The George Washington University

**Exploring Performance of Different Human-AI System Architectures**

Aditya Singh

Zoe Szajnfarber

August 9, 2024

**Introduction & Motivation**

The use of AI in safety-critical and lethal systems has prompted governments across the world including the United States [1] [2], European Union [3] [6], and China [4] to make calls for AI systems to be overseen by or partnered with human operators. Policy such as the Department of Defense’s directive on the use of autonomous weapons [5] has prescribed human control as a key method without providing substantial guidance on how oversight or partnership should be implemented. Generally, the hope is that Human-AI teams take advantage of AI’s “greater computational information processing capacity and an analytical approach,” which “can extend humans’ cognition when addressing complexity, whereas humans can still offer a more holistic, intuitive approach in dealing with uncertainty and equivocality in organizational decision making” [6, p. 577]. While the intended outcomes might be clear, the path to achieving them is not.

Much of the discussion around how to achieve this ideal team has centered around human-in-the-loop system architecture. The term human-in-the-loop was used at least as early as the 1950s to discuss how humans could control increasingly automated functions in aircraft [7] [8]. Despite its growing popularity or perhaps because of it, there is a lack of consensus on what human-in-the-loop control actually means. This is especially problematic since well-meaning policy is being written at this broad level, creating a great deal of uncertainty around how humans should be placed in the loop to achieve the intended policy outcomes. To meet this challenge, we have created a framework that decomposes different types of human-AI system architectures based on the roles that both humans and AI play in the system. We plan on applying this framework to the Silverfish Safe Passage problem to study the performance of different human-AI system architectures. This will enable us to understand the most effective way to partner AI and human operators under different operating conditions.

This paper first presents a systematic review of the ways that human(s) and AI(s) can be integrated in an operational decision loop. Drawing on insights from this review, we present our framework that lays out the human-AI system architectures we identified. Given the nature of the problem, we also apply an ethical analysis to each proposed system architecture to ensure that it is able to meet requisite policy requirements on the deployment of AI in safety-critical situations. We then apply the suitable architectures to the Silverfish Safe Passage system to illustrate how the system can incorporate humans and AI in different ways. We conclude with a discussion of anticipated results.

**Literature Review**

While humans are involved in the entire AI lifecycle from creating and preparing data to setting training objectives and verifying learning outcomes [9] [10], we are focused on human-in-the-loop architecture that defines how humans interact with autonomous systems during operations. Meng’s broad definition describes how human-in-the-loop “generally refers to the need for human interaction, intervention, and judgment to control or change the outcome of a process” in an autonomous system [11, p. 2]. This broad definition makes two key distinctions. First that ‘the-loop’ refers to a process, task, or function of a system rather than the entire system itself. Second, it stipulates that some form of human interaction and influence must be possible. While this broad definition is widely accepted, uses of the term are still inconsistent and mask critical nuances about humans are expected to team with and have control over AI systems. Other authors provide more nuance on different types of human-in-the-loop architecture.

Munir et al. decomposes human-in-the-loop into three applications: direct human control, system monitoring of humans, and a hybrid of these models [12]. Direct human control is further decomposed into *supervisory control* where a human monitors an autonomous system and affects the process by changing set points and *command control* where a human provides a command that is carried out by a system that later reports the results back to the human for feedback. The second application they describe, where a system monitors a human, flips the paradigm completely. Rather than the human having oversight and control over the system, it is the system that has oversight over the human. They decompose system monitoring into open loop systems, where the autonomous system takes no action based on the information it collects about the human, and close loop systems, where the autonomous system takes action to achieve some goal or avoid an accident. A hybrid system may incorporate aspects of both of these types of human-in-the-loop systems [12].

Lacher et al. describe how human decision makers and automated elements of a system operate with respect to the control loop of a system, which they define as “a series of control operations to carry out a task and/or perform a system function, including receiving reference inputs and system state feedback, deciding the desired system state, and taking action” [13, p. 2]. Their definitions of human-in-the-loop, human-on-the-loop, human-over-the-loop, and human-out-of-the-loop are described in Table 3. Similar to Munir et. al, they also lay out an alternate view which is more focused on the role of automation in the loop, as described in Table 3 [13].

Table 1: Lacher et al.'s Human & Automation Role Analysis

|  |  |
| --- | --- |
| Architecture | Description |
| Human-in-the-loop | Human must act for the system to function; automation serves as an assistant |
| Human-on-the-loop | Automation can control the system and act without human action; human monitors and guides automation, with the ability to takeover at any time |
| Human-over-the-loop | Automation can control the system and act without human action; human monitors and can attempt to change system behavior but has no ability to control or takeover |
| Automation-over-the-loop | Automation monitors a function and alerts human to act, may provide options of action, but cannot act |
| Automation-on-the-loop | Automation monitors a function and can act if needed without human approval |
| Automation-in-the-loop | Automation must act for the system to function; human can help |

*Literature Gap*

While Munir et al.’s and Lacher et al.’s distinctions are a great improvement beyond the widely accepted human-in-the-loop definition, they do not capture the full depth of possible human-AI control architectures. Their frameworks properly describe situations in which humans and AI collaborate and when they try to provide oversight or control on each other, but they fail to capture many of the nuances of how humans and AI implement collaboration or oversight. Many of these nuances are captured in different existing scales of automation and autonomy, but these scales are often too confined to specific domains or systems to fit into system architecture definitions of human-AI collaboration and oversight, which are more useful in describing and comparing systems broadly. We study these existing levels of automation frameworks to further decompose human-AI system architecture from the current high-level definitions prevalent in the literature today. The following section provides an overview of the frameworks we analyzed.

**Analysis: Levels of Automation**

Levels of automation have been a common framework to define how human oversight and control is implemented. Williams describes two types of levels of autonomy scales: linear and multi-dimensional [14]. Generally, linear scales do not define the actions that make up ‘the-loop’ explicitly. Instead, they focus on the roles of humans and automation, with the automation’s role increasing as the levels go up, generally corresponding to a decrease in the human’s role in the larger system. In contrast, multi-dimensional scales define levels of autonomy along a set of tasks a system is expected to perform. This set of tasks can be considered a generalization of ‘the-loop’. In multi-dimensional scales, increasing levels of autonomy may mean increased use of AI in only one, a few, or all of the tasks explicitly defined [14]. The analysis of levels of automation and autonomy analyzes and summarizes linear scales and multi-dimensional scales. In the following subsections, we focus on how the linear and multi-dimensional scales can be aggregated and analyzed to provide further granularity to the human/AI-in/on/over-the loop definitions that exist in the literature today. We use this analysis to inform our framework in the following section.

*Linear Scales*

We studied two types of linear scales: foundational scales that are not context or system specific and applied scales which are context / system specific. Table 1 shows the aggregated analysis for both types of linear scales. At the lowest levels of most frameworks, the human maintains continuous control of the task while the AI system acts as an advisor or as a tool the human operator can delegate subtasks to. The following levels give the AI control over the task but require humans to either select a plan of action from a list of options or to approve the AI system’s accepted plan of action. At these levels, the AI system cannot act without human direction. The next levels see the humans in a supervisory role, meaning that their action is not explicitly required for the AI system to act. In the lower level of supervisory control, the AI system gives the human a restricted time to veto the planned action before it is implemented. The higher level of supervisory control sees the human operator take control over the function of the AI system if there is an issue.

There is also a group of levels that define human-AI teams, where both the human operator and AI system perform tasks throughout the loop and are viewed as equal partners. We found two types of human-AI teams, one in which the subtasks each party is responsible for do not overlap and allocation is fixed and one in which the subtasks may be performed by either party with dynamic tasking. The penultimate level in our table of analysis does not allow the human to act but provides them with insight into how the system is performing. The final level flips the human control paradigm, instead giving AI the ability to override human actions.

Table 2: Summarized Linear Scales

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Architecture | Description of System | Corresponding Level in Frameworks | | | | | | | |
| **Sheridan** | **Endsley** | **Yang** | | **NHSTA** | **Riley** | | **Draper** |
| AI-out-of-the-Loop | All Human | 1 | 1 | 0 | |  | 1 | | 1 |
| AI Tool | **Human maintains continuous control; AI assists** |  |  | **1** | | **0, 1, 2** | **2-7** | | **2** |
| Assists as an advisor |  |  |  | |  | 2-6 | |  |
| Assists with task |  |  |  | |  | 7 | | 2 |
| Human Selector | **AI suggests alternative plans of action à**  **Human selects à AI implements** | **2, 3, 4** | **2** |  | |  |  | |  |
| *Complete set* | *2* |  |  | |  |  | |  |
| *Narrows down alternatives* | *3* |  |  | |  |  | |  |
| *Few alternatives* | *4* |  |  | |  |  | |  |
| Human Approver | **AI executes its decision if the human approves** | **5, 6** | **3** | **2** | |  | **8, 9** | |  |
| *Human must approve* | *5* |  |  | |  | *8* | |  |
| *Human has restricted veto time* | *6* |  |  | |  | *9* | |  |
| Human-AI Team | **Both human and AI perform tasks ‘in-the-loop’** |  |  |  | |  | | **10** | **3, 4** |
| *Human & AI share responsibility for tasks* |  |  |  | |  | |  | *4* |
| *Human & AI task responsibility does not overlap* |  |  |  | |  | |  | *3* |
| *Human and AI can override each other* |  |  |  | |  | | *10* |  |
| Human-on-the-Loop | **AI performs task; human has takeover capability** |  | **4** | **3 & 4** | | **3** |  | | **5** |
| *Humans decides tasks* |  |  | *3* | |  |  | |  |
| *AI decides tasks* |  |  | *4* | |  |  | |  |
| Human-over-the-Loop | **AI executes automatically; informs human** | **7, 8, 9** |  |  | |  |  | |  |
| *Necessarily* | *7* |  |  | |  |  | |  |
| *Only if asked* | *8* |  |  | |  |  | |  |
| *Only if computer decides to* | *9* |  |  | |  |  | |  |
| AI-on-the-Loop | **AI can override human during their operation** |  |  |  |  | | **11** | |  |
| Human-out-of-the-Loop | All Autonomy | 10 | 5 | 5 | | 4, 5 | 12 | |  |
| *Within environmental constraints* |  |  |  | | *4* |  | |  |
| *No environmental constraints* |  |  |  | | *5* |  | |  |

*Multi-Dimensional Scales*

Multi-dimensional scales use high-level task sequences to define levels of automation. Endsley and Kaber’s framework is based around four functions: monitoring, generating, selecting, and implementing [15], although the monitoring function is always shared between the system and the human. Beer, Fisk, and Rogers’ framework for robot autonomy is centered around three functions: sense, plan, and act [16]. Proud, Hart and Mrozinski from NASA’s Johnson Space Center developed a levels of autonomy scale using the observe, orient, decide, and act (OODA) loop [17]. To analyze across these scales, we focused on the different system architectures that were common rather than focusing on the specific ways each framework chose to define ‘the-loop’.

Many levels in the multi-dimensional scales correspond to those observed in linear scales, such as assisted teleoperation being the same as AI tool. While being functionally the same, multi-dimensional scale definitions are often more detailed in describing levels since they explicitly decompose ‘the-loop’. Batch processing, shared control, decision support, and rigid system are all different implementations of human selector architecture, but the multi-dimensional frameworks distinguish systems based on who can generate possible options. Shared control with human/robot initiative is the same as supervisory control in linear scales, but Beer makes a distinction between the two levels by whose initiative, robot or human, is required for the intervention to take place. While the majority of levels in the multi-dimensional scales correspond to levels in linear scales, executive control emerged as a new level. In executive control, the human operator gives a high-level goal to a system to carry out on its own. During operation, the human operator cannot intervene with operations, but there may be opportunities for feedback or evaluation after the fact. This is similar to but distinct from human-over-the-loop architectures observed in the linear scales.

Save and Feuerberg do not define levels of automation on a system level; rather, they define separate levels of automation for four functions: information acquisition, information analysis, decision and action selection, and action implementation [34]. The key distinctions between levels for information acquisition and information analysis levels of automation depend on whose initiative is required for the system to act. Action selection differs based on *who generates options and who selects the plan of action.* Finally, action implementation differs on whose initiative is required for the system to act, whether the system is providing information on how to do a task, assisting in the task, or preforming the task itself, and if the human can monitor and intervene at any time or only during certain times.

Table 3: Summarized Multidimensional Scales

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Description** | **Endsley** | **Beer** | **Proud** |
| Tool | | Human operates; system can intervene during task | --- | Assisted Teleoperation | --- |
| Human Selector | | Human selects plan of action | Batch Processing | Batch Processing |  |
|  | *Shared Control (SC)* | *Human & system both generate options; both act* | *Shared Control* |  |  |
|  | *Decision Support* | *System generates options; system acts* | *Decision Support* | *Decision Support* |  |
|  | *Rigid System* | *System generates options, human must select an option* | *Rigid System* |  |  |
| Human Approver | | System generates & selects plan of action |  |  |  |
|  | *Blended Decision Making (BDM)* | *System implements only if human approves or implements alternate selected by human* | *BDM* | *---* | *3* |
|  | Decision Making with Veto | Human has time to veto before action is implemented. | --- | --- | 4, 5 |
| Supervisory Control | | System acts on its own; human can override | Supervisory Control | Supervisory Control | 6 |
|  | *with Human Initiative* | *human may intervene with new goals / tasks if system encounters difficulty* |  | *SC with Human Initiative* |  |
|  | *with Robot Initiative* | *system can prompt human to intervene with new goals if it faces difficulty* |  | *SC with Robot Initiative* |  |
|  | Human Emergency Override | *System acts on its own, but human can override after action* | --- | --- | 6, 7 |
| Executive Control | | Human gives system a high-level goal; system acts | --- | Executive Control | --- |
|  | *Automated Decision Making (ADM)* | *Human and system generate options; system selects & acts* | *ADM* |  |  |

**Framework**

Our review of existing levels of automation and autonomy revealed key mechanisms that humans and AI have to exercise control over each other and also revealed more nuanced ways of working together. Using these insights, we created a framework, shown in Figure 2 in the appendix that decomposes system architecture much further than the literature does today. The framework decomposes human-AI systems by asking mutually exclusive questions about the roles of humans and AI. Our analysis of existing frameworks reveled many variables that are not shown in the framework that could be used to decompose this framework further, but we decomposed the broad array of human-AI system architectures to a level that is focused on mechanisms of control / partnership between humans and AI. The other variables identified are related to how exactly the system is implemented, such as remote monitoring vs in-person monitoring of a system, rather than providing insight as to how the human or AI is expected to exercise oversight or engage together as teammates. These implementation variables will be considered in later stages.

*Decomposing Human-AI Systems*

Human-in-the-loop vs on-the-loop is the most widely recognized distinction between human-AI systems. The exact definitions of both terms vary depending on the author, but generally, human-on-the-loop systems place the human in supervisory role where action is optional, while human-in-the-loop systems require some form of human action for the system to function. There is an often-overlooked class of system where AI serves as an optional supervisor called AI-on-the-loop. The question that differentiates these three overarching types of human-AI systems is whose participation is strictly required for the system to function. While these distinctions are helpful, the levels of automation literature provide another layer of resolution relevant to how these broad architectures function. Therefore, the framework further decomposes human-in-the-loop, human-on-the-loop, and AI-on-the-loop system architectures. We will first discuss human-on-the-loop and AI-on-the-loop system architectures since they both place humans and AI in parallel roles to each other, with humans supervising AI in the former and AI supervising humans in the latter. Human-in-the-loop systems are discussed at the end since they are a more special case.

*Human/AI-on-the-Loop*

Human-on-the-loop system architectures give humans the ability, but not the obligation, to act. AI-on-the-loop system architectures give AI the ability, but not the obligation, to act. In both, there are two distinct types of on-the-loop action. The first is supervisory control whereby the party that is on-the-loop can directly intervene directly in the operations. The second is over-the-loop control, whereby the party that is on-the-loop can only influence operations indirectly. We discuss the architectures for human supervisory control, AI supervisory control, and human/AI-over-the-loop control in the following sections.

*Human Supervisory Control*

There are two architectures for human supervisory control that differ based on when the human can intervene. One is command by veto in which humans may disapprove of the AI system’s planned course of action before execution, but if no negation is implemented, the system functions fully autonomously. The Patriot Missile Defense System, which defends a geographic area from missile attacks, is an excellent example of command by veto architecture. In its automatic control mode, the “system will fire unless human operator halts engagement,” meaning that the system is fully autonomous unless the human operators decide to act [18, p. 6]. The other means of supervisory control is humans taking over the system during operation, serving as a human supervisor. In contrast to command by veto where the AI is still tasked with the function after a human veto, the human takes over control of the system function for some time. Uber’s now-defunct self-driving car service provides a good example of human takeover architecture. Uber employed human operators who were “tasked with overseeing the system’s operation, monitoring the driving environment, and if necessary, taking control of the vehicle and intervening in an emergency” [19, p. 8]. This means that the main function of the system, driving, would be taken over by a human until the potential emergency was over or some other trigger point that would cause the human operator to relinquish control.

*AI Supervisory Control*

There are two ways for AI systems to intervene in human operations, which are differentiated by whose initiative is required for the AI system to act. One option is for an AI supervisor, which monitors the human(s) performing the task, to intervene, when necessary, by taking over control of the function. This may be in case of emergency or if human behavior deviates from some preset bounds. The Airbus Auto Pilot/Flight Director provides a classic example of an AI system that takeover for human operators. If a collision is imminent, the Auto Pilot “initiates and executes a sequence of actions to fly the avoidance maneuvers until the aircraft is clear of conflict” while the crew “cannot modify the ongoing action execution” [20, p. 52]. In this case, it is the system’s initiative that is required for it to act. The alternative is for the human performing the task to optionally delegate to an AI tool. The choice on when and what to delegate is up to the human and can be revoked to return to fully human operation. An example of an AI tool is GitHub copilot, provides suggestions as the human is working but the user must affirm a suggestion for GitHub to act and is not obligated to do so [21]. The question as to how much can be delegated to an AI system would be dependent on its capability and potential organizational constraints.

*Human-over-the-Loop*

Human-over-the-loop system architectures allow humans to influence the system but not to intervene. This influence can take place either before, during or after operation. Before operation, the human can set a high-level goal for the AI system to achieve, leaving the implementation to the AI system. The AutoNav system on the Mars Perseverance Rover “the rover to autonomously re-plan its route around rocks or other obstacles on its way to a pre-established destination” after getting an initial route and destination from the NASA crew on Earth [22]. Intervention during operation is not feasible given the radio delay between Earth and Mars [23]. During operation, humans can adjust set points or provide new goals if the AI system encounters issues or if there is a significant change in the operating environment. An example of this type of executive control would be the RQ-4 drone operated by the United States Air Force. The RQ-4 operators “can provide new guidance in terms of waypoints or heading but cannot directly provide inputs to the control surfaces” [13, p. 2]. After operation, humans serve as a feedback loop, providing feedback on system performance in order to change future behavior. Spam email filters like Gmail’s filter “are continuously updated with …the feedback from Gmail users about likely spammers” to continuously improve the system [24, p. 6].

*AI-over-the-Loop*

AI-over-the-loop architectures give AI systems the ability to influence humans but not to intervene. Just as AI supervisory control was differentiated by whose initiative is required, so too are AI-over-the-loop systems. AI systems may try to influence human behavior through warnings in case of emergency of or other behavior outside a determined set of bounds. Serving as an AI monitor, a system provides the alert or additional information on its own initiative. Traffic Alert & Collision Avoidance System (TCAS) provides an excellent example of a monitor architecture. If a collision is imminent, TCAS “triggers visual and aural indications to the flight crew to perform an avoiding vertical maneuver,” with the “execution of the maneuver itself [left] up to the flight crew” [20, p. 51]. Otherwise, humans may use AI as a decision aid whereby they call on an AI system to provide or analyze information to help them plan. We make a distinction between an AI decision aid and an AI tool by noting that a tool is performing a part of a task, while an aid only provides additional context or information for a human. In essences, an AI decision aid only augments human decision making while leaving the entirety of the task up to the human. En Route Air Traffic Organizer (ERATO) is an example of a decision aid architecture. ERATO is “activated on controller’s initiative” and “helps the controller in identifying all the potential intruders of a given flight in the medium-short term” [20, p. 52].

*Human-in-the-Loop*

Human-in-the-loop architecture requires active participation from both the human and autonomous system involved in the loop; however, within that broad definition, we found two distinct types of human participation. The first is humans serving as directors for AI systems, meaning that the human is responsible for deciding a plan of action for the AI system to carry out. The other type of participation is teaming, where both the human and AI participate throughout ‘the-loop.’ Without the action of both parties, the loop cannot be completed. We discuss the nuances of both of these types of human-in-the-loop systems in the following sections.

*Human-in-the-Loop: Director*

As the party responsible for deciding how an AI system should act, humans can be either implemented as approvers or selectors. The key differentiating factor is the number of potential plans of action that the human operator is assessing. If the AI system generates a single plan of action to be approved or denied, the human is serving as an approver. The system cannot act without approval. If denied, the system may propose a new plan of action or may cease function, depending on the implementation of the system. An example of a human approver system is the Patriot Missile Defense System’s semi-automatic mode, in which the “human operator must authorize engagement or system will not fire” [18, p. 6]. Under selector architecture, the AI system generates several plans of action for the human to choose among. Surgical robots in a wide range of practices including “bone milling in orthopedic surgery, prostate biopsy in urology, and hair follicle extraction in plastic surgery” employ selector architecture [25, p. 2]. Generally, these systems generate “potential strategies for [the] surgeon to select” from, after which the “system takes over control to execute the selected plan” [25, p. 3].

*Human-in-the-Loop: Team*

Human-AI team architectures involve humans throughout the loop, with two distinct types of teams. One is human-AI handoff where the subtasks that the human and AI system are responsible for fixed sub-tasks. For example, U.S. Customs and Border Control (CBP) uses facial recognition to match travelers to their photos, helping catch imposters, while officers perform behavioral / contextual screening [26]. Hypothetically, there should be no overlap between task responsibilities, but governance requirements can create the need for a human or AI backup. In the case of CBP, travelers have the right to request human verification, but in lieu of travelers exercising this right, the tasks are divided between humans and AI. The other type of teaming architecture is human-AI partnership where the task allocation is dynamic, meaning that many subtasks can be completed by either the human or the AI. Security robots have employed in a variety of municipalities and private businesses. These robots can actively patrol and record events, with some being able to step in and call for help if a disturbance is detected [27]. These robots act as partners to security guards since they can both do the same tasks, such as going on patrol, depending on environmental or other circumstances.

**Ethical Considerations for Human-AI System Architecture**

Meaningful human control is an ethical standard developed during international debate about the use of lethal autonomous weapons (LAWS) [28] [29] [30]. Horowitz & Scharre prescribe three requirements for meaningful human control: “human operators are making informed, conscious decisions about the use of weapons,” human operators have enough information to make determinations about the lawfulness of their decisions, and that the system is tested, with operators receiving training on usage [28, p. 4]. Adding to this perspective from a philosophical standpoint, Santoni de Sio and van der Hoven describe the need for two design requirements: tracking (system is able to incorporate moral reasoning of human designers and operators) and tracing (system decisions and outcomes can be traced back to the action of at least one person in decision chain) [31].

The United States Department of Defense (DoD) has its own policies on LAWS. DoD Directive 3000.09 on the use of autonomy in weapons systems requires that autonomous weapons systems are “designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force” [5, p. 2]. Congressional Research Service points out that human judgment can mean simply “involvement in decisions about how, when, where, and why the weapon will be employed” [32, p. 1]. 3000.09 also provides ambiguity on, appropriateness, reflecting the fact that “what is ‘appropriate’ can differ across weapon systems, domains of warfare, types of warfare, operational contexts, and even across different functions in a weapon system” [33, p. 2].

DoD Directive 3009.09 also lays out requirements for AI systems to incorporate governance. DoD defines governance as “possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior” [5, p. 6]. Adding to the requirements for human operators, the directive also requires that “a monitoring regime is in place to identify and address changes in operational environment, data inputs, and use that could contribute to … failures” [5, p. 16]. While vague, the directive clearly requires that operators be able to monitor the system for errors, whether they are caused by system malfunction, changes in the operating environment, or the emergent edge cases the system was not designed to handle.

Many have challenged the merit of current ethical standards that are used to govern safety critical AI systems. Challenges range from the vagueness of language to how they may fail to capture other relevant moral problems [34] [35]. Still, autonomous weapons and other safety critical systems are in use today [36] [37]. It is imperative to begin to analyze how to implement meaningful human control over these systems, even if by imperfect metrics. Meaningful human control still represents the moral foundation on which major militaries are willing to base decisions on the control and governance of LAWS [38]. Additionally, meaningful human control has been used to analyze other safety critical systems, such as autonomous vehicles [38] [39] [40].

We analyzed if the 11 architectures we found would be able to meet the tracing and tracking conditions, as shown in Table 4. We find that most architectures are able to meet the tracing and tracking requirements. Depending on the exact implementation of the system, the ability for tracing and tracking to be effectively met may be hindered, but we are focused on a general view of each architecture. The three human-over-the-loop architectures are unlikely to be able to meet these requirements since they are unable to directly act during operation, making it difficult to know who to trace decisions back to. Even tracking whether the human operator is able to apply moral reasoning is difficult since they have relatively little control over the system and the actions it takes. AI supervisor is also somewhat problematic in this sense since the human’s actions can be override by an AI system that cannot be held liable the same way a human can. This is also a violation of the human supervisory control requirement laid out by 3009.09 since it flips the role and has the AI as a supervisor that can step in and act without human oversight.

Table 4: Compliance with Tracking & Tracing

|  |  |  |
| --- | --- | --- |
| **Architecture** | **Tracking** | **Tracing** |
| *Moral reasoning applied through …* | *Decisions can be traced back to …* |
| Human-AI Partner | human’s action as part of a team | human teammate |
| Human-AI Handoff | human’s action as part of a team | human teammate |
| Human Approver | (dis)approval of AI’s plan of action | operator who makes decision |
| Human Selector | selection of plan of action | operator who makes decision |
| Executive Control | setting of overall goal | operator who set goals |
| Human Decision Aid | adjusting set points | operator if they are able to intervene |
| Human Feedback Loop | feedback after the fact | *no single person during operations* |
| Command by Veto | human’s decision to veto or not | operator who makes decision |
| AI Tool | human’s decision to takeover or not | operator who makes decision |
| AI Aid | human’s own action | human who is performing task |
| AI Monitor | human’s own action | human who is performing task |
| AI Tool | human’s decision to delegate or not | human who delegates task |
| AI Supervisor | human’s action, unless AI overrides | human performing task, unless AI overrides |

**Application to SilverFish**

There are a variety of ways that these architectures could be implemented in Silverfish. We do not currently consider the architectures that did not meet the current minimal ethical standards for employing AI in safety-critical and lethal systems, but we may consider them in later stages if these requirements are relaxed. Below we present an initial cut at some of the possible configurations Silverfish could employ according to the framework we developed. We will continue to develop and refine these implementations as the project progresses.

Both human-AI teaming architectures could be applied to the Silverfish Safe Passage System. Under a partnership, the command-and-control center would develop some heuristics as to when it is ideal to deploy the human SME or just use the AI system. Essentially, the decision on whether or not human SMEs should be employed on a link a dynamic decision, which might be modeled via a stochastic Monte Carlo process. Under a handoff scheme, the decision as to whether or not to employ a human SME would be fixed rules-based decision. If certain time, terrain, lighting, etc. conditions are triggered, only then the link would be automatically sent for review by a human SME.

Both human director architectures could also be applied to Silverfish. Under a human selector architecture, the role of AI in silverfish could extend to providing a series of options, including the option for additional human review if confidence is low, to the command-and-control operator to select from. Depending on the implementation, the human could be forced to select from only the AI-generated options or be able to generate and select their own option. If Silverfish employed a human approved architecture, the AI could develop its own optimized plan of action for routing including potential review for the operator to approve. If not approved, the system would present another option to the human. This cycle would run till a plan of action is approved.

As discussed in the ethical analysis section, implementing human-over-the-loop architecture is unlikely in a safety-critical system since humans cannot intervene or act in the system. This leaves the human supervisory control architectures. A command by veto architecture could be implemented whereby the system gives the human operator a fixed amount of time to veto after which its decision is locked in and must be implemented. A human takeover architecture might involve human operators having insight to a command-and-control AI’s decision-making process and being able to stop the process and assume command if there are inconsistencies or issues found.

**Anticipated Results**

Few papers have compared the performance of different system architectures. Those that do typically focus only on ‘in-the-loop’ participatory control vs ‘on-the-loop’ supervisory control [41] [42] [43] [44]. Additionally, since ‘in-the-loop’ and ‘on-the-loop’ have typically been ill defined, it is difficult to understand how findings of these papers should be implemented. Thus, it has been unclear what the tradeoffs of different architectures are. By implementing different human-AI architectures in the same reference system, we hope to understand the tradeoffs of implementing different system architectures. The key behavior we hope to understand through this study is the tradeoffs between *performance*, using measures such as total amount of time it takes to move troops across the different networks provided and *risk*, using measures such as the percentage of safe traversals, for each architecture.

For example, if we compare human approver and command by veto architectures, we can see some clear tradeoffs for both. In human approver architecture, there is a clear point where the human can intervene, reducing the mental load on the human. However, humans in this role have exhibited automation bias, meaning that they tend to blindly agree with the AI’s recommendation [45]. Calibrating human operators to have the correct amount of trust or distrust is a difficult balance. In command by veto architecture, the AI system is able to perform the entire loop without interruption, which may be preferable in high-intensity, fast paced situations. This also places a heavy load on the human operator who must continuously monitor the system while not having a clearly defined point where they should intervene.

We hope to discover more detailed tradeoffs for each architecture studied to provide better guidance on system architecting for AI-enabled safety-critical and lethal systems. Our breakdown of possible architectures, joined with our research on current ethical and legal standards enables us to more granularly define human-AI system architecture in this setting than existing studies. Combined with the Silverfish Safe Passage System as our sandbox case study, we can understand the tradeoffs between risk and performance will be heavily tested when non-lethality assumptions are relaxed in future stages of this project.

**Works Cited**

[1] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework,” U.S. Department of Commerce, Gaithersburg, MD, AI RMF NIST AI 100-1.0, Jan. 2023. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1

[2] J. Biden, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. 2023. Accessed: Nov. 18, 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[3] High-Level Expert Group on AI, “Independent High-Level Expert Group on Artificial Intelligence,” European Commission, Brussels, B-1049, Apr. 2019.

[4] China Academy of Information and Communications Technology and JD Explore Academy, “White Paper on Trustworthy Artificial Intelligence,” Ministry of Industry and Information Technology, Jul. 2021. [Online]. Available: https://cset.georgetown.edu/wp-content/uploads/t0390\_trustworthy\_AI\_EN.pdf

[5] Office of the Under Secretary of Defense for Policy, *Autonomy in Weapon Systems*, vol. 3009.09. 2023.

[6] M. H. Jarrahi, “Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making,” *Bus. Horiz.*, vol. 61, no. 4, pp. 577–586, Jul. 2018, doi: 10.1016/j.bushor.2018.03.007.

[7] A. B. Fontaine, “A Model for Human Tracking Behavior in a Closed Loop Control System,” Dissertation, The Ohio State University, Columbus, 1954.

[8] J. L. Decker, “The Human Pilot and the High-Speed Airplane,” *J. Aeronaut. Sci.*, vol. 23, no. 8, pp. 765–770, 1956, doi: 10.2514/8.3652.

[9] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, “A Survey of Human-in-the-loop for Machine Learning,” *Future Gener. Comput. Syst.*, vol. 135, pp. 364–381, Oct. 2022, doi: 10.1016/j.future.2022.05.014.

[10] C. Chai and G. Li, “Human-in-the-loop Techniques in Machine Learning,” *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, p. 16, Sep. 2020.

[11] X.-L. Meng, “Data Science and Engineering With Human in the Loop, Behind the Loop, and Above the Loop,” *Harv. Data Sci. Rev.*, vol. 5, no. 2, Apr. 2023, doi: 10.1162/99608f92.68a012eb.

[12] S. Munir, J. A. Stankovic, C.-J. M. Liang, and S. Lin, “Cyber Physical System Challenges for {Human-in-the-Loop} Control,” presented at the 8th International Workshop on Feedback Computing (Feedback Computing 13), 2013. Accessed: Aug. 03, 2023. [Online]. Available: https://www.usenix.org/conference/feedbackcomputing13/workshop-program/presentation/munir

[13] A. R. Lacher, L. Ren, D. R. Maroney, C. Schulenberg, and J. Daniels, “Dimensional Role Analysis: The Role of Humans and Automation for Increasingly Autonomous Aviation Systems,” in *2023 Integrated Communication, Navigation and Surveillance Conference (ICNS)*, Apr. 2023, pp. 1–6. doi: 10.1109/ICNS58246.2023.10124260.

[14] A. Williams, “Defining Autonomy in Systems: Challenges and Solutions,” in *Autonomous Systems: Issues for Defence Policymakers*, in Innovation in Capability Development, no. 2. , Norfolk: Headquarters Supreme Allied Commander of The North Atlantic Treaty Organization, 2015.

[15] M. R. Endsley and D. B. Kaber, “Level of automation effects on performance, situation awareness and workload in a dynamic control task,” *Ergonomics*, vol. 42, no. 3, pp. 462–492, Mar. 1999, doi: 10.1080/001401399185595.

[16] J. M. Beer, A. D. Fisk, and W. A. Rogers, “Toward a framework for levels of robot autonomy in human-robot interaction,” *J. Hum.-Robot Interact.*, vol. 3, no. 2, pp. 74–99, Jul. 2014, doi: 10.5898/JHRI.3.2.Beer.

[17] R. W. Proud, J. J. Hart, and R. B. Mrozinski, “Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach,” National Aeronautics and Space Administration, Houston, Texas, Sep. 2003.

[18] J. Hawley, “Patriot Wars,” Center for a New American Secuirty, Washington, DC, Jan. 2017. [Online]. Available: https://www.cnas.org/publications/reports/patriot-wars

[19] National Transportation Safety Board, “Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona,” National Transportation Safety Board, Washington, D.C., NTSB/HAR-19/03 PB2019-101402, Mar. 2018.

[20] L. Save and B. Feuerberg, “Designing Human-Automation Interaction: a new level of Automation Taxonomy”.

[21] “About GitHub Copilot Individual,” GitHub Docs. Accessed: Apr. 13, 2024. [Online]. Available: https://docs.github.com/en/copilot/copilot-individual/about-github-copilot-individual

[22] “How Mars Perseverance Rover AutoNav Avoids a Boulder,” SpaceNews. Accessed: Jul. 26, 2024. [Online]. Available: http://spacenews.com/how-mars-perseverance-rover-autonav-avoids-a-boulder/

[23] NASA, “NASA’s Self-Driving Perseverance Mars Rover ‘Takes the Wheel,’” NASA Jet Propulsion Laboratory (JPL). Accessed: Jul. 26, 2024. [Online]. Available: https://www.jpl.nasa.gov/news/nasas-self-driving-perseverance-mars-rover-takes-the-wheel

[24] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, p. e01802, Jun. 2019, doi: 10.1016/j.heliyon.2019.e01802.

[25] A. Lee, T. S. Baker, J. B. Bederson, and B. I. Rapoport, “Levels of autonomy in FDA-cleared surgical robots: a systematic review,” *Npj Digit. Med.*, vol. 7, no. 1, pp. 1–8, Apr. 2024, doi: 10.1038/s41746-024-01102-y.

[26] “Say hello to the new face of efficiency, security and safety,” U.S. Customs and Border Protection. Accessed: Apr. 13, 2024. [Online]. Available: https://www.cbp.gov/travel/biometrics

[27] C. Farivar, “Security robots expand across U.S., with few tangible results,” NBC News. Accessed: Jul. 31, 2024. [Online]. Available: https://www.nbcnews.com/business/business-news/security-robots-expand-across-u-s-few-tangible-results-n1272421

[28] M. C. Horowitz and P. Scharre, “Meaningful Human Control in Weapon Systems: A Primer,” Center for a New American Security, Working Paper, Mar. 2015. [Online]. Available: https://www.cnas.org/publications/reports/meaningful-human-control-in-weapon-systems-a-primer

[29] F. Santoni de Sio and J. van den Hoven, “Meaningful Human Control over Autonomous Systems: A Philosophical Account,” *Front. Robot. AI*, vol. 5, 2018, Accessed: Sep. 13, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2018.00015

[30] I. Verdiesen, F. Santoni de Sio, and V. Dignum, “Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight,” *Minds Mach.*, vol. 31, no. 1, pp. 137–163, Mar. 2021, doi: 10.1007/s11023-020-09532-9.

[31] F. Santoni de Sio and J. van den Hoven, “Meaningful Human Control over Autonomous Systems: A Philosophical Account,” *Front. Robot. AI*, vol. 5, 2018, Accessed: Sep. 13, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2018.00015

[32] Congressional Research Service, “Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems,” United States Congress, Washington, D.C., IF11150, Feb. 2024.

[33] United States, “Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems,” United Nations, Geneva, CCW/GGE.2/2018/WP.4, Aug. 2018.

[34] B. Green, “The flaws of policies requiring human oversight of government algorithms,” *Comput. Law Secur. Rev.*, vol. 45, p. 105681, Jul. 2022, doi: 10.1016/j.clsr.2022.105681.

[35] M. C. Canellas and R. A. Haga, “Toward meaningful human control of autonomous weapons systems through function allocation,” in *2015 IEEE International Symposium on Technology and Society (ISTAS)*, Nov. 2015, pp. 1–7. doi: 10.1109/ISTAS.2015.7439432.

[36] “Lethal Autonomous Weapons Exist; They Must Be Banned - IEEE Spectrum.” Accessed: Mar. 18, 2024. [Online]. Available: https://spectrum.ieee.org/lethal-autonomous-weapons-exist-they-must-be-banned

[37] G. D. Vynck, “The U.S. says humans will always be in control of AI weapons. But the age of autonomous war is already here.,” *Washington Post*, Aug. 13, 2021. Accessed: Mar. 18, 2024. [Online]. Available: https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/

[38] S. C. Calvert, S. Johnsen, and A. George, “Designing Automated Vehicle and Traffic Systems towards Meaningful Human Control,” Mar. 14, 2023, *arXiv*: arXiv:2303.05091. doi: 10.48550/arXiv.2303.05091.

[39] M. Christen, T. Burri, S. Kandul, and P. Vörös, “Who is controlling whom? Reframing ‘meaningful human control’ of AI systems in security,” *Ethics Inf. Technol.*, vol. 25, no. 1, p. 10, Feb. 2023, doi: 10.1007/s10676-023-09686-x.

[40] F. Ficuciello, G. Tamburrini, A. Arezzo, L. Villani, and B. Siciliano, “Autonomy in surgical robots and its meaningful human control,” *Paladyn J. Behav. Robot.*, vol. 10, no. 1, pp. 30–43, Jan. 2019, doi: 10.1515/pjbr-2019-0002.

[41] M. Gil, M. Albert, J. Fons, and V. Pelechano, “Designing human-in-the-loop autonomous Cyber-Physical Systems,” *Int. J. Hum.-Comput. Stud.*, vol. 130, pp. 21–39, Oct. 2019, doi: 10.1016/j.ijhcs.2019.04.006.

[42] M. Gil, V. Pelechano, J. Fons, and M. Albert, “Designing the Human in the Loop of Self-Adaptive Systems,” in *Ubiquitous Computing and Ambient Intelligence*, C. R. García, P. Caballero-Gil, M. Burmester, and A. Quesada-Arencibia, Eds., Cham: Springer International Publishing, 2016, pp. 437–449. doi: 10.1007/978-3-319-48746-5\_45.

[43] M. Gil, M. Albert, J. Fons, and V. Pelechano, “Engineering human-in-the-loop interactions in cyber-physical systems,” *Inf. Softw. Technol.*, vol. 126, p. 106349, Oct. 2020, doi: 10.1016/j.infsof.2020.106349.

[44] J. E. Fischer, C. Greenhalgh, W. Jiang, S. D. Ramchurn, F. Wu, and T. Rodden, “In-the-loop or on-the-loop? Interactional arrangements to support team coordination with a planning agent,” *Concurr. Comput. Pract. Exp.*, vol. 33, no. 8, p. e4082, 2021, doi: 10.1002/cpe.4082.

[45] N. Sharkey, “Staying in the loop: human supervisory control of weapons,” in *Autonomous Weapons Systems: Law, Ethics, Policy*, United Kingdom: Cambridge University Press, 2016, pp. 23–38.

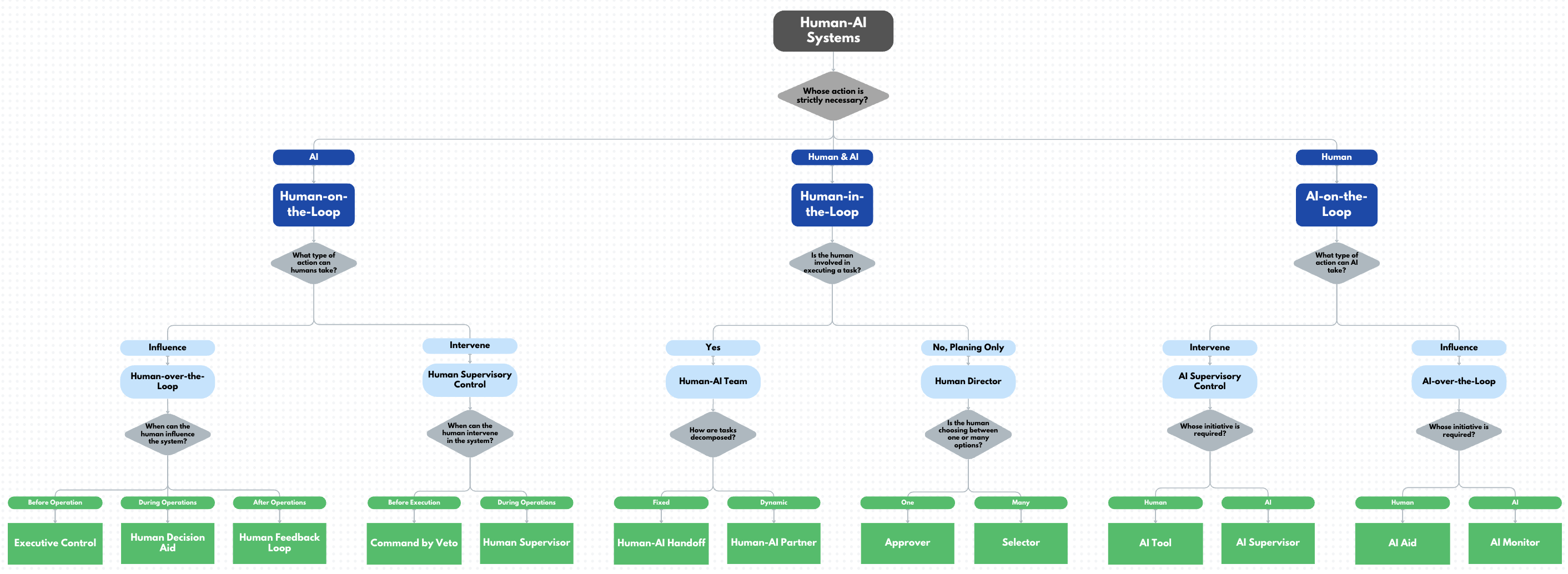


Figure 2: Framework