower expected value than the initial one according to the ordering among extended reals.

This translation guarantees that an acceptable improvement sequence results in locally optimal solutions. While this approach requires solving Λ LPs, the first LP will have only a sparse constraint matrix and can be solved very efficiently. The following LPs will only have to consider a very constrained solution-space, and therefore can also be solved more efficiently. In fact, since our main objective is to maximise values associated with higher ranking levels, and doing so will often restrict the space of possible policies to only a few candidates, we can limit our sequence of improvements only to a limited number of higher ranked levels and still expect to find close-to-optimal solutions. As we will show in Sect. 5, if we terminate our sequence of LPs when we find a non-zero coefficient for the value function, we obtain considerable saving in execution time without affecting the quality of the resulting policy. We refer to this approach as the *greedy* LP, and to the magnitude translation as the *magnitude* LP.

Algorithm 3 Algorithm for the Greedy LP optimization

```
Input: $b^t$, $\mathbf{a}$, $Q_0^t, \dots, Q_n^t$
Output: $\pi_0, \dots, \pi_n$
 1: initialise $\pi_0, \dots, \pi_n$ randomly.
 2: $V^\pi = eval(Q_0^t, \dots, Q_n^t, \pi_0, \dots, \pi_n)$
 3: repeat
 4:     $changed = false.$
 5:     for $ag = 0$ to $n$ do
 6:         for $l = 0$ to $\Lambda - 1$ do
 7:             $\langle \delta_l, \pi_{ag} \rangle = LP(V^\pi, Q_0^t, \dots, Q_n^t, \pi_0, \dots, \pi_n, ag, l)$
 8:             if $\delta_l > 0$ then
 9:                 $changed = true$
10:             end if
11:             $V_l^\pi = V_l^\pi + \delta_l$
12:             if $V_l^\pi > 0$ then
13:                 break
14:             end if
15:         end for
16:     end for
17: until $changed = true$
```

Algorithm 3 formalises the greedy LP optimization. The algorithms takes as input the belief against which we are evaluating our policy b^t, the candidate joint action $\mathbf{a}$, and the set of possible sub-policies Q_i^t for each agent $1 \leq i \leq n$, and it returns, for each agent $1 \leq i \leq n$, a function π_i that maps its local observations to a probability distribution over Q_i^t. The algorithm starts by initializing the mappings randomly and evaluating these mappings over the belief b^t using Eq. 4. It then considers each agent in turn in order to improve their local policy (Lines 5–16). For each agent, the algorithm applies the LP of Fig. 9 to improve each component of the expected value function V^π (7) starting from the one associated with the highest ranked (least norm-compliant) level. After each call to the LP, the value V^π is updated accordingly, and, if the value for the level being considered is greater than 0, the improvement for the current agent is terminated (Line 13). This procedure is repeated until we complete an iteration without any change in the value function (Lines 3–17).

4.3 The most-critical-states heuristic

In PBPG, heuristics are used to identify relevant belief states against which to optimize the policy. The intuition behind PBPG is that, if the agents act in a way that is close to the optimum, only a subset of states will be reachable. In building policies in a bottom-up fashion, therefore, we can optimize them only against those states that are most likely to be encountered during execution, increasing the scalability of the algorithm.

In finding a policy for an N-Dec-POMDP, our objective is to minimize the probability of visiting states associated with a higher ranking level (lower compliance). The approach we take, therefore, is to restrict the belief states we select so that they include only those reachable states that are likely to lead to more severe violations. In so doing, we focus planning effort on developing agent strategies to *avoid* these undesirable states. This has the effect of improving the scalability of the algorithm without affecting the performance of resulting policies. Inspired by the MDP and Dec-POMDP heuristics (we refer to these by MDP and Dec) proposed by Seuken and Zilberstein [25] we exploit information that is easily obtainable by computing the optimal policy for the underlying MDP and its value function. The value function $V_{MDP}(s_i)$ represents the expected utility that we would obtain if the execution had started from s_i, assuming the agents were able to fully observe the current state of the system and to coordinate their decisions at each step. V_{MDP}, thus, represents a good candidate to heuristically assess the importance of a state, taking into account the capability of a coalition to remain compliant or to recover from current violations. We define $R^t_{MDP}(s_i) \subseteq S$ to be the set of reachable states from a state s_i if the coalition follows the MDP policy for t time steps, and we define $pr_{s_i}(s_j)$ to be the probability of reaching state $s_j \in R^t_{MDP}(s_i)$. The values of these can be estimated using standard sampling techniques. Given a threshold $\tau \in \Psi$, $mc^{s_i}_\tau$ is the subset of states, among those reachable from s_i, such that the product of their value function and their probability is less than τ. Formally:

$$mc^{s_i}_\tau := \{s_j \in R^t_{MDP}(s_i) : V_{MDP}(s_j) * pr_{s_i}(s_j) < \tau\}$$

Intuitively, if we accept $V_{MDP}(s_i)$ to be a good approximation of the value that we can expect to obtain from a state s_i in the decentralized case, $mc^{s_k}_\tau$ includes those reachable states that have a potential impact on the expected ranking that is higher than τ.

Given an initial state s_i, and a threshold τ, the MDP Most-Critical-States (MDPMcs) heuristic is the heuristic that returns a belief state b such that:

$$b(s_j) = \begin{cases} \frac{pr_{s_i}(s_j)}{\sum\limits_{s_k \in mc^{s_i}_\tau} pr_{s_i}(s_k)} & \text{if } s_j \in mc^{s_i}_\tau \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

These definitions can be easily generalized for an initial belief state. An improved version of MDPMcs (denoted as DecMcs) can be obtained by using a previously obtained policy $\mathbf{q}^H$ to sample the set of reachable states and to evaluate them. We can simulate the execution of the system assuming that the coalition follows $\mathbf{q}^H$ for t steps. For each of a number of simulations we obtain a state s_i and a sub-policy $\mathbf{q}^{(H-t)}$, which we use to evaluate $V_{Dec}(s_i)$. Note that a state can be reached following different paths, potentially resulting in different sub-policies and values of V_{Dec}. For each state we take the minimum among those values. In our experiments, a mixed approach (denoted as MixMcs), using $\mathbf{q}^H$ to find the reachable states, and V_{MDP} to evaluate their criticality led to better policies. Moreover, this mixed approach does not require us to evaluate additional joint policies. While the best value for τ depends on the scenario, a typically good value is $\tau = c\varepsilon^{s+1}$, where s is the order of $V_{MDP}(b^0)$, and c a small real. This threshold captures all those states that might potentially to lead to the worst (highest) ranking level reachable by an MDP-based execution.

With our method of mapping a normative system specification into a ranking over possible worlds, the magnitude and greedy approaches to solving an N-Dec-POMDP, and the Most Critical States (MCS) heuristic, we can move on to evaluate our model. Existing benchmark problems for (Dec-)POMDPs do not include normative, or soft constraints, and so we use the multi-agent harbour protection scenario that involves both CTD structures and varying severity. We can, however, directly compare the magnitude and greedy approaches to solving an N-Dec-POMDP, and compare MCS with standard PBPG.

5 Evaluation

In the harbour protection scenario introduced in Sect. 2, there are restricted and unrestricted areas, and both agents (UAV, helicopter and patrol boat) and unauthorized boats can move between them. The UAV and the helicopter can perform the action *monitor* in order to start monitoring their current area; this action always suceeds, but monitoring does not guarantee detection of an unauthorised boat. Each agent is able to observe the location of unathorized boats in the same area with probability 0.15. By monitoring an area, an agent increases this probability to 0.75. Each agent can perform $intercept_i$ in order to intercept the ith unauthorized boat, and an action *report* to report an incursion. Each of these actions will succeed with a probability of 0.8, and the agent must commit to this task over two time steps to have an effect. Each agent is able to observe its own location and, with a certain probability, the location of agents in the same area. By monitoring an area, an agent increases its probability of correctly observing other agents' locations. The behaviour of unauthorized boats is controlled by the simulation. Throughout the simulation, each boat will move from the unrestricted area to the restricted area with probability 0.11 or return to the unrestricted area with probability 0.3. Initial exploration of possible values for these probabilities indicated that these gave a good level of dynamism and indeterminism in the simulation, and hence represent a good level of challenge for the planning problem. We chose a horizon, $H = 20$, for all simulations; preliminary experiments showed that a horizon greater than 20 offered no additional benefit to the quality of the plans computed for any of the algorithm/heuristic combinations. For the same reason, for each simulation $maxTrees = 2$ (the number of policies retained at each step during plan generation). Each experimental condition was repeated 20 times with identical initial conditions: all unauthorized boats in the unrestricted zone, and all agents in the restricted zone with no agent monitoring.

We used three different instantiations of the harbour protection scenario, chosen in order to investigate both under- and over-constrained conditions. The first consists of two agents (one UAV and one helicopter) and three unauthorized boats. This is over-constrained because if all three unauthorized boats are in the restricted area at the same time, there are insufficient agents to intercept them all. In this case, the best choice for the agents is for one of them to intercept a boat and for the other to issue a report to headquarters. The second case consists of three agents (one UAV, one helicopter, and one patrol boat) agents and two unauthorized boats. In this case, even if both unauthorized boats are in the restricted area at the same time, the agents can intercept both (one by the patrol boat and one by the helicopter) and maintain surveillance (the UAV). The agents can remain fully compliant. In the third case there are three agents and three unauthorized boats. If all three unauthorized boats enter the restricted zone, the agents are then faced with the choice between intercepting all three and maintaining surveillance. The three cases are, therefore, designed to maximise the challenge with respect to agents making the most compliant joint action choices.

In Table 3 we report the average and standard deviation of the execution time and the quality of the resulting policy for our planning algorithms and PBPG. Specifically we compare execution of our planner using the standard PBPG heuristic and the MCS heuristic, and the magnitude and greedy linear programs. The columns **Ag** and **B** specify the number of agents and unauthorized boats in each scenario. The two columns **Standard** and **MCS** compare the results for the two different heuristics. In particular, the MCS executions use a combination of the MDPMcs and random heuristics[3] and the standard executions use a combination of

[3] The random heuristics sample for reachable states by simulating agents that choose a random action at each step.

Table 3 Planning results for standard PBPG heuristics and MCS with the magnitude and greedy linear program planners

Ag	B	Standard		MCS	
		Value	Time	Value	Time
Magnitude LP					
2	3	−5.88e−20	32.05	−5.88e−20	27.10
		±3.76e−23	±2.89	±3.59e−23	±2.10
3	2	−3.68e−20	185.25	−2.69e−20	146.75
		±3.70e−23	±15.69	±8.27e−23	±14.61
3	3	−4.61e−20	7609.10	−4.76e−20	5375.60
		±3.58e−23	±484.27	±2.64e−22	±578.30
Greedy LP					
2	3	−5.88e−20	31.15	−5.88e−20	23.65
		±3.02e−23	±2.96	±3.04e−23	±2.78
3	2	−3.28e−20	112.20	−2.61e−20	79.85
		±2.86e−21	±7.38	±4.56e−22	±7.66
3	3	−4.58e−20	4984.80	−4.70e−20	3236.25
		±3.04e−22	±483.49	±5.94e−22	±392.69

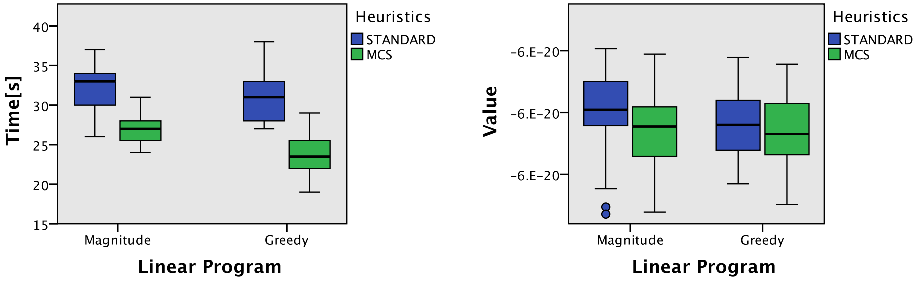


Fig. 10 Execution time and policy quality (value) for the 2 agents and 3 boats case

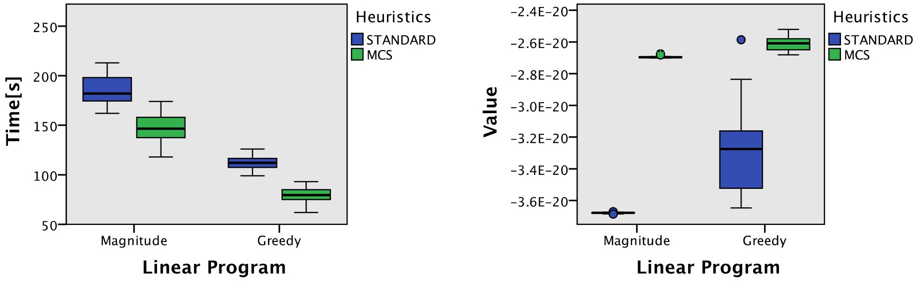


Fig. 11 Execution time and policy quality (value) for the 3 agents and 2 boats case

MDP and random heuristics. The first part of the table presents results obtained using the magnitude LP, which optimises the inverse of the magnitude of the reward. The second part of the table presents obtained using the greedy LP, interrupting improvement of the solution as soon as we find a level with value lower than −0.001. The results are also summarised in the box-plots presented in Figs. 10, 11 and 12.

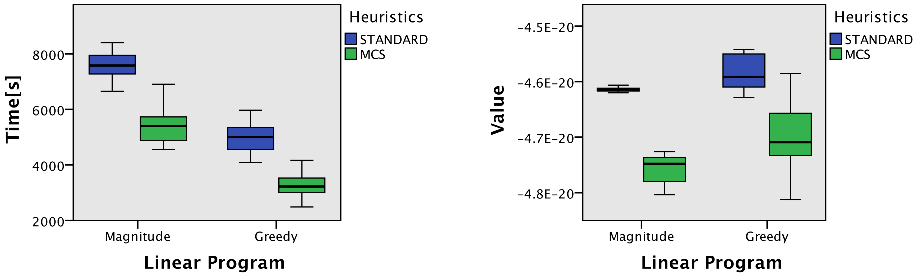


Fig. 12 Execution time and policy quality (value) for the 3 agents and 3 boats case

Table 4 Pair-wise differences in execution time: *p*-values

Comparison of LPs				Comparison of Heuristics			
Scenario		Heuristics		Scenario		LPs	
Ag	B	Standard	MCS	Ag	B	Magnitude	Greedy
2	3	1.000	0.133	2	3	0.000	0.000
3	2	0.000	0.000	3	2	0.048	0.034
3	3	0.000	0.000	3	3	0.003	0.003

The results are not normally distributed, and so we tested them for significance using the Kruskal Wallis (analysis of variance) test, which does not assume a normal distribution of residuals. There are no significant differences for the value obtained (policy quality): asymptotic *p*-value of 1.000. We do, however, observe significant differences in execution time. In order to better understand where these between-groups differences lie, we performed a post-hoc analysis consisting of Bonferroni-corrected pairwise Mann–Whitney tests. Table 4 summarises the *p*-values for all the pairwise tests comparing the execution times of different algorithms. Using this conservative method, we found that all pair-wise differences are significant with the exception of two.

We further explored the differences in performance of the LPs in order to isolate the effect of choosing the greedy LP over the magnitude LP. There is no significant difference in the execution time of the magnitude and greedy LPs with both Standard and MCS heuristics in the 2-agents, 3-boats case (Fig. 10, *p*-values being 1.000 and 0.133, respectively). We believe that this is due to the fact that this is an over-constrained problem, such that we cannot easily discount those strategies that are more likely to lead to the most severe violations. The strategy of solving multiple, smaller LPs at different ranking levels and terminating when no significant improvement can be found, therefore, has little effect. There are, however, significant differences in all other comparisons between the LPs, such that the greedy LP significantly out-performs the magnitude LP.

We then explored the differences in performance of the heuristics in order to isolate the effect of choosing the MCS heuristic over the Standard heuristic. The positive effect of using the MCS heuristic over Standard is significant in all cases for either LP. It is interesting to note that, although significant, the benefit in using the MCS heuristic is more marginal in the 3-agents/2-boats case. This is the least constrained of the problems considered, given that even if both unauthorised boats enter the restricted area at the same time, there are sufficient agents to intercept both and maintain surveillance. This is expected because the MCS heuristic was specifically designed to provide additional guidance to decision-making in more challenging, over-constrained scenarios.

Fig. 13 Comparison of MCS+Greedy and random play

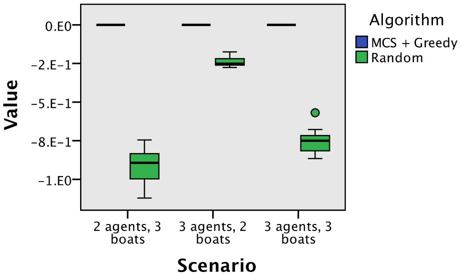


Table 5 Comparison of MCS+Greedy and random play

Ag	B	MCS + Greedy	Random
2	3	−5.88e−20 ±3.04e−23	−9.10e−1 ±1.03e−1
3	2	−2.61e−20 ±4.56e−22	−2.40e−1 ±2.69e−02
3	3	−4.70e−20 ±5.94e−22	−7.59e−1 ±7.17e−02

It is important to note here that the effects on execution time due to the use of the greedy LP and the MCS heuristic are *additive*. When the problem is over-constrained, the use of the MCS heuristic provides significant execution time improvements, and continues to provide some benefits in less constrained scenarios. The greedy LP algorithm is able to exploit the structure of the domain in all but over-constrained scenarios. In combination, the greedy LP algorithm and MCS heuristic significantly improves execution times across *all* norm-governed scenarios, regardless of how they are constrained with respect to resources. Further, as the complexity of the scenario itself increases, the combined effect is more marked. In the 3-agents, 3-boats case, execution times are more than halved: mean execution time for magnitude/standard being 7609s, and for greedy/MCS being 3236s.

In common with all other algorithms for planning in Dec-POMDPs, our N-Dec-POMDP planner does not provide guarantees for the quality of the solution. No existing model is able to utilise the qualitative reward function that is necessary to reason about levels of norm compliance, but comparison with the alternatives above, including an adaptation of the standard heuristic, provides the most objective assessment available. There is the question, however, of whether our approach (or, indeed, whether any of the others considered) finds *good* policies. We may do this by comparing the quality of policies computed by our greedy algorithm using the MCS heuristic with the expected value of a set of randomly computed policies. In Fig. 13 and Table 5 we present the results of this comparison. Our policies behave consistently better, with the difference being more pronounced in more constrained situations. Again, we tested these results for significance using the Kruskall-Wallis test, and obtained a *p*-value of 0.000 in each case. We can, therefore, claim that our planning algorithm is effective in finding good policies for decentralized, collective planning problems under uncertainty.

6 Discussion

In placing this research in context, we first discuss the current research landscape on single and multi-agent planning under normative (or equivalent) constraints. We then move on to explore in more detail the model of norms used in this research, alternative modelling approaches and discuss some avenues for future investigation. Our conclusions follow this extended discussion.

6.1 Related research in norm-governed planning

Models of practical reasoning where action is both constrained by causal dependencies and guided by ideals of behaviour (soft constraints or preferences) have been studied in a range of contexts. Gerevini and Long [17] extended the Planning Domain Description Language, PDDL, to include preferences. These are represented as boolean formulae that are satisfied or violated by a plan. Even in the presence of preferences, however, PDDL requires the domain designer to specify a quantitative metric function to be optimised by a planner. These metrics may, or may not, depend on the satisfaction of each constraint. For example, it is possible to assign a real-valued weight to each constraint violation. Our approach is different because we specify *qualitative* preferences among constraint violations. Bienvenu et al. [5] do consider qualitative priorities over constraints. Their formalism allows alternative plans to be evaluated according to a lexicographic ordering over the set of constraints. Using this approach, we can specify that avoiding the violation of a single norm is more important than avoiding the violation of any other norms that follow in the lexicographic order. Our approach is different, however: we do not simply consider norms in a lexicographic order, rather we compute severity levels that depend on all norms that are violated in a state, and then consider severity levels in a lexicographic order. Moreover, Bienvenu et al. [5] do not take into account how many times each constraint is violated. They do model the specification of temporally extended preferences; that is, preferences that regulate execution paths, instead of single states. While we do not discuss this issue in this paper, our model can be directly extended in order to represent and reason about temporally extended norms, as discussed in prior research [16]. These models only capture deterministic, single agent scenarios, and do not support the specification of contrary-to-duty constraints.

In norm-aware practical reasoning, a number of different methods for reasoning about norm compliance have been proposed. Meneguzzi and Luck [20], for example, extend the BDI architecture to take into consideration norms. They describe algorithms that enable agents to react to the activation and expiration of norms by modifying their intentions; i.e. by introducing plans for the fulfilment of obligations and removing plans that violate prohibitions. Dignum et al. [12] discuss the introduction of a preference relation over norms to solve normative conflicts. This preference relation is taken into account only in situations where it is not possible to comply with all norms. Our decision theoretic approach is different in the sense that an agent might decide to violate a less severe norm even in the absence of a conflict if, in doing so, the probability of violating a more severe norm in the future decreases. Moreover, Dignum et al. only consider single agent scenarios and with a simplified (state-based) representation of norms. While these approaches support the specification and reasoning about contrary-to-duty obligations, they only consider single agent scenarios where the environment is fully observable and actions are deterministic.

Fagundes et al. [13] use Markov Decision Processes (MDPs) [23] to model a self-interested agent that takes into account norms, and the possibility of violating them, in deciding how to act. Violations are associated with sanctions, which result in the modification of the transition probabilities, or of the agent's capabilities. Agents consider the effects of sanctions on their expected utility and weigh these against the potential benefits of violating norms in order to decide upon a course of action. This model does not, however, explicitly capture the relative severity of norm violations, and the representation of sanctions relies on the assumption that the norm enforcement authority has the power to affect the agent's capabilities and the probabilities of transitions.

A more appropriate representation of severity levels of norm violation could be obtained by representing the problem as a multi-objective MDP; that is, an MDP where the reward is a vector, rather than scalar quantity, and where each component of the vector may represent a different objective. A number of researchers have focussed on methods to efficiently solve multi-objective POMDPs (Partially Observable MDPs) [24,26,30]. Some of these methods attempt to find a set of policies that maximise the expected value for a set of possible scalarizations. A scalarization essentially gives a weight to each component of the reward value and is formally defined as a linear function that takes a vector and returns a scalar. Roijers et al. [24], for example, present OLSAR, a point-based algorithm based on Perseus [27] that efficiently finds a set of approximately optimal policies for different scalarizations. Reasoning about norm compliance may be seen as a particular case of a multi-objective POMDP, where each component represents the degree of compliance. From this point of view, different scalarizations may be used to represent different degrees of severity for norm violation. Such an approach, however, does not avoid the sort of fallacies in reasoning illustrated in the introduction in terms of "fair labelling", or with different classification levels in a security setting. An alternative would be to take the approach proposed by Soh and Demiris [26], where genetic algorithms are used to find the set of Pareto optimal solutions for a multi-objective POMDP. A solution is Pareto-optimal if it is not possible to improve any component of the expected reward vector without decreasing the value of another component. This approach assumes that the different components of the reward are of incomparable importance and requires the user to decide which of the Pareto optimal solutions to adopt. We propose a method that exploits the knowledge of the relative importance of each component to improve the efficiency of the planning algorithm. A similar approach is taken by Wray and Zilberstein [30], where they propose an algorithm to solve multi-objective POMDPs where the reward components are ordered according to their degree of importance. This work only deals with the single agent case, however, without considering the issue of coordination among agents.

Problems in which a coalition of agents collaborate in order to maximise a joint reward are often modelled as Decentralized Partially Observable MDPs (Dec-POMDPs) [2]. In a Dec-POMDP, each agent has a local and partial view of the environment, and must take a decision on what action to perform based only on its local observations. Finding a γ-approximation of an optimal policy for a Dec-POMDP[4] has been proven to be intractable [4]. For this reason, a substantial amount of research has focused on algorithms that can efficiently find sub-optimal solutions without providing guarantees on the solution quality. The vast majority of this focuses on quantitative models, where the joint reward is real-valued.

Wu et al. [32], for example, propose Point Based Policy Generation (PBPG), an algorithm for solving finite horizon Dec-POMDPs with real-valued rewards. The algorithm relies on a set of heuristics to find belief states (probability distributions over possible states) that are likely to be reachable after a given number of steps. Given an execution horizon H, the algorithm starts by finding the best one-step policies (a one-step policy consists of a single action) and evaluating them from the beliefs that are reachable at time $H - 1$. It then uses these policies as sub-policies to build a set of candidate two-step policies, which are evaluated from the beliefs reachable at time $H - 2$. The algorithm proceeds in this way until it builds the set of candidate policies for time 0. At each step, only the best *MaxTrees* policies are retained and used as possible sub-policies, resulting in bounded memory complexity and time

[4] Given a real γ, and an optimal policy with value V^*, a γ-approximation of this policy is a policy with value $V' \geq V^* - \gamma$.

complexity linear in the execution horizon. Because of this pruning, however, the algorithm does not provide guarantees on the quality of the solution with respect to the optimum.

To the best of our knowledge, the problem of qualitative decision making in decentralized, stochastic scenarios has been previously addressed only by Brafman et al. [9]. Their work is different in spirit, however. The authors build upon a simplified Dec-POMDP, where only qualitative statements about the possible transitions and observations are available, and a set of goal states is defined in place of a reward. They show that this problem can be solved using classical planning techniques. Their formalism does not permit the specification of different degrees of preferences among goals, however. Norms are often seen as constraints over the behaviour of (groups of) agents. From this point of view, our work is related to research on constrained Dec-POMDPs by Wu et al. [31]. Wu et al. consider a Dec-POMDP with a single reward function, but multiple cost functions. The objective is then to maximise the reward function, subject to constraints over cumulative costs. Rather than trying to minimise the number of constraint violations, their algorithm excludes all solutions that violate one or more constraints. Our aim is to find solutions that minimise the qualitative level of violation severity that occurs, and minimise the number of violations at each level.

Before moving on to present the two key contributions of this research (in Sects. 3 and 4), we outline a scenario that both illustrates the normative concepts that are core to the model and gives an intuition of the practical reasoning problem we address.

6.2 Discussion of alternatives and future research

Our starting point in this research is the extensive body of research in normative systems specification and reasoning. We model classical contrary-to-duty structures, which capture the important notion of reparation. We also argue for the related, but complementary notion of violation severity levels. There are, of course, assumptions we make in this research, and, as discussed in Sect. 6.1, practical reasoning under normative constraints is closely related to preference-based planning. In this discussion, therefore, we briefly explore alternative approaches to model violation severity, and alternative reward functions for N-Dec-POMDPs. We discuss some of the limitations of the mechanisms proposed and indicate some avenues for future research.

First we discuss *ceteris paribus* networks (CP-nets) [7] in more detail, which is a common means to capture preferences in planning domains. Given a set of variables, V, each of them with a domain of possible values, preferences may be expressed: given a variable $X_i \in V$, and a (possibly empty) set of variables $Y \subseteq V \setminus X_i$, we can specify a preference over different outcomes of X_i conditioned on a given assignment for the variables in Y. The fact that these preferences are valid only when all the other assignments are equal makes them weaker than the norm-induced preferences considered in our model. They do not allow us to represent CTD structures; consider, for example, the adaptation of the CTD structure concerning surveillance where we make explicit the idea that we only want one agent (the UAV or the helicopter) monitoring the restricted area at any time (see Fig. 14):

1. It ought to be that the UAV is monitoring: $O_1 = \mathbf{O}(m_u \mid \top)$.
2. If the UAV is monitoring, it ought to be that the helicopter is not monitoring: $O_2 = \mathbf{O}(\neg m_h \mid m_u)$.
3. If the UAV is not monitoring, it ought to be that the helicopter is: $O_3 = \mathbf{O}(m_h \mid \neg m_u)$.

The most preferred world is one in which the UAV is monitoring, but the helicopter is not. The two worlds where only O_2 is violated $(\neg m_u \wedge m_h)$ or only O_1 is violated $(m_u \wedge m_h)$

Fig. 14 Deontic preference model for the surveillance example

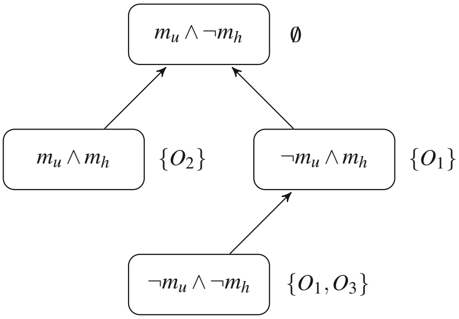
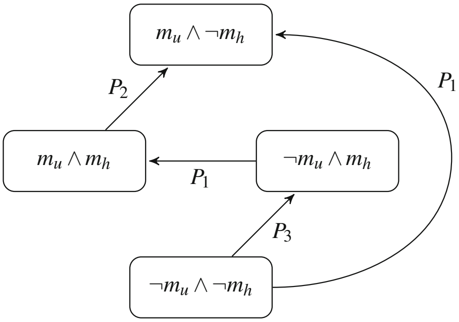


Fig. 15 CP-Net model for the surveillance example



are incomparable, and worlds in which neither agent is monitoring the area (O_1 and O_3 are violated) is least compliant.

We may attempt to capture this as a CP-Net, the preferences for which are illustrated in Fig. 15, thus:

1. $P_1 : m_u \prec \neg m_u$
2. $P_2 : \neg m_h \prec m_h$ if m_u
3. $P_3 : m_h \prec \neg m_h$ if $\neg m_u$

Each arrow is labelled with the statements that induce the corresponding preference, and the result is a complete ordering over possible worlds. While this ordering is consistent with that obtained in our model, it introduces an additional constraint: $m_u \wedge m_h \prec_w \neg m_u \wedge m_h$, which is not entailed by the normative specification. We can introduce this additional constraint by stating that a violation of O_1 is more severe than a violation of O_2, but that may not be what is intended. In fact, CP-Nets implicitly consider an unconditioned preference over a variable X_i as being more important than another preference that is conditioned on the value of X_i. We may remove the preference P_3, making the two worlds $m_u \wedge m_h$ and $\neg m_u \wedge m_h$ incomparable, but this would also result in worlds $m_u \wedge \neg m_h$ and $m_u \wedge \neg m_h$ being incomparable. These two worlds differ in all their variables, and CP-Nets do not offer a way to specify a direct preference between them.

In subsequent research, Brafman et al. [8] extend CP-Nets by introducing preference relationships among variables. If, for example, variable X_1 is more important than X_2, we should always prefer an improvement in X_1 to one of X_2. This addresses some of the limitations of CP-Nets, but the two variables concerned must be mutually preferentially independent. In other words, the preference over the values of one variable must not depend on the value of the other variable. Since, in our example, the preference over m_h depends on the value of m_u, importance relationships are not sufficient. Thus, CP-Nets cannot be used to express contrary-to-duty obligations.

It may be argued that the severity relation does not add to the expressiveness of a model of norms that already includes contrary-to-duty structures. Given a desired ranking of worlds, it is always possible to define a normative system that uses only CTD norms, and that would result in the ranking required. If we say that L_i is the boolean expression that identifies all the worlds at the ith level, we could define a normative system as:

- $\mathbf{O}(L_1 \mid true)$
- $\mathbf{O}(L_2 \mid \neg L_1)$
- …
- $\mathbf{O}(L_n \mid \wedge_{i=1}^{n-1} \neg L_i)$

In order to do this, however, it would be necessary to know *in advance* the desired ranking of worlds, which is not trivial. Moreover, our approach enables a more straightforward and natural formalization for the same normative system. The mechanisms presented in Sect. 3 can then be used to efficiently compute exactly this ranking.

Our severity specification is defined as a strict partial order over single obligations; i.e. $P_o \subseteq OS \times OS$. This allows us to specify that the violation of one obligation is more severe than any number of violations of another. Moreover, if we define $O_1 \succ_o O_2$ and $O_1 \succ_o O_3$, worlds violating both O_2 and O_3 will be preferred to worlds that violate O_1. It would be interesting to consider an alternative relation: $P_o \subseteq 2^{OS} \times 2^{OS}$. We could then express relationships such as $\{O_1\} \succ_o \{O_2\}$, $\{O_1\} \succ_o \{O_2\}$ and $\{O_2, O_3\} \succ_o \{O_1\}$. An interesting direction for future research would to be study how to compute a meaningful, acyclic preference relation over possible worlds, given this richer severity specification. Assuming such an ordering can be reliably computed, this alternative domain analysis could be directly used as input to our N-Dec-POMDP solver.

The reward function of an N-Dec-POMDP favours histories where states associated with an higher ranking level are visited less often. This approach may lead to unexpected results in some situations that involve independent[5] norms of incomparable severity. Consider two histories, h_1 and h_2. History h_1 consists of a sequence of states, $(s_{1.1}, s_{1.2})$, such that in state $s_{1.1}$ there are no violations, but in state $s_{1.2}$ both obligations O_1 and O_2 are violated. History h_2 consists of a sequence of states, $(s_{2.1}, s_{2.2})$, such that in state $s_{2.1}$, O_1 is violated, in $s_{2.2}$, O_2 is violated. Assuming that violations of O_1 and O_2 are incomparable with respect to their severity, we might expect these two histories to be equally good (or bad). In the model proposed here, state $s_{1.2}$ in history h_1 would lie at a higher ranking level than either states $s_{2.1}$ or $s_{2.2}$ in h_2, and hence h_2 will be preferred to h_1. The reason for this is that our objective is not to minimise the sanctions received as a result of norm violation, but to minimise the possible consequences of these violations. The goal of the norm analysis phase is to ensure that more severe consequences are associated with higher ranked states. Of course, this does not guarantee that any increase in ranking is associated with more severe consequences.

An alternative reward function for an N-Dec-POMDP may be defined that would result in histories h_1 and h_2 being assessed as equally good. We could, for example, rank all the obligations according to their severity using an adaptation of Algorithm 2, applied to the set of norms rather than the set of possible worlds. We may then give rewards to states that equate to, for each ranking level l, $-n.\varepsilon^{\Lambda-l}$ where n is the number of violations at level l. Histories h_1 and h_2 would then have the same reward. It is not clear, however, how we could capture the fact that violating contrary-to-duty norms should be considered less desirable than violating the corresponding primary norms, which is an important aspect of our model.

[5] Two norms are independent if neither is a contrary-to-duty obligation of the other.

In our model, norm compliance is necessarily evaluated on single states. On the face of it, this restricts the types of norm that can be represented. In many domains, for example, obligations may include a deadline for fulfilment: temporally-extended norms. Norms may also link individual actions, such as in separation of duty constraints where two actions must be performed by two different agents. In order to evaluate compliance with such norms, we must take into account sub-histories rather than individual states. It is possible, however, to directly extend our model to consider such norms by keeping track of the evolution of norm instances (activation, expiration, etc.) in each state, as discussed in previous research [16]. The cost is an (potentially significant) increase in the number of states, placing additional burden on the planner.

In this paper, we have focussed exclusively on normative motives. These are social drivers of action, but autonomous agents may also be driven by individual goals. Individual goals may be encoded as obligations, but this would be to combine compliance to social expectations and individual drives. Severity could be used to capture the relative importance of individual goals and norms if goals are expressed as obligations. It may, however, be more appropriate to make explicit the distinction between social norms and individual goals. We could then employ multiple-objective optimisation methods to manage the trade-off between remaining compliant with social expectations and satisfying individual goals. This is an avenue for future research, given that a suitable approach would need to account for the naturally qualitative nature of the reward function for norm-governed planning.

7 Conclusions

In the introduction, we claimed contributions both to modelling and practical reasoning in normative multi-agent systems, and to algorithms for decentralised planning under uncertainty. For the former, we have presented what we believe to be the first end-to-end model from the analysis of a domain where the behaviour of agents is governed by norms, through to a decentralised planning mechanism for multiple agents to act in concert such that they maximise their compliance with these norms. We consider normative system specifications that include guidance for recovering from violations (contrary-to-duty obligations) and avoiding critical levels of failure (severity). The domain analysis mechanism proposed is guaranteed to generate a transitive and acyclic preference relation over possible worlds. This preference relation enables possible worlds to be ranked from the most to least compliant. This is then used to guide collective decision making in the presence of uncertainty, with the goal of maximising the expected compliance of states in an execution history.

The N-Dec-POMDP planning mechanism is an adaptation of Dec-POMDPs for use with a qualitative reward function. Our greedy LP algorithm approximately solves an N-Dec-POMDP by starting with the problem of optimising against the highest levels of the reward function, adding additional constraints associated with lower levels until no significant improvement can be found. The most-critical-states (MCS) heuristic also exploits the qualitative structure of the reward function to guide planning effort. From the results obtained from evaluating this planning mechanism, we may reliably conclude that both the greedy LP and the MCS heuristic provide significant and considerable savings in terms of execution time without affecting the quality of policies computed.

Acknowledgements This research was funded by Selex ES. The software developed during this research, including the norm analysis and planning algorithms, the simulator and harbour protection scenario used during evaluation is freely available from doi:10.5258/SOTON/D0139.

References

1. Alechina, N., Dastani, M., Logan, B. (2012). Programming norm-aware agents. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, (pp. 1057–1064).
2. Amato, C. (2014). Cooperative decision making. In M. J. Kochenderfer (Ed.), *Decision making under uncertainty: Theory and application*. Cambridge: MIT Press.
3. Ashworth, A. (2006). *Principles of criminal law* (5th ed.). Oxford: Oxford University Press.
4. Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, *27*(4), 819–840.
5. Bienvenu, M., Fritz, C., McIlraith, S.A. (2006). Planning with qualitative temporal preferences. In: Proceedings of the 10th International Conference on Knowledge Representation and Reasoning, (pp. 134–144).
6. Bonet, B., Pearl, J. (2002). Qualitative MDPs and POMDPs: An order-of-magnitude approximation. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, (pp. 61–68).
7. Boutilier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., & Poole, D. (2004). CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, *21*, 135–191.
8. Brafman, R. I., Domshlak, C., & Shimony, S. E. (2006). On graphical modeling of preference and importance. *Journal of Artificial Intelligence Research*, *25*, 389–424.
9. Brafman, R.I., Shani, G., Zilberstein, S. (2013). Qualitative planning under partial observability in multi-agent domains. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence, (pp. 130–137).
10. Castelfranchi, C. (2003). Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic*, *1*(1–2), 47–92.
11. Chisholm, R. M. (1963). Contrary-to-duty imperatives and deontic logic. *Analysis*, *24*(2), 33–36.
12. Dignum, F., Morley, D., Sonenberg, E.A., Cavedon, L. (2000). Towards socially sophisticated BDI agents. In: Proceedings of the 4th International Conference on Multi-Agent Systems, (pp. 111–118).
13. Fagundes, M. S., Billhardt, H., & Ossowski, S. (2010). Normative reasoning with an adaptive self-interested agent model based on Markov decision processes. In A. Kuri-Morales & G. R. Simari (Eds.), *Advances in artificial intelligence—IBERAMIA 2010. Lecture notes in computer science* (Vol. 6433, pp. 274–283). Berlin: Springer.
14. Forrester, J. W. (1984). Gentle murder, or the adverbial Samaritan. *The Journal of Philosophy*, *81*(4), 193–197.
15. Gasparini, L., Norman, T. J., Kollingbaum, M. J., & Chen, L. (2015). Severity-sensitive robustness analysis in normative systems. In A. Ghose, N. Oren, P. Telang, & J. Thangarajah (Eds.), *Coordination, organizations, institutions, and norms in agent systems X. Lecture notes in computer science* (Vol. 9372, pp. 72–88). Berlin: Springer.
16. Gasparini, L., Norman, T.J., Kollingbaum, M.J., Chen, L. (2016). Decision-theoretic norm-governed planning. In: Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems, (pp. 1265–1266).
17. Gerevini, A., & Long, D. (2005). *Plan constraints and preferences in PDDL3* (Vol. 75). Brescia, Italy: Department of Electronics for Automation, University of Brescia.
18. Kagan, S. (1988). The additive fallacy. *Ethics*, *99*(1), 5–31.
19. Kahn, A. B. (1962). Topological sorting of large networks. *Communications of the ACM*, *5*(11), 558–562.
20. Meneguzzi, F., Luck, M. (2009). Norm-based behaviour modification in BDI agents. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, (pp. 177–184).
21. Prakken, H., & Sergot, M. (1996). Contrary-to-duty obligations. *Studia Logica*, *57*, 91–115.
22. Prakken, H., & Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations. In D. Nute (Ed.), *Defeasible deontic logic. Synthese library* (Vol. 263, pp. 223–262). Berlin: Springer.
23. Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. Hoboken: Wiley.
24. Roijers, D.M., Whiteson, S., Oliehoek, F.A. (2015). Point-based planning for multi-objective POMDPs. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, (pp. 1666–1672).

25. Seuken, S., Zilberstein, S. (2007). Memory-bounded dynamic programming for DEC-POMDPs. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, (pp. 2009–2015).
26. Soh, H., Demiris, Y. (2011). Evolving policies for multi-reward partially observable Markov decision processes (MR-POMDPs). In: Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, (pp. 713–720). ACM.
27. Spaan, M. T., & Vlassis, N. (2005). Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, *24*, 195–220.
28. Tijs, S. (2006). Lexicographic optimization on polytopes is linear programming. Discussion paper, Tilburg University, Center for Economic Research
29. van der Torre, L., & Tan, Y. H. (1999). Contrary-to-duty reasoning with preference-based dyadic obligations. *Annals of Mathematics and Artificial Intelligence*, *27*(1–4), 49–78.
30. Wray, K.H., Zilberstein, S. (2015). Multi-objective POMDPs with lexicographic reward preferences. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, (pp. 1719–1725).
31. Wu, F., Jennings, N., Chen, X. (2012). Sample-based policy iteration for constrained DEC-POMDPs. In: Proceedings of the 20th European Conference on Artificial Intelligence.
32. Wu, F., Zilberstein, S., Chen, X. (2010). Point-based policy generation for decentralized POMDPs. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, (pp. 1307–1314).