THE ROLE OF THE HUMAN IN AN AUTONOMOUS SYSTEM

R.D. Alexander*, N.J. Herbert†, T.P. Kelly*

*University of York, UK, robert.alexander@cs.york.ac.uk, tim.kelly@cs.york.ac.uk

†BAE Systems Military Air Solutions, UK, Nicola.Herbert@baesystems.com

Keywords: autonomous, unmanned, human

Abstract

To analyse a system, we must define it. We must identify what's inside the system, and what's outside in the environment. This distinction has been a source of controversy for some time; for example, is the operator part of the system, or external to it? The issue is tied in with several other contentious topics, such as the relationship of operational personnel with rules and regulations (scripted actors, or creative improvisers?). Growing interest in Autonomous Systems (AS) looks set to force the issue by creating new ambiguities. Of particular concern is the role that human operators, supervisors and peers play with respect to the AS. There are no easy answers here, but we will identify a simple, pragmatic option, and show how a truly system-theoretic modelling process avoids the worst of the issue. We then look to future prospects.

1 Introduction

Interest in autonomous systems (AS) is growing, and with this comes the need for safety engineering techniques. The authors have defined (in [1]) a basic approach to deriving safety requirements for AS. However, in doing so we have run into a problem; a problem that has always been present, but is doubly challenging when AS are involved.

The heart of the issue is the need to define a system boundary. To analyse a system, you must define the system – what's in the system, and what's outside (in "the environment"). This is well known, and we have many ways of dealing with it.

When we're analysing AS, it is not obvious how humans fit into the system. In theory, humans could be taken out of "the system" entirely, and described as "the operator" or put in "the environment". In practice, AS need humans for a whole variety of reasons, and thus force us to answer the question of how the two fit together.

2 Possible Human Roles

There are as many potential roles for humans as there are system concepts. Here, we will identify some of the most salient ones. All of these roles have safety implications – if we ignore any of them, we will fail to address some aspect of safety risk.

2.1 Human as a Safety Function

One role for the human is to stand by, and then intervene when problems arise. Effectively, the human provides a safety function. This can work, and has a long history (in human supervision of other humans and human supervision of low-autonomy machines). It does, however, require us to argue that this function will be delivered *reliably*. This aspect is often ignored or downplayed in discussions of AS.

A subset of this is the use of a human because it's legally mandated or a political necessity. The latter is no minor consideration, especially for novel systems such as autonomous vehicles (AVs). Many existing systems (e.g. most national military forces, and the civil aviation industry) are built around the assumption that in any situation someone (human, singular) is in charge, and is by default responsible for whatever goes on.

2.2 Human Does What AS Can't Do

It is clear that many in the autonomous systems community want to make the leap to high autonomy. There is an obvious attraction in building systems that take very general instructions (like Clough's "Go Scud hunting" [2]) and execute them without intervention, or even without supervision. It is clear that such systems have substantial benefits over the interactive, inhabited and remote-control systems that dominate at the moment. Unfortunately, however, high autonomy is very difficult to achieve.

To perform tasks that we can't automate, a human can be brought into the system as a provider of a particular function or capability. This may be a safety function (as in 2.1, above), or it may be a function that's needed for the AS's primary mission. Sometimes it will be both. In this role, the human has been incorporated into the *mechanism* of the AS, and therefore we must analyse them as such.

2.3 Human as Necessary General Intelligence

Many people, including some safety engineers, have a model of human operators as mere followers of procedures. They assume that we can define procedures that will keep the system safe. This view is long-established and convenient for analysis (humans can be treated as procedure-following machines that vary somewhat in their performance). It is also quite wrong, as recent research has shown (see Wright in [3], and Dekker in [4]). Real-world procedure-following requires significant leaps of interpretation and judgement. It is not something that our current machines and computers can do.

Some authors go further, and suggest that safety in complex systems is achieved only by creative improvisation wholly beyond any possible procedures. This idea is at the core of Resilience Engineering (see Hollnagel et al in [5]). If we remove a human then we remove this key component of safety.

It is clear that current proposals for autonomous systems do not provide human-level intelligence. Most technologies being developed to increase autonomy are very specialised – they provide a single new capability, such as object identification or visual localisation. Even the most advanced planning and learning algorithms require extensive configuration and specialisation before they can be used for a given task. They do not provide “general intelligence”.

Many authors have claimed that human-level AI will never be achieved – examples include Dreyfus and Dreyfus [6], Searle [7] and Penrose [8]. All of these arguments are controversial – see McCarthy [9] for responses to those cited here. Ultimately, we don’t know what will be possible in the future. It is clear, however, that after 40 years of ‘AI’ research we are nowhere near a general solution. Human thinking still requires human thinkers.

2.4 Human as Mediator and Liaison

A human can act as a translator between the AS and other humans. This has two aspects – human as translator of communication, and human as interpreter of AS state. In practice the two go together – we need more than just command, more than just information – we need either control or collaboration. Both of these require both aspects.

This role is central to the unmanned air vehicle (UAV) strategy adopted by the Civil Aviation Authority (see their position in [10]). It is likely that many AS will need a human to provide this function. Indeed, even if we have safety-critical speech synthesis and interpretation we will still need the human ability to understand their interlocutor and update their mental model. Intelligent communication is not a simple mechanical process.

This “mental model update” may be harder than we’d like – AS have the potential for incredibly complex state. It is clear that when operators misunderstand AV state they create a safety risk - see Johnson [11] for a simple example this leading to an accident.

2.5 Human as Maintainer

Machines need maintenance, and it’s humans that provide that maintenance. In the long term, “high autonomy” may come to include a measure of self-repair, but for the near future AS will need humans to look after them. Maintenance is, of course, a critical safety issue – Johnson, in [12], describes two errors in maintenance that caused UAV crashes.

2.6 Human as Peer

Regardless of their explicit control strategy, most AS will have humans (and low-autonomy systems under their direct control) as peers. A few specialised AS (e.g. those performing space exploration) will operate alone, but in all other cases humans and AS will have to share resources, cooperate, and collaborate. It follows that they will have to predict each other’s behaviour, at least in safety-critical aspects.

This need for peer predictability can require an extra set of bounds on an AS’s behaviour: as well as the bounds required to achieve its primary mission, and the bounds required to avoid direct accidents, there can be bounds required to let peers can safely interoperate with you.

We can see examples of interoperability bounds in the UK Rules of the Air [13] and the UK Highway Code [14]. Both achieve overall system of systems safety by prohibiting some behaviour that might, on its own, be perfectly safe. For example, it is safe for a car travelling on empty roads to align itself with either side, or to the centreline, but in order to allow safe simultaneous traffic in both directions the Highway Code mandates that all vehicles drive on the left.

Because we need to support humans as peers, we may find that we have to bound AS behaviour in ways that move us away from optimum flexibility and performance. It is quite possible to have an AS that is *too* clever and flexible to be safe.

2.7 Human as Rescuer

If an AV becomes immobilised (or an immobile AS needs to be retrieved) then it falls to a human to perform this task. This is part of the AV operational lifecycle that is often ignored, and at cost – Johnson, in [15], describes a situation in Afghanistan where a British Army officer was killed attempting to retrieve a downed UAV.

3 Challenges Driving the Issue

3.1 Managing Complexity of Analysis

In order to keep their work manageable, safety engineers need to minimise the architectural assumptions they make. In particular, safety analysis should strive towards the “naked man” situation – we should start by analysing the system without any pure safety features (such as an operator who is only needed for safety). It follows that we should keep the human roles as simple as possible, giving them responsibilities only when it is clear that we need to.

On a related note, we want to minimise the size and complexity of the "system" that we analyse. But we don't want to miss any phenomena (any hazards, any casual pathways) that are important. Getting the right boundary is a key part of resolving this tension.

3.2 Using Humans Intelligently

As ever, we must avoid false panaceas. The most salient in this context is "It's ok, the operator will prevent that" as a solution to every hazard. Humans are as often unreasonably glorified as they are unfairly maligned – we need a sober (but not pessimistic) view.

We need to understand what humans are good at, and what they find difficult. The "control-room problem" in the nuclear and chemical industries is well-established – plant operators struggle with vigilance during quiet periods, then suddenly have to intervene but don't have the understanding or practical experience they need. Reason comments in [16] – *"… if a group of human factors specialists sat down with the malign intent of conceiving an activity that was wholly ill-matched to the strengths and weaknesses of human cognition, they might well have come up with something not altogether different from what is currently demanded of nuclear and chemical plant operators."*

The control-room problem is a serious concern for UAVs, specifically, because the "safety function" role from Section 2.1 is much more like the role of plant operators than that of pilots and ATC operators. If "supervision" means doing very little most of the time, then springing into action when the situation degrades, then supervisors will suffer from the problems noted above. By contrast, pilots and ATC operators have highly *interactive* roles which avoid those problems. We may be ill-prepared for the sea change in operator roles that high autonomy will bring.

There is also the question of how much we understand of what humans do now, which is crucial when we want to replace them with AS. There is reason to believe that we don't understand much, and that we systematically fail to understand. For example, Dekker [17] discusses work on the role of flight progress strips in air-traffic control. One empirical study [18] reported few problems working without strips in the short term, but failed to address the diverse ways that operators used the strips in the medium term (these uses were uncovered by Ross [19] in a qualitative study). Dekker claims that this is a consequence of a systematic ideological prejudice – we are obsessed with "real science" (numbers and experiments) to the exclusion of the qualitative anthropology we need.

3.3 Keeping on Top of Risk

Unmanned vehicles reduce safety risk by not having a crew on board. Against a fixed standard of tolerable risk to life, this gives us some "spare budget" that we can spend. For example, we can sacrifice operator control or component reliability and achieve the same overall safety as an equivalent manned vehicle. Once we have done that, however, the risk budget is "spent". We cannot then "spend" it on anything else, such as reducing overall process rigour. In any case, if we operate UAVs over populated areas the "spare budget" may be small in the first place. It seems that many people have dubious intuitions in this field, and seek to spend the risk budget again and again.

3.4 Role Boundaries and Role Transitions

The roles described in Section 2 do not have hard boundaries. In practice, the same human may be asked to fulfil several roles (either sequentially or at the same time). Under pressure, humans will take creative action and will (not may) step outside the remit of explicitly-defined roles. We need ways of reasoning about human behaviour in the face of this flexibility.

3.5 Comprehension Problems

Autonomous systems, and the technologies used to create them, provide new ways for machines to have increasingly complex state and (crucially) ever-more complex *response* to that state. This makes human comprehension of an AS increasingly difficult. More than ever before, we will be limited not by what we can create but what we can *control*.

3.6 Making the Right Tradeoffs

Finally, we want have AS that are as capable as possible, not pale shadows of their manned equivalents. This means that we need to exploit the best of both the humans and the machines. We have to do this without the long-standing exemplars we have for other systems, because much of what we are doing is new and those exemplars do not exist yet.

This is particularly crucial when we consider the military roles where AS are of most current interest. In those situations, every performance failure has potential to cause loss of life. People do not just die from accidents, so we must not let safety automatically trump performance.

4 Ways Forward – the Adequate, the Good, and the Future

4.1 The Adequate – Combined Autonomous Systems

A basically adequate approach is to use expected scenarios as your unit of analysis, and to define "the system" as the AS itself (including any external control interface) and a responsible operator. We call this definition of the system the "Combined Autonomous System" (CAS) (after Hollnagel's "Joint Cognitive System" concept [20]). The scenario bounds the environment (crucially, it shows you what stimuli the system must respond to), and the CAS concept greatly constrains what counts as "mechanism" to be analysed. One benefit of this is that the CAS is externally indistinguishable from an equivalent manned system – its autonomous nature becomes a concern for architecture and design, not for top-level requirements.

As noted in Section 2.1, having a single operator is often legally and politically pragmatic. However, the danger with this approach is that it leads you towards one operator, one site of blame, and a generally poor approach to safety (see Dekker [21] and Leveson [22] for further discussion of this).

4.2 The Good – Whole-System Models

If you have a true system-theoretic model (e.g. as promoted by Leveson in [22]), then the man-machine distinction blurs. Both are parts of the system, both have roles and responsibilities, capabilities and constraints. The hard global issue of human role is replaced by lots of smaller local ones. Challenges remain, but they are smaller and more manageable.

A barrier to whole-systems modelling is unavailability of methods and mental tools for performing it. Although there are undoubtedly "whole-systems engineers" out there, it is hard to point at a textbook that guides a beginner through this kind of modelling. For example, Leveson's book [22] remains unfinished.

Whole-systems modelling is inherently difficult. It requires knowledge of diverse practical and engineering domains and a wide technical skill set. Many engineers have, by virtue of working only in one area, a narrow range of skills and experience. Some of this can be tackled by cultivating good systems-engineering teams (where the members have complimentary skills). Specific skills aside, systems thinking is cognitively difficult and is likely to remain so (see Ring in [23] for a pessimistic view of the potential for developing a whole-systems capability).

Similarly, many engineers are not familiar and fluent with the models and notations that can help to make this modelling comprehensible. Often, they are stuck with functional breakdowns, fault trees and network diagrams, all of which are an awkward fit for the dynamics of complex systems. The rise of MODAF[1] is hopeful, but the difficulties encountered in producing good MODAF models bear witness to the difficulty of the systems-theoretic undertaking.

4.3 The Future – the Role of Humans

In the long run, we can expect to get better idea of the role of human intelligence in maintaining system safety. Certainly, it is an area getting increasing attention (e.g. the recent book by Reason [24]).

We need models of human capability that are neither the ever-popular "humans are stupid and irrational" (e.g. as in Sutherland [25]), nor the similarly enduring "good people don't need procedures or machines". We need models that capture cognitive capability rather than just describing physical dimensions and the situations that reduce performance (this latter is a crude but reasonable characterisation of Def Stan 00-25 [26]). We need models that can draw out all the differences between a human, a computer, a trained animal and a rock.

4.4 The Future – Advanced Analysis Techniques

Better analysis techniques will make it easier to model complex human-AS interaction. Three areas of great potential are constructive simulation, model-checking and synthetic environments. All of these allow us to perform virtual experiments without the cost and risk of field trials, and model-checking and constructive simulation can also explore a much larger set of scenarios than we otherwise could. They are all good for creating thought experiments – for "animating assumptions". However, none of them can guarantee correspondence to reality – they cannot guarantee that events in the model will match events in the real situation.

Of course, not even the simpler equation-based models used throughout the natural sciences since Newton's day are immune to this problem. For example, many equations that describe natural phenomena assume that the output is linear (or a simple curve) with respect to some input. These assumptions cannot always be validated.

These advanced computer-based techniques, however, introduce a new danger because of their flexibility. A programmer can build a computer model of "anything" to any level of notional detail. Often, the set of factors modelled in a simulation is limited only by programmer time (and their ability to mentally manage an ever-growing model). They can easily create a model that is incomprehensible because of its complexity.

Computer models are also rife with psychological 'traps' for the unwary; aspects of their nature that give us misleading cues about their validity. Programmers can easily create attractive visualisations, leaving a non-technical audience vulnerable to Roman's "Garbage In, Hollywood Out" (GIHO) [27]. For technical workers who are not modelling experts, there are traps like the individual-based fallacy – the assumption that we can identify the "entities" or "objects" in the modelled system, derive their properties, then combine them and see "what will emerge".

Synthetic environment models can use human participants to explore human behaviour. The other forms of model cannot,

[1] Ministry of Defence Architecture Framework

by design, and the challenge there is capturing aspects of human variability (and capability) that are useful for analysis.

Hopefully, we are past the heyday of boundless enthusiasm for computer models. Perhaps we can look forward to progress in these areas with a practical but positive attitude.

4.5 The Future – New AS Control Concepts?

Perhaps the most important future development would be finding new control concepts for AS, concepts that keep the human continuously in the loop and thereby let them gain the skills they need to tackle emergency situations. These would have the added bonus of keeping the operators awake.

It is likely that the human role in this case would not be something we have described here – it would be wholly new.

5 Conclusions

There is no escaping the fact that we need boundaries. Practical limits mean that some things will forever be outside the system, relegated to “the environment” and modelled in a simplistic way. Nevertheless, there remain some thorny questions and one of them is “Where do the humans go?”

The question can’t be dodged, but we can offer a pragmatic starting point in the form the CAS concept. This has a number of benefits, and embeds an assumption (that of a single responsible operator) that is useful for safety and is politically and legally expedient.

One upside is that if your modelling and analysis is good, in that you truly take a “systems” view, then this issue becomes much easier. This is hopeful for the future. Nevertheless, it won’t just evaporate the wider constellation of problems, so it is likely that growing use of AS will force some hard thought and hard research.

Acknowledgements

The work in this paper was funded by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence.

References

[1] R. D. Alexander, N. J. Herbert, T. P. Kelly, "Deriving Safety Requirements for Autonomous Systems," in *Proceedings Of the 4th SEAS DTC Technical Conference*, Edinburgh, (2009).

[2] B. T. Clough, "Metrics, Schmetrics! How The Heck Do You Determine A UAV’s Autonomy Anyway?," in *Proceedings Of The 2002 PerMis Workshop*, NIST, Gaithersburg, MD, (2002).

[3] P. C. Wright, J. M. McCarthy, "An analysis of procedure following as concerned work," in *Handbook of cognitive task analysis*, E. Hollnagel, Ed.: Lawrence Erlbaum, (2003).

[4] S. Dekker, "Why Don't They Follow the Procedures?," in *Ten Questions About Human Error: A New View of Human Factors and System Safety*: CRC Press, (2004).

[5] E. Hollnagel, D. D. Woods, N. Leveson, *Resilience Engineering: Concepts and Precepts*: Ashgate, (2006).

[6] H. L. Dreyfus, S. E. Dreyfus, *Mind Over Machine*: Free Press, (1988).

[7] J. Searle, "Minds, Brains and Programs," *Behavioural and Brain Sciences 3*, pp. 47--424, (1980).

[8] R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*: Penguin Books, (1989).

[9] J. McCarthy, *Defending AI Research: A Collection of Essays and Reviews*: Cambridge University Press, (1997).

[10] UK Civil Aviation Authority, *CAP 722 — Unmanned Aerial Vehicle Operations in UK Airspace: Guidance*: The Stationery Office, (2004).

[11] C. W. Johnson, C. Shea, "The Hidden Human Factors in Unmanned Aerial Vehicles," in *Proceedings Of the 26th International Conference on Systems Safety*, Vancouver, Canada, (2008).

[12] C. W. Johnson, "Act in Haste, Repent at Leisure: An Overview of Operational Incidents Involving UAVs in Afghanistan (2003-2005)," in *Proceedings Of the Third IET Systems Safety Conference*, Birmingham, UK, (2008).

[13] "ICAO Annex 2: Rules of the Air," International Civil Aviation Organization, (2007).

[14] "The Highway Code," Road Safety Directorate/Driving Standards Agency, (2004).

[15] C. W. Johnson, "Military Risk Assessment in Counter Insurgency Operations: A Case Study in the Retrieval of a UAV Nr Sangin, Helmand Province, Afghanistan, 11th June 2006," in *Proceedings Of the Third IET Systems Safety Conference*, Birmingham, UK, (2008).

[16] J. Reason, *Human Error*: Cambridge University Press, (1990).

[17] S. Dekker, "Will the System Be Safe?," in *Ten Questions About Human Error: A New View of Human Factors and System Safety*: CRC Press, (2004).

[18] C. A. Albright, T. R. Truitt, A. B. Barile, O. U. Vortac, C. A. Manning, "How Controllers Compensate for the Lack of Flight Progress Strips," Oklahoma University Norman Dept of Psychology, (1996).

[19] G. Ross, "Flight Strip Survey Report," The Australian Advanced Air Traffic System Operations Instructor, Air Traffic Services Australia, (1995).

[20] E. Hollnagel, D. D. Woods, *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*: CRC Press, (2005).

[21] S. Dekker, *Ten Questions About Human Error: A New View of Human Factors and System Safety*: CRC Press, (2004).

[22] N. Leveson, *System Safety Engineering: Back To The Future*, (2002).

[23] J. Ring, "Strategy for a Whole System Modelling Capability," in *Proceedings Of the 3rd IEEE Systems Conference*, Vancouver, Canada, (2009).

[24] J. Reason, *The Human Contribution: Unsafe Acts, Accidents and Heroic Recoveries*: Ashgate (2008).

[25] S. Sutherland, *Irrationality*: Pinter & Martin Ltd, (2007).

[26] "MoD Defence Standard 00-25 - Human Factors for Designers of Systems," Ministry of Defence, (2000).

[27] P. Roman, "Garbage In, Hollywood Out!," in *SimTecT 2005*, (2005).