frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 22 July 2022
DOI 10.3389/fphy.2022.942245

Building trust and responsibility into autonomous human-machine teams

Tony Gillespie*

Electronic and Electrical Engineering, UCL, London, United Kingdom

OPEN ACCESS

EDITED BY
William Frere Lawless,
Paine College, United States

REVIEWED BY
Dioneia Motta Monte-Serrat,
University of São Paulo, Brazil
Victoria Hailey,
VHG-The Victoria Hailey Group
Corporation, Canada
David Short,
BAE Systems, United Kingdom

*CORRESPONDENCE
Tony Gillespie,
Anthony.gillespie@ucl.ac.uk

SPECIALTY SECTION
This article was submitted to Interdisciplinary Physics, a section of the journal Frontiers in Physics

RECEIVED 12 May 2022
ACCEPTED 27 June 2022
PUBLISHED 22 July 2022

CITATION
Gillespie T (2022), Building trust and responsibility into autonomous human-machine teams.
Front. Phys. 10:942245.
doi: 10.3389/fphy.2022.942245

Harm can be caused to people and property by any highly-automated system, even with a human user, due to misuse or design; but which human has the legal liability for the consequences of the harm is not clear, or even which laws apply. The position is less clear for an interdependent Autonomous Human Machine Team System (A-HMT-S) which achieves its aim by reallocating tasks and resources between the human Team Leader and the Cyber Physical System (CPS). A-HMT-S are now feasible and may be the only solution for complex problems. However, legal authorities presume that humans are ultimately responsible for the actions of any automated system, including ones using Artificial Intelligence (AI) to replace human judgement. The concept of trust for an A-HMT-S using AI is examined in this paper with three critical questions being posed which must be addressed before an A-HMT-S can be trusted. A hierarchical system architecture is used to answer these questions, combined with a method to limit a node's behaviour, ensuring actions requiring human judgement are referred to the user. The underpinning issues requiring Research and Development (R&D) for A-HMT-S applications are identified and where legal input is required to minimize financial and legal risk for all stakeholders. This work takes a step towards addressing the problems of developing autonomy for interdependent human-machine teams and systems.

KEYWORDS

human machine team, trust, autonomy, legal liability, artificial intelligence, risk, interdependence

Introduction

Achieving Artificial Intelligence's (AI) full potential for any application will require considerable research and engineering effort [1]. New AI-engineering techniques will need to be developed, especially when AI-based systems interact with humans [2]. Technology has evolved to the point where Human Machine Teams (HMTs) can dynamically and automatically reallocate tasks between human and machine team members to optimise workloads and resource usage, an Autonomous Human Machine Team System (A-HMT-S). However, interdependence between team members with very different capabilities raises serious system challenges to ensure the safe, trusted transfer of authority between human and machine.

When the human user of a Cyber-Physical System (CPS) has given it an aim, and its subsequent actions are guided by AI, questions arise about the roles of the human, the AI, and that of the people responsible for its autonomous behaviour. Who was responsible for its actions and any harm caused by those actions? The legal position is evolving, with no clear consensus. Reference [3] covers the current legal position for AI and suggests likely developments.

The use of an A-HMT-S to achieve an aim implies complexity, requiring reasoning to achieve it. Although a team approach may be efficient, there are legal complications when the aim is to take an action, or to provide information for someone or something to take an action that could cause harm. Assignment of responsibly for the consequences of machine-made decisions is becoming an important issue now that CPS such as "autonomous" cars have already caused serious injury to humans. Even in this case, there is divergence between national jurisdictions [4].

Singapore is exploiting its unique geography and legal system to advance the use of Autonomous Vehicles (AVs) through road trials with close interaction between the government, regulators, and industry [5]. They expect to continue this collaboration as technology, public opinion, and law develop.

The United States government also see legal issues arising now and in the future. The US Department Of Transportation's latest autonomous vehicles guidance document [6] states that jurisdictional questions are likely to be raised by Automated-Driving-System (ADS) enabled vehicles which they need to address as a regulatory approach is developed.

The Chinese legal system may also need urgent revision to meet the needs of AVs [7].

The English and Scottish Law Commissions, on behalf of their governments, formally review important societal developments to provide a basis for new legislation. The final report [8] of their AV Project [9] concludes that the problem of assigning legal responsibility and hence liability for harm is unclear and, additionally, that this lack of clarity applies across all autonomous products. Their view is that using autonomy levels to describe a system is legally meaningless; an automated vehicle is either autonomous or it is not, with different laws applying in the two cases. AVs require a new regulatory authority, with responsibility and hence liability lying with the organizations responsible for the supply and maintenance of an automated driving system; in all other cases the driver is responsible. Data must also be recorded, stored, and provided for use in accident enquiries. Their recommendations directly affect all aspects of autonomous system design.

Analogous principles cover lethal autonomous weapon systems [10], so it can be assumed that most, if not all, A-HMT-Ss will provoke similar ones with responsibilities on all participants in the design cycle.

The legal views can be summarized in one system requirement which must be used in deriving more detailed system requirements:

> Responsibility for all decisions and actions of an A-HMT-S must be traceable by an enquiry to an identifiable person, or role-holder, in the organization using or supplying it.

The core problem with meeting this requirement for AI-based actions is their non-deterministic nature and consequent uncertainties in a system's behaviour. Considerable Research and Development (R&D) work will be needed to allow risk management of these legal issues in A-HMT-S lifecycles, as is the case with current safety-related systems. This paper identifies three key questions which are addressed, giving methods for acceptable risk management in meeting the requirement, and identifying the areas for R&D when AI is introduced into an interdependent A-HMT-S.

Assumptions and terminology

An A-HMT-S comprises at least one human and one or more CPS, with continual interaction between them, reallocating tasks as necessary. Only one human can be the Team Leader with responsibility for the actions of the A-HMT-S. Their interaction with the A-HMT-S is through the Human Machine Interface (HMI) which has an important place in an A-HMT-S as emphasised by [11, 12].

It is assumed that any A-HMT-S can cause unacceptable harm to a person or property if its behaviour is not controlled. This gives a requirement for trust which is defined as [13].

> The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.

Trustworthiness, the property required to be trusted, is defined as [14]:

> The demonstrable likelihood that the system performs according to designed behavior under any set of conditions as evidenced by characteristics including, but not limited to, safety, security, privacy, reliability and resilience.

Dynamic human machine teaming and trust

The simplest non-adaptive HMT has a human using automated, deterministic subsystems to meet their aims by delegating tasks to single or multiple subsystems. The human issues instructions, updating them based on either their responses, sensor information or a change in aims, i.e. all adaption is by the human. Safety is assured by a combination of testing and mechanical, electrical or software limits. When the subsystem responses are deterministic, human users have a trusted mental model of the system and will accept

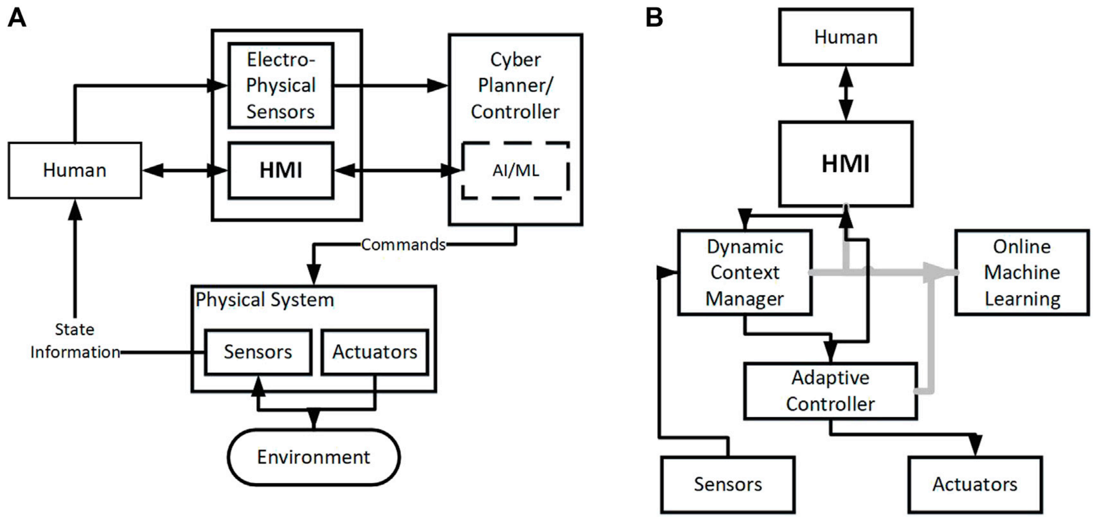

FIGURE 1
(A) An A-HMT-S with AI and ML embedded in the cyber planner/controller (Adapted from Madni & Madni 2018). **(B)** An A-HMT-S with its control elements drawing on results from separate on-line ML systems that are not in the control chain. (Adapted from Madni *et al.* 2018).

responsibility for the consequences of their instructions. It is assumed that the human is authorised to operate the system, indicating a level of trust in them by others, i.e., the HMT is trustworthy.

When resources and tasks need rapid, multiple reassignments, these must be automated and dynamic for optimum performance. "Optimum performance" will be system-dependent but must include ensuring that human workloads allow them to make considered decisions. The HMT is then an interdependent A-HMT-S, responding to external changes by internal reorganizations to meet its aims, with a human Team Leader.

Trust can be seen as a problem of ensuring that the Team Leader knows that the system is reliable, and well tested, with bounds to its action. It then follows that the trustworthiness of each element of the A-HMT-S is known. We take the view here that, in addition to the definition of trustworthy given in the Assumptions and Terminology section, human trust requires that the A-HMT-S responses can be understood and accepted as reasonable even if they are not necessarily the expected ones. The Team Leader is likely to develop trust in the CPS elements if they reliably achieve their aims but report back if there are problems.

The HMI provides the Team Leader with information about team status and task progress. The Team Leader will have a mental model of the A-HMT-S, with varying levels of detail and accuracy of its subsystems and resources, which provides expectations of system behaviour as conditions change. The control problem then becomes one of compatibility between the Team Leader's expectations and the information presented by the HMI. The HMI is taken to be a control station with a pre-determined range of user controls and displays that change depending on predetermined variables. The variables will include the user's workload and situational awareness as measured by the HMI, supplemented, if necessary, by other sensors. This implementation is an adaptive HMI as described by Blakeney [15]. Using the information provided by the HMI, the Team Leader decides if new system instructions are needed, checks that the system is trustworthy if changes are necessary, and issues the instructions through the HMI.

The subsystem implementing the instructions by making dynamic system control decisions has a crucial role. This role has been demonstrated for simulated environments using either a cyber planner/controller [11], or splitting it into a dynamic context manager and an adaptive controller [12] as shown in Figure 1. This shows that Machine Learning (ML) can be used in different places to support CPHS performance. However, ML is likely to introduce non-deterministic inputs into the A-HMT-S′ control system, giving the potential for instability. This makes it essential to identify the role ML plays in control decisions and which node has the authorisation to initiate the consequent actions. Unless this is known, the Team Leader cannot justify the system's decisions or be responsible for its consequent actions.

The preceding arguments show three issues when ML plays a role in decisions and actions in an A-HMT-S:

Issue 1. An adaptive HMI which learns and adapts its outputs, based on its own model of the Team Leader, must still present the essential information for the human to accept responsibility for the actions of the A-HMT-S;
Issue 2. Automation in the cyber planner/controller means that the Team Leader is not choosing the subsystems for a task at any given time. The introduction of ML into this choice will

lead to the system changing task and resource allocation according to circumstances as judged by a non-deterministic subsystem in the control chain;
Issue 3. ML in subsystems may change their behaviour due to its own understanding of circumstances, not necessarily that of the higher-level systems. The higher levels could request an action from a subsystem whose behaviour has changed and will respond in a manner not expected by a higher level. Both will then try to understand the situation and remedy it but, without close feedback, confusion is highly likely. This is an example of the wicked problem. a well-known one in systems engineering [16].

These issues must be addressed before a human can trust an A-HMT-S and accept responsibility for its actions.

Trust in any AI system depends on many characteristics. Alix et al. [17] give: reliability; robustness; resistance to attack; transparency; predictability; data security; and protection against incorrect use. They then propose that an AI-based system has to implement the three following features:

- Validity *to make sure that the AI-based system must do what it is supposed to do, all what it is supposed to do, and only what is supposed to do. It is crucial to deliver reliable, robust, safe and secure critical systems.*
- Explainability *to make the team leader confident with the AI based system through human-oriented and understandable causal justifications of the AI results. Indeed, the end-team leaders' trust cannot be neglected to adopt AI-based systems.*
- Accountability *in respect of ethical standards and of lawful and fair behaviours.*

These assume that an AI system will act on its decisions without human intervention, implying that the Team Leader is comfortable taking responsibility for all its actions, a very high threshold for trust. The threshold can be lowered if the system makes effective predictions about the consequences of its actions but if there is doubt about the effect of the action, or if it will exceed a predetermined limit, then the action and its justification is referred to the human Team Leader first. This behaviour is analogous to a member of an all-human team referring to the team leader for confirmation of an action or requesting an alternative course of action.

Summarizing, an A-HMT-S must be trustworthy by design and only take actions that are limited and authorised through its organizational structure, with reference to the Team Leader if necessary. The important questions that must be answered before an A-HMT-S can be trusted and used are:

Q1. Can a dynamic A-HMT-S with AI be designed so that the liability for the consequences of every action are clearly assigned to an identifiable human or organisation?
Q2. What guidance can be given to all stakeholders, including regulators, to ensure clear identification of responsibility for actions by the A-HMT-S?
Q3. How will the potentially liable individuals develop sufficient trust to carry out their work?

These questions must be resolved for a new design by setting requirements with possible design solutions. The resolution for an existing system will concern its actual performance and setting limits on its behaviour. An architectural approach is taken as it is a well-known methodology for both new and existing systems The architecture and the views used to describe it must be precise, internally consistent, and describe the system to the level of detail needed to answer the questions.

Architectures for an A-HMT-S

Architecture aim

Every A-HMT-S must have a consistent and coherent structure which can be described by an architecture which drives its design and upgrades by decomposition of high-level requirements into verifiable system and subsystem requirements and behaviours. Every examination of the system will use architecture views to describe the particular aspects required for a specific aim. The views are drawn up and analysed using standard engineering processes to achieve that aim.

The aim in this paper is to demonstrate that a dynamic A-HMT-S with AI, including ML, can be trustworthy; it must answer the questions at the end of Section 3 and meet the top-level requirement given in the *Introduction*. It follows that the architecture must separate decisions from actions and embed clear authorisation of actions before they are taken. It is assumed that the A-HMT-S will have to achieve its goals in environments with varying levels of complexity and associated uncertainties.

The architecture aims should be achieved by:

1. using a model of human cognition and action to describe all subsystems in the architecture;
2. having a clear line of control and action authorisation from the Team leader down to the lowest level subsystem;
3. enabling rapid referral up the control chain if a node does not have the authority to act
4. giving the Team leader visibility of the automated subsystems' options in making decisions if needed; and by
5. providing or establishing clear limits to actions which can be taken by every subsystem in the architecture.

The 4D/RCS architecture

The 4D/RCS Reference Model Architecture for Unmanned Vehicle Systems V 2.0 [18], is used here as it meets the five criteria set out in Section 4.1. It has been demonstrated with human levels of intelligence in its subsystems [19] and for

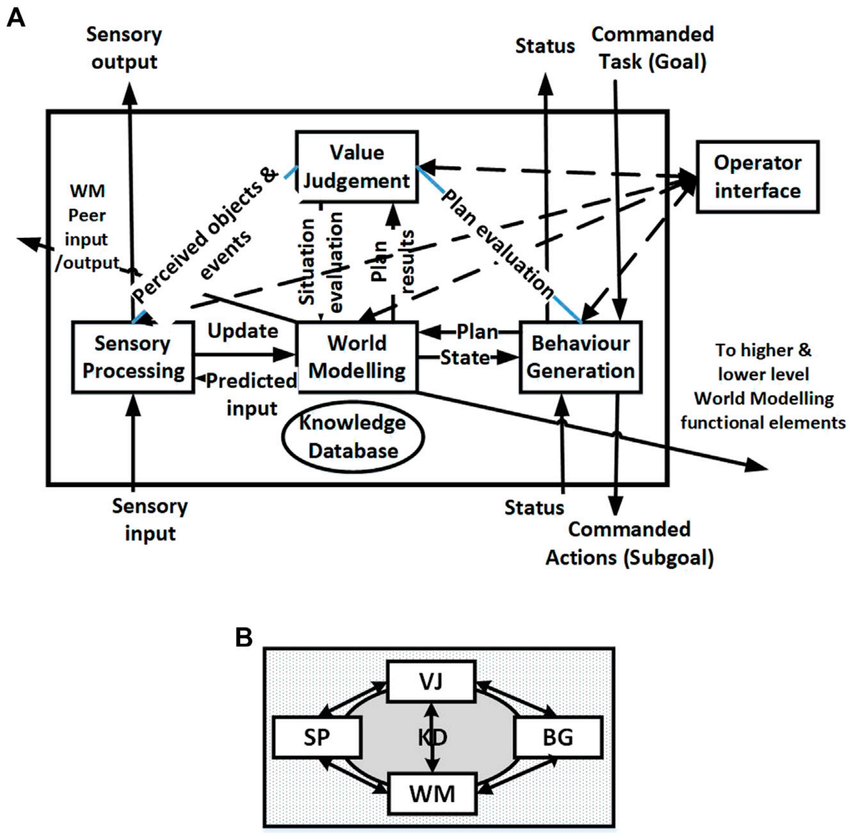

FIGURE 2
(A) A single 4D/RCS node, taken from NSTIR6910. **(B)** A schematic representation of a node used in later figures. Key: Value Judgement (VJ), Behaviour Generator (BG), Sensory Processing (SP), World Modelling (WM), Knowledge Database (KD).

identifying legal responsibilities in autonomous systems [20]. A full description of, and application as a hierarchical control structure for road vehicles is given in Ref. [21]. Other hierarchical architectures could be used provided they clearly identify where decisions are made and where the authorisation of these actions occurs.

The 4D/RCS architecture was devised for military command structures from high command to vehicle actuators. It defines responsibility for actions made by nodes which may be either human, machine or a mixture of the two. A node is defined as an organizational unit of a 4D/RCS system that processes sensory information, computes values, maintains a world model, generates predictions, formulates plans, and executes tasks.

Processes to apply the architecture are described in [22]. Descriptions of its use to identify legal responsibilities for the control of unmanned weapon systems and for autonomous cars is given elsewhere [23–25]. It is applied here to address the problems of trust and responsibility for the human Team leader by consideration of the three questions at the end of previous section.

Figure 2A is from the standard and shows a single node. Figure 2B is a schematic representation of its principle functions used later for simplicity. These functions are:

- the knowledge database which is the common repository for information for all nodes at that level;
- sensory processing which interprets sensor data and reports it to higher levels;
- a dynamic world model at every level with the resolution appropriate to that level. It is continually updated, based on information from the sensory processing function at that level. The distinguishing feature of 4D/RCS is that the world model makes predictions about the consequences of potential actions;
- the value judgement function assesses the predictions from the world model against the node's success criteria and ranks options for action; and

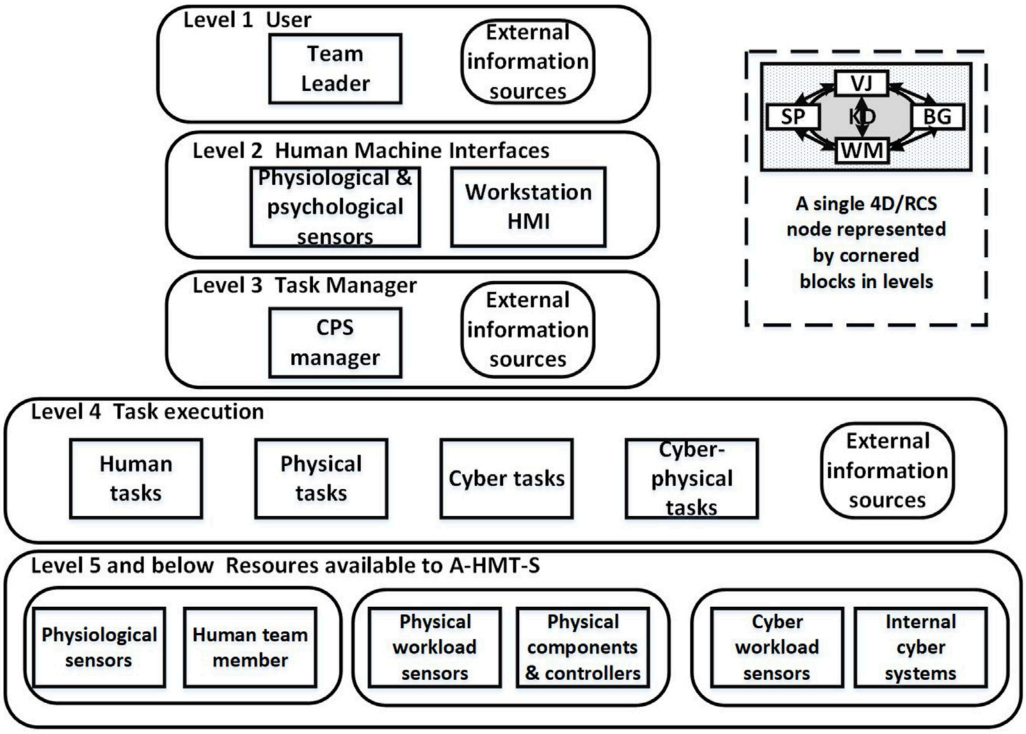
Level 1 User
Team Leader
External information sources
Level 2 Human Machine Interfaces
Physiological & psychological sensors
Workstation HMI
Level 3 Task Manager
CPS manager
External information sources
Level 4 Task execution
Human tasks
Physical tasks
Cyber tasks
Cyber-physical tasks
External information sources
Level 5 and below Resoures available to A-HMT-S
Physiological sensors
Human team member
Physical workload sensors
Physical components & controllers
Cyber workload sensors
Internal cyber systems
VJ
SP
KD
BG
WM
A single 4D/RCS node represented by cornered blocks in levels

FIGURE 3
An interdependent A-HMT-S structured as a 4D/RCS architecture with a 804 human as Team Leader. Acronyms in node are in the key for Figure 2.

- the behaviour generator takes the value judgement's outputs and acts, setting goals and success criteria for lower levels; if there is no safe action, the behaviour generator makes its part of the system execute a fail-safe mode, informing other nodes of its action.

The sensory processing, value judgement and behaviour generator functions form the three-part model of human decision-making and behaviours as described by Rasmussen [26] and the Observe, Orient, Decide, and Act (OODA) loop [27]. This model enables a common representation of both human and automated nodes. A node can have non-deterministic behaviour provided its actions are limited to its level of responsibility. Authority for decisions and responsibility for the consequences of their actions is determined through the commands and responses in the hierarchy of behaviour generators. The concept of authorised power has been introduced recently which sets the limits to a node's freedom of action. It is defined as [28]:

> The range of actions that a node is allowed to implement without referring to a superior node; no other actions being allowed.

This restriction allows hard limits to be set on a node's behaviour. Their sum for all nodes restricts the overall A-HMT-S′ behaviour, giving a basis for specifying trustworthiness in engineering terminology.

4D/RCS applied to an A-HMT-S

Figure 3 gives the broad characteristics of a 4D/RCS architecture for an A-HMT-S, with the user as Team Leader at Level 1 and the plurality of resources needed to complete the system's overall tasks at Level 5. For clarity, individual nodes are shown as blocks, each one representing a 4D/RCS node as shown in the dashed box in the figure. External information sources will be available at many levels, and indicated where appropriate.

Nodes at every level report to only one node in the next higher level, with clear responsibilities and limits to their actions based on their fixed position in the architecture. Sensory processing information is shared across levels in the hierarchy and can be passed up to the highest level. All information is shared between nodes at the same level as they have a common knowledge database.

Common response times, or other characteristics, across a level allows simplification of the data structure and world model at that level. They also enhance detection of differences between the real and expected world at any level, with a rapid escalation of the awareness of a problem. The time divisions for an A-HMT-S are not as straightforward as in the original concept for 4D/RCS. It should be possible to construct a set of timescales for any given application, recognising that there will be a range of timescales for completion of activities at any given level.

ML has been incorporated into 4D/RCS. Initially Aldus *et al.* put ML solely in the world models at different node levels, using it to assimilate data in many formats [29] but later incorporated ML in more node functions in the system [30] during the DARPA Learning Applied to Ground Robots (LAGR) programme; in particular the authors state that:

> The learning in each of the modules is not simply added on to the process that implements the module. It is embedded as part of the module, and operates in accordance with its location in the hierarchy.

An adaptive HMI at Level 2 with the sensors monitoring the Team Leader meets this criterion for an A-HMT-S. Elsewhere ML can be introduced into functions within any node in the hierarchy provided there are clear authorised powers set for each node.

The node hierarchy of 4D/RCS ensures precise specification of every node's individual role and hence its responsibility. It builds in a well-structured control chain that allows tasks to be transferred between nodes provided that this transfer is authorised by the next higher level in the hierarchy. The nodes at any given level are interdependent, but this interdependence is managed at the next higher level. The problem here is that task allocation is dynamic, based on node workload at any given time. Task reallocation can be rapid for the completely automated nodes, but the human nodes must be given enough time and information to make considered decisions. Assessing and quantifying human cognition times in a dynamic system will be a problem requiring a model of the Team Leader.

Inside the architecture

It is necessary to examine each level in Figure 3 in more detail to establish the feasibility of an A-HMT-S and the key problems requiring solutions. An A-HMT-S must assess available options for actions and their consequences by comparing plans with the current "real" world as reported by the sensory processing function. The world model at each level is a key part of this process, with its role brought out in the following sections.

It is likely that the problems highlighted by the analysis here will be common to all architecture frameworks, so they could become potential research topics if not already developed.

Level 1, the human team leader

The Team Leader will have direct access to the functions in the HMI and indirectly to lower-level functions through the behaviour generator chain. Team-Leader visibility of all parts of the system is made available through the HMI sensory processing module.

Level 1 functions are specified to ensure the owner's business priorities are met, monitoring the current team status, predicting future events, and resolving conflicts. Although a team member, the Team Leader's role must be the highest hierarchical level, instructing lower levels. Instructions are given as team goals and success criteria, with priority weightings for the A-HMT-S to interpret. The Team Leader must also trust the CPS to flag up all those problems requiring their attention through the HMI.

It is essential that the Team Leader's workload is manageable so there is time to understand the options considered by lower levels and the issues they cannot resolve. It is assumed that the Team Leader's workload can be monitored at Level 2 supplemented by other sensors if necessary. Potential overloads will be presented to the Team Leader with Level 2's recommendations for their removal. The Team Leader will then decide what new instructions must be issued.

A smaller A-HMT-S may have the Team Leader also carrying out some Level 4 functions in parallel with Level 1 functions. This structure does not fit in an ideal hierarchical architecture and would need detailed attention in system design. Potential solutions might include applying a temporary surrogate chain of command at Levels 1 or 4 whilst the Team Leader concentrates on the higher priority functions, or delaying the Level 4 task and letting the low-level consequences be managed automatically.

Level 2, the HMI

The interaction between the Team Leader and the CPHS will be through the HMI at Level 2. It is put in Figure 3 as a specific function, following Madni & Madni and Madni *et al.* [11, 12] as it plays a key role in any human-machine system. The Team Leader will probably have access to other information sources such as phones, direct visual checks and independent access to the internet.

The HMI's first role is to translates Team-Leader-defined aims or changes into goals for the system with priorities and other necessary information. The information is passed through the behaviour generator chain to Level 3. The Team Leader must have both cognition of the A-HMT-S task status and the detail required to issue effective instructions. Although this is a normal human factors problem, it does not help solve the problem of translating human-language queries or goal changes into team instructions in the machine language used at Level 3.

The HMI's second role is the separation of functions between the Team Leader and the dynamic task manager so that the Team Leader does not become overloaded by involvment in actions which can be handled automatically. Part of this role is to monitor the Team Leader's own workload through indicators such as response times and other indicators of their cognitive and physical state. If the workload is excessive, the HMI must present the Team Leader with options for reducing it. With ML, the A-HMT-S can learn an individual Team Leader's behaviour and overload signatures, but it must recognize individuals and variations in their performance.

The HMI can only direct changes at lower levels, through the behaviour generator chain, if it sees a problem and has the authority to implement a solution. Specific system designs will need to address which actions it can take, and how the reasoning is presented to the Team Leader when action is taken.

The HMI's third role is to ensure that the Team Leader is presented with clear statements of problems which it cannot resolve, backed up with relevant information and options considered for action. It checks that the current and predicted operations are being managed correctly at Level 3, flagging actual and potential problems to Level 1. Problems can be identified by both the HMI and Level 3. The Team Leader may then wish to access further data from Level 3 and add new information into the HMI knowledge database to increase the range of options. It may be necessary for some of this information to be passed to lower levels for more detailed analysis, but kept separate from measured sensor data.

These three roles are fixed, so the HMI is not one of the nodes that can have its tasks changed by the dynamic task manager. However, if it determines that the Team Leader's workload or situation awareness is likely to be outside a safe and efficient level it will inform both the Team Leader and the dynamic task manager. The dynamic task manager can make suggestions to the Team Leader through the HMI but not act on them. The Team Leader can change his or her tasks and workload through the HMI's behaviour generator chain.

The HMI world model requires a model of human capabilities, the human's state and warning signs of overload based on available sensor mechanisms. The model may be supplemented with information about individual Team Leaders if this is permitted. The use of the three-part model of cognition in 4D/RCS will facilitate this interface.

This HMI world model will require all the information from the Level 3 model, and set it in the wider context of external factors acting on the A-HMT-S. The wider factors included at Level 3 will have been filtered for reasons given in the next section; the HMI can use the Level 3 internet access to overwrite its constraints whilst deriving its own options and selections for presentation to the Team Leader. The Team Leader will also have this option through the sensory processing chain for their own mental model.

Level 3, dynamic task manager

Level 3, the dynamic task manager, has only one node, the CPS manager and its external information sources. The external sources may include the internet, but this must be well controlled. The use of external AI engines to search for and select information may not be reliable so, as a minimum, information will need to be tagged with its source and an estimate of reliability. Unquestioning acceptance and use of external search engine results will expose the Team Leader to unacceptable risk as a court may decide later that the information was clearly unsuitable for the A-HMT-S's use.

The CPS manager's role is to provide efficient use of resources at Level 4 and below. It specifies the tasks required to meet the goals and priorities from Level 2, their success criteria and other instructions, then issuing them to Level 4 through the behaviour generator chain. It draws on timely information about task status from Level 4 and allowed external information sources; these form its sensory processing functions. Decisions to assign and reassign resources are taken by its behaviour generator either autonomously or after referral to Level 2 and possibly Level 1. Level 3 is the lowest level at which there is an overview of all tasks.

The Level 3 world model includes: all current tasks and their status; available resources; and their allocation to tasks, both current and future. It will not have all the detailed task information in the Level 4 world model. The Level 3 world model will include the wider activities which do not form part of a task but do affect them. Examples are maintenance and staff holidays.

Comparison of Level 3's sensory processing function output, workload plans and task success criteria will identify potential problem areas for action by Level 2 if it cannot resolve them itself. The system architecture must mandate whether all changes at Level 4 are dictated by Level 3 or if Level 4 nodes are allowed to negotiate due transfer of resources or parts of tasks between themselves at a local level. This transfer could be advantageous as it removes work from Level 3 but could create problems if the Level 3 world model is not aware that these changes have been made. The use of surrogate chains of command may provide a solution to these problems.

Level 4, individual task management

Individual tasks are managed at Level 4 by drawing on the human, physical and cyber resources at Level 5 and below which have been allocated to the task by Level 3. The names for the Level 4 nodes in Figure 3 simply reflect the types of task required, and do not imply a separation of task types based on their required resources. It is unlikely that a human will manage tasks at Level 4, although there may be parts of many tasks which require human resources at Level 5.

The world model and knowledge database common to all Level 4 nodes include resources and their availability for each task as a function of time. Time resolution and resource detail will be lower than that required at Level 5. The world model predicts the effects of changes due to instructions from above or responses from lower levels. Task-related problems will become known at Level 4, giving it the ability to solve many of them. However, each node must have clear limits on its authority to authorize actions with consequences outside its own task. The Level 4 nodes must have the ability to flag problems for attention by higher levels. An example might be when two tasks require the same resource at the same time in the future which Level 3 could resolve by changing the time criteria on one task or by redeploying resources across several tasks.

Level 5, resources

Functions below Level 4 are not considered in any detail here as their structure depends on the specificA-HMT-S, recognising this treatment as a necessary simplification. However the 4D/RCS architecture and the structure of Level 4 do set some constraints on Level 5 nodes and the functions they perform.

The complexity of the nodes at this level will depend on the A-HMT-S under consideration. A node may include all the dedicated resources for one task which are always allocated to that node, the resources being used for other tasks only when that task is not needed. On the other hand, nodes may be subsets of physical or computing resources suitable for a range of tasks; their allocation at any time being under the control of Level 4 task managers. It is unlikely that nodes at this level will be able to negotiate reallocations between them.

Any given Level 5 node may be a complex system in its own right; for example it could be an electro-mechanical system embodying complex adaptive control systems using advanced methods [31]. These systems could easily include AI-based techniques provided their freedom of action is limited by a suitably framed authorised power covering cyber and physical outputs.

There are workload monitors for every resource at Level 5. These may be discrete components such as thermometers for motor drives, or they may be a part of a resource's software. Combinations of individual resource sensors may need to be reconfigured when resources are reallocated to determine the workload being used for current tasks.

The Level 5 world model will be centred on resources and their current and future allocation to tasks on the shortest timescales. It will be based on the structures below Level 5 and their requirements as the tasks evolve. However, it will be visible to higher levels through the sensory processing chain which enables the Team Leader to request information about every resource in the system. The higher levels may consider that changes in resource allocation or task parameters are necessary at Level 5 or below, but they can only make these changes using the behaviour generator chain which will identify the consequences of such requests and then report back.

Decision making process and action authorisation in a node

Each node in Figure 3 fits in the 4D/RCS hierarchy as shown in Figure 4. (For clarity, lower nodes are only shown for the middle node). Every node's aim is to execute its task whilst managing workloads for the resources under its control. It is given tasks and success criteria from its superior node; these are interpreted, and subordinate nodes are given their tasks and success criteria through its behaviour generator. The knowledge database is shared across its level. Every node's actions are constrained by node-specific authorised powers.

Figure 5 shows the information flows inside the node. The four principal node functions are indicated by the shaded areas. For simplicity, it is assumed here that the A-HMT-S is already executing a task and that the new instructions will change its plans. Instructions are aims for the revised task and, if necessary, revised success criteria to assess task completion. The node checks that the task is within its authorised power and then derives one or more workload plans for comparison with the current world.

The current world model covers the timeframe relevant to this level in the hierarchy and is derived from the sensory processing function. Predictions are made for workloads and compared with the available resources to give the N task consequences shown in Figure 5. It is assumed that the node has some freedom in planning its own and its subservient nodes' instructions and that there will be a range of success criteria for different parts of the task. A number of plan options M, which will be less than or equal to N are assessed in the value judgement function and ranked according to criteria set by either the higher node or from its knowledge database. A check is made in the behaviour generator that the node is authorised to implement the chosen plan. If it is, the plan is accepted, if not, another option is chosen. If none are allowed, a fail-safe plan is implemented and the superior node informed. Authorisation of action is still within the node and its own task.

The node's authority will, among other factors, allow it to use resources that are not assigned to other nodes for the period required for an acceptable option. If it does, the change is accepted as a new task, instructions are sent to lower levels as revised success criteria, and the revised plan is incorporated into the knowledge database for that level. The other nodes at that level will compare the revised plan with their plans; should there be a conflict due to their own replanning, then the nodes will cooperate to resolve them with the results passed through the behaviour generator chain to the next higher level. If the problems cannot be resolved, for instance if one node's authorised power will not allow it to act, then the next higher node is informed through the behaviour generator chain. Revised instructions, generated as success criteria, will be created at that level by the same process and the lower nodes will respond accordingly.

The decision-making process described above is generic with differences in the information used at any point in the process at different levels. Table 1 describes the type of information at key points in Figure 5 when applied to Levels 2, 3 and 4 in Figure 3.

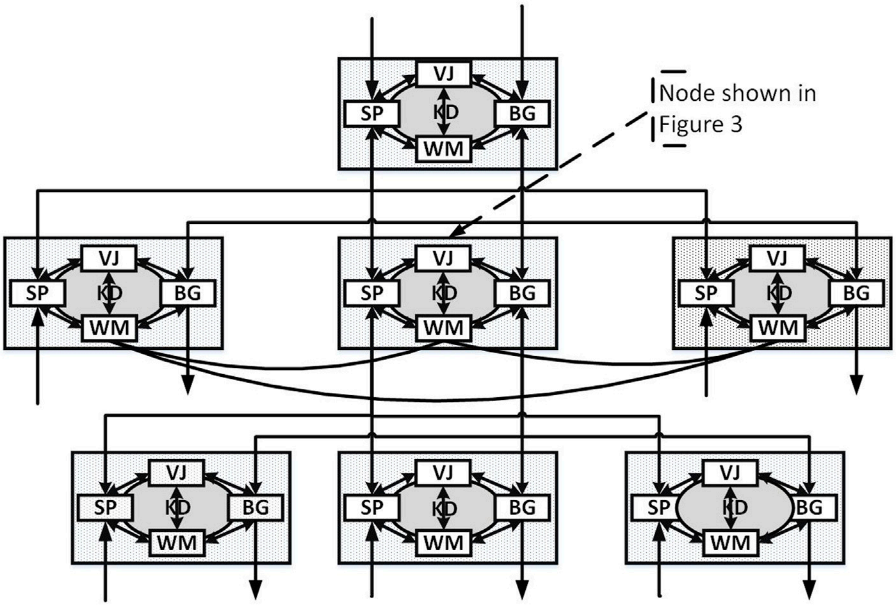

FIGURE 4
Showing the connections between nodes in each position in the hierarchy in Figure 3A 4D/RCS node and connected nodes. Key: Value Judgement (VJ), Behaviour Generator (BG), Sensory Processing (SP), World Modelling (WM), Knowledge Database (KD).

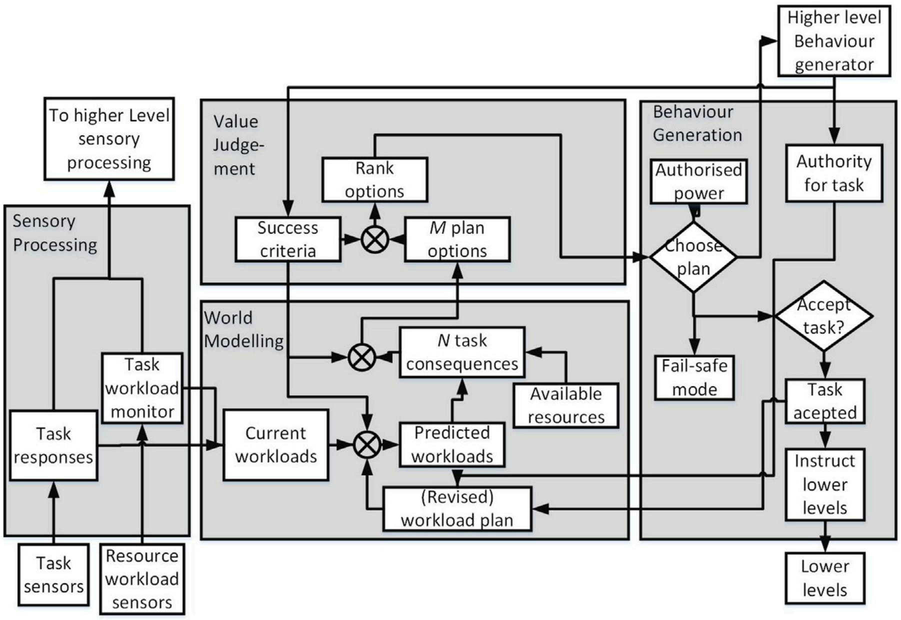

FIGURE 5
Information flows within and between the functions in a node. Each function is a shaded box. Information processing is in the white boxes and comparisons in the circles.

TABLE 1 Description of information used at key points in Figure 5 for different architecture levels.

Term	Level 2	Level 3	Level 4
Inputs from higher level behaviour generator	• A-HMT-S tasking from Team Leader	• New or revised tasks and success criteria	• New or revised task • success criteria • resource allocation
Success criteria	• Business priorities for tasks or groups of tasks	• Match of all proposed options against business criteria	• Cost, time and quality for task
Outputs from Sensory Processing to World Modelling	• Match of future activities and plans against Team Leader's criteria • Human activity and stress level	• Match of task progress and resource allocation against Team Leader's success criteria	• Progress reports on task progress • Workload on resources currently or planned to be used by node
Output from Current Workloads	• Need for extra or fewer resources	• Workload across all resources	• Workload for one node's task
Plan options	• Look for and secure external resources if possible • Present options to Team Leader	• Reassignment of resources across tasks • Slips in progress allowed for lower priority tasks if overall success criteria are met.	• Changes to current resource plans • Slip in task completion deadline
World model horizon	• Current and predicted operations under current plans • External world as it affects current operations	• Resource use across all tasks plus likely new ones • Overall costs	• Detailed task plans with current and predicted progress • Options to reduce costs in individual tasks
Default authorised power	• Limited ability to draw on external resources • Cannot exceed fixed criteria when considering options	• Can reallocate resources across tasks at Level 4 • Can only instruct restricted set of available resources	• Only use previously assigned resources • All systems to follow safety protocols

ML, decisions, and actions in a node

Non-deterministic processes must have strictly limited behaviours when included in a control system requiring any level of safety. AI has been included in a 4D/RCS architecture in [19] and applied to autonomous ground vehicles traversing rough terrain. In these applications AI is used mainly for interpreting sensor data to recognise obstacles in building up a representative world model, although Albus and Barbera [32] propose using AI to adjust parameters in the equations used to decompose goals into tasks and further decomposition.

Human trust has been incorporated into the A-HMT-S by using a dynamic world model that replicates a human mental model and having strict limits on each node's actions. However, although necessary, these are not sufficient. The human may not be able to easily understand why a learning algorithm made a decision, but they can accept it if they think that it is reasonable; i.e., the human perceives the decision as sensible and that fits with their own mental model of the problem. The A-HMT-S must be able to present relevant information about the options considered when choosing an action so that the human can understand and assess the choice.

Restricting every AI-algorithm's operating domain to be within a node limits its effect. The aim of every node is to complete its task by meeting its success criteria, solving problems within the limits of the authority it has in the architecture. The learning system will decide its best option for solving the node's problem by using information available at that level in the heirarchy. This solution can then be considered as one option of the *N* options generated in the world model. It will then be assessed with those not generated by the learning algorithm in the value judgement module, and the result passed to the behaviour generator. Whichever option is chosen, the behaviour generator checks if consequent actions are within its authority, ensuring that the Team Leader and any regulatory authority know that actions arising from the learning algorithm cannot exceed predetermined safe limits.

The sensory processing chain can pass information directly from any level to all higher levels. The Team Leader can interrogate the data and other information used by any node in its decision making, giving the potential safeguard of human-initiated enquiries. However it will be impossible for the Team Leader to query every decision before its consequent action, so triggers should be included in the value judgement and behaviour generator modules to initiate a Team Leader enquiry into the decisions leading to defined types of critical action before their execution.

Examples of learning systems which can improve efficiency in the A-HMT-S include:

- tasking workload monitors to pick up warning signs of overloads due to variations and tolerances on workload data, timing *etc.*;
- adding to task consequences based on historic data;
- assessing changes in the business environment that may alter task success criteria;
- ranking of options, but rank order may need confirmation by the next higher level before acting;
- using AI to identify potential problems at the level below the node it is in; and
- monitoring external conditions to give early warnings e.g. an approaching snowstorm probably changing delivery times for materials.

Some nodes at Level 4 may be people acting in accordance with their task and management instructions. They will not be fully autonomous as their authorised power will be set by their organization's management processes. Their workload sensors may be more subjective than for other parts of the A-HMT-S and so will need explicit inclusion in the sensory processing chain. They will almost certainly use network support tools for their work, and their use may provide a suitable mechanism to monitor their workloads.

Example architecture applications

The architecture presented in Figures 3–5 is generic in nature. It needs to be applied to some sample scenarios both to check their practical validity and to identify more precisely the topics that warrant further R&D effort. One vignette is taken from each of the three classes of HMT used at a recent conference on human-machine teaming [33].

Recommendations to a human for their immediate action

Automated identification of targets to a pilot who is about to release a weapon is a well-known A-HMT-S problem and the subject of international debate [10]. The 4D/RCS architecture has been applied to it in Chapter 12 of [20] with architectures similar to Figure 3 shown in Figures 12.8 and 12.9 in that reference and Figure 5 similar to the one shown in Figure 13.4. The architecture took an incremental evolution of current systems by replacing human nodes with automated ones whilst maintaining the necessary response speed for human assessment of action and the consequent changes in military tactics and rules of engagement.

[25] shows how legal responsibilities for the driver and vehicle can be derived for autonomous vehicles at all autonomy levels.

The consequent changes to responsibilities in the design chain for military and civilian products are discussed in Ref. [34]. It is shown there that a hierarchical architecture is essential for the design of an autonomous system so that safety-related decisions can be identified with the legal responsibility for the system's actions assigned to individual organizations and role-holders. The principal issues are link-integrity to ensure continuous control of the weapon, and reliable identification of both targets, non-targets, and the civilian objects which should not be attacked. Similar issues will apply for vehicles.

Carebots

We take the case of a robot caring for an elderly person in their own home which has one floor. The carebot is leased from a health care provider who are responsible for its maintenance and updates. Figure 6 gives the broad characteristics of a carebot HMT architecture equivalent to Figure 3.

The Team Leader is the elderly person giving instructions to the CPS part of the team. Mutual trust and interdependence is critical. The CPS can provide facilities or resources such as medication but cannot force the person to take them as this legally is assault; similarly, the elderly person may be critically dependent on the CPS for provision of medication and their regular supply. The person will have normal interaction with other people and resources using the non-carebot resources that they are capable of using; these may be restricted but could be extensive for a mentally agile but physically infirm person.

The HMI at Level 2 will be safety-related as a minimum standard if it provides calls to emergency services on behalf of the Team Leader. This places high demands at Level 2, making an adaptive HMI essential with a sophisticated model of the Team Leader and voice recognition for a range of human emotions. The adaptive HMI will be very different from that assumed in earlier sections with considerable scope for AI-based development here. There is only one human to model, and scope to incorporate intelligent analysis of physiological sensors looking for precursors of serious medical conditions. Actions will be requested from the Team Leader and passed, as necessary, to Level 3 to alert necessary medical or social services or relatives. This may raise the software standard to safety-critical with

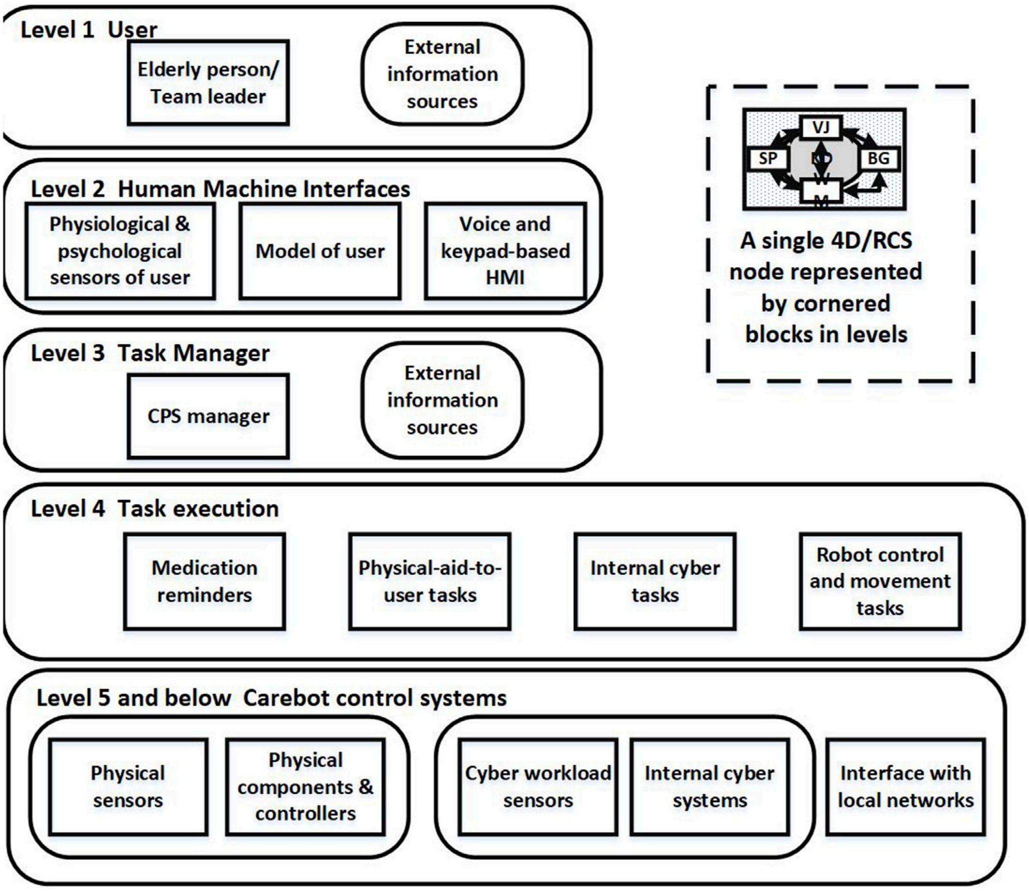
Level 1 User
Elderly person/ Team leader
External information sources
Level 2 Human Machine Interfaces
Physiological & psychological sensors of user
Model of user
Voice and keypad-based HMI
Level 3 Task Manager
CPS manager
External information sources
Level 4 Task execution
Medication reminders
Physical-aid-to-user tasks
Internal cyber tasks
Robot control and movement tasks
Level 5 and below Carebot control systems
Physical sensors
Physical components & controllers
Cyber workload sensors
Internal cyber systems
Interface with local networks
VJ
SP
KD
BG
WM
A single 4D/RCS node represented by cornered blocks in levels

FIGURE 6
Carebot as an A-HMT-S with the elderly person as the Team Leader.

associated standard and regulatory requirements, a very high standard for an evolving model of an individual.

The dynamic task manager at Level 3 will perform approximately the same as those described for Level 3 in the Inside the Architecture section, but there will almost certainly be mandated external interfaces for medical and emergency services. Medical records, medication and related data will need to be in the knowledge database; their location at Level 2 or 3 or a split between these levels will be a design decision, as will the method of updating them. The task to alert external organizations must decide the type of aid sought and be able to communicate with them effectively. The decision will be based on a comparison of the person's current status compared with their expected status, the level and type of difference, and the confidentiality of information in its database. There will probably be a need for a medical professional to talk to or visit the Team Leader so arrangements may need to be made for this. This interface represents a large R&D challenge.

The carebot will need to continuously monitor the Team Leader's well-being through signatures such as movement and heart beat as well as external environmental conditions such as a sudden cold snap or thunderstorm which may necessitate precautionary measures in the house or changes to the Team Leader's diet, for example, by offering more hot drinks.

The CPS aspects of controlling, maintaining, and upgrading the functions and resources at Level 5 will be similar to any other A-HMT-S system. The main difference will be the notifications and revised instructions given to the Team Leader in a way that they are understand, possibly with prior warning and a familiarisation session before installation, based on the Team Leader's specific needs.

A system which operates alone for long periods then reforms as an A-HMT-S

An example of this type of system is a robotic planetary explorer that is visited periodically by humans who rely on it for support while they are on a planet. Levels 4 and 5 will be similar to most robotic applications, but the higher levels will have major

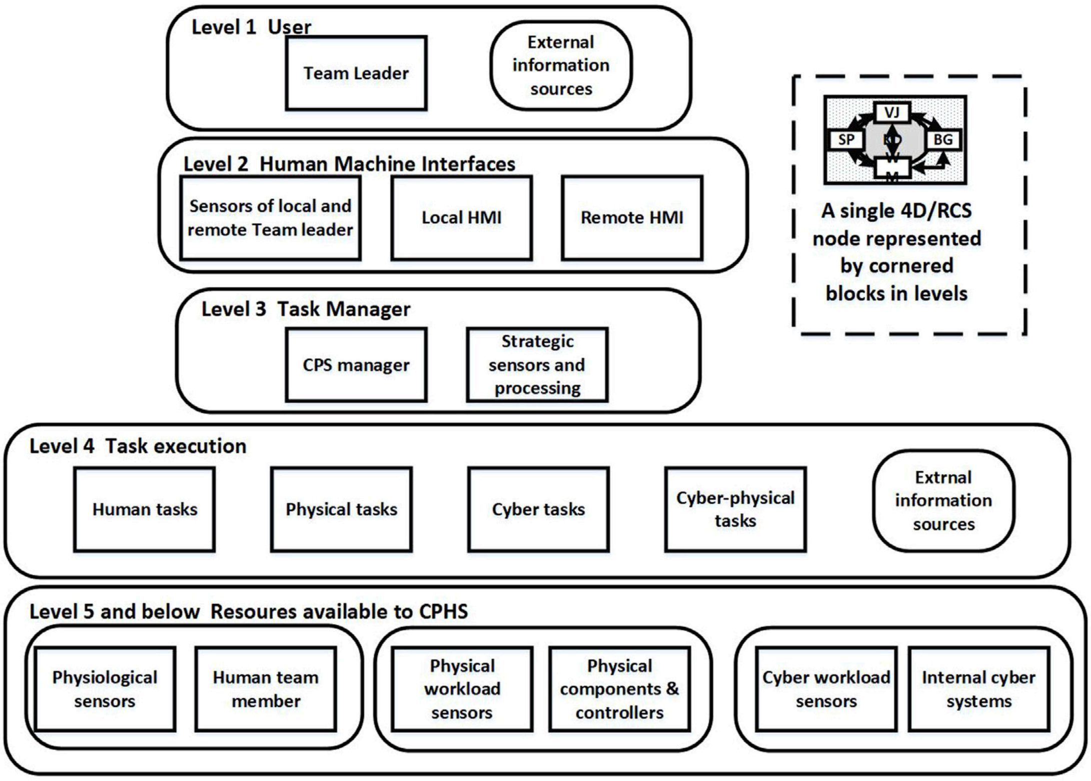
Level 1 User
Team Leader
External information sources
Level 2 Human Machine Interfaces
Sensors of local and remote Team leader
Local HMI
Remote HMI
Level 3 Task Manager
CPS manager
Strategic sensors and processing
Level 4 Task execution
Human tasks
Physical tasks
Cyber tasks
Cyber-physical tasks
Extrnal information sources
Level 5 and below Resoures available to CPHS
Physiological sensors
Human team member
Physical workload sensors
Physical components & controllers
Cyber workload sensors
Internal cyber systems
VJ
SP
KD
BG
WM
A single 4D/RCS node represented by cornered blocks in levels

FIGURE 7
A planetary rover as an A-HMT-S showing the two Team Leaders.

architectural problems. There will be two types of human interaction: remote monitoring, with instructions and updates sent from Earth or a relay satellite, and local interactions by visiting astronauts using an embedded HMI. These are shown in Figure 7.

Level 1 is shown with one Team Leader as it assumed that there will be protocols to prevent a remote person sending instructions when the rover has a local Team Leader. The Level 2 sensors will be for data transfer using remote links, checking for errors and missing messages, and will operate at all times. It is assumed that the astronauts' state of health will be monitored by other means such as their personal life-support equipment, reducing the HMI requirements considerably from the general case for an A-HMT-S.

Extensive fault detection systems will be necessary due to long periods without human attention in a hazardous environment with, for example, high radiation levels increasing the chance of semiconductor failure in the narrow-track high-frequency processors needed for advanced AI systems. Contingency reconfiguration of functions and tasks will need to be chosen based on probably incomplete diagnosis of apparently random failures and clear symptomatic information passed to the Team Leader, another area for R&D.

The strategic sensors at Level 3 will monitor local planetary conditions and provide assurance that software updates are not only received and installed, but will also run the required performance tests, sending the results back to Earth and to the local astronauts before and after their arrival for checking. This is to ensure the vital mutual trust between Team Leader and machine when restarting an interdependent relationship. The information will be held in the Level 3 database and its world model compared with the Levels 3 and 4 sensor processing outputs. The CPS manager will play a similar role to that in all the other A-HMT-S.

Discussion

Trust for an A-HMT-S

Three important questions were posed in at the end of the third section:

Q1. Can a dynamic A-HMT-S with AI be designed so that the liability for the consequences of every action are clearly assigned to an identifiable human or organisation?

Q2. What guidance can be given to all stakeholders, including regulators, to ensure clear identification of responsibility for actions by the A-HMT-S?
Q3. How will the potentially liable individuals develop sufficient trust to carry out their work?

Questions 1 and 2 can be answered by the use of a hierarchical architecture. It can be used to identify important and critical issues for each stakeholder based on the following points:

- The architecture can and must give clear separation of human nodes and automated ones in the hierarchy. This separation ensures that liabilities can be clearly assigned to the Team Leader or the organization responsible for the design or upkeep f the automated node. The architecture in this paper has taken Level 1 to be exclusively human, interacting through an adaptive HMI at Level 2. Other humans will be at Level 4, participating as Team Leaders of tasks, again with separation by levels within the tasks. These Level 4 humans receive their tasking from Level 3 and are monitored as part of the overall A-HMT-S.
- The architecture should separate decisions from actions with an assessment of the reliability of the decision. The Level 3 dynamic task manager is automated and should refer uncertain actions to the Team Leader through the Level 2 HMI when problems cannot be resolved within its authority level. Action is based on the choice of an option from those arising from the comparison of the world model with physical reality, a difficult task for a complex environment. The decision to refer to a higher level is critical as the false alarm rate must be low in order to maintain trust. This decision will require intelligent AI analysis based on mainly uncertain data.
- A bounded system, such as a distribution network or airport where tasks and progress can be readily quantified, will make the comparison of the real world and its world model easier than with subjective information. Additionally, the range of actions and their authorisation node can be defined uniquely. Developing such an A-HMT-S with humans at several levels would give opportunities for R&D progress in developing Level 3 CPS techniques for both complex (Level 3) and simpler (Level 4 or 5) scenarios.
- The use of predictive models in 4D/RCS and the information flow model used here ensures that the consequences of an action are assessed and authorised before it happens. Auditable authorisation of actions by the system enables consequent identification of responsibility for the consequences of every action. The choice between automated or human authorisation becomes a part of the design process as it is recognised that ultimately any human authorisation of an action must be legal and follow local and national requirements.
- The use of authorised power as part of the behaviour generator in every node ensures that no unauthorised actions can be carried out without reference to a higher node and ultimately the human Team Leader. This does place the onus for safety on the person who specifies what must be raised to the next architectural level. However, when the specification is for a function within one node with defined authority, its implementation becomes a tractable problem which can be addressed by CPS designers. They will also require clear directions about local changes, regulations, and processes for system upgrades. Every A-HMT-S will be designed, or tailored, for specific applications so explicit considerations of authority levels and the allowed options for action at each node should give answers to questions 1 and 2 above.

The third question should be answered by the following points

- Limiting the behaviour of every node by setting and applying limits to actions based on a comparison of the real world and predicted consequences of a range of actions leads to it being trustworthy for defined conditions. Defining the conditions becomes a design and procedural issue which can be addressed by current engineering processes.
- Careful specification of the adaptive HMI so that it presents clear information about problems, whilst allowing the Team Leader to see the options and consequences that the lower-level nodes considered. This transparency should allow trust to develop. If it does not, the Team Leader can alter the authorised power of specific nodes so that actions that appear untrustworthy will be highlighted for further human action.

It is possible to set up a trustworthy A-HMT-S that satisfies the three critical questions and has little or no AI in it for specific applications. In these cases the A-HMT-S would have limited flexibility because most of its decisions would be made using deterministic processes with well-understood uncertainties. It could be argued that these are not teams but are adaptive control systems that change their behaviour in defined ways, triggered by pre-determined thresholds. AI is needed to achieve flexibility, autonomy and interpretation of uncertain inputs.

Trust-specific R&D

It was noted earlier in this paper that the authors of [30] found that learning processes must be embedded in nodes and not across them. That work was for one specific system and mainly concerned the sensory processing chain. A more general approach is to consider a node’s functions in detail. Figure 5 gives more detail than the NSTIR standard, allowing an examination of the processes to identify which will benefit from AI and the type of R&D work that is needed. The following sections highlight the important areas for A-HMT-Ss without giving a review of current research, which is beyond the scope of this paper. There will be uncertainties at all points in the architecture, arising from many sources, making AI-based solutions attractive, but they must not detract from the CPS's trustworthiness or the HMT becomes untrustworthy. This places large requirements on all AI processes in the A-HMT-S and hence on the R&D work for every application.

Sensory processing chain

Levels 2 to 4 are mainly concerned with creating task instructions from human-defined aims, workload issues in the A-HMT-S, and selecting problems which require human cognition and authority to resolve. Level 3 has restricted access to outside networks so the access limits can be tailored to the ability of the CPS manager to interpret the information.

Level 5 and below will have the application-specific sensors for the outside world such as imagers and collision warning systems. Many of these already have some AI and some will be safety-critical.

At Level 4, the sensory processor outputs from Level 5 are interpreted by the task manager as progress on the tasks required for the A-HMT-S to achieve its aim. An accurate interpretation will be impossible if the world models at Levels 4 and 5 are incompatible or in conflict. Comparison will be difficult as they have different levels of detail with different time horizons, so checking their consistency may be a better approach, carried out in the "current workloads" process in the world modelling function in Figure 5.

World models, world modelling and value judgement

All nodes at a given level have a common world model in their knowledge database. The world modelling function uses it for multiple comparisons and predictions. The results are assessed by a node against its success criteria, ready for decisions and action. The success criteria may include non-interference with higher priority tasks. The world models will need regular and intermittent updates for two reasons: real-time changes in the environment; and the detection of incompatibilities between world models at different architectural levels. All world models must be under configuration control, with a process for updates and the knock-on effects in other nodes and levels. Authorisation of a change to a model can only come from the next higher level as that has an overview of all the lower level's nodes and, with AI, will develop a model of each node's behaviour.

World models at all levels must be consistent, even though they have different time horizons. The model at any level must include the available resources, their current allocation into the future, and the authority vested in lower nodes to change an allocation. Figure 5 illustrates that each node will create its own set of $M \leq N$ predicted world models based on its interpretation of its workload plan and its success criteria. The N predictions could be based on multiple simulations representing the uncertainty range in the world model at that level. Alternatively, it may be straightforward to introduce AI into nodes performing well-bounded tasks and then to generate one preferred option. Each option will affect resource usage and the environment at different times due to their interdependencies, so each option's affects must be assessed before implementation of any action.

Behaviour generator

The behaviour generator function in each node has limits on its actions set by the system design. These may be temporarily changed by the next higher level if that level's predictions allow it. When the task is within a node's authority, a chosen option is created and offered to the behaviour generator by the value judgement function. This choice includes the plans and tasks for lower levels.

The final check before action is taken is to compare the chosen option with the node's authorised power. This includes not only what the node can do, but also what it cannot do. Prohibitions may come from higher levels, including higher priority levels of other tasks on available resources, and effects on the wider world. At the highest A-HMT-S levels (Level 3 and above) this will include the societal issues such as interpretation of laws and regulations. An example for the carebot is a lower level offering of an approved medication, the Team Leader refusing, which is their legal right, and Level 2 issuing instructions to re-offer in 5 min; several refusals could trigger an alert as an external human medical judgement would be needed, the fail-safe mode. The behaviour generator could include comparison tools developed using AI techniques, utilizing the power they bring to the assessment function. However, they must be thoroughly tested to ensure they do not evolve after installation to ensure that they have deterministic behaviour.

The check against a node's authorised power is effectively asking if the consequences of choosing the offered option are reasonable. If they are, the option is chosen and action taken. If not, and no other option is acceptable, the task is rejected, the higher node's behaviour generator informed and the higher-level node must reconsider its options. If no choice is acceptable to the higher node then the problem is escalated, eventually to the Team Leader for human assessment. This guards against the build-up of errors or large uncertainties producing an unexpected and unreasonable action which must requires human assessment. The Team Leader has access to information at all levels in the 4D/RCS architecture, enabling them to make a more-targeted assessment of the problem and potential solutions than the unaided CPS can make.

The definition of reasonable is crucial as it is a societal and legal term, not an engineering one. At lower levels limits can be set by their design as clear technical bounds can be set for most tasks, based on avoiding interference with other higher-priority tasks and preventing physical harm. At higher levels the limits become softer making an AI-based approach attractive, but this jeopardises its role as a safeguard because of the non-deterministic nature of AI. The interpretation of the softer issues and translating them into the engineering terminology of deterministic limits will require iterations between lawyers, social scientists and engineers. It is possible that eventually robust ML algorithms, their training data, and their automated reasoning approaches may develop to the stage of meeting legal challenges but this is unlikely at the current state of technology.

Conclusion

The use of a hierarchical architecture improves the effectiveness of A-HMT-S design and development. The analysis presented here gives approaches to solving problems in R&D for autonomy in interdependent A-HMT-Ss in three ways:

i) Specific ML tools can be introduced into a task where it will produce clear benefits as part of the world model at that node's level, yet all consequences of its decisions will still be bounded by that node's authorised power;
ii) ML can be introduced into all parts of a node, *except in the behaviour generator function.* This design decision will ensure that actions cannot happen based on unexpected decisions without authorization by the human Team Leader; and
iii) Introducing ML into the node's value judgement function highlights the often-subjective nature of assessing the value of tasks when setting priorities. Recognising the associated risks before introducing ML in this function should explicitly raise, and help resolve, the complex questions in these applications.

An underlying problem with the use of AI is that of uncertainties in the interpretation of input information for comparison with world models which are themselves incomplete or inaccurate in some respects. Solving this problem is Research Objective 2-2, *AI Uncertainty Resolution* in the 2022 NAS report [1] for the general case: the approach presented here allows the uncertainties to be identified and their effects limited for specific cases. The offering of the alternatives considered by the system to the human goes some way to addressing Research Objective 5-5, *Explainability and Trust.*

AI will always generate a solution, so there must be a safeguard against unreasonable action, as interpreted by society or an accident inquiry. Setting limits using authorised power, and their use for deterministic testing of reasonable behaviour in every node provides a potential safeguard, although it does create its own design problems. However, locating authorized power in the behaviour generator function of every node bounds the problems, and provides a clear context for the essential cross-disciplinary and societal agreements before an A-HMT-S can be considered trustworthy.

Decomposition of A-HMT-S requirements using a hierarchical architecture into requirements for nodes comprising functions, with limited authority to act, allows targeted introduction of AI into the areas where it will bring maximum benefit, and will also identify the R&D needs before its safe introduction. This goes some way to meeting the 2022 NAS Report's Research Objective 10-1, *Human-AI Team Design and Testing Methods* and Research Objective 10-2, *Human-AI Team Requirements.*

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

The author is sole contributor to this work except where referenced in the text.

Acknowledgments

The author is grateful for discussions with and comments from Bill Lawless, Sarah Spurgeon and attendees at the AAAI Spring Symposium 2022, *Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams.*

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. National Academies of Sciences. *Engineering, and medicine. Human-AI teaming: State-of-the-Art and research needs*. Washington, DC: The National Academies Press (2021).

2. Llinas J. Motivations for and initiatives on AI engineering. In: WF Lawless, J Llinas, DA Sofge, R Mittu, editors. *Engineering artificially intelligent systems*. Springer (2021). p. 1–18.

3. Barfield W, Pagallo U. *Advanced introduction to law and artificial intelligence*. Cheltenham UK, Northampton MA USA: Edward Elgar Publishing (2020).

4. Baker S, Theissen CM, Vakil B. Connected and autonomous vehicles: A cross-jurisdictional comparison of regulatory developments. *J Robotics, Artif Intelligence L* (2020) 3(No. 4):249–73. https://heinonline.org/HOL/LandingPage?handle=hein.journals/rail3 div=38 id= page=.

5. Tan SY, Taeihagh A. Adaptive governance of autonomous vehicles: Accelerating the adoption of disruptive technologies in Singapore. *Government Inf Q* (2021) 38(2):101546. doi:10.1016/j.giq.2020.101546

6. USDOT. Autonomous vehicle guidance AV 3.0, "preparing for the future of transportation" (2022). Available at: https://www.transportation.gov/av/3 (Accessed June 16, 2022).

7. Sun Y. Construction of legal system for autonomous vehicles. In: In4th International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2020), 19. Atlantis Press (2020). p. 598–602.

8. Law Commission and Scottish Law Commission (2022). Automated vehicles: Joint report, (25 January 2022) HC 1068 SG/2022/15.

9. English and Scottish Law Commissions. Automated vehicle Project (2022). Available at: https://www.lawcom.gov.uk/project/automated-vehicles/ (Accessed March 31, 2022).

10. United Nations. Annex III to the final report on the 2019 meeting of the high contracting parties to the convention on prohibitions or restrictions on the use of certain conventional weapons which may Be deemed to Be excessively injurious or to have indiscriminate effects (2019). Available at: CCW/MSP/2019/9-E-CCW/MSP/2019/9-Desktop(undocs.org) (Accessed March 18th, 2021).

11. Madni AM, Madni CC. Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. *Systems* (2018) 6(4):44. doi:10.3390/systems6040044

12. Madni AM, Sievers M, Madni CC. Adaptive cyber-physical-human systems: Exploiting cognitive modeling and machine learning in the control loop. *Insight* (2018) 21(3):87–93. doi:10.1002/inst.12216

13. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manage Rev* (1995) 20(3):709–34. doi:10.5465/amr.1995.9508080335

14. Griffor E, Wollman DA, Greer C. NIST special publication 1500-201; framework for cyber-physical systems. *Working Group Rep* (2017) 2. doi:10.6028/NIST.SP.1500-202

15. Blakeney RA. *A systematic review of current adaptive human-machine interface research (doctoral dissertation)*. Embry-Riddle Aeronautical University (2020).

16. SEBoK Editorial Board. The guide to the systems engineering body of knowledge (SEBoK), v. 2.5. In: RJ Cloutier, editor. Hoboken, NJ: The Trustees of the Stevens Institute of Technology (2021). BKCASE is managed and maintained by the Stevens Institute of Technology Systems Engineering Research Center, the International Council on Systems Engineering, and the Institute of Electrical and Electronics Engineers Systems Council Available at: www.sebokwiki.org (Accessed December 7, 2021).

17. Alix C, Lafond D, Mattioli J, De Heer J, Chattington M, Robic PO. Empowering adaptive human autonomy collaboration with artificial intelligence. In: 2021 16th International Conference of System of Systems Engineering (SoSE). IEEE (2021). p. 126–31.

18. Albus J, Huang HM, Lacaze A, Schneier M, Juberts M, Scott H, et al. 4D/RCS: A reference model architecture for unmanned vehicle systems version 2.0. In: NIST Interagency/Internal Report (NISTIR). Gaithersburg, MD: National Institute of Standards and Technology (200). [online]. doi:10.6028/NIST.IR.6910

19. Albus JS, Barbera AJ. Rcs: A cognitive architecture for intelligent multi-agent systems. *Annu Rev Control* (2005) 29(1):87–99. doi:10.1016/j.arcontrol.2004.12.003

20. Gillespie T. Good practice for the development of autonomous weapons: Ensuring the art of the acceptable, not the art of the possible. *RUSI J* (2021) 165(5-6):58–67. doi:10.1080/03071847.2020.1865112

21. Albus J, Barbera T, Schlenoff C. *'RCS: An intelligent agent architecture'. Intelligent agent architectures: Combining the strengths of software engineering and cognitive systems*. Palo Alto, California: AAAI Press. no. WS-04-07 in AAAI Workshop Reports 2004. Available at: https://web.archive.org/web/20110324054054/http://www.danford.net/boyd/essence.htm. (Accessed July 8, 2022).

22. Madhavan R, Messina ER, Albus JS. *Intelligent vehicle systems: A 4D/RCS approach*. New York, NY, USA (2007). Available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=823578 (Accessed April 24, 2022).

23. Gillespie T. *Systems engineering for ethical autonomous systems*. London: SciTech, Institution of Engineering and Technology (2019). p. 512.

24. Serban AC, Poll E, Visser J. A standard driven software architecture for fully autonomous vehicles. In: IEEE International Conference on Software Architecture Companion. ICSA-C (2018). p. 120–7.

25. Gillespie T, Hailes S. Assignment of legal responsibilities for decisions by autonomous cars using system architectures. *IEEE Trans Technol Soc* (2020) 1(3): 148–60. doi:10.1109/tts.2020.3014395

26. Rasmussen J. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans Syst Man Cybern* (1983) 13(3):257–66. doi:10.1109/tsmc.1983.6313160

27. Boyd JR. The essence of winning and losing. *Unpublished lecture notes* (1996) 12(23):123–5. Slides also available at http://pogoarchives.org/m/dni/john_boyd_compendium/essence_of_winning_losing.pdf (Accessed April 22, 2022).

28. Gillespie T. *Systems engineering for autonomous systems*. Stevenage: SciTech, Institution of Engineering and Technology (2019). p. 292.

29. Shneier M. Learning in a hierarchical control system: 4D/RCS in the DARPA LAGR program. *J Field Robot* (2006) 23(11-12):975–1003. doi:10.1002/rob.20162

30. Chang T. Integrating learning into a hierarchical vehicle control system. *Integr Comput Aided Eng* (2007) 14(2):121–39. doi:10.3233/ica-2007-14202

31. Ding Z. *Nonlinear and adaptive control systems*. London, United Kingdom: Stevenage: The Institution of Engineering and Technology (2013). p. 287.

32. Aldus J, Barbera A. 4D/RCS reference model architecture for unmanned ground vehicles chapter 1 in R Madhavan, ER Messina, JS Albus. In: *intelligent vehicle systems: A 4D/RCS approach*. New York, NY, USA (2007). Available at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=823578 (Accessed April 24, 2022).

33. Linas J, Gillespie T, Fouse S, Lawless W, Mittu R, Sofge D, et al. *AAAI spring Symposium 2022 proceedings. Putting ai in the critical loop: Assured trust and autonomy in human-machine teams*. Elsevier (2022). in preparation.

34. Gillespie T. Risk reduction for autonomous systems. In: WF Lawless, J Llinas, DA Sofge, R Mittu, editors. *Engineering artificially intelligent systems*. Springer (2021). p. 174–91.