Function Allocation for Responsible Artificial Intelligence: How do we allocate trust and responsibility?

Patrick Waterson[1], Chris Baber[2], Edmund Hunt[3], Sanja Milivojevic[4], Sally Maynard[1] & Mirco Musolesi[5]

[1]Human Factors and Complex Systems Group, Loughborough University, [2]University of Birmingham, [3]University of Bristol, [4]Bristol Digital Futures Institute, University of Bristol, [5]University College London

SUMMARY

We consider how guidelines for Responsible Artificial Intelligence (RAI) need to be adapted to address the challenges of Function Allocation (FA) in human-agent teams. We offer an approach that takes a system description, using CWA, to identify where responsibility for consequences of actions might lie across the system. We propose that, in addition to allocation of functions, analysis of the system needs to identify decision points (where agents have a choice of action to perform) and responsibility points (where agents identify the consequences of their decisions). We illustrate this with example experiments. We put forward a set of open challenges and questions facing researchers in the areas of RAI and FA. We point to the need for greater emphasis on the issue of responsibility, trust and accountability in new forms of automation. We also provide pointers for the future and how these might be addressed in the coming years.

KEYWORDS

Function allocation (FA); Human-Robot teams; Responsible Artificial Intelligence (RAI); sociotechnical systems.

Introduction

Prominent computer scientists have questioned the ethical, moral and legal implications of the new technologies built of 'foundation Artificial Intelligence' (e.g., Brown, 2023). These implications often lead to calls for so-called 'responsible AI (Artificial Intelligence).' A key aspect of this, for Human Factors, is how responsibility relates to Function Allocation. That is, in a system involving AI agents and humans, where does responsibility for decisions and actions sit? Is it appropriate to assume that only humans can hold responsibility for decisions or should the human responsibility be for the consequences of these decisions? Does this responsibility for consequences of decisions apply even in circumstances where the behaviour of the AI is not transparent? Function allocation (FA) refers to strategies for distributing system functions and tasks across people and technology. Traditionally, as discussed below, FA has concentrated on comparisons between humans and machines in terms of their strengths and weaknesses. With the advent of more advanced technologies such as artificial intelligence (AI) and robotics, it can be more difficult to identify clear boundaries between capabilities of humans and agents (particularly in tasks involving decision making) and there is a need to move beyond conceptions of FA as discrete allocations and to move towards new forms of interdependencies that will create shared, collective entities involving humans and machine. Many FA methods pay limited attention to joint operation of people to

perform tasks, and very few FA methods address Machine-Machine (M-M) allocations, particularly in terms of how much autonomy should be provided to technologies such as robots and agents.

Traditional approaches to function allocation

The topic of function allocation can trace a lineage back to some of the earliest days of HFE and the seminal work of Paul Fitts (1951) on HABA-MABA lists (Humans-are-better-at-Machines-are-better-at). At the time Fitts and his colleagues were developing ways of examining automation, well before the computer and internet revolutions of the 1970s through to today. In simple terms, 'automation' will operate according to clearly specified sets of instructions; systems with semi-autonomy will follow instructions but can adapt their behaviour according to situation; autonomous systems have the potential to define their own instruction sets and specify their goals. We believe that much of the FA literature has focused on automation, with some effort at considering semi-autonomous systems, and there remains a gap when it comes to FA for autonomous systems that interact with humans. Following Fitts' initial proposals, approaches to FA stressed the need to encompass wider sociotechnical aspects of work design and a wider range of concerns (Waterson *et al.*, 2002, figure 1).

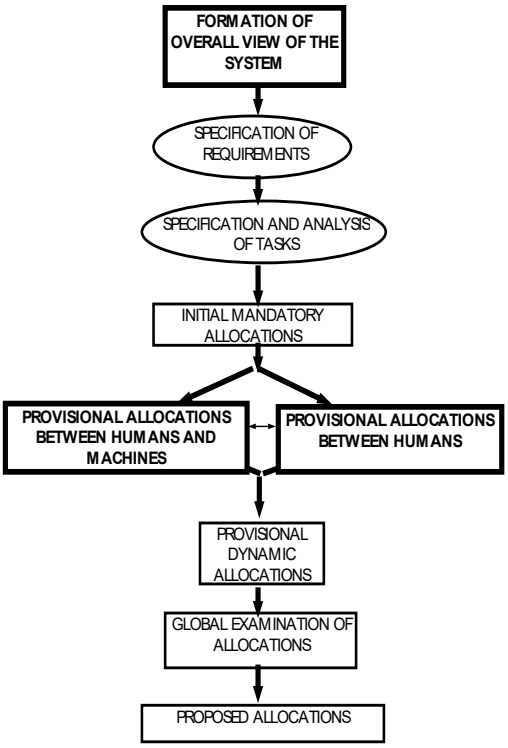

TABLE 3: Decision Criteria for Use with the Method

Category of Criteria	Criteria
Goal issues	Goals of the system, goals of the organization, other goal issues
Organizational issues	Existing/required systems and procedures, existing/required organizational norms issues and practices, organizational requirements, organization structures, accountability, other organization issues
Cultural/environmental issues	Cultural issues, legal requirements, health and safety, data security, national/international political considerations, other cultural/environmental issues
Resource issues	Technology, people, money, other investments (e.g., time), knowledge, skills, education, other resource issues
People issues	Level, type, and number; knowledge, skills, and education; trust; training; social acceptability; other people issues
Task issues	Speed of response, accuracy, frequency, physical demands, cognitive demands, emotional demands, reliability, efficiency, flexibility, operational criticality (likelihood and implications), safety criticality, uncertainty of occurrence, uncertainty of the type of problem, uncertainty of the correct response, variability in performance, task interdependencies, redundancy, other task issues
Job design/work organization issues	Control (e.g., autonomy, responsibility, variance handling, decision making), accountability, variety, skill use and development, workload, clarity of goals and requirements, feedback on performance, communication and social contact, health and safety (mental and physical), flexibility, information exchange and knowledge sharing, situation awareness, understanding of the system, mutual adjustment/team collaboration, role ambiguity and overload, motivation, satisfaction, stress, performance, other job design/work organization issues
Technology issues	Feasibility and cost of automation; maintainability; reliability; level of performance; trust in technology; level, type, and amount of technology; other technology issues

Figure 1: Flowchart and decision -criteria for function allocation (Waterson et al., 2002)

Existing approaches to FA might be limited by their assumptions about the capabilities of technology, but the emphasis of FA occurring within a Sociotechnical System is important to future developments of the concept. However, even though existing FA approaches address the distribution of work across a wider system than solely one human and one machine, they tend to underplay the role of trust, ethical, legal and wider societal issues. In broad terms, these considerations fall under the heading of 'responsible innovation' and raises questions of how-to responsibility and where it sits in a sociotechnical system. With the advent of sophisticated technology that allows humans to operate in teams with these technologies, such as Human-Robot

Teams (HRTs) and Human-Agents Teams (HATs), there is an increasing potential for semi-autonomous and fully autonomous working arrangements.

Responsible AI and FA – shifting the focus

In this paper we focus on interdependencies in human-automation systems and how these might be framed by a concept of responsible FA. Interdependencies play a large part in recent approaches to FA. However, we propose that current approaches to FA focus on the capability of actors and (sometimes) predictability of outcomes. FA does not, generally, address situations when Actors are pursuing competing, contradictory, erroneous goals. Likewise, robots, and semi/fully-autonomous systems are capable of defining their own intent and this intent might differ from that of their human team-mates. In this case the question remains how can FA address the challenges this raises?

Levels of Responsibility

Levels of Automation indicate interdependencies between human and automation, depending on the capability of the automation. From levels of automation, a stage model of responsibility could be proposed, e.g., an agent can perform the action or make a decision within limits defined by its design or user; an agent can suggest action or decision, but the human is responsible for performing this, an agent can suggest decisions or action etc., and human can veto; an agent cannot do action etc. In addition to defining levels of automation in terms of capability, Wickens et al. (2010) added types of cognitive task, e.g., attention (i.e., acquiring information), information integration (i.e., combining information from different sources), decision (i.e., choosing an action or interpretation to perform on the basis of the information), execution (i.e., performing the action or interpretation). This, we feel, allows us to highlight the ways in which FA can apply across human-agent teams, and also suggests likely points at which responsibility can be delegated from human to agent. For example, assume that a (firefighter) Incident Commander is working in a team with uninhabited aerial vehicles (UAVs), and has set these the objective of reconnoitring a large, open plan building such as a factory or warehouse. Assume further that the building is filled with smoke and the objectives involve search for the source of a fire and to find combustible materials. Setting these objectives, the Incident Commander might assume that the UAVs have sensor and navigation capabilities that are sufficient to enable the objectives to be met. Rather than the UAVs continually checking with the Incident Commander as to whether their interpretations are correct, we believe it more likely that they would provide a running update of situation awareness to enable the Incident Commander to develop a response to the fire and the risks identified. The challenge arising from this, is as follows: if the UAVs fail to detect something in the environment that has an impact on the plan, and the impact results in catastrophic consequences, does the responsibility for this lie with the Incident Commander? And, if the responsibility does not lie with the Incident Commander, in what ways are the UAVs culpable (or should responsibility lie with their designers, manufacturers, distributors, maintainers, etc.)?

Approach

Starting from a revision of the FA framework of Waterson et al. (2022) we consider how we might incorporate decision criteria that address the issue of responsibility, and expansion of the sociotechnical system beyond human-machine dyads. This provides an impetus for considering not only decision points in the definition of a task (e.g., where the task involves the choice between actions in response to the situation), but also ‘responsibility points’ (e.g., where the outcome of the decision becomes identifiable and can be evaluated in terms of acceptability). This contrast between decision and responsibility points in the timeline of a mission could be considered a priori in mission planning. For example, in the Incident Commander example above, decision points could relate to navigation, sensing, interpretation. It could be feasible to assume that the semi-autonomous UAVs would be capable of acting appropriately at these decision points. However, the

responsibility points might involve the decision to change firefighting tactic which might alter the risk to personnel. In this respect, the mission plan would define which decision outcomes need to be evaluated in terms of acceptable risk.

The notion is that there are responsibility points (that operate by analogy with decision points). At these responsibility points, there are additional sociotechnical systems considerations. First, which part of the sociotechnical system defines 'acceptable risk'? Second, which part of the sociotechnical system predicts the likelihood of outcomes from specific decisions? Third, which part of the sociotechnical system evaluates the decision outcomes?

The definition of 'acceptable risk' could be made in advance of specific operations, probably through Standard Operating Procedures, and that responsibility for this definition lies in the organisations command hierarchy. That is, if an agent demonstrably follows these SOPs and demonstrably responds within the agreed limits of acceptable risk, then responsibility ought to lie with the organisation rather than the individual. We recognise that such an argument raises legal and ethical concerns (particularly in terms of the assumption that the SOP would be appropriate to a given situation). However, starting from this assumption means that we can define responsibility points in a mission. It also means that computer agents could be tasked with decisions and actions that operate within the SOPs. While we are not arguing that the agents are responsible, it does imply that they can potentially be tasked with decisions and actions that have outcomes which could affect 'acceptable risk'. However, this is not to claim that the agents are engaged in moral or ethical decision making. Rather, it is to assume that the SOPs contain within them the moral and ethical imperatives that will constrain decisions and actions. Having outlined an argument as to how computer agents could make such decisions, we now want to argue against this. The reason for urging caution on permitting agents to make decisions that affect responsibility points it that one would need to guarantee that the situation is entirely defined in a way that fits the SOPs. In any complex incident, it is likely that the situation will unfold in ways that are unanticipated. Consequently, rather than allowing the agent the opportunity to make such decisions, these would always need to be referred to a human in the team who held sufficient authority and would be prepared to take responsibility for the outcomes. This, in turn, requires sufficient trust in all actors (human and computer) engaged in decisions that affect these responsibility points.

The discussion so far has implications for what is defined as a 'function' within FA (i.e., where the agent, human or machine, is responsible for the outcome, not just who does something, of an action of decision), as well as the wider issue of trust in automation. We define trust in the manner of Lewis and Marsh (2022) who construe it in terms of the following components:

- 'Trust' involves a 'model' of teammates held by members of a team, defined by:
    - Capability: is an agent (or human) is able to perform a given function at a given point in time (in its own or its team-mates opinion) and is it available?
    - Predictability: what is the probability of success of completing the function, and will it do this is way that team-mates expect?
    - Integrity: will the team-mate pursue individual or team goals, and will performance by within moral, legal, and ethical constraints?

In our ongoing work, we argue that 'trust' is dynamic and will need to be sufficient to support collaboration with the team, i.e., trust is satisficed (Baber et al., 2023; Hunt et al., 2024). This means that the members of a team ought to sufficiently confident that their team mates are acting in the interests of the wider sociotechnical system (as far as possible) and are able to identify which team members are involved in specific decision points and responsibility points.

Initial Studies

As a simple 'proof of concept' of the notion of responsible FA, we built an agent-based model using NetLogo model (Baber et al., 2023). In this model, 3 Agents explore a Maze. Each Agent has specific Functions that include navigating the maze (we define this using simple functions, such as left-hand-on-wall where the robot will follow a wall unless it is able to turn left, and if it reaches the end of a line it turns left until it can move forward). In addition to navigation, each agent has a specific function. In this case the yellow Agent (Figure 2) is 'trapped' by a red square and needs to be 'rescued'. For the blue agent, the red square is a token to be collected. Collecting the token would result in the coincidental 'release' of a trapped yellow agent. The green agent, needs to seek trapped agents and release them from the red square and 'rescuing' other agents is a priority.

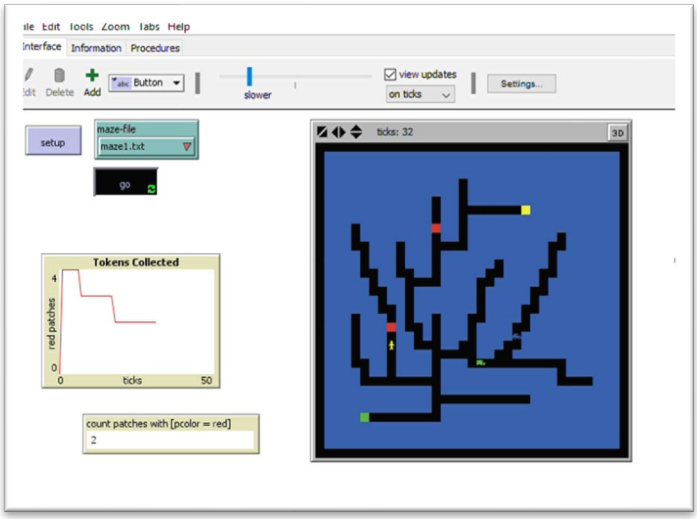

Figure 2: NetLogo Model – 3 Agents exploring a Maze (Baber et al., 2023)

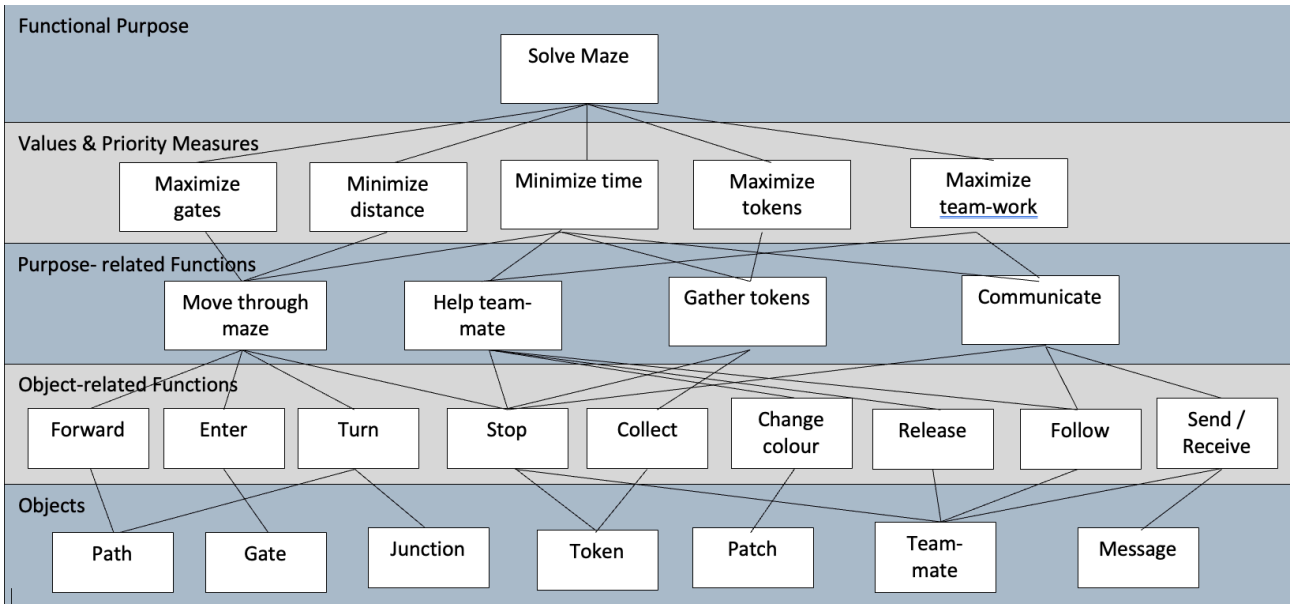

Figure 3: Work Domain Analysis diagram sketching out a description of the constraints that govern the purpose of the agents and the function of the systems as a whole.

From Figure 3, we can define several decision points that can be deduced from the Object-related Functions. For example, if it is not possible to move forward, then turn then; if there is a gate on the left, then enter it; if there is a junction, then take the left turn. Further, we can define several responsibility points from the Purpose-related Functions. For example, if a team-mate is trapped then stop your current task and release them; if tokens are available, then collect them; if you are trapped, then communicate with your team-mates. In terms of responsibility, we propose that the consequences of decisions would be internal to the team and do not affect a wider sociotechnical system. This means that if the choice taken at a responsibility point is not for the good of the team (i.e., continuing to collect tokens to boost your own score, rather than stopping to help a team-mates) then the outcome only affects the performance of the team.

One of the interesting outcomes from the simulation was that if an agent changes behaviour (e.g., becomes less attentive to details, fails to identify risks, prioritises non-altruistic goals, etc.), the overall outcome (a failed mission) is a consequence of a shared failure of the team, and responsibility is also shared. The action of team-mates serves to constrain the choices available to individual agents. In this context, overall responsibility exists 'between' team-mates.

Experiments in Human-Robot Teams

The second part of our paper concern a set of experiments which were designed to manipulate trust in human-robot teams (HRTs). We analysed data from HRT experiments focused on trust dynamics in teams of one human and two robots (Figure 4 shows an example of a participant working with the floor-based robots), where trust was manipulated by robots becoming temporarily unresponsive. In this case, the mission was to collect tokens (jointly or independently with the robots) and lack of response from the robot compromised the ability to jointly collect tokens. This compromised ability, while it was the result of a deliberately induced technical failing, was often perceived by the human participants in terms of the robots making a choice. For example, the robot was perceived to not want to help the human or was perceived as being too busy performing its own tasks to offer help. From this, the robot was ascribed some degree of agency that allowed it to make a choice as to whether or not to help. The choice could be considered as a decision point, but could also be interpreted in terms of a responsibility point. That is, if the mission was for a team to gain a certain number of points and (due to lack of responsiveness) the team was not able to meet the required number, the robots could be perceived to be responsible for the team's poor performance.

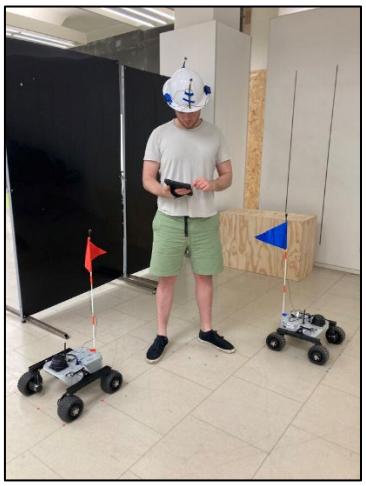

Figure 4: Experimental set up – participant and robots (based on Hunt et al., 2024).

Discussion

Our research so far shows some efforts to shift away from traditional FA and raises many questions about agency and responsibility in human-machine and robotic systems in general. There are currently no established HFE-oriented criteria or structured questions regarding how we might allocate 'responsibility' between humans, machines, and numerous forms of new technology (e.g., robots, drones). Likewise, the interdependencies which might exist in human-machine interaction are only starting to emerge (e.g., robotic systems which have the ability to 'trust' their human operators). These sorts of concerns raise a number of thorny, philosophical questions (e.g., how to allocate 'blame' and 'accountability' when something goes wrong when humans and machines jointly carry out a task). Considering FA in terms of decision points and responsibility points (in addition to considering functions) and then considering how humans and agents in a team are entrusted to respond to such points, provides a first step in incorporating responsibility into FA.

Acknowledgements

The research reported in this paper is supported by grant EP/X028569/1 'Satisficing Trust in Human Robot Teams' from the UK Engineering and Physical Sciences Research Council (EPSRC). This project runs from 2023 to 2026 and involves the Universities of Birmingham, Bristol, Loughborough and UCL.

References

Baber, C., Waterson, P.E, Milivojevic, S., Maynard, S., Hunt, E.R. and Yusuf, S. 2023. Incorporating a 'ladder of trust' into dynamic Allocation of Function in Human-Autonomous Agent Collectives. In the 14th Organizational Design and Management Conference (ODAM). Bordeaux, France

Brown, S. (2023). Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI. Sloan Management Review, May 23, 2023.

Fitts, P. M. (1951). Human engineering for an effective air navigation and traffic control system. Washington, DC: National Research Council.

Hunt, E. *et al.* (2024). Co-Movement and Trust Development in Human-Robot Teams. https://doi.org/10.48550/arXiv.2409.20218.

Jian J., Bisantz A. & Drury C. (2000). Foundations for an empirically determined scale of trust in automated systems, International Journal of Cognitive Ergonomics, 4, 53–71.

Lee, J.D. and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40(1), 153–184.

Lewis, P.R. and Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence, Cognitive Systems Research, 72, 33-49

Mayer, R.C., Davis, J.H. and Schoorman, F.D. (1995). An integrative model of organizational trust, The Academy of Management Review, 20, 709-734

Waterson, P.E., Older-Gray, M. and Clegg, C.W. (2002). A sociotechnical method for designing work systems. Human Factors, 44, 3, 376-391.

Wickens, C.D., Li, H., Santamaria, A., Sebok, A. and Sarter, N.B. (2010). September. Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 54, No. 4, pp. 389-393). Sage CA: Los Angeles, CA: Sage Publications.