Dynamic Human-Machine Teams
Trust & Responsibility

Tony Gillespie PhD CEng FIET FREng
Visiting Professor
Electronic & Electrical Engineering
anthony.gillespie@ucl.ac.uk

1

Trust and Responsibility

- Trust
  - "...an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of the designer's intentions, be relied on to achieve the design goals" *
- Responsibility
  - The state or fact of being accountable or to blame for something. (Oxford English Dictionary)
- **The leader of a trusted automated system accepts responsibility for its actions**

* Moray N. & Inagaki T. 1999, Laboratory studies of trust between humans and machines in automated systems. Trans of the Insti of Measurement & Control **21**(4–5), 203–211

2

Trust and Responsibility

- Trust
  - "...an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of the designer's intentions, be relied on to achieve the design goals" *
- Responsibility
  - The state or fact of being accountable or to blame for something. (Oxford English Dictionary)
- **The User of a trusted human machine team accepts responsibility for its actions even with no direct human involvement in decisions**

* Moray N. & Inagaki T. 1999, Laboratory studies of trust between humans and machines in automated systems. Trans of the Insti of Measurement & Control **21**(4–5), 203–211

3

Responsibility for Harm to Others

- Restrict considerations to systems with mix of physical and human components
  - Cyber-Physical Human System (CPHS)
  - Human Machine Team (HMT) with human user
- Human user sets high-level aims
  - Aims met by task allocation to subsystems
- Work overload, especially of user, not allowed
- Desire to use Artificial Intelligence and Machine Learning (AI/ML) in multiple places

4

A Potential Example

5

Dynamic resource planner Technical problems

- Need manageable workload on human(s)
  - Predictive so user has time to take action
- Dynamic task reallocation
  - User changes aims
  - Task outcomes
  - Uncertainties/Risks
- User must understand task allocation so he/she can take over

6

UCL

The Lawyer's View

- Responsibility = liability
- A machine cannot be held responsible for its actions
- Who is responsible?
- Same problem addressed by UN discussions on bans on Lethal Autonomous Weapons (LAWS)
- Non-military lawyers now see the same problem:
  - **Design responsibility**

7

UCL

English and Scottish Law Commissions' Joint Report on Automated Vehicles 2022

Recommendations 71, 73 and 74

- Product liability law should be reviewed … over all product liability, not confined to automated vehicles.
- The (new) authorisation authority should require specified minimum data to be collected and stored to process insurance claims. ...
- It should be a criminal offence if a commercial practice uses: the terms "self-drive", "self-driving", "drive itself", "driverless" and "automated vehicle"; …

Law Commission
Reforming the law

Scottish Law Commission
promoting law reform

8

Responsibility - Three Questions

1. Can a dynamic CPHS be designed with the liability for the consequences of every action assigned to identifiable humans?

Use a hierarchical architecture to drive design

2. What guidance is to be given to stakeholders to ensure clear responsibilities for actions?

Each node to have unambiguous authority

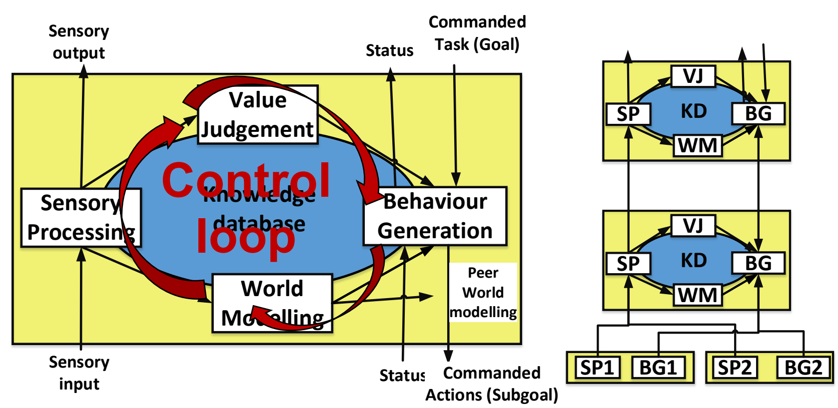
3. How will the potentially liable individuals develop sufficient trust to carry out their work?

Node responses to mimic human behaviour

9

Use 4D/RCS Architecture

Based on Hierarchical Delegation of Responsibility

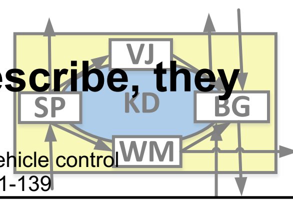


Awareness Understanding Deliberation
Observe Orient Decide Act

10

Architecture Aim Using 4D/RCS

- Ensure a human can trust dynamic decisions made by a human-machine team and take responsibility for the consequences
- Dynamic HMT will have AI and learning as part of the decision-making process
- ML works most successfully when introduced at different levels in the hierarchy and separately in specific functions in its nodes*
- **Architectures don't just describe, they drive the drive design**



* Albus et al. 2006, Integrating learning in a hierarchical vehicle control system. *Integrated Computer-Aided Engineering*. 14(2), pp121-139

11

Role of AI in a Node

- Observe
  - Input workloads will have uncertainties
  - Some subjective predictions in inputs
- Orient - Predictive, not reactive, process
  - Comparison – Probably subjective
  - Consequences will be subjective and uncertain
- Decide and act
  - Ranking will be uncertain
- Authorisation
  - May be subjective
  - Has accumulated uncertainties in inputs

12

UCL

Requirements for AI-based system

(Based on Alix et al. 2021*)

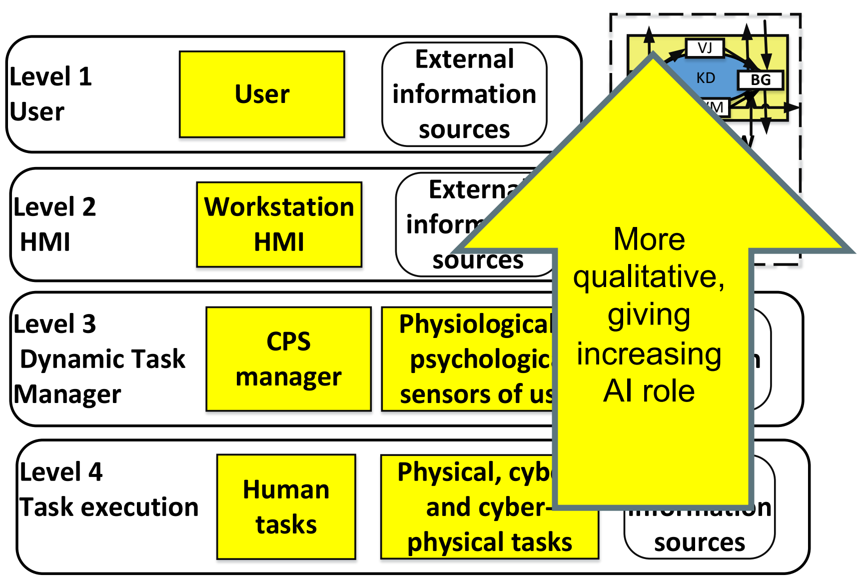
- Validity
  - The system must do what it is supposed to do, all it is supposed to do and only what is supposed to do
- Explainability
  - Ensure user confidence through human-oriented and understandable justifications of the AI results
- Accountability
  - Meet ethical standards and exhibit lawful and fair behaviours
- **Test each AI node against these with different weight depending on authority**

* Alix *et* al, 2021, Empowering adaptive human autonomy collaboration with AI, Int, J, Conf, Syst Of Syst Eng pp. 126-131, IEEE

13

UCL

CPHS Higher Architecture Levels - 1



14

Human Machine Interface (HMI) Functions

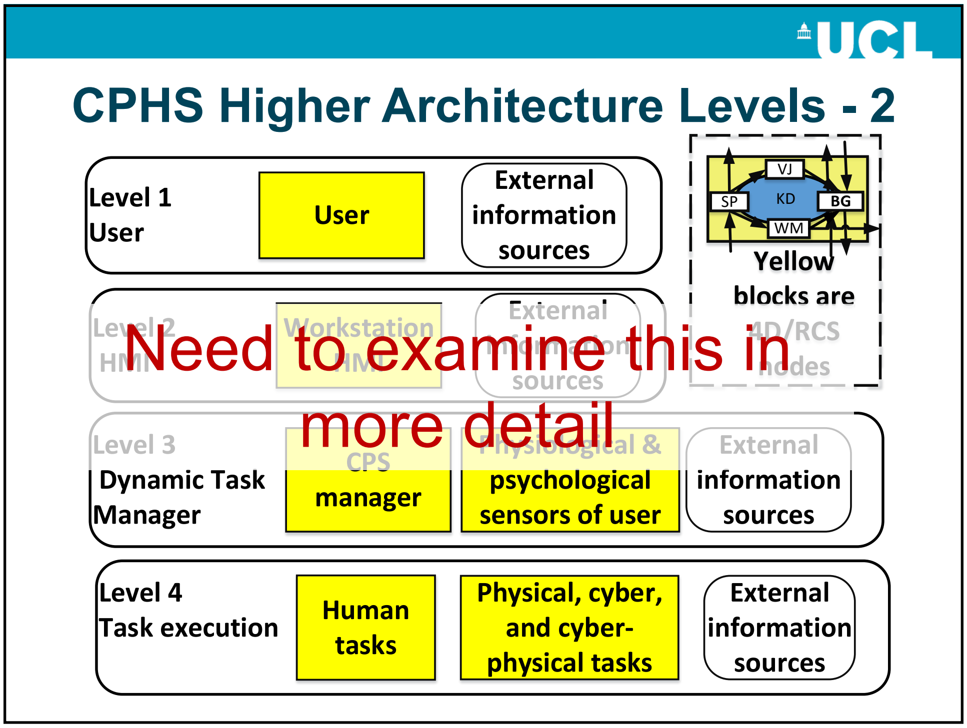
- Present management information to user at business timescales
- Allows user to interrogate information
- Monitors external information and warns user of likely increased workloads arising
- Predicts task manager and resource workloads
  - Seeks extra resources via user
- Converts user instructions into success criteria
  - Issued through behaviour generator chain

15

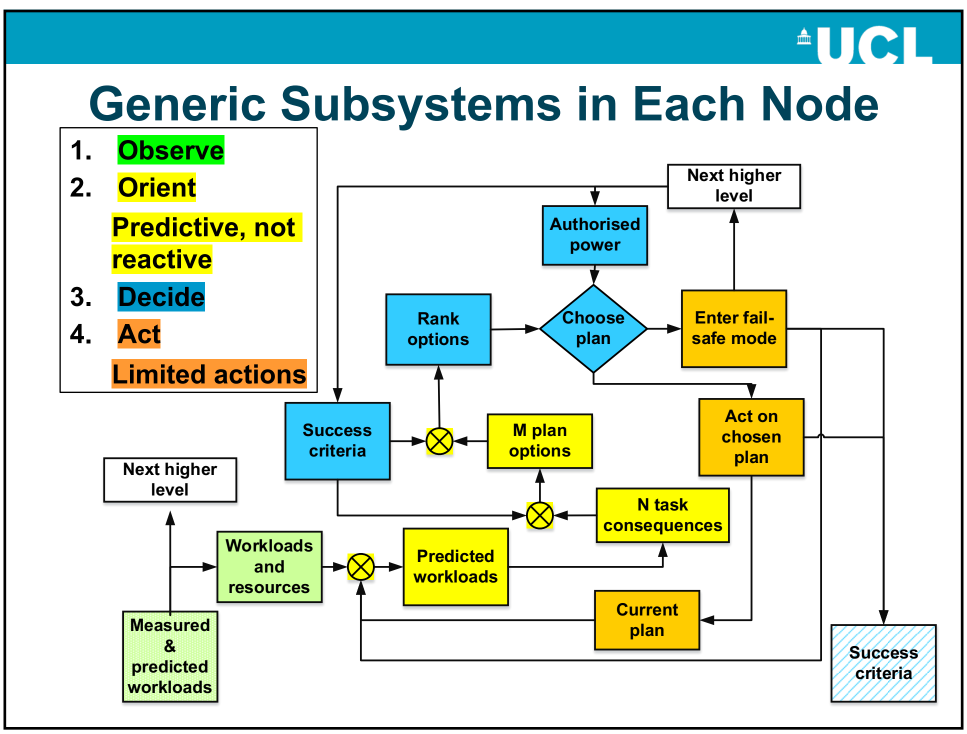
Task Manager Functions

- Dynamic management of lower levels in task timescales
  - Predictive not reactive system
- Deals almost exclusively with internal HMT information
- Converts input success criteria from HMI into success criteria for next level down
- Warns if human workloads at any level will be high
- Flag up problems to HMI

16


CPHS Higher Architecture Levels - 2
Level 1
User
User
External information sources
VJ
SP
KD
BG
WM
Yellow blocks are 4D/RCS nodes
Level 2
HMI
Workstation HMI
External information sources
Need to examine this in more detail
Level 3
Dynamic Task Manager
CPS manager
Physiological & psychological sensors of user
External information sources
Level 4
Task execution
Human tasks
Physical, cyber, and cyber-physical tasks
External information sources

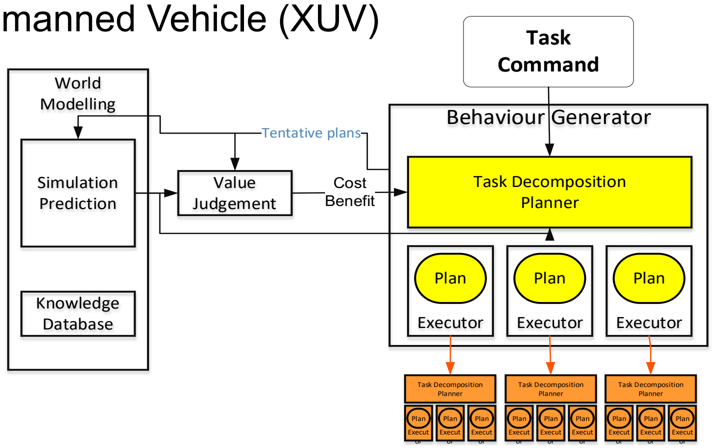
17


Generic Subsystems in Each Node
1. Observe
2. Orient
Predictive, not reactive
3. Decide
4. Act
Limited actions
Next higher level
Authorised power
Rank options
Choose plan
Enter fail-safe mode
Success criteria
M plan options
Act on chosen plan
N task consequences
Next higher level
Workloads and resources
Predicted workloads
Current plan
Measured & predicted workloads
Success criteria

18

Comparison with proven approach

- Army Research Lab Demo III eXperimental Unmanned Vehicle (XUV)



* Based on Chapter 1, Figure 3 in Madhavan, R., Messina, E.R. and Albus, J.S., 2007. Intelligent vehicle systems: a 4D/RCS Approach. *New York, NY, USA*

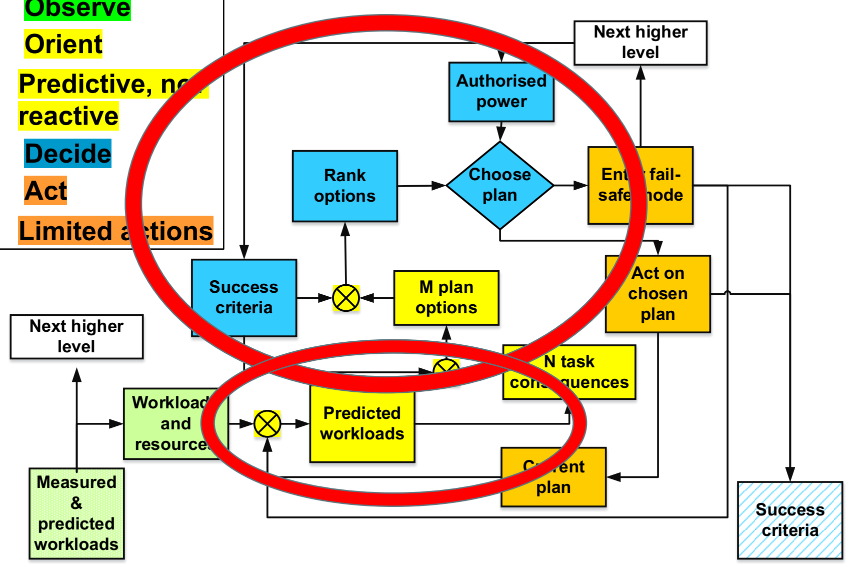
19

Generic Subsystems in Each Node

1. **Observe**
2. **Orient**

   Predictive, no reactive
3. **Decide**
4. **Act**

   Limited actions

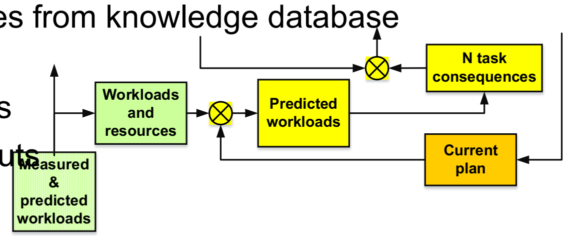


20

UCL

Workload Prediction - 1

- Predicts workloads for sum of lower level nodes if current plan is followed
  - Current plan is in knowledge database
- Inputs:
  - World model from sensory processing
  - Workload predictions from its lower level nodes
  - Peer sensor processing node for HMI node
  - Available resources from knowledge database
- Uncertainties:
  - Real-time changes
  - AI/ML-based outputs

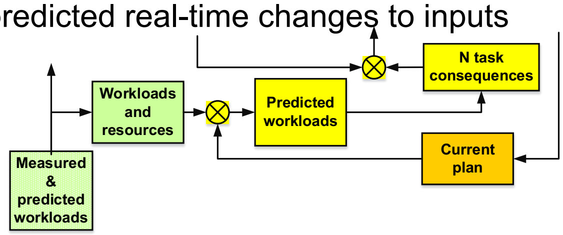


21

UCL

Workload Prediction - 2

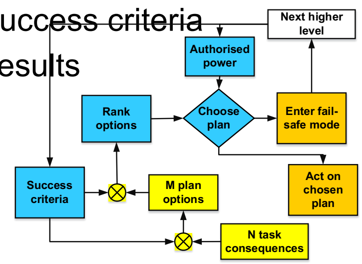
- Outputs:
  - Set of predicted workloads following current plan
  - Consequences of each set member
- Problems:
  - Measurement of human and CPS workloads
  - Comparison of plan with real world
  - Assessment of consequences
  - Predicted and unpredicted real-time changes to inputs at lower levels
  - Fail-safe mode



22

Planning Deciding and Acting - 1

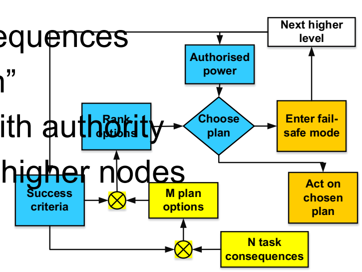
- Planner input:
  - Set of predicted workloads following current plan
  - Consequences of each set member
- Generates one plan for every consequence in set
  - Workload planner techniques well-developed
- Decide
  - Consequences compared with success criteria
  - Plans ranked for best/optimum results



23

Planning Deciding and Acting - 2

- Act
  - Check first choice and consequences are authorised
  - If not, do at least one of:
    - Refer to higher level node and warn HMI
    - Enter fail-safe mode
- Problems
  - Identifying uncertainties in consequences
  - Setting criteria for "best/optimum"
  - Comparison of consequences with authority
  - Ensuring low false alarm rate to higher nodes



24

Problems Identified

- Predictor problems:
  - Measurement of human and CPS workloads
  - Comparison of plan with real world
  - Assessment of consequences
  - Predicted and unpredicted real-time changes to inputs at lower levels
  - Fail-safe mode
- Decide and act problems
  - Identifying uncertainties in consequences
  - Setting criteria for “best/optimum”
  - Comparison of consequences with authority
  - Ensuring low false alarm rate to higher nodes

These become manageable when tackled for each node’s limited authority and timescale as a system evolves

25

Conclusions

- Three aims from Aix *et al* can be met
  - Validity. Explainability, and accountability
- Need to test each AI node against these with different weight depending on authority
  - Nine principal problems identified. Main ones are:
    - Comparison of real world and plans
    - Assessing uncertainties in predictions
  - All nine can be solved for each node as AI is steadily introduced to replace human or automated actions
- Possible to introduce AI **and** meet legal liability problems

26

Questions?

27