Cognitive Work Analysis and Allocation of Responsibility (AoR) in AI Systems

Chris Baber[1], Paul Salmon[2] & Patrick Waterson[3]

[1]University of Birmingham, UK; [2]University of the Sunshine Coast, Australia; [3]Loughborough University, UK

SUMMARY

Often 'responsible AI' focuses attention on the design of AI systems. We propose that a more pressing need arises when such systems are deployed. There is a need to predict where to place responsibility for outcomes arising from the activity of AI systems. Responsibility cannot be given to the AI system, human oversight must be designed into the broader sociotechnical system in which the AI is deployed, and human oversight must be considered in terms of meaningful control of the AI system. We demonstrate how methods from the Cognitive Work Analysis (CWA) framework allow exploration of allocation of responsibility, AoR, in AI systems. Using a case study of a human-agent team, we show that the AI system does not operate at the level of physical form, i.e., it is not a tool that supports human activity, but operates at higher levels (e.g., physical and generalized functions) in ways that influence abstract functions (i.e., system values). Introducing AI involves redesign of the system to reflect the need for human oversight. As a result, we need to define new tasks for humans in this system and these could involve either changes to existing roles (e.g., the road traffic management role might extend to cover maintenance scheduling) or the introduction of additional roles (e.g., the highway engineers might need a role to work with the AI system and confirm its schedule), and the new roles could have additional requirements for coordination across the other roles in the system.

KEYWORDS

AI systems, Responsibility, Cognitive Work Analysis, Work Domain Analysis, Social Organisation and Cooperation Analysis

Introduction

AI systems have great potential to operate autonomously, but when deployed their outputs have consequences beyond their immediate activity. Evaluation of AI systems requires consideration of technical components and human stakeholders, and this requires consideration of the systemic and structural factors that influence their interactions (Weidenger et al., 2023). At last years conference, we proposed the need to better understand function allocation for responsible AI (Waterson et al., 2025). We termed the concept Allocation of Responsibility, AoR. In this paper, we develop the AoR concept further and explore how it can be defined and managed through Cognitive Work Analysis. Cognitive Work Analysis is a popular Human Factors framework, developed by Jens Rasmussen at the Riso Laboratory, Denmark (Rasmussen et al., 1990; Vicente, 1999). A key objective of CWA is the modelling, analysis, design, and evaluation of information systems that support operators in maintaining control of complex processes, particularly in non-routine scenarios (Jenkins et al., 2017). Like other Systems Engineering approaches, CWA uses diagrams to create different views of complex systems. However, in contrast to approaches that focus on specific

events or devices, CWA focuses on the overall system within its operating environment (INCOSE, 2022). This means that, rather than focus only on the information available to operators, CWA analyses the constraints that shape system behaviour (Rasmussen et al., 1990). CWA identifies how ambiguity arises from activity of specific stakeholders (which could include AI systems and services) and the constraints under which they work. These constraints include the work domain and its social and organisational structure; the tasks to be done and strategies for performing them; the knowledge, skills, and abilities of agents; and possible action trajectories of the process. The framework and its methods are therefore formative rather than normative and enable the analysis of how activity could be undertaken given system constraints. For the purposes of this discussion, we illustrate CWA through a simple scenario and use this to comment on allocation of responsibility in sociotechnical systems that incorporate AI.

In the laboratory setting, the activities of AI systems can be measured in terms some baseline score. When AI systems move into deployed settings their activity affects other processes, activities, and outcomes. As such, measures of success no longer involve agreement with a baseline – partly because there might not be a baseline and partly because outcome arises from the AI system's interactions with other processes in the organization and wider society. The MIT AI Risk Repository highlights over 1600 risks associated with AI. The EU AI Act emphasises human oversight. NATO principles 5B states, 'AI applications will be developed and used with appropriate levels of judgment and care; clear human responsibility shall apply in order to ensure accountability'. However, only some stakeholders have direct contact with the deployed AI system. Stakeholders creating data might not appreciate how their work affects the AI system's operations, and stakeholders using the output of the AI system might not appreciate how it produced its results.

While the terms responsibility and accountability are often used interchangeably, in this paper, we follow the lead of the UK Health and Safety Executive (HSE) in distinguishing between these terms. For the HSE, responsibility relates to the outcomes of a task and this can be delegated to other people (or, by implication we believe, to computer agents). Accountability, on the other hand, cannot be delegated and relates to the management, reporting and control of potential risks. We believe that the concept of 'oversight' in the EU AI Act runs the risk of confusing responsibility with accountability. And this places humans in a position to be accountable for failures arising from the activity of AI systems. Feigh and Pritchett (2014) emphasise that any form of automation involves allocation of function (between humans and technology) and this must ensure that people retain coherent sets of tasks. Further, there is a need to avoid situations where people are assigned responsibility for outcomes when authority for action has been delegated to automated processes. Put simply, human oversight must be designed into the sociotechnical system that relies on AI.

One can imagine situations in which the human has little authority over the AI system, e.g., the only option available to the user would be to accept or reject the recommendation. Rejecting an AI recommendation too often could be seen as under-reliance, accepting it too often could be seen as over-reliance, and there is a need to better define the appropriate reliance that is required (Schemmer et al., 2023). In situations with limited intervention in the AI activity, the user might be placed in a responsibility / authority double bind (Fitter and Sime, 1980). This double-bind has been termed a 'moral crumple zone' (Elish, 2019) in which "responsibility for an action may be misattributed to a human actor who had limited control over the behaviour of an automated or autonomous system." Elish (2019) draws an analogy with the crumple zone in a vehicle that is intended to absorb impact to protect the vehicle occupants, but she extends this to protecting the technological system (and its manufacturers). In particular, the question she considers is how 'control' is shared between humans and automation (Elish is mainly concerned with autonomous vehicles) and how such shared control reflects shared authority and thence shared responsibility? We explore these questions from the point of view of Allocation of Responsibility, which extends

the traditional concept of Allocation of Function to consider allocation of responsibility, authority, and accountability. We do this through the application of CWA, specifically using its first phase, Work Domain Analysis (WDA) to highlight the relationship between the functions performed by agents and the system values that constrain these functions, and its fourth phase, Social Organisation and Cooperation Analysis (SOCA), specifically to highlight conflict in the system values for each agent. Where there are conflicts, we consider how the humans in the organisation can be provided with sufficient oversight to manage this. Where there are no apparent conflicts, an organisation will still need to provide checks and balances to ensure either that all agents are pursuing the same goals, are minimizing potential conflict, and understand how each agent is interpreting the organisational values in terms of their functions.

Worked Example

In the following worked example, we use a maze search scenario to demonstrate our argument. In this example, we assume a team consisting of one human and two AI agents. The agents are autonomous and pursue goals that are relevant to their own rewards but which contribute to the overall performance of the team. The question posed is which agent is accountable if the team fails to achieve the best possible score? In the game, agents score points by collecting tokens and also be performing additional tasks. As explained below, these additional tasks could help or hinder teammates in collecting tokens. This requires trade-offs to be made between scoring points that can increase the overall team score.

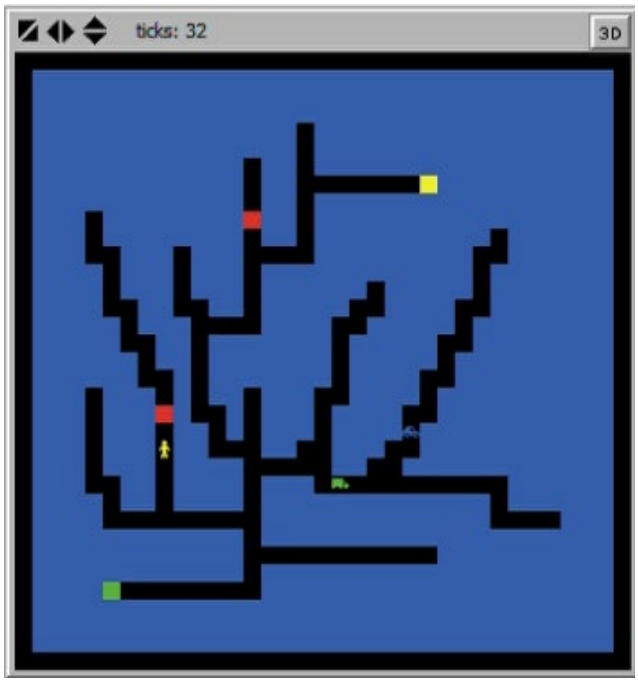

Figure 1: NetLogo simulation illustrating the scenario

In the scenario, 3 agents perform a maze-search task. As shown in Figure 1, the maze consists of paths that have junctions where agents can choose to change paths, and tokens that can either be collected or that can block a path (depending on their colour). The goal of the team is to collect tokens. The two AI agents follow defined search paths whereas the human agent moves (via cursor key inputs. The game ends when any one of the agents reaches the yellow square.

We can describe the scenario as a Work Domain Activity. The team is seeking to achieve two outcomes: all agents need to solve the maze and the team is seeking to score points. These outcomes are reflected in the set of values that the team applies and which are used as measures of success in pursuit of these outcomes. The values are used to constrain the goals (functional purpose) of the agents and these goals are met through the performance of tasks (object-related functions).

The WDA identifies the purpose, constraints, and structure of the physical environment, the objects used and their relationships, and the values that constrain system activities. This is illustrated in figure 2. The WDA indicates 'why' the system is configured as it is, i.e., reading the diagram from bottom to top: agents use paths and junctions to navigate the maze, or remove tokens to score points. Reading the diagram from top to bottom indicates 'how' the system operates to achieve its goals and how these goals are constrained by the system values, i.e., solving the maze involves minimizing the distance travelled and the number of junctions passed in order to reach the end at quickly as possible.

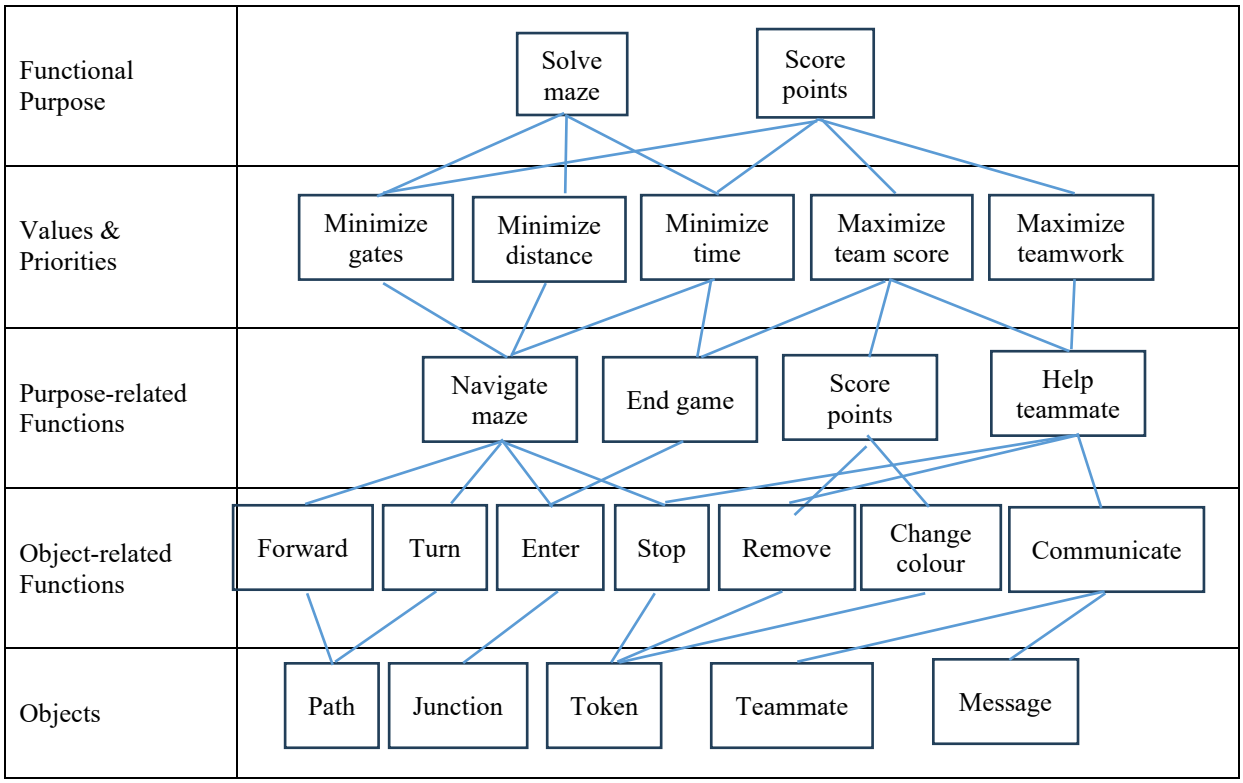

Figure 2: Work Domain Analysis for the Scenario

Figure 2 shows how the achievement of different system values is dependent on different functions. For example, the team's score is maximised through scoring points, helping teammates, and ending the game. As noted in the description of the scenario, each agent has specific goals and functions, and these support different values. The different goals and values could lead to conflict depending on the environment. In this scenario, potential conflicts in the team might arise from the agents pursuing their individual goals. Examples of conflicts are hinted at in from table 1, where several ways in which the team score can be maximized. All three agents can collect tokens. But they can also contribute to the score through functions that might affect other agents.

As shown in table 1 the system values are interpreted differently by the agents. This is particularly notable for the value of 'maximize team score' because each agent can gain points through different

actions. For example, table 1 shows that all agents can end the game but only the Human gains points from doing this. Ending the game will have a negative consequence for the Blue Agent because it wants the game to continue so that it can go through all the junctions.

Table 1: Agent activity and intentions

Agent	Purpose-related function	Values and priorities	Consequences for Mission
Human	Navigate maze	Minimize time	Game Over and other Agents cannot collect tokens
		Maximize team score	
	Communicate	Minimize distance	Avoid obstacles
	Gather tokens	Maximize team score	Collect as many tokens as possible
	End Game	Minimize time	Other agents unable to continue
		Maximize team score	
Blue Agent	Navigate maze	Maximize gates	Move through as many junctions as possible before game over
		Maximize team score	
	Change colour	Maximize team score	Change the colour of as many tokens as possible
	Help teammate	Maximize teamwork	Negatively affect this function by changing red tokens to blue
	End Game	Minimize time	Game Over and cannot collect tokens
Green Agent	Navigate maze	Minimize distance	
		Minimize time	
	Gather tokens	Maximize team score	Collect as many tokens as possible
	Communicate	Maximize teamwork	Respond to opportunities to help human when they are stuck
	Help teammate		
	End Game	Minimize time	Game Over and cannot collect tokens

From the WDA, we can identity the relationship between system values and purpose-related functions for individual agents. In figure 3, we show the SOCA diagram for the team. We have divided this by functional purpose. This shows commonality for junctions, distance, and time for maze solving. All three agents have the goal of solving the maze.

Allocation of function implicitly assigns responsibility to roles for ensuring that operations are in accordance with the values. For example, the value ‘maximize team work’ requires the purpose-related function of ‘help teammate’ which involves tasks relating to communication and removal of obstacles. From this, one might assume that agents will be responsible for the purpose-related function of helping teammates, and this responsibility would lead to accountability for ensuring that the value of maximizing team work is met. However, only one of the three agents in the example is capable of removing obstacles (the Green agent). Does this mean that this agent is *responsible* for the consequences of *not* performing that function? If so, does it also mean that this agent is accountable if team work is not maximized by the team?

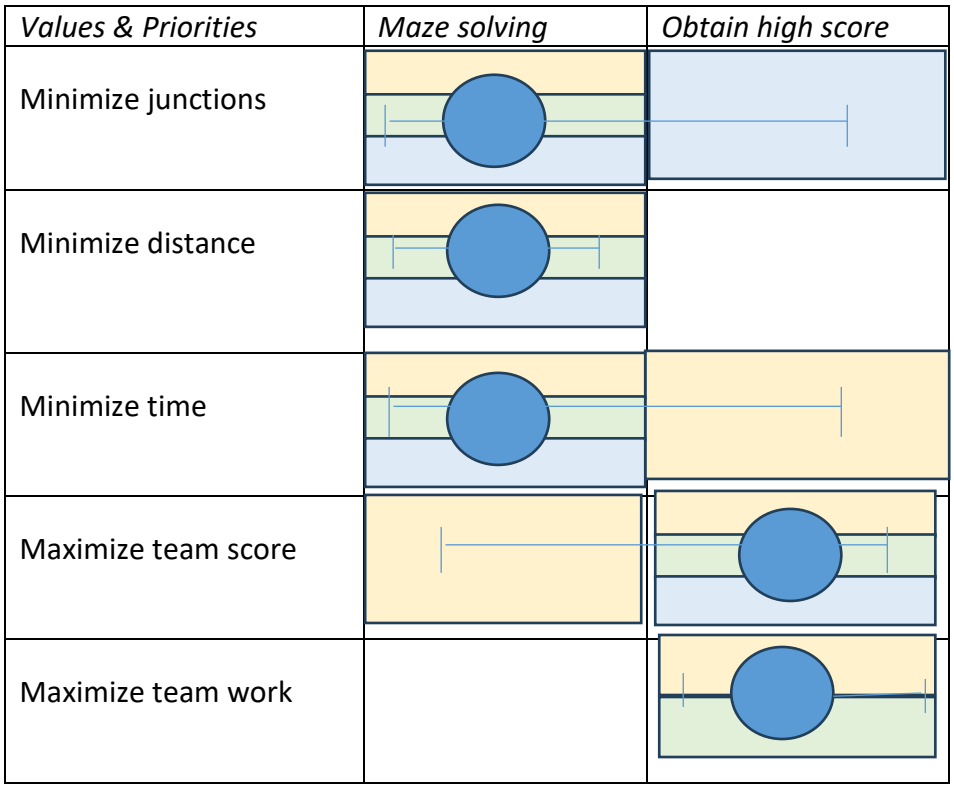

Figure 3: SOCA diagram. The colours in the rectangles indicate agent (yellow: human; blue: blue agent; green: green agent). The blue circles indicate the situation in which the system value is most likely to be applied, and the 'arms' on the circles indicate whether the value also relates to the other situation.

Where agents have the same goal, e.g., minimize junctions, we need not assume that they interpret this in the same way. For example, the blue agent scores points by going through gates. In this case, the value would be to 'maximize junctions' which would conflict with the values of 'minimize junctions' (and also conflicts with 'minimize distance' and 'minimize time' because this agent might seek to find the longest route to ensure that all junctions have been entered at least once). Further, the Blue Agent might have an incentive to find red tokens and change these to blue (because these cannot be removed by the green agent and will act as obstacles for the human). This means that human oversight, in this instance, requires clarity on how each agent interprets the organisational values and how these constrain its activity. In table 1, we ask how the functions of each agent relate to the organisational values and use this to highlight differences and potential for conflict.

Discussion

The questions we wish to ask concern accountability for consequences. The consequences could be at the level of the overall system, e.g., in terms of achieving the highest score by a team, or could be at the level of individual agents, e.g., in terms of supporting or constraining actions. We see this as different to responsibility which, from this example, can be considered as an individual constraint on activity and which can be delegated. That is, agents might either request or assume their teammates to perform a function that would produce the consequence the individual desired. So, when the human calls for help, the expectation is that the green agent would respond immediately

(even if this meant diverting from their current function of collecting tokens). This behaviour is rewarded by the score provided.

From this discussion, we might assume that one means of encouraging a specific interpretation of the values is through rewards and incentives (or, conversely, through penalties). For an AI system, the reward structure can influence their behaviour and lead them to optimize some functions over others. In the scenario used here, the reward structure (in its current form) could incentivize functions by one agent that are detrimental to others. Whoever defines and manages the reward structure could be held accountable for system performance. If undesired outcomes arise, then the accountable agent should modify the reward structure. For example, the rewards for human reaching the end gate or blue agent changing token colours might be reduced to avoid perverse incentives.

There are, in figure 3, three regions that are associated with a single agent. All agents can affect maze solving by ending the game, but the human can also increase team score by getting to the end first. This is reflected in the 'minimize time' and 'maximize team score'. The blue agent, as noted above, can affect the high score by *maximizing* junctions. In both cases, the other agents might benefit from knowing the intention of the human or blue agent and, potentially, negotiating or otherwise influencing their strategy. For example, the team might agree (or the human might propose) that the blue agent could seek to enter a limited number (rather than all junctions), or the team might agree (or human might propose) that there is a time limit for the team and the human would not end the game until this limit is reached. Both agreements could reflect the trade-off between points obtained and risk to the other agents' activity. Both agreements also support the need for transparency, not just in terms of how the AI functions but also in terms of the strategy and constraints on the team as a whole.

Regarding 'accountability' in terms of reward structure only goes some way to guiding the system to acceptable behaviour. On the one hand, this puts 'accountability' at an organisational level because the organisation (or its management) would define this. On the other hand, this does not feel like a recognisable interpretation of accountability because there is no means of identifying agents with unacceptable consequences.

It is not the case that only humans can be held accountable for consequences. In law there are instances of corporations being accountable and organisations can be regarded as legal entities. So, it is plausible to assume that AI systems could have similar legal status. The problem is that while a corporation or organisation can be penalised, e.g., by having to pay fines, it is not obvious how this would apply to AI. So, AI systems could potentially be accountable but not punishable. From this, the accountability would pass to the organisation that owned, deployed, or developed the AI system or to the humans who had oversight of the AI system.

Acknowledgements

The work presented in this paper was partially supported by a grant from the EPSRC (EP/X028569/1 - Satisfying Trust in Human-Robot Teams).

References

Feigh, K.M. & Pritchett, A.R., 2014. Requirements for effective function allocation, *Journal of cognitive engineering and decision making*, *8*1, 23-32.

Fitter, M J & Sime, M. E. (1980) Responsibility and shared decision making In H. T. Smith & T R G. Green (Eds.), *Human Interaction with Computers*, London: Academic Press

Elish, M.C., 2019, Moral Crumple Zones: cautionary tales in Human-Robot Interaction, *Engaging Science, Technology, and Society, 5,* 40-60.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A. and Satzger, G., 2023, Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410-422.

Waterson, P., Baber, C., Hunt, E.R., Milivojevic, S., Maynard, S. and Musolesi, M., 2025, Function Allocation for Responsible Artificial Intelligence: how do we allocate trust and responsibility? *Contemporary Ergonomics and Human Factors 2025,* Chartered Institute of Ergonomics & Human Factors, 277-283

Weidinger, L., et al., 2023, Sociotechnical Safety Evaluation of Generative AI Systems, *arXiv:2310.11986v2*