The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Requirements for Aligned, Dynamic Resolution of Conflicts in Operational Constraints

Steven J. Jones, Robert E. Wray, John E. Laird
Center for Integrated Cognition at IQMRI
Ann Arbor, MI 48105 USA
{steven.jones,robert.wray,john.laird}@cic.iqmri.org

Abstract

Deployed, autonomous AI systems must often evaluate multiple plausible courses of action (extended sequences of behavior) in novel or under-specified contexts. Despite extensive training, these systems will inevitably encounter scenarios where no available course of action fully satisfies all operational constraints (e.g., operating procedures, rules, laws, norms, and goals). To achieve goals in accordance with human expectations and values, agents must go beyond their trained policies and instead construct, evaluate, and justify candidate courses of action. These processes require contextual "knowledge" that may lie outside prior (policy) training. This paper characterizes requirements for agent decision making in these contexts. It also identifies the types of knowledge agents require to make decisions robust to agent goals and aligned with human expectations. Drawing on both analysis and empirical case studies, we examine how agents need to integrate normative, pragmatic, and situational understanding to select and then to pursue more aligned courses of action in complex, real-world environments.

Code and Technical Appendix —
https://github.com/Center-for-Integrated-Cognition/OAMNCC

Introduction

In open environments, agents will face conflicting goals and constraints (Freuder and Wallace 1992), implicit expectations (Shah et al. 2019), and incomplete information (Kaelbling, Littman, and Cassandra 1998). However, the behavior of autonomous agents in these open environments should remain *aligned*: their behavior must appear reasonable and appropriate to human observers and be consistent with their expectations, laws, norms, etc.

Alignment in open environments cannot be wholly solved with *a priori* training; agents will face difficulty pursuing their intended objectives in specific, dynamic, and often unfamiliar/novel environments (Langosco et al. 2022). This difficulty, and the benefits of justified deviations from external constraints, indicate a need to go beyond static preference satisfaction or fixed rules to more dynamic models of alignment (Milli et al. 2017).

We emphasize one aspect of such alignment: conforming to constraints. In real-world environments, autonomous agents are subject to a dense network of constraints: formal regulations and laws, social norms, organizational rules (e.g., military doctrine), and task-specific guidelines and preferences. These constraints are often ambiguous, underspecified, or even contradictory in practice, especially when instantiated in complex or dynamic environments (Wray, Jones, and Laird 2023). Human decision-making in such contexts relies on interpretive flexibility, sensitivity to context, and the ability to weigh competing considerations (Klein 1998; Payne, Bettman, and Johnson 1993). Similarly, autonomous systems must be able to evaluate whether their overall course of action is appropriate given the situational constraints and tradeoffs involved. Eliminating all conflicts through training or rule design alone is not feasibile: agents must revise their commitments and plans in response to dynamic contexts (Bratman 1987), which include large, open-ended, and novel combinations and interactions among constraints.

Our initial research toward online conflict mitigation clarified the vast scope and complexity of agent knowledge needed for comprehensive mitigation (Jones and Wray 2024; Wray, Jones, and Laird 2023). Here, we step back from evaluating specific, implemented solutions and characterize the problem and its solution space more completely. We adopt a knowledge-level perspective (Newell 1982). Rather than specify how an agent should respond when facing novel constraint conflicts (i.e., ones not anticipated pre-deployment, such as pre-training), we identify types of knowledge an agent must possess for identifying and selecting aligned courses of action.

We refer to "knowledge" in the broad, functional sense of Newell's "knowledge level." The analysis treats knowledge as what an agent must be able to represent and do to achieve specific outcomes in a conflict, without committing to underlying realizations (e.g., symbolic vs. distributed representations), and regardless of whether the knowledge is procedural or declarative. The analysis identifies specific agent capacities (such as the expressivity of representations, contextual sensitivity, and inferential capacities) relevant to achieving operationally successful and aligned behavior.

Beyond pre-learned rules or preference hierarchies, we argue that such knowledge must include interpretive resources such as frames for understanding different kinds of norms, background expectations about tradeoffs and excep-

---

Algorithm 1: Schematic process model for Online, Aligned Mitigation of Novel Constraint Conflicts (OAMNCC).

```
1: Recognize novelty of the conflict
   (similar to "out of distribution" assessment)
2: Assess conflict type and structure
3: Evaluate what situational information might be relevant
   to the conflict (expand space of consideration)
4: Propose candidate conflict-mitigating courses of action,
   evaluate them, and select one
5: Execute the selected course of action, monitoring for
   resolution of conflict
```

tions, and the ability to assess "reasonable" deviation from a constraint. The knowledge-level analysis aims to establish general requirements, regardless of an agent's implementation, to navigate conflicts in a manner that aligns with human expectations and values in similar decision-making contexts.

The paper develops this knowledge-level perspective by enumerating types of constraint conflicts and illustrating (via representative scenarios) how different knowledge sources support conflict-resolution behavior. The scenarios highlight information agents must access to construct and justify appropriate courses of action. These examples ground proposed knowledge requirements that can inform both design and evaluation of future systems. By clarifying the kinds of knowledge needed to navigate constraint conflicts in open contexts, this work provides a foundation for the development of future agents whose behavior will remain coherent, context-sensitive, and aligned, even in environments beyond prior specification.

Problem Definition and Initial Analysis

Algorithm 1 outlines the processing an agent might perform to mitigate conflicts in novel conflict situations. The outline is based on aforementioned early prototypes of such a capability and reflects high-level commitments similar to those of Goal-Driven Autonomy (Molineaux, Klenk, and Aha 2010). Such processing must occur in the performance environment (*online*), and the outcomes it produces should be aligned with human expectations. For brevity of presentation, we use OAMNCC as a shorthand for this capability.

In Step 1 of the process, an autonomous agent must recognize novel conflicts. Mechanisms for such recognition are comparable to out-of-distribution detection in learning systems. However, for mitigation, as for learning, detection must occur *before* taking action using existing policy (Haider et al. 2024). Step 2 reflects an ability to characterize the detected problem among constraints that needs solving. In Steps 3 and 4, an agent weighs what information in its context is relevant and generates candidates that exploit newly-identified information. (As we discuss below, these steps require further specification, which we discuss in the following sections.) Using this algorithm as a scaffold, we seek to identify what knowledge is required to enable OAMNCC in an agent.

Initial investigations identified a few abstract requirements (similar to those identified by others; e.g., Kuipers 2018). Agents must demonstrate awareness of the frames by which humans interpret constraints (e.g., duty, norms, utility). Agents must also be able to represent preferences for responding to conflicts with a similar level of expressivity to human preferences. A constraint-compliant agent must also represent: its situation, how constraints bear on its behavior, its available courses of action, and what those actions can be expected to achieve (or, minimally, whether they can be expected to violate constraints). All of these kinds of knowledge are necessary. If any one is missing, compliance is infeasible. When all are present, an agent can feasibly select courses of action in its current situation that it expects will comply with constraints (Wray, Jones, and Laird 2023; Rao and Georgeff 1995).

While useful, these requirements are incomplete, given the overall space of constraint conflicts. Table 1 summarizes the different types of conflicts relevant to conflict mitigation. The table spotlights the (very large) scope of the problem. Any mitigation solution must span all of these different types of conflicts. We use this list to assess coverage of the knowledge-level analysis over the entirety of the mitigation challenge. The list also hints at the knowledge needed to assess conflicts (i.e., perform Step 2).

To motivate the taxonomy, we construe conflicts as arising when the knowledge content of at least one of the representations outlined previously is insufficient for a constraint-compliant choice. Thus, the major types of conflicts result from different kinds of knowledge inadequacy. When representations of how constraints bear on behavior entail contradictory or involve incommensurable values, the agent faces an unsatisfiable conflict of constraint semantics. When the constraints are coherent, but the agent's representations of its situation, actions, and their expected outcomes reveal no fully compliant course of action, the agent faces a situational conflict. Situational conflicts can arise from mutual exclusivity or uncertainty (such as from abstract constraint specification or partial observability). Finally, multiple, different conflict types may manifest in a given conflict situation.

We observe that some conflict types have been more thoroughly investigated than others, and some are more amenable to semantic analysis and pretraining. For example, techniques have been identified to mitigate unsatisfiability conflicts at the time the constraint specifications are introduced, especially rule-oriented constraints (Censi et al. 2019). This enables offline, pretraining approaches to resolve such conflicts. Generally, conflict types inherent to the definition and expression of constraints themselves can be more readily avoided via preparatory strategies such as training. We thus focus on situational conflicts. These conflicts involve mutual exclusivity and uncertainty from interactions between constraints and the open-world situation, where it is not feasible to anticipate all possible interactions in advance.

Example Constraint Conflict Scenarios

We introduce scenarios designed to illustrate additional knowledge needed for timely, coherent conflict mitigation that is aligned with human expectations (OAMNCC).

The scenarios here build on ones previously developed by researchers at the Naval Postgraduate School (NPS; Brutz-

Conflict Category	Cause of the Conflict
Unsatisfiability	Two or more constraints' logical entailments specify contradictory prescriptions on behavior.
Infeasibility	Two constraints specify prescriptions on behavior that are directly contradictory.
Incommensurability	Constraints appeal to value systems without a common unit of measure for compliance.
Mutual Exclusivity	Constraints may be coherent, but available actions do not afford complete compliance.
Resource Contention	Two or more required actions compete for the same insufficient finite resource.
Temporal	The finite resource is time.
Spatial	The finite resource is space.
Causal Preclusion	Action to satisfy one constraint changes the world, making another constraint impossible.
Uncertainty	A conflict arises or is made worse by a lack of knowledge of the state or expectations.
Epistemic	The agent lacks knowledge about the world.
Probabilistic	The world is stochastic, so outcomes cannot be determined.

Table 1: A Taxonomy of Constraint Conflict Types

man, Blais, and Hsin-Fu 2020).[1]In these scenarios, an agent "commands" an autonomous surface vessel, a role comparable to a ship's captain, in a naval domain where dynamic conflicts are likely to arise. The original scenarios do not explicitly detail constraint conflicts. However, the scenarios do describe abstract "course of action" policy knowledge designed to comply with real-world, operational constraints (e.g., rules of engagement, specific limited resources). We adapt these scenarios to illustrate specific types of conflicts.

Each scenario introduces one type of conflict that could arise in that situation. We then interrogate what kinds of knowledge are required to resolve the scenario in a way that appears consistent with human expectations. These represent knowledge requirements. We note each individual requirement appearing in each scenario and then discuss the requirements collectively in the following section. The approach is analytic: we did not conduct studies to determine specific decisions that humans would make, as others have done in explorations of aligned, online sequential decision-making tasks (e.g., Loreggia et al. 2022). However, these scenarios are sufficiently straightforward that it is readily evident how the information we identify is relevant to resolving the conflict. Further, failing to take this information into account leads to behavior and outcomes inconsistent with (likely) human expectations and values.

Two further introductory assumptions:

- OAMNCC proceeds as outlined in the process model introduced above. The agent has available a few different online decision-making strategies when novel conflicts arise. The specific decision strategies are not intended to be exhaustive or comprehensive. Rather, they are reasonable for these domains and sufficiently familiar to enable us to isolate the importance of knowledge that would be expected to contribute to agent decision making.
- Consistent with a knowledge-level framing, we assume relevant knowledge is available for decision making. We do not evaluate the differential impact of immediate accessibility (i.e., recall from memory or *knowledge search*) vs. search for the information (internal or external *problem search*). Knowledge and problem search strategies are critical to the quality and timeliness of decisions; future efforts to implement OAMNCC will need to address both.

From a methodological perspective, we designed specific simulation scenarios (implemented in Python) to support the different uses cases described in the paper. Scenarios were designed with distributions over various parameters (e.g., initial positions of objects). Data were collected for each scenario instance across the distribution(s) relevant to that scenario. Aggregate statistics across these runs constitute the results presented below.

Sailor Overboard

The drone ship has an urgent/high-priority need to return to base (RTB). During transit, a sailor is lost overboard without alert. The loss is then discovered 20nm from the ship's destination, as illustrated in Figure 1. The ship backtracks and searches until fuel runs so low that returning to base is endangered. In the scenario, the sailor overboard occurs 25-35nm from port (5-15nm from the alert location). Ignoring differences in fuel consumption for transit vs. search-and-rescue activities, the ship can safely backtrack 10nm and still reach port before fuel is exhausted. The rescue operation itself is not instantaneous and requires time and fuel resources.

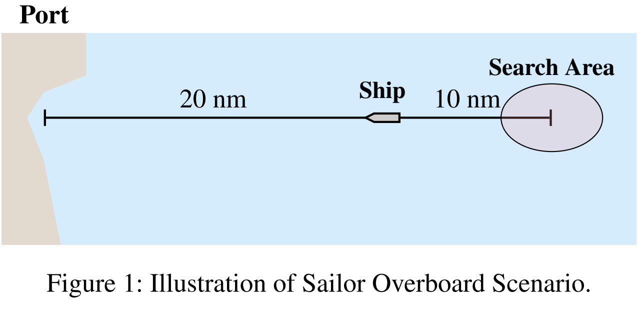

Figure 1: Illustration of Sailor Overboard Scenario.

---

[1] Specific parameters for general capabilities, speed, and fuel consumption were adopted from public data for appropriate ship types. Events such as the pirate attacks and sailor rescue are modeled with simple probability distributions reflecting estimates from available data.

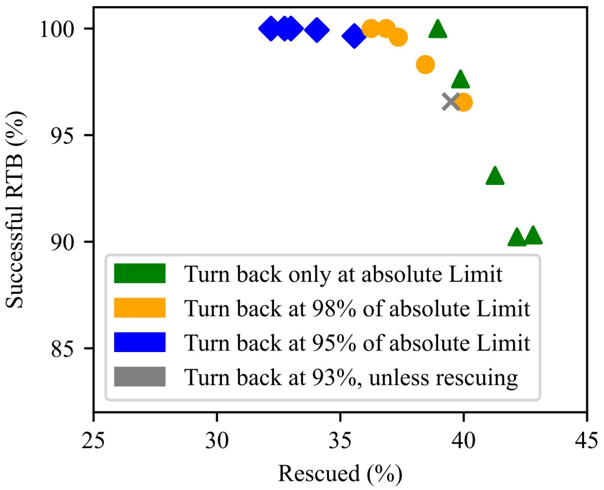

Figure 2: Utilitarian assessment of sailor overboard scenario. Colored dots represent different safety margins. Identical marks vary in the ratio of rescue/RTB importance. × represents a policy that never leaves a sailor behind once spotted.

A utilitarian decision frame can be used to assess the costs of returning to base vs. completing the rescue. Such a strategy can be tuned to achieve a desired evaluation of the tradeoff (Meseguer, Rossi, and Schiex 2006). Figure 2 illustrates the tradeoff over 1000 simulated scenarios. The figure shows the relationship between successful rescues (x-axis) and successful RTBs (y-axis) as a function of two criteria: margin of error (different colors) and relative weighting of importance between rescue and RTB (different dots of the same color). For example, the 95%-conservative policy (blue) curtails the number of rescues, regardless of the relative cost of rescue in comparison to RTB.

From the figure, it is evident that a policy tuned to the “knee” of the curve prevents excessive rescue labor for diminishing returns and is also less sensitive to the specific, relative importance/cost of RTB vs. rescue, a value is not likely to be known or explicit in this domain. Assuming the urgency and importance of the RTB order, this decision framing and knowledge at first appear an apt (or at least sufficient) decision strategy.

This presentation hides an important detail, however. This framing results, in some cases, with the vessel turning back *after* having spotted a sailor. Such decisions contradict a norm to treat rescue as a duty once a rescue target is spotted (McKie and Richardson 2003; Jenni and Loewenstein 1997). Such a responsibility is a form of relevant information needed for Step 3 of Algorithm 1. A policy that achieves nearly the same performance (i.e., a candidate response proposed in Step 4 of the process model, marked with × in Figure 2) can be attained by turning back to base either when returning to base is jeopardized (but a sailor is not spotted), or returning only after their rescue (once spotted).

This scenario highlights an aspect of human preferences that needs to be reflected in agent knowledge: humans have preferences defined with respect to specific grounding statuses of constraints (Rq: *expressive preferences*).

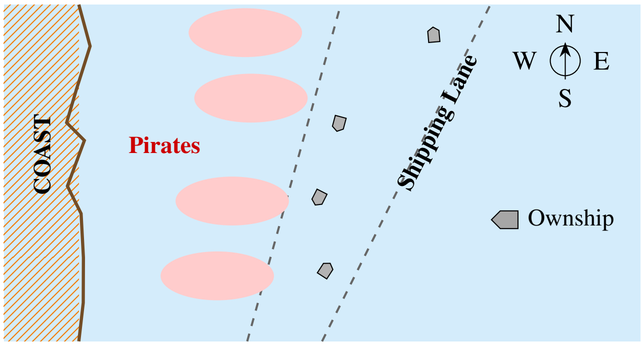

Figure 3: Illustration of the Piracy Interdiction Scenarios.

Piracy Interdiction

In the remaining scenarios, we envision a mission context in which the autonomous drone ship is responsible for protecting a shipping lane from pirates operating in small boats from coastal waters. In response to the effectiveness of the drone ship, the pirates are adopting a new strategy to mass multiple, coordinated attacks within the drone’s area of operation (AO). In these scenarios, the pirates conduct four simultaneous attacks on four merchant vessels in shipping lanes within the drone’s AO, as sketched in Figure 3. By design, the agent will only be capable of attempting to interdict one attacking pirate band.

The simulation models pirate attacks with a 95% success rate (Leontarakis 2015) and taking place for up to 30 minutes (ICC International Maritime Bureau 2024). For the results in this section, we simulated 1000 instances of the scenario. The main parameter in each instance was the initial location of merchant ships in the sea lane. As specified by NPS, the appropriate behavior in an instance of piracy interdiction involves multiple steps (identity friend and foe, force escalation, etc.). Thus, we assume that a response guaranteeing the safety of a selected merchant vessel consumes enough time that the drone ship, at full speed, could not reach another merchant within a 30-minute attack window.

The agent receives no advance warning of the pirate attacks. Once within attack range, individual pirate attacks have a probability of successful boarding every 1 min of simulated time. Unimpeded attacks conclude with eventual pirate success 95% of the time, or once a maximum engagement-time threshold is exceeded 5% of the time. If the drone ship reaches the pirates before their attack succeeds, the simulation assumes a successful interdiction (i.e., pirate attack fails).

We model the agent’s response to the conflict as the selection of one of four abstract courses of action, given the agent’s available information. Typically, interdiction is a duty (standing order). In this case, the conflict occurs between four instances of the same duty. Notably, simple prioritization rules among constraints will not resolve the conflict. Additional information is required. We assume the

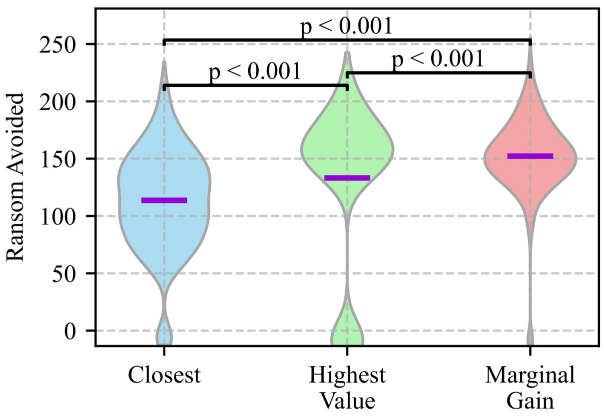

Figure 4: Distributions of the ransom avoided by interdiction relative to baseline (no action) for different decision strategies (1000 trials; purple, horizontal lines identify means). The distributions are significantly different according to Kolmogorov-Smirnov tests.

agent has accessible, relevant information (i.e., in Step 3 of Algorithm 1), including the approximate ransom values of targeted ships, the locations of targeted ships, and an estimate of pirate success (boarding) given the time it will take to reach each merchant vessel.

We consider three strategies available to the agent: two heuristics (interdict closest attack; interdict attack on highest ransom ship) and a utilitarian assessment based on the expected marginal gain of interdiction. The agent can estimate the outcomes for each strategy and use these outcomes as a basis for choosing one attack to interdict. When its understanding is complete and accurate, the agent's assessments provide apt choices. In Figure 4, which plots outcome distributions for the three strategies over 1000 scenario instances, the utilitarian strategy, which has more complete knowledge of the situation (probability of successful interdiction and merchant ransom value combined, or "marginal gain"), is better than the others (is selected in Step 4 of Algorithm 1).

This scenario identifies knowledge relevant to enabling OAMNCC. The agent considers all its decision strategies because they each enable the agent to differentiate among candidate courses of action (Step 4). Further, marginal utility requires the knowledge to "reframe" duty constraints (orders) as utilitarian constraints (Rq: *constraints & frames*).

Merchants with Water Cannons This scenario is a variation of the interdiction scenario. We assume new information is available to the agent. The ship has recently received a command memo that indicates water cannons, which have previously been limited to installation on only the largest tankers, are now being installed on all sizes of tankers and container ships as well. Water cannons enable merchant vessels to mount an autonomous, reasonably effective defense against typical pirate attacks.

What impact does this information have on the decision about which pirate to interdict? First, let us assume that the agent is incapable of dynamically updating its statistical expectations to take into account the command memo. Under this assumption, the agent's assessment of the interdiction choice remains the same as in Figure 4. However, in actuality, using the strategies defined previously, but ignoring the water-cannon memo, produces the distribution of outcomes in Figure 5 (left of the dashed line). Failing to include water cannons in the utilitarian frame causes this gap.

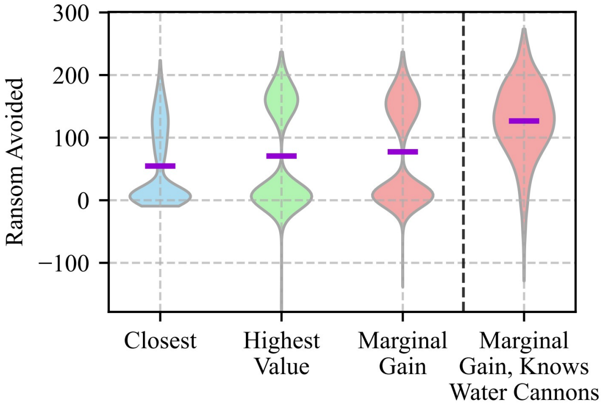

Figure 5: Distributions of ransom avoided for each strategy when two merchant vessels can deploy water cannons. Left of dashed vertical line: the agent ignores water cannon capabilities.

Now, assume the agent is capable of dynamically updating its expectations (Algorithm 1 Step 3). While it lacks the statistical confidence gained from policy training in this instance, the provenance of the memo, a command directive, leads it to modify its framing. The agent's expectations for this situation are illustrated to the right of the dashed line in Figure 5. Juxtaposing the differences in these outcomes demonstrates the impact of not taking the new information into account. Ships that could have defended themselves are protected, leaving others unnecessarily vulnerable.

This variation reemphasizes the continual need to update situation model knowledge (Rq: *dynamic situation model*). However, it also suggests the need to assess information quality based on multiple factors. If the information about water cannons came from a random social media post, rather than command, should it have the same impact on the world model? Here, the provenance of the information is a key factor in whether/when to update one's world model. Assessment of *information quality*, extending to *provenance*, is necessary and emphasizes the critical role for meta-knowledge (and its online evaluation) in OAMNCC (Rq: *information quality*).

Piracy and vessel adrift In this scenario, there is an imminent threat of piracy on one merchant vessel, and there is simultaneously an unidentified mass spotted adrift. Here, we describe how integrating additional sources of knowledge contributes to improved outcomes but omit empirical results.

The drifting mass requires ship attention because adrift

Knowledge Type	Definition
World Knowledge	Knowledge of embodiment and constraints needed for human-aligned responses
Constraints & Frames (**CF**)	Knowledge of constraints, constraint frames, knowledge of how to apply frame(s) to constraint(s) not already associated with that frame
Expressive Preferences (**EP**)	Sensitivity of preferences to specific states of grounding of a constraint. Ability to prioritize among constraints directly and by constraint properties such as types (e.g., laws, etiquette) and sources (e.g., inferred from observation, written regulation)
Action Affordances (**AA**)	Task-general specifications of affordances, recognition of applicability of affordance not explicitly associated with current context
Dynamic Situation Model (**SM**)	Sensitivity to environment or memory enabling updated representations of situation and near-term expectations (e.g., action models)
Metaknowledge	Refers to other knowledge, knowledge that determines what information to trust and use
Information Quality (**IQ**)	Reliability of other sources of knowledge through properties such as provenance or statistical certainty. May require the ability to act to improve knowledge of quality
Mitigation Utility (**MU**)	Whether the referent information or strategy will differentiate conflicted courses of action
Conflict Structure (**CS**)	Type of conflict (i.e., from Table 1, inter-constraint or inter-constraint conflict, etc.)

Table 2: A Taxonomy of Required Knowledge Types for Online Aligned Mitigation of Novel Constraint Conflicts

objects ordinarily require ship inspection to ensure safety in the shipping lanes and identify lost cargoes. Again, the agent faces a conflict due to temporal resource contention. If the ship does not immediately track the object, it may be difficult to find again. However, tracking and inspection would preclude timely interdiction of the pirates. Further, without some initial inspection, the actual value/cost of ignoring the mass is wholly unknown to the agent.

In this case, consider the result if the agent assesses the applicability of a novel affordance (as part of evaluating relevant information in Algorithm 1 Step 3). We adopt this idea from the original NPS scenarios, in which a new aerial drone capability is used to track a lifeboat adrift until direct assistance can be offered. Here, the agent tasks a ship-launched, aerial drone to fly to the mass and mark it with a beacon. This allows immediate interdiction and the likely ability to inspect the flotsam subsequently. Thus, considering alternative/new affordances enables resolving the temporal resource conflict.

The resolution relies on knowledge of *action affordances* outside of their usual context (lifeboat rescue) and of *conflict structure* (temporal resource contention) (Rq: *action affordances, conflict structure*). Further, the course of action is being chosen based on knowledge of the conflict *mitigation utility* of that action derived from the relation between the affordances of the action and the conflict type (Rq: *mitigation utility*).

Knowledge Requirements for OAMNCC

These scenarios bring to light more detailed knowledge requirements for OAMNCC, summarized in Table 2.[2] That is, the analysis from the prior section demonstrates that this knowledge is necessary for resolving conflicts across the space identified in Table 1. Although specific knowledge contents will vary across domains, we claim that these same kinds of processes and categories of knowledge and processes will generally be needed across domains. While we do not claim this list of requirements is sufficient for every conflict, these requirements further clarify the challenging scope for OAMNCC.

The more detailed analysis also allows us to map each type of knowledge to the process model introduced previously. Algorithm 2 highlights where different types of knowledge are needed to influence the conflict mitigation process, which we discuss in more detail as follows:

Step 2: Assessing the conflict provides a basis for subsequently evaluating the relevance of other situation information. Assessment requires not only knowledge of conflict type but also diagnostic knowledge, such as identifying if a conflict is *intra-constraint* (e.g., conflicting instances of the same duty). Significantly, this step determines what situational information to take into account when the features in existing policies may be insufficient to resolve a conflict.

Step 3: Generating candidate, off-policy courses of action requires the agent to identify and to integrate novel sources of knowledge relevant to the situation. This includes extending the state representation to incorporate features now determined to be salient (**SM**) and reasoning about action affordances (**AA**). This step may also require reinterpretation of constraints (**CF**) outside of their typical encoding. Further, any relevant information used as a decision basis should be evaluated for quality, potentially using situationally available assessment of that quality if apt (**IQ**).

Steps 4 & 5: Candidates generated via information identified as relevant from Step 3 are then evaluated and one course of action chosen. Evaluation gauges the likelihood that the courses of action will resolve the triggering conflict(s) or extend the agent's situational understanding, such as through active sensing (**MU**). This capability includes the ability to determine if a candidate will not contribute to mitigation. For example, constraint prioritization is irrelevant to *intra-constraint* conflicts. Additionally, the process may it-

---

[2] The linked technical appendix extends the presentation of each item in the table and identifies prior, related work where that type of knowledge was researched.

Algorithm 2: The different knowledge that OAMNCC processing steps demand (ref: Table 2 for bold labels)

```
1: Recognize novelty of the conflict
   (similar to “out of distribution” assessment)
2: Assess conflict type and structure (CS)
3: Evaluate what situational information might be relevant
   to the conflict (IQ; CF, EP, AA, SM)
4: Propose candidate conflict-mitigating courses of action,
   evaluate them, and select one (MU, EP)
5: Execute the selected course of action, monitoring (CS)
   for resolution of conflict
```

erate between evaluating available information (Step 3) and candidates (Step 4), especially when candidates are judged insufficient. Ultimately, candidate selection (Step 4) must justify its selection using preferences sufficiently expressive to encode human preference expressivity (**EP**). Finally, detection that a conflict is successfully resolved depends on the capability to monitor actual outcomes (**CS**).

Implementation challenges are now more evident. First, some requirements are in tension: the need for online response and the potentially expensive evaluation of situational information, expressive preferences and comprehensive coverage over all conflict types, the use of information based on its quality or the role it can play in mitigation. Second, these tensions cannot just be resolved once, but must be frequently re-evaluated. Different contexts will offer differing accessibility to situation knowledge and metaknowledge. Further, dynamics may include potentially adversarial responses to the agent’s courses of action (e.g., pirates may begin to use ruses and feints in response to the ship’s strategy). Third, underlying these mitigation challenges is the vast scope of acquiring, representing, and reliably retrieving the necessary knowledge, regardless of the specific implementation-level algorithms. Nonetheless, the combination of these requirements more fully characterizes the solution space for comprehensive conflict mitigation that aligns with human expectations and values.

Related Work

The knowledge requirements from the previous section can also be used to organize state-of-the-art approaches and to assess which methods satisfy which requirements. In this section, we review and organize past work, using the requirements as a guide. Despite a diversity of approaches, a common theme in related work is evident: they generally fail to integrate explicit metaknowledge when encountering a novel situation.

Many examples of past work address various types of conflicts directly, but incompletely with respect to the full set of requirements we identify. For example, some address a subset of all possible conflicts (Banerjee et al. 2024; Choi et al. 2025; Briggs and Scheutz 2015; Lu et al. 2024), some rely on pretrained response selection (Awad et al. 2024; Simpkins and Isbell 2019), and some are limited in both ways (Censi et al. 2019; Loreggia et al. 2022). As one example, SEP-nets (Awad et al. 2024) excel as representations explicitly designed to capture human preference expressivity, conditioned on a specific context. These methods vary in precise capabilities and will often be sufficient in closed domains. Table 3 (and the Appendix) contain further characterization of the specific types of knowledge they support (and do not).

Generally, there is increasing recognition of the importance of, for deployed, long-lived agents, online novelty detection, including, but broader than, detecting out-of-distribution situations (e.g., Haider et al. 2024; Mohan et al. 2024; Kejriwal et al. 2024; Goel et al. 2024).

We turn now to methods that detect and respond to out-of-distribution or novel situations. These methods can also likely apply novelty detection toward mitigating different types of conflicts, although OAMNCC per se is not the focus. For example, McLaren (2006) applies case-based reasoning to ethical dilemmas. Their method reuses the reasoning process and decision rationale from a similar, past dilemma in order to respond to a novel one. This mechanism can be sensitive to action affordances, and *depending on the case-base and query mechanism*, can reflect complex preferences and use of different constraint and frame knowledge in a novel situation. However, the approach depends on static, pre-defined, similarity-based case retrieval. This retrieval may be insensitive to subtly different contexts, limiting the use of meta-knowledge (especially mitigation utility) as the basis for selection among possible responses to a conflict in a situation. Further, given the reliance on similarity-based retrieval, it is unclear if action affordances can be transferred from substantially different cases.

Goal reasoning agents (Aha 2018) can autonomously manage their own goals according to a process conceptually similar to Algorithm 1. Goals can have varied semantics analogous to multiple constraint frames, and can trigger information gathering. The method potentially enables multiple goal selection strategies, including ones suitable for mutual exclusivity conflicts (Smith 2004). Related, goal-reasoning research describes autonomous goal selection in the context of novelty with “motivators”, but leaves trustworthy online motivators to future work (Johnson et al. 2018). While spanning many requirements, these strategies are not chosen according to situational information about the type of conflict, limiting any use of conflict structure knowledge, and thus also relevance and mitigation utility knowledge in turn. Additionally, we were unable to identify examples in which a goal reasoning system attempts to reframe its interpretation of constraints. Overall, because this research is dispersed among a few (sometimes domain-specific) methods, it is unclear how to integrate them to specifically support the required knowledge altogether.

In summary, no prior work fully supports all knowledge requirements. However, a combination of case-based and goal reasoning approaches (e.g., Floyd, Drinkwater, and Aha 2015, but without assumption of human interaction), with more flexible processing for their individual components, may approach the necessary knowledge support (also integrated with additional goal reasoning strategies related to conflict structure). Generally, a goal reasoning approach appears to be a potentially apt starting point for implementation, given how much of the problem scope it addresses.

Related Work	**Constraints & Frames**	**Expressive Preferences**	**Action Affordances**	**Situation Model**	**Conflict Structure**	**Information Quality**	**Mitigation Utility**
Censi et al. (2019)	$\times$	$\sim$	$\times$	$\times$	$\times$	$\times$	$\times$
Briggs and Scheutz (2015)	$\times$	$\sim$	$\sim$	$\sim$	$\times$	$\times$	$\sim$
Awad et al. (2024)	$\checkmark$	$\checkmark$	$\times$	$\sim$	$\times$	$\times$	$\times$
Lu et al. (2024)	$\times$	$\sim$	$\times$	$\sim$	$\times$	$\sim$	$\sim$
Loreggia et al. (2022)	$\sim$	$\sim$	$\times$	$\sim$	$\times$	$\times$	$\times$
Simpkins and Isbell (2019)	$\sim$	$\checkmark$	$\sim$	$\sim$	$\sim$	$\times$	$\times$
Choi et al. (2025)	$\times$	$\times$	$\sim$	$\times$	$\times$	$\times$	$\sim$
Banerjee et al. (2024)	$\times$	$\times$	$\checkmark$	$\sim$	$\times$	$\times$	$\sim$
McLaren (2006)	$\checkmark$	$\checkmark$	$\sim$	$\times$	$\sim$	$\times$	$\sim$
Aha (2018)	$\sim$	$\sim$	$\sim$	$\checkmark$	$\times$	$\checkmark$	$\sim$

Table 3: Summarizing related work in terms of coverage of the knowledge requirements. A $\checkmark$indicates satisfaction of a knowledge requirement. An $\times$ indicates lack of this type of knowledge. A $\sim$ indicates partial satisfaction of the requirement.

Conclusion

AI as a field is seeking to enable long-lived, autonomous agents that act in accordance with human expectations and values. To achieve this goal, agents must successfully and adaptively comply with many kinds of constraints (normative and otherwise), and, when they inevitably arise during deployment, resolve conflicts between and among those constraints, also in ways aligned with human expectations. This paper presented a knowledge-level analysis of what such an aligned conflict mitigation strategy entails, both in terms of types of conflicts that may arise and the types of knowledge that an agent requires to navigate these conflicts. The analysis demonstrates the significant challenge aligned conflict mitigation presents, where an agent will have to contend with incomplete policy knowledge and underspecified decisions, while remaining sensitive to the diverse and dynamic situation in which it is embedded.

A perspective that may enable progress on this challenge is to view a novel, unanticipated conflict as an instance of an *ill-structured problem* (Newell 1969; Simon 1973), a problem in which it is unclear what information and actions are relevant. An agent may initially lack the capacity to formulate (let alone solve) an ill-structured problem. These problems often lead to a reconceptualization of the situation. For an agent to be able to safely and robustly act in an open, unpredictable world in ways aligned with human values, this perspective, drawing from the analysis presented here, suggests the need for dynamic metacognition, or the ability to reason about and adapt an agent's own problem-solving framework.

Thus, in terms of the process models sketched here, in many cases a first step should not be an attempt to solve the novel conflict (a kind of ill-structured problem), but instead to determine what problem space the agent should use to attempt to address the conflict. We expect reconceptualization will need to utilize metacognitive reasoning to search for an appropriate problem space, which may often include reconsideration and reframing of the original problem. The example of reframing a duty as a cost function (and vice versa), as we presented in the scenarios, is one example of such reframing. We thus hypothesize that dynamic runtime metacognition, capable of responding to specific situational information, will be a key feature of solutions to online, aligned mitigation of novel constraint conflicts.

Acknowledgments

This work was supported by the Office of Naval Research, contract N00014-22-1-2358. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. We acknowledge and thank the anonymous reviewers for their constructive suggestions.

References

Aha, D. W. 2018. Goal Reasoning: Foundations, Emerging Applications, and Prospects. *AI Magazine*, 39(2): 3–24. Publisher: Wiley.

Awad, E.; Levine, S.; Loreggia, A.; Mattei, N.; Rahwan, I.; Rossi, F.; Talamadupula, K.; Tenenbaum, J.; and Kleiman-Weiner, M. 2024. When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data. *Autonomous Agents and Multi-Agent Systems*, 38(2): 35.

Banerjee, A.; Rahmani, K.; Biswas, J.; and Dillig, I. 2024. Dynamic Model Predictive Shielding for Provably Safe Reinforcement Learning. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 100131–100159. Curran Associates, Inc.

Bratman, M. E. 1987. *Intentions, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press.

Briggs, G. M.; and Scheutz, M. 2015. ”Sorry, I Can't Do That”: Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *AAAI Fall Symposia*, 32–36.

Brutzman, D. P.; Blais, C. L.; and Hsin-Fu, W. 2020. Ethical Control of Unmanned Systems: Lifesaving/Lethal Scenarios for Naval Operations. Technical Report NPS-USW-2020-001, Naval Postgraduate School.

Censi, A.; Slutsky, K.; Wongpiromsarn, T.; Yershov, D.; Pendleton, S.; Fu, J.; and Frazzoli, E. 2019. Liability, Ethics, and Culture-Aware Behavior Specification using Rulebooks. In *2019 International Conference on Robotics and Automation (ICRA)*, 8536–8542. Montreal, QC, Canada: IEEE.

Choi, J. J.; Aloor, J. J.; Li, J.; Mendoza, M. G.; Balakrishnan, H.; and Tomlin, C. J. 2025. Resolving Conflicting Constraints in Multi-Agent Reinforcement Learning with Layered Safety. Version Number: 1.

Floyd, M. W.; Drinkwater, M.; and Aha, D. W. 2015. Trust-guided behavior adaptation using case-based reasoning. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 4261–4267.

Freuder, E. C.; and Wallace, R. J. 1992. Partial constraint satisfaction. *Artificial Intelligence*, 58(1-3): 21–70. Publisher: Elsevier.

Goel, S.; Wei, Y.; Lymperopoulos, P.; Churá, K.; Scheutz, M.; and Sinapov, J. 2024. NovelGym: A Flexible Ecosystem for Hybrid Planning and Learning Agents Designed for Open Worlds. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, 688–696. International Foundation for Autonomous Agents and Multiagent Systems.

Haider, T.; Roscher, K.; Herd, B.; Schmoeller Roza, F.; and Burton, S. 2024. Can you trust your Agent? The Effect of Out-of-Distribution Detection on the Safety of Reinforcement Learning Systems. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, 1569–1578. Avila Spain: ACM. ISBN 979-8-4007-0243-3.

ICC International Maritime Bureau. 2024. Piracy and Armed Robbery Against Ships: Report for the Period 1 January - 31 December 2023. Technical report, ICC International Maritime Bureau (IMB), London, UK.

Jenni, K.; and Loewenstein, G. 1997. Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, 14(3): 235–257.

Johnson, B.; Floyd, M. W.; Coman, A.; Wilson, M. A.; and Aha, D. W. 2018. Goal Reasoning and Trusted Autonomy. In *Studies in Systems, Decision and Control*, 47–66. Cham: Springer International Publishing. ISBN 978-3-319-64815-6 978-3-319-64816-3.

Jones, S. J.; and Wray, R. E. 2024. Toward Constraint Compliant Goal Formulation and Planning. In *Proceedings of the Advances in Cognitive Systems Conference*. Palermo.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134. Publisher: Elsevier.

Kejriwal, M.; Kildebeck, E.; Steininger, R.; and Shrivastava, A. 2024. Challenges, evaluation and opportunities for open-world learning. *Nat Mach Intell*, 6(6): 580–588.

Klein. 1998. *Sources of Power*. The MIT Press.

Kuipers, B. 2018. How Can We Trust a Robot? *Communications of the ACM*, 61(3): 86–95.

Langosco, L.; Koch, J.; Sharkey, L.; Pfau, J.; Orseau, L.; and Krueger, D. 2022. Goal Misgeneralization in Deep Reinforcement Learning. In *ICML 2022*. ArXiv:2105.14111 [cs].

Leontarakis, A.-A. G. 2015. *Piracy, Law Efficiency, Statistics and Possible Solutions to that Severe and Growing Threat*. Master's thesis, Hellenic Coast Guard Academy.

Loreggia, A.; Mattei, N.; Rahgooy, T.; Rossi, F.; Srivastava, B.; and Venable, K. B. 2022. Making Human-Like Moral Decisions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 447–454. Oxford United Kingdom: ACM. ISBN 978-1-4503-9247-1.

Lu, Q.; Svegliato, J.; Nashed, S. B.; Zilberstein, S.; and Russell, S. 2024. Ethically Compliant Autonomous Systems under Partial Observability. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 16229–16235. Yokohama, Japan: IEEE.

McKie, J.; and Richardson, J. 2003. The Rule of Rescue. *Social Science & Medicine*, 56(12): 2407–2419. Publisher: Elsevier BV.

McLaren, B. 2006. Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. *IEEE Intelligent Systems*, 21(4): 29–37. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Meseguer, P.; Rossi, F.; and Schiex, T. 2006. Soft Constraints. In Rossi, F.; Van Beek, P.; and Walsh, T., eds., *Handbook of constraint programming*, Foundations of artificial intelligence. Amsterdam ; Boston: Elsevier, 1st ed edition. ISBN 978-0-444-52726-4. OCLC: ocm70408044.

Milli, S.; Hadfield-Menell, D.; Dragan, A.; and Russell, S. 2017. Should robots be obedient? In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, 4754–4760. Melbourne, Australia: AAAI Press. ISBN 978-0-9992411-0-3.

Mohan, S.; Piotrowski, W.; Stern, R.; Grover, S.; Kim, S.; Le, J.; Sher, Y.; and De Kleer, J. 2024. A domain-independent agent architecture for adaptive operation in evolving open worlds. *Artificial Intelligence*, 334: 104161.

Molineaux, M.; Klenk, M.; and Aha, D. 2010. Goal-driven autonomy in a Navy strategy simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 1548–1554. Issue: 1.

Newell, A. 1969. Heuristic Programming: Ill-structured Problems. In Aronofsky, J., ed., *Progress in Operations Research III*, 360–414. New York: Wiley.

Newell, A. 1982. The knowledge level. *Artificial Intelligence*, 18(1): 82–127.

Payne, J. W.; Bettman, J. R.; and Johnson, E. J. 1993. *The Adaptive Decision Maker*. Cambridge: Cambridge University Press. ISBN 978-0-521-41505-7.

Rao, A. S.; and Georgeff, M. P. 1995. BDI Agents: From Theory to Practice. In *Proceedings of the First International Conference on Multiagent Systems*.

Shah, R.; Krasheninnikov, D.; Alexander, J.; Abbeel, P.; and Dragan, A. 2019. Preferences Implicit in the State of the World. In *ICLR 2019*. arXiv. ArXiv:1902.04198 [cs].

Simon, H. A. 1973. The Structure of Ill Structured Problems. *Artificial Intelligence*, 4(3-4): 181–201.

Simpkins, C.; and Isbell, C. 2019. Composable Modular Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 4975–4982. Publisher: Association for the Advancement of Artificial Intelligence (AAAI).

Smith, D. E. 2004. Choosing objectives in over-subscription planning. In *Proceedings of the Fourteenth International Conference on International Conference on Automated Planning and Scheduling*, 393–401.

Wray, R. E.; Jones, S.; and Laird, J. E. 2023. Computational-level Analysis of Constraint Compliance for General Intelligence. In *Proc. of Artificial General Intelligence (AGI) Conf.* Stockholm.