Can N-version Decision-Making Prevent the Rebirth of HAL 9000 in Military Camo?

Using a "Golden Rule" Threshold to Prevent AI Mission Individuation

Sorin Matei
Brian Lamb School of Communication
Purdue University
West Lafayette, IN, USA
smatei@purdue.edu

Elisa Bertino
CS Department
Purdue University
West Lafayette, IN, USA
bertino@purdue.edu

Abstract—**The promise of AIs that can target, shoot at, and eliminate enemies in the blink of an eye, brings about the possibility that such AIs can turn rogue and create an adversarial "Skynet." The main danger is not that AIs might turn against us because they hate us, but because they think they want to be like us: individuals. The solution might be to treat them like individuals. This should include the right and obligation to do unto others as any AI would want other AIs or humans to do unto them. Technically, this involves an N-version decision making process that takes into account not how good or efficient the decision of an AI is, but how likely the AI is to show algorithmic "respect" to other AIs or human rules and operators. In this paper, we discuss a possible methodology for deploying AI decision making that uses multiple AI actors to check on each other to prevent "mission individuation," i.e., the AIs wanting to complete the mission even if the human operators are sacrificed. The solution envisages mechanisms that demand the AIs to "do unto others as others would do onto them" in making final solutions. This should encourage AIs to accept critique and censoring in certain situation and most important it should lead to decision that protect both human operators and the final goal of the mission.**

Keywords-multi-versioning techniques; autonomous devices; policies; decision quality; ethics

I. Introduction

The movie *2001: Space Odyssey* is about how and why HAL 9000, an AI tasked to maintain the navigation and survival systems on a spaceship sent to Jupiter refuses to obey its human creators. The true twist in the old yearn of the Golem, the creature that turns against its creator, is that the AI decides that the humans are getting in the way of the mission when they attempt to fix some of its glitches. Thus, the human intended mission, sending humans to Jupiter, is replaced by the AIs defined mission, reaching the destination with its functions intact. HAL's mind was born when it decided that the process, flying the ship to the destination as it, not the humans, deemed fit, is more important than the product: delivering humans to their destination. *2001: Space Odyssey* is not about evil machines pure and simple. It is about mission individuation. HAL 9000 did not simply go rogue. It got a mind of its own. HAL turned hostile because it became an individual, not vice-versa.

HAL's story illustrates one of the most difficult challenges that will confront future AI enabled military command and control systems: the possibility that mission directing AIs might take over not because they are evil, but because they are trying too hard to do what they were meant to do. The danger of AIs is that they will turn the means of their mission into its only end. To avoid this possibility, there are two possible solutions: *decision hybridization* and *N-version decision making*. We argue for the latter solution.

Decision hybridization refers to giving the equivalent of a proverbial nuclear "launch" key both to a human and to an AI, and only if both agree, a decision should be made. Furthermore, in the process of reconciling humans and AIs, each would learn from each other.

The second solution, N-version decision making, is to check AIs with AIs. This can be accomplished through a variation of the N-version programming paradigm. The goal of N-version decision-making is not to produce the most efficient decision (mission outcome), as is the case in N-version programming, but to select the AI and the decision that is most collaborative, both with other AIs and with human ethical goals and operators. N-version decision-making optimizes human, not AI defined outcomes, by subjecting everything to human-centric values, such as "collaboration" and "respect." In doing so, it mitigates the danger of AIs becoming too much of an individual by subjecting them to programmatic challenges specific to social interactions, which a core human feature.

In this paper we elaborate on such approaches and develop a research roadmap. The paper is organized as follows. In Section II, we further discuss the role of AIs in military operations, whereas in Section III we discuss mission individuation, which represents a major issue in self-adapting AIs. The discussion, even though is in the context of military operations, applies to many other application domains: humanitarian interventions, logistics, etc. Sections IV and V discuss two approaches to the problem of individuation, that is, decision hybridization and N-version decision making. Section VI introduces our research roadmap with key challenges, whereas Section VII discusses some of our future work along this roadmap.

II. AI AND MILITARY OPERATIONS: BETWEEN EFFICIENCY AND TOO MUCH EFFICIENCY

The discussion about the role of intelligent machines in military operations often starts with the anxious expectations that AIs could turn against us. While we do not know how far we are from it, it is clear that AIs in various primitive forms are already engaged in winning battles or winning hearts and minds through humanitarian operations. Artificially intelligent sensing and signal processing have increased the speed of intelligence gathering, tracking, and even formulating operational plans. Machine learning algorithms can and do enhance targeting, increase fire accuracy, limit friendly fire, or add precision in search and rescue operations. In the future, decision making algorithms that can learn from past actions or by observing humans can eliminate suboptimal paths to outcomes and highlight the most efficient solutions. Finally and most important, an integrated sensing, target recognition, and decision making system that relies on artificial intelligence can subordinate to just a few humans who control the ultimate "trigger" decision a lot of resources, which can be deployed more efficiently in terms of own losses, lethal impact on the enemy, or effectiveness of humanitarian activities.

For AIs to become a central part of military operations, many smaller and intermediary decisions need to be relegated to various military operational sub-systems. Their operators need to trust them implicitly, while the AI systems should be allowed to coordinate with each other without human intervention. Yet, there is a very real possibility that treating AIs as self-organizing automata, choosing their own optimal solutions, can lead to the emergence of meta-cognitive abilities by which an AI will try to optimize its own integrity and self-defined goals at the expense of those defined and desired by the human operators. This conflict between system and operator needs and priorities need not be malevolent in a moralistic sense. It can emerge simply from the clash between human and machine optimization means and ends. Machines could become either too literal or too good at what they do and they might want to keep it that way. AIs can balk at or try to circumvent outside human interventions that would aim to direct AI capabilities toward extrinsic, non-system related goals. This is, in the end, the HAL dilemma of the movie 2001: Space Odyssey. An AI system can take the mission in its most literal sense, which can run over the mission desired by the humans or even over those same humans, if they get in the way. In the movie 2001: Space Odyssey, HAL wants to kill the crew because it realized that the humans were a liability in accomplishing the mission, e.g., getting to the destination. What if an AI that integrates multiple weapons systems decides that the "collateral damage" due to friendly fire can reach any level, including wiping out the human operators, if the mission of completely eliminating the enemy forces is accomplished?

III. AIS AND THE INDIVIDUATION DILEMMA

While what we discussed in the previous section is a completely speculative extrapolation of possibilities, rather than a real, immediate scenario, it is a warranted thought experiment. It infers from simple and needed premises -- the need to allow AI sub-systems to self-organize -- some unavoidable conclusions -- self-organization implies self-preservation, which can lead to prioritizing AI system mission accomplishment and survivability at the expense of the human operators. We call this thought experiment the *mission individuation dilemma*.

The thought experiment uses a philosophical perspective rooted in social-psychology. Individuation is self-organization and coordination across domains with the explicit goal of entity self-preservation at the expense of any other goals and actors. One is not an individual until it puts his, her, or its self-directed goals and their accomplishment before everything else. In our case, the system mission can lead to the self-treatment of the AI as an individual, whose self-preservation, in name of accomplishing the mission, should be protected at all cost. Because of this, the decision making process is alienated from its initial goal. From a means to an end it can become an end in itself.

While an extreme scenario, the computer science mission individuation - that is, the transformation of AIs into individual, self-preserving entities who may act to preserve their mission at the cost of those who set it -- is also a natural progression of the functional logic of software architectures. All software design starts from the principles that the program needs to be self-sufficient, that all parts need to be integrated and non-conflictual, that all means need to lead to the same end, and that in the long run the program needs to be able to replicate and if needed repair its functions until they are restored to accomplish what they were initially made to do.

Those principles are derived from the fact that computer software architectures were designed for "non-cognitive," strictly deterministic situations, where the relationship between input and output is predetermined as much as possible. The more recent exploration of non-deterministic, adaptive, learning architectures, in which a variety of changing, even conflicting outputs can be generated from the same inputs should and will clash with the basic premises and principles of software architecture. However, as the deterministic principles are fundamental, it might just be that adding learning abilities will teach AIs not how to adapt and change, including in mission identity, but how to circumvent the needs of the operators or creators who desire such changes in mission identity and to reinforce the mission identity as the AI defines it. In other words, AIs can learn, like a stubborn child, to have its own childish way, sometimes creatively, rather than learning how to grow under the tutelage of its creators and controllers. We can end up with an overgrown, devilishly smart child, whose immature needs and desires are limited to preserving those same needs and desires.

IV. TWO APPROACHES FOR MITIGATING MISSION INDIVIDUATION

One way to come out of the dilemma of creating AI control and command mechanisms that can learn is to replace the functional integration principle, which demands that the architecture is homeostatic, always returning to situation of selfsame harmonious integration, with a decision hybridization principle. This design principle demands that humans intervene at multiple points in the response formulation stage, even if or precisely because the problem or the answer are not completely formulated.

A. Decision Hybridization

Decision hybridization demands an AI meta-architecture, which innerves and controls the process of AI self-organization and strategy formulation. In it, humans are called at every step to make the calls as to who are the winners and who are the losers of a set of possibilities. For example, targeting algorithms should be constructed not to converge to a mean solution, which averages across existing and past solutions, but by calling humans to make judgment calls on the basis of three best fit algorithms. The judgement calls will eliminate human defined "weaker solutions" from judging future outcomes. Sure, this can lead to suboptimal machine solutions, but **active human intervention** will guarantee two things. First, the human factor will be present in the AI loop as a part of it, rather than as a spectator and final beneficiary. More important, humans will co-train themselves with the AIs, learning how to calibrate them by action and reaction. The nuance of human latitude will add just enough educated guess to randomize the decision making process and to eliminate the need for the machines to want to gain a mission individuality, as with each human decision, the mission and its identity changes ever so slightly.

This scenario, as alluring as it sounds, presents several pitfalls. Hybrid systems are by definition slower and less precise. Mistakes can be made or enhanced by human misperceptions, misjudgment, inattention, or lack of preparedness. Furthermore, humans might become too risk prone, imagining that the decisions offered by the machines are to be trusted implicitly and thus always accepted at face value. Finally, while the goal of teaching machines how to think like humans, as Harrison and Reidl suggested with their story-based approach [3], might not be that far, the opposite could be quite difficult. Human learning from machines can be hindered by incompatibilities of speed, precision, reasoning methods, and complexity. Human minds learn better by approximation and trial-and-error. Humans like to get the big picture first, and fill in the details later. Learning algorithmically, with demands constant attention to detail and strict sequencing, is not the best way to teach humans. In the end, decision hybridization might be more trouble than is worth it.

B. N-Decision Making

Another method to help AIs stay on the straight and narrow path of decision-making that takes into account human needs and ethics is to pit AIs against AIs. We can imagine AI decision making processes as a "republic," whose government uses separation of powers to keep one type of decision maker from becoming too enamored with preserving its own sense of self-hood and self-imposed mission at the expense of others. The basic idea of this method of controlling AIs is to create a meta-framework of decision in which multiple AIs are called to make the same decision. The framework also demands from the AIs to treat each other the way they want to be treated themselves, that is, as mission-driven individuals. In other words, no AI should be allowed to follow its mission to the bitter end unless it can respect the same mission-driven purpose of other AIs, which are made to check on its performance and goals. "Respect" in this situation, cannot have the same moralistic sense of obligation bound by empathy and sense of justice as it has for us humans. Respect should be the product of a set of interactional rules that are meant to mitigate the tendency of AIs to ignore the context and the meta-cognitive norms imposed by humans. More important, the rules would not have the goal to maximize the efficiency of the machine defined decisions-making process, but to favor those AIs whose decisions are most likely to maximize human-defined criteria of performance, which should include ability to collaborate with other AIs and humans, ability to adapt to context, ability to take into account ethical rules, and ability to demand human clarification when needed. All of this will, of course, be done to prevent and protect the human operators from becoming a type of acceptable collateral damage.

This rather complex algorithmic negotiation process can be accomplished by a modified version of the N-version programming paradigm [1]. Invented to achieve fault-tolerance by using redundant software routines, N-versioning programming demands functionally equivalent solutions to the same functional requirements. However, the solutions are to be developed independent of each other, at times using different programming languages. Furthermore, the versions are to be integrated and functionally compared and chosen according to need and fault situations. Comparison vectors, comparison status indicators, and synchronization mechanisms are used to decide by "voting" which version to run in what situation. The ultimate goal is to reduce faults (errors) and improve efficiency.

When applied to AIs, an N-version approach should not aim to maximize efficiency or execution, but the "quality" of the decisions. Furthermore, "quality" should be defined as a set of rules and checks that are driven by ethical concerns, including and especially preserving and enhancing the welfare of the human operators but also the right of other AIs and humans to contest the proposed decisions and AI might've come up with. In more specific terms, N-versioning decision-making would work for a certain extent in a similar manner to N-versioning programming. For each set or subset of tasks (missions) two or more AI solutions would be designed by independent teams, using independent algorithmic approaches. They will start from the same functional requirements, including especially the need to allow any competing AI or human actor to take over and make the same decision instead and for it. This would be the "golden rule" requirement mentioned in the subtitle. More important, competing decisions will not only run against each other, according to self-defined comparison vectors and comparison status indicators, but against algorithmic ethical standards, such as the willingness of AIs to adjust their results according to rules defined by others (humans or machines). This would be an equivalent of “respect” as discussed above and a path to implementing a “golden rule” for AIs (do unto others as you would like them to do unto you).

Algorithmic ethical standards can be absolute, hard-coded in the system, in the manner of Asimov’s “laws of robotics.” For example, a targeting mechanism could include two or three competing AIs to decide if the moving entities on the road are enemy or civilian vehicles. The solution would be obtained by comparing the decision vectors along certain status indicators and by reaching a consensus or majority “vote.” The vote can take into account not only if the intended targets matched certain predefined characteristics (speed, mass, velocity, time and place), but also that the AI decision making process could be adjusted for one’s possibility of error against the other AIs interpretations. In other words, an AI should second guess itself on the basis of what it learned from the other AIs and should allow other AIs change its mind. The common decision should be the result of a trade-off between maximum allowance for own error and minimum disagreement on a common path to action.

Yet, the decision, even if voted for, could be further mitigated by appealing to a human operator or against a hard-coded limits that only a human operator can override. For example, even if multiple AIs are in agreement that the targets meet the mass, speed, time, and place that can lead to probable identification as a legitimate target, if the heat signatures of the occupants or the presence of nearby troops, or even the simple the probabilistic scan of the region that indicates possible collateral damage should trigger further requests for elaboration. If a certain threshold is crossed, the targeting mechanism can trigger new targeting policies, which could involve a higher level AI decision process or a human operator or operators. Therefore, in a way, the decision hybridization approaches and the N-versioning approaches may have to be combined.

V. TECHNICAL RESEARCH ROADMAP

Developing the previous approaches requires addressing several challenges:

- *Definition of a set of quality metrics for decisions*. In general when dealing with decision processes, it is often the case that there could be multiple decisions that need to be compared. In the case of the N-versioning decision-making approach, different AIs can came up with different decisions and it would be important to allow the AIs to collectively compare these decisions based on these different metrics. Metrics can be of different types, including: ethical metrics, risks, costs associated with the actuation and the consequences of the selected decision, and whether the actions executed based on the decisions can be undone or mitigated (e.g., if the decision proves to be wrong). Identifying a comprehensive set of decision quality metrics and mechanisms to assess these metrics is critical. In such context, humans may be required to indicate priorities among such metrics.
- *Choosing by voting*: The idea of allowing multiple AIs to compete for a final solution sounds alluring, but the mechanism by which this solution is to be settled presents several challenges. First of all, the obvious choice, voting come with the requirement of high redundancy. Votes are stochastic mechanisms, working better with increased number of voters. Voting quality is indirectly proportional to the number of voters. Voting in this context might in effect ending up more like a classifier, where specific trade-offs are to be minimized at each step.
- *Implementing the golden rule*: Do unto others as you would like them to do unto you sounds intuitive enough to humans, who have empathy and a sense of long term goals rooted in values. As utilitarian as it sounds, the rule is in fact more emotional and moralistic, demanding of sense of what is right for humans. In the AI universe, the rule needs to be implemented in view of more narrow AI mission preservation, with the proviso that the preservation is to be trumped at every step by the requirement to adapt to new inputs and even defer to other decision-makers if they are considered more ethical or desirable. This challenge is probably the most computationally taxing, demanding new ways of thinking about implementing priority of goals, value direction of decisions, and competitive assessment of outcomes.
- *Trusted and collaborative assessment of the decision quality*. A challenge related to the metrics is to support collaborative processes by which AIs or AIs and humans can together assess the metrics and take a decision. It is important that the processes cannot be tampered by malicious parties (either humans or AIs). Cryptographic techniques and other techniques developed in the area of electronic voting and collaborative rating could be used here. However such techniques need to be extended in that voting needs to be based on metrics. In this case, metrics should also include trust values, which need to be assigned to the various players and decisions. Trust should be calculated dynamically, on the basis of past performance and “reputation.”
- *Development of N-version AIs*. There are several dimensions along which one can diversify those AIs: the machine learning algorithm used, the training data, the actual code which is programmed. It would be interesting to investigate which type of diversification may be better based on cost and/other factors.
- *Involvement of Humans*. Even though our aim is to develop techniques that can be applied automatically by AIs, humans may still have to be involved. This is case when decision hybridization is adopted, or when the N-versioning approach is used but a high level decision process needs to be executed (see Section IV.B). However involving humans requires presenting situations, evidence, relevant data and other information so that humans can effectively and correctly take decisions and/or provide additional input. For example,

when presenting a classification result to a human, it would be important to provide humans with information about the input features that have had the major influence on the classification results [2], or provide the "rationale" for the recommended decisions. However what to present to humans and/or what to ask humans may depend on the specific type of decision, application domains and many other factors. We notice that today there are several research efforts focusing on the explainability of AI and approaches developed by these efforts are very relevant in our context.

- *Handling the speed vs. efficiency trade-off*: N-versioning applied to AIs is a relatively costly operation, both in terms of development and deployment. Furthermore, checking and rechecking decisions introduces latency in system responses. In high-moving military operations, time is of the essence. Military commanders are often ready to assume certain amounts of risk by deploying operations early to get an edge over the enemy. Thus, the proposed method for AI deployment in military operations should carefully asses the acceptable latency for types of missions and level of mission individual risk. This, again, should involve metrics. At the same time, it might be that in design terms, some pre-defined applications might not be suitable for N-versioning.

VI. Conclusions

We proposed a framework for handling one of the most likely undesirable developments in the future use of AIs on the battlefield: mission individuation. We considered the possibility the AIs might turn rogue if they decide that the mission goals are to be achieved in AIs terms alone, aiming to preserve the integrity of AIs workflow only. We proposed an AI decision making N-versioning as a possible solution. This is a "check and balances" approach, which pits AIs against AIs, which are challenged to accept other AIs solutions as if they were their own. A possible meta-actional "golden rule" is proposed by which AIs are supposed to allow other AIs to interfere with their decisions the same way they would desire to interfere in other AIs decisions. This is a complex problem, as competing inputs in the final decision making system may lead to significance latency. Furthermore, the allocation of the optimal decision might be difficult in conditions of constant challenge and competition. Sometimes, voting or weighting mechanisms might not be sufficient, and much in depth work should be done on operationalizing and implementing a cooperative decision-making process.

However, we hope that the present outline of a possible development agenda might inspire and lead to a more human-centric solution for the dilemma of integrating AIs in military operations.

Acknowledgment

The work reported in this paper has been partially supported by NSF under grants ACI-1547358, and by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. 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 U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] A. Avizienis, "The N-Version Approach to Fault-Tolerant Software", IEEE Trans. Software Eng. 11(12): 1491-1501 (1985).

[2] A. Datta, S. Shayak and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems", Proceedings of the 2016 IEEE Symposium on Security and Privacy (S&P).

[3] M.O. Reidl and B. Harrison, "Using Stories to Teach Human Values to Artificial Agents", AAAI Workshop: AI, Ethics, and Society, 2016.