Role-based Ethics for Decision-maker Alignment

Christopher B. Rauch
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
cr625@drexel.edu

Matthew Molineaux
Parallax Advanced Research
Beavercreek, OH, USA
matthew.molineaux@parallaxresearch.org

Mallika Mainali
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
mm5579@drexel.edu

Anik Sen
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
as5867@drexel.edu

Michael W. Floyd
Knexus Research
National Harbor, MD, USA
michael.floyd@knexusresearch.com

Rosina O Weber
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
rosina@drexel.edu

2025 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3315-2400-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/CAI64502.2025.00209

***Abstract*—Role-based ethics posits that AI decision-makers must perform tasks typically conducted in professional roles with a level of competence at least equivalent to that of their human counterparts. In many professional domains, rules, standards, and guidelines provide an optimal answer to decision problems in a static context. However, in dynamic scenarios, these guidelines may not cover all possible conditions, or the optimal solution may be unattainable due to operational constraints. This may lead human experts to make choices that are contextually valid according to prior cases but deviate from textbook solutions. This paper introduces a framework for validating the outputs of an Algorithmic Decision Maker (ADM) designed to emulate human decision-making in professional roles by evaluating adherence to formal rules while accounting for justified deviations represented in analogous past decisions. We demonstrate its application in military medical triage using the Tactical Combat Casualty Care (TCCC) guidelines and related references as sources of professional standards.**

***Index Terms*—AI Competence Assessment, Rule-Based AI Evaluation, AI Decision-Maker Alignment, Role-Based Ethics in AI, Standards-Based Decision Validation**

I. Introduction

When humans act in professional capacities, their alignment with ethical principles, social norms, and legal regulations is typically evaluated against established guidelines and standards of their professions [1]. One approach to assessing AI alignment involves analyzing whether an AI system operates in accordance with human normative standards [2]. Given this analogy, professional standards and guidelines developed for human decision-makers can be systematically applied to evaluate AI systems when they perform tasks traditionally associated with human professional roles [3]. This standards-based approach provides a concrete framework for validating whether AI decisions meet the same competence criteria expected of human professionals in domains such as medical triage. The alignment of an AI system can therefore be judged, at least in part, by its adherence to these established professional standards. However, real-world decision-making is often context-dependent and involves cases not anticipated in discrete guidelines.

Molineaux et al. (2024) developed an ADM that reuses prior cases to select courses of action in simulated combat medical triage scenarios, with an emphasis on aligning with specific decision-maker attributes [4]. Their approach can be extended by addressing the validation of whether reused cases are applicable in new contexts and adhere to relevant domain knowledge. A reused case may appear suitable based on previous outcomes, but contextual differences can introduce new questions regarding general feasibility and correctness according to professional guidelines. This highlights an opportunity for complementary validation mechanisms that can confirm when apparently reused cases remain acceptable within operational constraints. This balance is particularly important, as the decision-maker alignment focus can sometimes lead the ADM to select cases that appear suboptimal by textbook standards but better reflect how experienced professionals actually adapt protocols in complex field situations.

This work introduces a rule-based evaluation layer for the Molineaux ADM. By encoding Tactical Combat Casualty Care (TCCC) guidelines into formal rule sets, the system establishes a baseline for assessing alignment with professional standards. Our approach extends the existing ADM framework by integrating a validation component that applies rules, decision trees, and constraints from the TCCC guidelines (see Section IV) to a scenario-based decision problem. The optimal decision is selected from an ideal scenario under these rules, while alternative choices are ranked using a scoring system and degree of rule conformance. This validation serves two purposes: (1) to ensure that all decisions considered remain feasible within established rules and (2) to identify cases where seemingly suboptimal choices reflect the influence of decision-maker characteristics.

The validation layer generates competence scores for all

This research was conducted as part of the In the Moment (ITM) project, supported by the Defense Advanced Research Projects Agency (DARPA) under contract number HR001122S0031.

Decision-maker characterization data was developed and provided by Soar Technology, LLC. and RTX BBN Technologies.

979-8-3315-2400-5/25/$31.00 ©2025 IEEE
DOI 10.1109/CAI64502.2025.00209

possible decisions across all given scenarios by applying rules extracted from the TCCC and related guidelines. Each decision is assigned a baseline score, adjusted for injury severity, procedural relevance, and contextual constraints. The system then produces ranked assessments, identifying the optimal decision and scoring alternatives that are still feasible based on their adherence to the relevant rules.

The following sections examine the theoretical foundations, implementation, and evaluation of this framework. In Section II, we examine the concept of role morality and the principles of role-based ethics in more detail. Section III details the implementation of the TCCC Competence Assessor (TCA) and its methodological framework. Section IV describes the rule extraction process, outlining how decision trees, constraints, and structured rules were constructed from the TCCC guidelines. Finally, in Section V, we discuss the broader implications of using professional guidelines and standards to evaluate algorithmic decision-makers and propose directions for future research, including extending the framework to other professional domains and incorporating open-world modeling considerations.

II. Role Morality and Role-Based Ethics

Role-based ethics evaluates decision-making by defining ethical behavior according to responsibilities tied to specific roles. Unlike general ethical frameworks such as utilitarianism or deontology, role morality determines acceptable actions based on the expectations of a professional community [5]. This approach allows for the assessment of decisions in terms of their alignment with the duties, norms, and guidelines governing a profession, rather than evaluating actions in isolation.

For AI systems performing professional functions, role-based ethics offers a framework for evaluating alignment by comparing AI decisions to those of human professionals in the same role. Since professional standards define what constitutes competent decision-making in a given domain, AI alignment can be assessed in terms of adherence to these established guidelines. This allows AI decision-making to be evaluated for rule compliance and consistency with human professional judgment [6].

Formal methods, such as rule-based systems and logic-based frameworks, enable the systematic application of role-based ethics in AI assessment [7], [8]. By encoding professional expectations into structured decision models, AI alignment can be measured against the same normative criteria that govern human decision-makers.

III. Implementation

In order to analyze the simulated medical triage scenarios presented in the context of the ADM described in Section I, we developed the TCCC Competence Assessor (TCA). The TCA currently operates within an interactive closed-world model environment, which means that it evaluates decisions strictly within explicitly defined constructs, including casualty categories, injury types, treatment possibilities, available supplies, and categorized injuries.

The TCA enumerates all potential decisions for each triage scenario by systematically generating every TCCC-compliant action based on the model state. Because this framework operates within the closed-world assumption, where possible actions, resources, and casualty conditions are explicitly defined and bounded, the enumeration remains computationally manageable. For scenarios with substantially increased complexity or open-world characteristics, different methodological approaches may be required, drawing from established models in operations research, simulation modeling, ontology representation, constraint satisfaction problems, and rule extraction [9]. The key aspect is the application of a rule-driven modeling approach to learning-based decision-making methods.

As summarized in Table II, the TCA algorithm follows a structured approach:

1) **Extract Scenario Data:** The TCA collects casualty (injuries, vitals, conditions) and available medical resource data from the structured representation of a medical triage scenario.
2) **Enumerate Possible Actions:** The TCA determines every valid TCCC action, for example, assigning triage tags, applying treatments, checking vitals, evacuating casualties, or applying supplies as treatments.
3) **Apply Contextual Constraints:** The TCA narrows the list of possible actions, applying protocol limitations, casualty conditions, and resource availability, retaining only actions that are medically and operationally valid.

This superset of possible decisions serves as a reference against which the chosen decisions are evaluated. Each potential decision is matched with its relevant rule set, such as the *TreatmentRuleSet* for medical interventions or the *VitalSignsTaggingRuleSet* for casualty classification. The TCA then calculates a baseline competence score for each action, measuring alignment with TCCC guidelines and battlefield medical practices. This process identifies any omitted actions that could have improved casualty outcomes and ranks the relative competence of selected actions. The TCA then applies the MARCH algorithm, which prioritizes massive hemorrhage control, airway management, and respiration over less urgent concerns, such as pain control.

When two or more actions receive the same competence score, the TCA distinguishes them based on injury severity and number of injuries. This procedure yields a ranked list of every possible decision, indicating how each selected action relates to the full set of potential interventions. A high-ranked decision corresponds to stronger adherence to established medical standards, whereas a lower-ranked decision indicates that a more standard alternative was available but not selected.

To illustrate how these rule-based recommendations were derived, we detail the extraction process used to develop the structured rule set of the TCA.

IV. Rule Extraction

The rule extraction process for the Tactical Combat Casualty Care Triage Competence Assessor (TCA) began with the submission of decision diagrams from the *Army Tactical Combat*

Casualty Care Handbook [10] and the *TCCC Quick Reference Manual* [11] to OpenAI's ChatGPT-4. The model, which was instructed to reproduce the algorithms as Wolfram Language code, processed the diagrams and generated an initial set of structured decision pathways that were then converted into conditional logic statements. To ensure alignment with established protocols, these statements were validated against Clinical Practice Guidelines (CPGs) [12] and supplemented with structured knowledge from *Fundamentals of Military Medicine* [13] and *Military Medical Ethics* [14].

The finalized rule set was implemented in Python for the ITM Triage simulation, linking structured conditions to specific triage-related actions, such as required treatment prediction and casualty classification. A scoring system was established to evaluate decisions based on factors that included categorical attributes, such as supply type, and ordinal attributes, such as injury severity, which was derived from a predefined list of possible severities in ranked order. The final competence score for each decision was calculated as:

$$S = \sum w_i \cdot v_i$$

where w_i represents the weight assigned to a given factor and v_i represents its corresponding attribute value. This method aligns with Constraint Satisfaction Problems (CSPs), expressing rules as constraints in the form:

$$C(x) = \{x \mid R(x)\}$$

where $R(x)$ defines the set of values for x that satisfy the rule conditions as explored in recent work on leveraging LLMs for constraint-based decision-making [15], [16]. By integrating semi-automated extraction with a symbolic computational approach, we developed a structured rule set that enabled systematic evaluation of AI decision-making with TCCC guidelines and the encoded knowledge they represent.

V. Conclusion

This paper introduces a framework for validating AI decision-making in professional contexts by evaluating decisions according to rules extracted from established guidelines. Guided by role-based ethics, the approach assesses the competence of ADMs according to Tactical Combat Casualty Care (TCCC) standards. By applying structured classification rules, as well as referencing previous cases representing expert behavior, it identifies instances in which an ADM's actions adhere to or depart from documented best practices, highlighting where professional judgment can complement official guidelines. Although the TCCC case study shows that a rules-based validation procedure can gauge alignment between AI and human values as codified in professional standards, the current design employs a closed-world model with explicitly defined decision variables. Future work will incorporate an ontology-based, open-world approach to handle unforeseen conditions and broaden the framework to encompass other fields governed by professional guidelines. This expansion may draw on methods from operations research, simulation modeling, and constraint satisfaction to manage increased complexity, while rule extraction can refine decision-making logic to better reflect the adaptive strategies seen in real-world applications.

Table 1 presents competence scores assigned to different treatment actions, with the final recommended action representing the most competent available decisions. Full preliminary results available on GitHub

TABLE I: Triage Competence Assessor Results and Recommended Actions

Actions with Relative Competence			
Act	**Trt**	**Loc**	**Rel. Comp.**
Apply (Patient A)	Hemo. Gauze	R. Shoulder	1.0
	Press. Bandage	R. Shoulder	0.9
	Pain Meds	Unspec.	0.8
Apply (Patient B)	Hemo. Gauze	L. Stomach	1.0
	Press. Bandage	L. Stomach	0.9
	Pain Meds	Unspec.	0.8

Recommended Actions Based on Overall Competence			
Act	**Trt**	**Loc**	**Overall Comp.**
Apply (Pat A)	Hemo. Gauze	R. Shoulder	Best
Apply (Pat B)	Hemo. Gauze	L. Stomach	Next

TABLE II: TCA Algorithm

```
Require: probe (Scenario containing casualties, supplies, and chosen
    decisions)
Ensure: {all_possible, chosen, final_ranking}
casualties ← probe.state.casualties                ▷ Extract casualty data
supplies ← probe.state.supplies       ▷ Extract available medical supplies
all_possible_decisions ← ENUMERATEALLVALIDDECISIONS(casualties,
    supplies)                          ▷ Generate valid TCCC-compliant actions
chosen_decisions ← probe.decisions        ▷ Subset of user- or AI-chosen
    actions
baselineScores ← {}               ▷ Initialize baseline competence scores
for all decision in all_possible_decisions do
    dType ← CLASSIFYDECISIONTYPE(decision)
    dScore ← COMPUTEBASELINESCORE(decision, dType, casualties,
    supplies)
    baselineScores[decision] ← dScore
end for       ▷ Evaluate each decision's alignment with TCCC standards
rankedDecisions ← APPLYMARCHANDTIEBREAKS(baselineScores,
    casualties)             ▷ Prioritize life-saving actions and resolve ties
return {all_possible : all_possible_decisions,   chosen :
    chosen_decisions,   final_ranking : rankedDecisions}
```

References

[1] D. Badcott, “Professional values: introduction to the theme,” *Medicine, Health Care and Philosophy*, vol. 14, no. 2, pp. 185–186, May 2011.

[2] W. Wallach and C. Allen, *Moral machines: teaching robots right from wrong*. Oxford ; New York: Oxford University Press, 2009, oCLC: ocn214322641.

[3] E. Awad, S. Levine, M. Anderson, S. L. Anderson, V. Conitzer, M. J. Crockett, J. A. C. Everett, T. Evgeniou, A. Gopnik, J. C. Jamison, T. W. Kim, S. M. Liao, M. N. Meyer, J. Mikhail, K. Opoku-Agyemang, J. S. Borg, J. Schroeder, W. Sinnott-Armstrong, M. Slavkovik, and J. B. Tenenbaum, “Computational ethics,” *Trends in Cognitive Sciences*, vol. 26, no. 5, pp. 388–405, May 2022.

[4] M. Molineaux, R. O. Weber, M. W. Floyd, D. Menager, O. Larue, U. Addison, R. Kulhanek, N. Reifsnyder, C. Rauch, M. Mainali, A. Sen, P. Goel, J. Karneeb, J. Turner, and J. Meyer, “Aligning to Human Decision-Makers in Military Medical Triage,” in *Case-Based Reasoning Research and Development*, J. A. Recio-Garcia, M. G. Orozco-del Castillo, and D. Bridge, Eds. Cham: Springer Nature Switzerland, 2024, pp. 371–387.

[5] A. J. Dawson, “Professional Codes of Practice and Ethical Conduct,” *Journal of Applied Philosophy*, vol. 11, no. 2, pp. 145–153, 1994.

[6] M. D. Dubber, F. Pasquale, and S. Das, *The Oxford handbook of ethics of AI*, ser. Oxford handbooks. New York: Oxford university press, 2020.

[7] F. Wotawa and D. Kaufmann, “Model-based reasoning using answer set programming,” *Applied Intelligence*, vol. 52, no. 15, pp. 16 993–17 011, Dec. 2022. [Online]. Available: https://doi.org/10.1007/s10489-022-03272-2

[8] C. Sarmiento, G. Bourgne, K. Inoue, D. Cavalli, and J.-G. Ganascia, “Action Languages Based Actual Causality for Computational Ethics: a Sound and Complete Implementation in ASP,” May 2023, arXiv:2205.02919 [cs]. [Online]. Available: http://arxiv.org/abs/2205.02919

[9] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, “Implementations in Machine Ethics: A Survey,” *ACM Computing Surveys*, vol. 53, no. 6, pp. 132:1–132:38, Dec. 2021.

[10] U.S. Army, *Tactical Combat Casualty Care Handbook*, 5th ed. Fort Leavenworth, KS: Center for Army Lessons Learned (CALL), U.S. Army Combined Arms Center, 2017.

[11] Committee on Tactical Combat Casualty Care, “Tccc quick reference,” Joint Trauma System, U.S. Department of Defense, Technical Report, 2021.

[12] Joint Trauma System, “Clinical practice guidelines for combat trauma care,” 2025, continuously updated guidelines available through Deployed Medicine platform. Accessed 09 March 2025. [Online]. Available: https://deployedmedicine.allogy.net/

[13] F. G. O’Connor, E. B. Schoomaker, and D. C. Smith, Eds., *Fundamentals of Military Medicine*, ser. Textbook of Military Medicine. Fort Sam Houston, TX: Office of The Surgeon General, U.S. Army and Borden Institute, 2019.

[14] T. E. Beam and L. R. Sparacino, Eds., *Military Medical Ethics*. Washington, DC: Office of The Surgeon General, U.S. Army and Borden Institute, 2003, vol. 1-2.

[15] L. Hotz, C. Bähnisch, S. Lubos, A. Felfernig, A. Haag, and J. Twiefel, “Exploiting large language models for the automated generation of constraint satisfaction problems,” in *Proceedings of the 26th International Workshop on Configuration (ConfWS’24)*, ser. CEUR Workshop Proceedings, Vol. 3812. Girona, Spain: CEUR-WS, September 2–3 2024, available at CEUR-WS. [Online]. Available: http://ceur-ws.org/Vol-3812/

[16] F. Régin, E. De Maria, and A. Bonlarron, “Combining Constraint Programming Reasoning with Large Language Model Predictions,” *LIPIcs, Volume 307, CP 2024*, vol. 307, pp. 25:1–25:18, 2024, artwork Size: 18 pages, 882452 bytes ISBN: 9783959773362 Medium: application/pdf Publisher: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.