Human Factors 2024
8th – 9th October 2024
Wageningen, Netherlands

PERILS OF THE HUMAN OPERATOR IN MARITIME AUTONOMOUS SYSTEMS

A Frizell BMT, UK
A Weir, UK

SUMMARY

This technical paper explores some of the Human Factors (HF) challenges surrounding the transition to the deployment of highly autonomous vessels, with an emphasis on the Defence domain. The paper explores the significance of the human operator as a safety influencing factor, and the importance of allocation of function. The issues raised will be of relevance to all maritime vessels that incorporate a high degree of autonomy, but still depend on some degree of onboard crewing or offboard control. This paper considers how we might define the levels of autonomy within a complex system, using a warship as an example, and identifies human-autonomy relationships to be avoided. It further considers the cultural changes required to ensure that the challenges of autonomy can be overcome.

1. UNDERSTANDING THE PROBLEM SPACE

The Highly Autonomous Warship

Over the next 20 years, BMT expects to see the increased adoption of offboard systems, coupled with higher levels of automation and high trust systems on board Naval vessels. However, human interaction will still be required to perform certain operational and decision-making functions due to ethical or operational complexity reasons. To support this, highly autonomous naval vessels will need to enable a small number of personnel to conduct operations and make key decisions while the vessel itself will largely operate autonomously with minimum operator interaction. This paper investigates some of the HF aspects of the development of highly autonomous naval vessels.

Defining Automation and Autonomy

Autonomy and Automation are two terms often used in similar contexts. At a simple level, 'automation' simply describes the use of machines and computers that can operate without needing human control. However, this covers a broad array of capability. The ability to operate free from human control need only to refer the specific process to which that automation is applied, defined by the start and end point of the automated process. The process may be short and simple or long and complex, but it will typically exist as part of a larger task or system with which the process must interface, receiving inputs and providing outputs. An automated process will still require periodic maintenance and may still require a degree of oversight from a human operator, even under normal operating conditions. The automated process will also be subject to external forces and pressures from the environment in which it operates, which may require human intervention to maintain system function.

Autonomy, in general parlance, is the ability to act and make decisions without being controlled. At first glance this appears similar to automation, but the key element here is to 'make decisions'. In the context of an automated process, autonomy therefore denotes that it incorporates sufficient Artificial Intelligence (AI) to be able to manage environmental and contextual variability on its own, responding and adapting without human assistance in order to maintain operation within acceptable performance and operating parameters.

Another key aspect of autonomy is the extent to which it is free from human control, which can be considered in two dimensions: time and scope. An automated process will be time bound, with a defined start and end point outside of which external management is required. Autonomy implies a greater longevity of independence, incorporating a level of machine learning such that the system is able to make decisions about system behaviour in future scenarios based on learned experiences, adapting over time. With regards to scope, an automated process may only manage a subset of a much larger and complex overall task. The author argues that true autonomy implies that we are describing more than a simple process; rather we are describing a system that performs a complete task in its own right. To illustrate this latter point, consider automation in a road vehicle. Basic automation may manage a particular control parameter (such as speed control) but this is meaningless outside of the context of the overall driving task. More complex automation may manage all control parameters, but only under certain conditions. A fully autonomous vehicle would need to manage the driving task as a whole.

For the purposes of this paper the terms automation and autonomy are defined as follows:

- Automation is a general term for where a machine carries out a defined process as programmed, where human intervention is not required under normal operating conditions.

- Autonomy is a subset of automation incorporating a level of Artificial Intelligence (AI) such that the machine is able to adapt to changes in the environment in order to maintain system function and make decisions about system behaviour in future scenarios based on previous experiences; and where the automation performs a complete task in its own right.

Defining the overall task in which automation sits may appear simple, but in practice it is often not straightforward. Any system exists to some extent within a larger system, which itself will be part of a yet larger system. Defining a complete task is therefore somewhat subjective and may require defining a boundary that later shifts. This paper explores some of the reasons for potential ambiguity and the challenges associated with this. It draws upon examples of automation within the automotive domain and transposes lessons learned to the context of a highly autonomous naval warship. It further explores the issues surrounding the definition of the boundaries of system capability and the implications for when systems deemed to be operating without human intervention do, in fact, require human intervention.

Crucially, automation relates only to the specified process under control and makes no assertion that external assistance is not required periodically or at the start and end point of the process. Systems will also operate without human assistance only insofar as they remain within tolerable system boundaries. Autonomous systems may be more resilient to changes in the environment or operating context but will still be governed by limits to their adaptive ability, whether or not these limits are known in advance, and will still be bound by the extent of their remit of authority and capability. Should the system/environment/context stray outside of these conditions then the autonomy may not function as intended (if at all) and external intervention will be required for acceptable performance to be maintained. The degree to which human intervention will be possible in such a scenario is not straightforward and will be explored in more detail in the following sections. The critical aspect is the boundary of capability, whether for automation or autonomy, and the context of the wider system in which it sits.

The Goals of Autonomy

It is important to establish the purpose of autonomy in order to provide a frame of reference for determining the success of its implementation. Warships are, by their nature, potentially subject to hostile action, with the accompanying potential for loss of human life. First and foremost, autonomy (and automation) offers the potential to reduce personnel numbers on board and therefore reduce the exposure of personnel to hazardous situations. In that regard, reduction in head count is a goal of autonomy in its own right. This may also lead to reduced costs, although cost reduction may or may not always be seen as a primary driver.

Beyond that, the goals of autonomy become more specific to the particular system in which they are introduced, although it is perhaps universally true that autonomy and automation will be expected to perform at least as well as, or ideally better, than an equivalent system with human operator(s). In that regard the implementation of autonomy will (or certainly should) be designed to play to the strengths of machines whilst minimising the exposure of the system to human weaknesses.

Allocation of Function

An allocation of function analysis is often used to determine how to allocate jobs, tasks, functions, and responsibilities for the system or capability under analysis. The advantages and disadvantages of the task being performed by a human machine is central to the analysis. There have long been efforts to classify human and machine capabilities according to their strengths and weaknesses. This is sometimes referred to as 'humans are better at, machines are better at' (HABA-MABA). Perhaps the most commonly cited of these is the Fitts List, published in 1951 (Fitts, 1951).

Table 1. Fitts List

HABA	MABA
1. Ability to detect a small amount of visual or acoustic energy	1. Ability to respond quickly to control signals and to apply great force smoothly and precisely
2. Ability to perceive patterns of light or sound	2. Ability to perform repetitive, routine tasks
3. Ability to improvise and use flexible procedures	3. Ability to store information briefly and then to erase it completely
4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time	4. Ability to reason deductively, including computational ability
5. Ability to reason inductively	5. Ability to handle highly complex operations, i.e. to do many different things at once.
6. Ability to exercise judgment	

It is not the intention of this paper to go into the details of HABA-MABA in the current era, but suffice to say that, despite the developments in AI, the Fitts List is remarkably relevant today considering how long ago it was conceived and the

subsequent development in machine capability. As much as the HABA list is still broadly true, it may be tempting to think that as machine learning and AI become ever more capable it is inevitable that machines will eventually supersede humans in all capacities. Whilst this may or may not be true, the context of this paper is the highly autonomous warship, not the entirely autonomous warship, and as such considers the shorter-term future where humans are still a critical element, requiring their strengths to be brought to bear. Human strengths and weakness are of course (at least in the absence of human augmentation) unchanging. For example, humans are universally poor at performing repetitive, mundane tasks over extended periods of time. Equally, humans are not suited to monitoring such processes. It may sometimes seem as though humans are viewed as nothing more than an inconvenience by some designers, a source of potential error to be marginalised and protected against lest they be allowed to cause any harm. One of the perils of automation is often that the operator is excluded from the process, but not from the larger task. As previously discussed, processes do not exist in a vacuum, but rather as part of a wider system and capability. The fact that humans are still involved in this wider system testifies to the dependence we still have on human strengths such as ingenuity, adaptability and problem solving. Humans are currently a critical component of most systems, perhaps all systems when viewed at a sufficiently broad level, and without them those systems would surely fail. We depend on humans to keep systems working, effectively and safely; and must ensure that humans are integrated in such a way as to maximise their ability to do so. Automated and autonomous systems that exist as part of a larger socio-technical system must be designed on the basis of this wider context.

Levels of Autonomy

As mentioned previously, ostensibly automated processes rarely operate entirely without human intervention, be that maintenance, error correction, or by providing the link with the wider task/environment/context at the boundaries of automation. It is perhaps useful to first consider autonomy and automation as regarded within the automotive domain. Vehicles provide a context in which there is a (relatively) simple relationship between a single human operator (the driver) and a machine (the vehicle). Automation in the automotive domain is commonly categorised using the 6-level scale devised by SAE International (SEA, 2016). Level 0 represents the lowest level of automation, the driver is in full control with minor assistance. Level 5 represents full autonomy, with the vehicle in full control and the driver essentially just a passenger. Levels 1-4 involve varying degrees of the sharing of control. Strictly speaking Level 5 may be regarded as full automation rather than autonomy necessarily, insofar that the driving task may be considered to be a single journey only. Full autonomy would imply that all the user's driving needs are met.

- Level 0 – Driver is in control. Automation provides limited warnings and momentary assistance.
- Level 1 – Driver in control. Automation provides single control input assistance.
- Level 2 – Driver in control. Automation provides multiple control input assistance.
- Level 3 – Automation in control but driver monitoring required.
- Level 4 – Automation in control. Full vehicle automation, but only in selected environments.
- Level 5 – Automation in control. Full vehicle automation in all environments.

The key point here is the transition between levels 2 and 3, from the driver being in control to the vehicle being in control. Within the automotive world, level 3 is hugely problematic. If we think back to the discussion of the Fitts List (Table 1), humans as monitors is an extremely poor use of human capability. We are essentially saying that the vehicle is theoretically capable of driving itself but is not sufficiently reliable to be depended upon; or that there are known elements of the operational context that the automation is unable to handle. The driver must therefore be on standby to step in at such time as the automation is asked to exceed the limit of its capabilities. This two fundamental problems:

1. Failure of the automation is likely to happen rapidly and at short notice. The driver is therefore asked to step in without any opportunity to reacquaint themselves with the driving task, nor with the benefit of having been fully engaged in the development of the situation in which they are now asked to take control.
2. Failure of the automation is likely to happen in complex, worst-case situations. The driver is therefore being asked to step in at exactly the time where they will be least-plausibly capable of doing so.

For short, discrete tasks where driver engagement can be maintained (such as conducting a parking manoeuvre) it may be practicable for the driver to act as a monitor of the automation. However, Level 3 is pretty much a no-go for any extended driving activity. Any vehicle ostensibly operating at Level 3 in the real world, during normal driving, is likely doing so without the human failsafe it claims to have. Unfortunately, this is not a view seemingly shared by vehicle manufacturers! Good practice instead requires the jump from Level 2 to Level 4. Level 4 still requires periodic driver control, but without the need for the driver to monitor the automation and step in at short notice. Instead, the vehicle and the driver will conduct a formal handover as the vehicle recognizes it is reaching the boundary of its remit. This is not without its challenges in the automotive world, but it is practicable in principle.

2. DISCUSSION

Transposing to a warship context

The SAE system of classifying levels of autonomy has been introduced here as it is generally well known and understood, but it does require some modification when applying to the context of a warship. The European Defence Agency's Safety and Regulations for European Crewless Maritime Systems (SARUMS) group has produced a similar 6-point scale that does this transposition, but it does so on the basis of classifying an autonomous ship in its entirety. This may be best suited to smaller vessels rather than a larger warship and, for reasons that will be discussed later, it is not necessarily suited for the purposes of this paper. As such the SAE scale is carried forward.

The SAE scale is constructed explicitly around the idea of automation of vehicular control inputs. This may be directly applicable to the role of helmsman but little else, whereas there are a multitude of different roles on a warship when autonomy may be introduced. For the purposes of carrying over to a warship context we might therefore redefine the levels as follows:

- Level 0 – Operator is in control. Automation provides alerts and information only.
- Level 1 – Operator in control but automation manages minor elements of the task.
- Level 2 – Operator in control but automation manages significant elements of the task.
- Level 3 – Automation in control but Operator monitoring required.
- Level 4 – Automation in control. The system can be run in fully autonomous mode, but only in selected environments.
- Level 5 – Automation in control. The system is fully autonomous in all environments.

This is presented graphically in Figure 1.

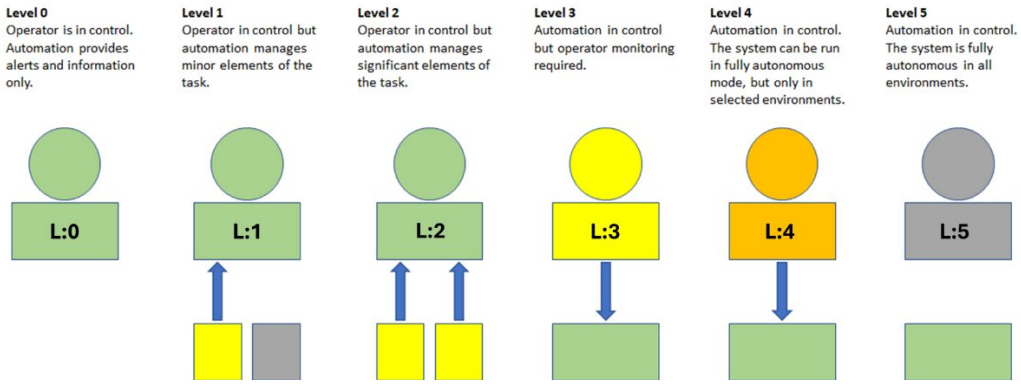

Figure 1. Levels of Automation

Once again, the key transition is from level 2 to level 3, where primary control of the system switches from the operator to the automation. We have already stated that level 3 is generally to be avoided, but in the automotive world level 4 is feasible. When operating at Level 4 the operator is temporarily uninvolved in the process (referred to as being 'out of the loop') whenever the automation has control and must be brought back in the loop before taking back control. The formal handover between vehicle and driver will be directed by the vehicle leaving a defined geographic area or road type for which the automation is certified. This can be planned in advance and will typically present an opportunity for the driver to either get back up to speed with the driving task in advance of taking over, or for the vehicle to pull over if the driver is not ready or not capable of taking over control.

On a warship there are likely to be short notice handovers required in the event of enemy action, and in particular the sustaining of battle damage. Not only may this handover be at short notice, but it likely involves a more complex situation for the operator to understand than a driver might be faced with upon resuming control on a motorway. The operator is therefore asked to step in at the worst possible time, without the benefit of having built up an understanding of the process and its state. This understanding is referred to as situation awareness (SA) and is often cited as having three levels, as defined by Mica Endsley (Endsley, 1995):

1. Level 1 – the perception of environmental elements and events with respect to time or space;
2. Level 2 – the comprehension of their meaning; and
3. Level 3 – the projection of their future status.

SA is not a property of the system, it exists only in the mind of the operator (though is undoubtedly affected by system design). It generally takes time to develop SA, and this is variable dependent on operator experience, state of mind and situation complexity or novelty. Asking an operator to step in at short notice, for a task that is normally automated, means that at least one, and probably several, of the following will be true:

- The scenario is highly complex (given that the automation can't handle it);
- The scenario may be such that previously held assumptions have been violated, and there is no defined correct response;
- The operator may not have been involved in the task prior to stepping in, and lacks even basic baseline SA;
- The operator is de-skilled from lack of practice and experience; and
- This may well be the operator's secondary duty / capability.

The latter two points potentially apply to any situation where there is a level 4 relationship and is something of a paradox as to the benefits of level 4. A system that can act independently for only a short amount of time or in a few select situations limits the potential to reduce operator workload or to allow for operator redeployment. Conversely, a system that can act independently most of the time leads to the operator receiving limited on-the-job experience. This can lead to the operator becoming de-skilled.

In much the same way as in Level 3 for the vehicle, asking an operator to step in under these conditions, especially when that situation is complex and potentially fraught with danger, is perhaps not a realistic prospect. For the purposes of designing systems on a highly autonomous warship then, levels 3 and 4 should be considered as potentially dangerous and to be avoided if possible.

Assessing levels of autonomy in multi-operator systems

The SAE scale (SEA, 2016) is based on a binary relationship of a single human operator in partnership with automation. In multi-operator setups, however, there various relationships that may exist, with different parts of the system operating at different levels of automation. For example, one subsystem may be operating at the equivalent of SAE level 2, whereas another may be at level 4. Take for example, the hypothetical system shown in Figure 2. This would be a fairly typical hierarchical team structure, with a team leader supported by human operators each assigned to the control of individual subtasks. There are multiple operator-automation relationships at play, but none exceeds level 2.

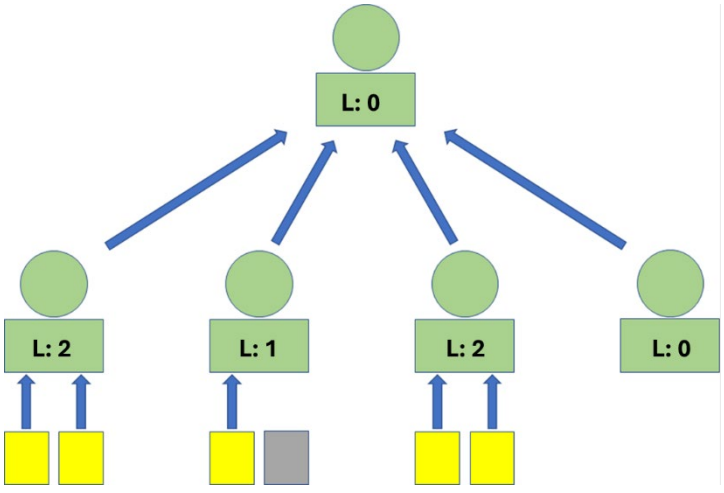

Figure 2. Multi operator- automation relationships at levels 0-2

Consider now the situation in Figure 3. Here, two of the previous roles have been automated, with a single operator monitoring both. Whilst a simplistic example, this represents a possible way that automation may be implemented to reduce crewing levels. The issue in this example is the level 3 relationship that has been created. It is obvious in this case, but may be harder to spot in more complex arrangements. This is particularly so if the designers create what they think is a level 4 or 5 autonomous relationship without due consideration of what the actual task boundary is. As stated previously, full autonomy is not achieved unless the automation can handle the entirety of the task.

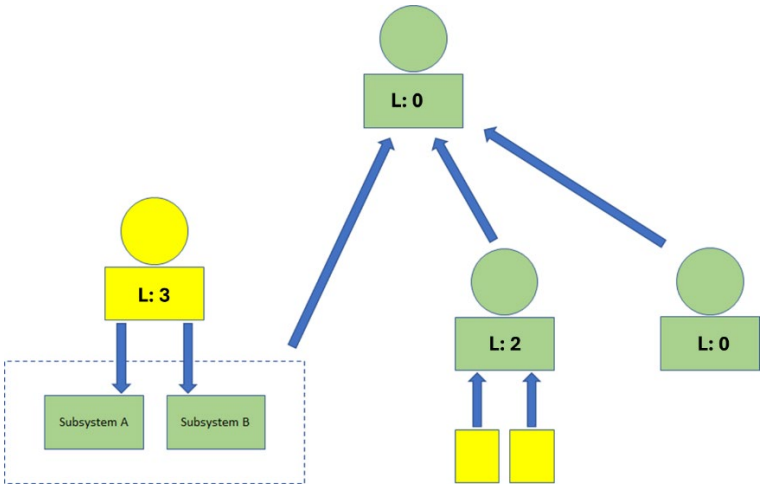

Figure 3. Levels 2 and 3 within the same system

Humans as the third line of defence

The way that humans intervene in a level 3 or level 4 relationship is also potentially altered by automation if the human effectively becomes the third line of defence. Imagine a typical operator intervention scenario where an otherwise automated system goes offline. This operator is likely to be lacking SA to at least some degree and may be under time pressure and/or stress. Not only that, but the operator may first need to spend time understanding where and how they are even needed. As discussed already, this is not an ideal situation. Potentially there could be a separate autonomous recovery system introduced to perform this role in place of the human. If it works then it may solve some of the challenges mentioned, but there is always the possibility that the secondary autonomy is simply a patch that also becomes overwhelmed by the escalating situation. When a human operator eventually has to step in, they do so later in the timeline, with even greater system damage, and lacking the SA they may have developed had they stepped in sooner. Once again, if this is a realistic prospect, the capacity for such an individual to genuinely rescue the situation must be questioned.

Avoiding levels 3 and 4 in complex systems

If levels 3 and 4 are to be avoided, this leaves only two options. Either the system in question remains at level 2 or we make the jump to level 5. On a platform with human operators this jump to level 5 must clearly exist at a lower level than the overall platform. This requires identifying tasks that can meaningfully be automated in their entirety and therefore fully autonomously. This is not to say that the system must be infallible, only that we consider the negatives of any potential failure of the autonomy to be outweighed by the positives of the autonomy. Consider the autonomous car; evidently these will still fail occasionally, and as there is no driver to intervene such failure may be expected to be catastrophic, but the overall benefits are considered to be worth the risk. In the case of the autonomous vehicle a failure may result in a crash. For an autonomous system that exists as part of a wider platform the outcome is not so clear cut. It may be that an autonomous system is sufficiently expendable that the overall ship can continue to function (if potentially in a degraded state). Conceivably it could be the case that a system is so critical that its failure renders the whole ship inoperable, but its failure is determined to be so unlikely, and the benefits of the automation so great, that the cost-benefit trade-off is acceptable. In such an arrangement it is likely that some form of reversionary mode would be planned for, but there must be acceptance that the ability of a reduced crew, with limited experience of this state, to perform in a reversionary mode may be limited at best.

Comparing the performance of humans and autonomy

As stated previously, whether or not it is considered to be the primary reason for introducing autonomy, there will be an expectation that the autonomy performs at least as well as a human undertaking the role. This, however, is not always a straightforward comparison to make. Firstly, because humans and computers have different strengths and weaknesses, they tend to fail in different ways. A good example of this again comes from the world of automobiles and self-driving cars. When a human driver is well trained, engaged and concentrated on the driving task, they can perform to an extremely high standard; capable of responding effectively to highly complex and potentially unfamiliar and unusual situations. Humans tend to fail at the driving task through a lack of attention, either through distraction or fatigue, to the extent that a human driver may even fall asleep at the wheel and therefore have zero attention at all. A self-driving car will never do this. Human failure is often context dependent. A human driver may momentarily fail at a task they would normally be able to do, as a result of their temporary state of mind. Machines, however, tend to be more binary; either they can do something or they can't. Failure of automation will therefore occur when the system exceeds the bounds of its capability, and this can often be dramatic.

There are currently no true level 5 cars, but there are ostensibly level 4 self-driving cars that can at least operate semi-autonomously in defined environments. Statistically, even these relatively early iterations of self-driving cars are safer than human drivers when defined as the rate of crashes per distance travelled. When they do crash though, they tend to make the news. This may partly be explained by them being held to higher levels of scrutiny but also because they often seem to be involved in crashes that any experienced and engaged driver would not be expected to have, often involving a collision with an object that might have been immediately obvious as a hazard to an engaged human driver but simply wasn't detected by the system. Crucially, these objects will typically be unusual in some way. The human brain is incredibly good at recognising patterns and objects and can often tolerate great variability and complexity in the scene. We are typically unaware of what cues we are taking from a scene in order to make our judgments and there may be countless pieces of information our brains are utilising, stitched together from the entirety of our life experiences and understanding of the world. Computers, however, recognise objects by referencing characteristics of that object against previously seen examples of similar objects, using only the (often limited) information channels available to them. Current systems are typically constrained to relatively few recognisable characteristics compared with human recognition. Sometimes, even relatively subtle differences may mean the system fails to recognise that which seems obvious to a human (although this will undoubtedly improve in the future with greater processing capability). In these situations there can be a mismatch between our expectations and reality. We are told that the system is safer than a human driver, and yet it has failed to recognise a hazard that we consider would be obvious to even a novice driver. For many people the overall safety record of the vehicle is then disregarded and only the performance in that specific situation is used to make an appraisal, from which the vehicle does not emerge well. We do not see all the times that the vehicle avoided a crash that a human may not have, only the time it did crash. As long as autonomy fails in ways that are different from typical human failure, there will be a tendency to apportion greater blame in those situations.

Understanding why autonomy can fail

We have so far talked about human and automation 'failure'. This may be interpreted as an implicit apportioning of blame but is intended simply to highlight that the end goal has not been achieved, or has been achieved through unacceptable means or performance parameters. In the case of machines, in particular, this failure to achieve the end goal may not be through any 'fault' of the machine itself. System failures have traditionally been understood through the mechanism of component failure, which propagates through the system into an overall system failure. Modern, complex, computer-based systems may often 'fail' despite working exactly as they were programmed to do. The failure has instead occurred in the design of those systems such that there are scenarios where the programmed behaviour leads to unacceptable outcomes. This may be because of the unexpected emergent properties of complex interactions, or simply because designers made assumptions that proved to be false. Much interesting work has been done in this field by Nancy Leveson of MIT and her development of Systems-Theoretic Accident Model and Processes (STAMP). The point is that complex systems can fail in varied and unexpected ways. If defining a particular system capability as being autonomous we must anticipate that things still can and will go wrong. In such circumstances it will be a human that we depend upon to step in and rectify the situation. The challenge is in planning for this in scenarios that had not been predicted in the first place. The alternative is to accept the consequences when automation failure occurs, but to make this so rare that when it does occur, it is accepted on the basis that the automation has, statistically, outperformed a human operator in the long run. This may require us to constrain the autonomy to applications where we can conceive of a failure being allowed to persist (at least in the short term) without the expectation of human intervention to bring it back online. It will certainly require a change in mentality whereby we judge an autonomous system on its overall performance, not based on specific situations.

Acceptance of human error versus autonomy error

Autonomy may sometimes fail in ways that seem rudimentary to us and may be blamed disproportionately for it as a consequence when viewed purely in the context of that specific situation. But autonomy may also be held to higher standards than human operators, even if failures occur in situations accepted to be challenging even for an expert human operator. When humans make mistakes there is often a tendency to seek out someone to blame, but there is also a degree of understanding and empathy. Provided we feel that the persons involved were genuinely doing their best we at least recognise that people do make mistakes and that challenging circumstances can sometimes prevail despite our best laid plans. Depending on the outcome we may be able to investigate the situation retrospectively to understand what the person was thinking, to understand the decisions they made. This can serve to make the outcome feel more acceptable but also allows us to learn lessons for the future. Sometimes we make changes to the system to convince ourselves (rightly or otherwise) that such an outcome won't happen again. Other times we modify training procedures as mitigation. Indeed, the very presence of established mechanisms for people to be trained and to gain experience and accreditations over time gives us confidence that the system is safe and that failures are isolated cases. This is often not so with autonomy. Machine learning is an immensely powerful tool for allowing computers to generate solutions to otherwise unsolvable problems, but the process can be opaque. We assess the performance of the computer based on the outcome, but we may not necessarily know why it has done what it has. This leads to two issues of trust:

When things go wrong, we may not be able to easily explain why. We cannot ask it what it was thinking or explain it away simply as 'human error' (however inappropriate that may be).

It may be difficult to know how much we should trust the system. There is no time-honoured process for training and easing into service. There is no process for promotion and appraisal; we cannot relate to a time when we were in a similar position, with similar questions and thinking

With regards to point 2, we can go further in saying that we cannot put ourselves inside the mind of a machine. Outside of any specific trade or role, we broadly understand people and how they think, despite our differences. There are so many characteristics of what it is to be human and how we perceive the world. We trust that our colleagues have the ability to adapt and work with us in times of adversity; and that as humans we have at least some ability to display ingenuity in the face of the unexpected, drawing upon a vast life experience outside of the immediate task context. We don't have that with a computer. We can trust it only insofar as we can set it tasks and measure its performance. We cannot know how it will deal with a situation against which it has not been tested and there is no bank of life experience to draw from, only task-specific learning. This is not to say that an autonomous system can never be trusted, only that our frame of reference must be shifted.

3. KEY RECOMMENDATIONS FOR HIGHLY AUTONOMOUS WARSHIPS

The key aspect of the above discussion relates to the importance of correctly defining the system boundary in relation to the task boundary. True autonomy (level 5) can only exist if the system can handle all elements of the task, in all situations. A failure to handle all situations (level 4) means the operator must step in to take control, possibly when under difficult conditions, with limited situation awareness and with potential skill fade. Putting the automation in control but not trusting it (level 3) requires the operator to act as a monitor, which (for anything other than relatively short durations) is something humans are not suited to do.

For a highly autonomous warship, there will be multiple relationships between humans and automation, most likely at varying degrees of human involvement. This all comes together into the complete socio-technical system of the ship and its operating context. The answer is to ensure that the socio-technical system is designed as a whole and not broken up to be designed as individual parts. Human Factors and Engineering must go hand in hand in a coordinated approach. This will be how we ensure that the role of the human within the system is fully accounted for and we are not inadvertently creating situations that the human operators are not equipped to face.

The highly autonomous warship will rely on humans to provide the resilience and adaptability that cannot be fully automated. To do this we must provide those humans with the means to recognise and respond to such scenarios. This means that the system must be designed to provide operators with a means of observing the system state and to recognise when the system is approaching a hazardous state; operators must be able to develop and maintain the situation awareness necessary to understand the implications of that state and what can be done to resolve it; and they must have the skills and capability to enact those interventions. Systems thinking and the associated engineering approaches, adopting multi-disciplinary coordination throughout the CADMID lifecycle, are our best tools to achieving this.

4. CONCLUSIONS

Wherever we still have humans as part of a wider socio-technical system, as is the premise for this paper, there must be acceptance that they are there for a good reason; the system depends upon them to function. Any automation or autonomy that is introduced into that wider system must do so for the benefit of the wider system and the people who are part of it. The automation does not exist in isolation and can be regarded as operating without human control only within the boundaries of the specific function it performs.

Unless a vessel is entirely autonomous it is not useful to classify the entire vessel according to levels of autonomy. Differing levels of autonomy and automation will be achieved for the various systems onboard.

When classifying a system according to its level of autonomy and automation it is necessary to consider the boundary of the system, such that it incorporates the entirety of the meaningful task or function that the system performs. If the system boundary is drawn too narrowly it risks masking a need for operator engagement, situation awareness, experience and skilling.

It is rarely a good idea to reduce a human operator to the role of monitor, based on the previous HABA-MABA discussion. Systems designed on the assumption that a human operator will be there to step in at short notice if required, may be setting the operator up to fail. This will not be the fault of the operator. Systems thinking is required to ensure that system

boundaries are drawn appropriately and that potential risks are assessed on a system scale. Human Factors and engineering considerations must be considered together as part of a complete socio-technical system, throughout the Concept, Assessment, Design, In-Service, Disposal (CADMID) lifecycle.

Automation should ideally be introduced only where it supports overall operator control (levels 0-2) or where it is trusted to perform fully autonomously such that no human intervention is assumed, even in the case of failure. It is important to ensure that any relationship between autonomy and humans plays to the strengths of each. The human should not be required to monitor the autonomy, instead the system must be trusted to operate without constant human oversight. If a human will ever be required to take over some or all capability of the autonomy, the human will need to have both the training/experience necessary to perform that role, and must have sufficient time and resources to develop the necessary situation awareness to step in.

The main strength of humans (and likely the reason they continue to be part of the system) is their potential to adapt to unusual circumstances by drawing upon their wider experiences and understanding of the world. To do so requires experience and proficiency; the overall system must facilitate the development of these.

There may need to be a shift in culture as to how we appraise the performance of autonomy. There needs to be an understanding that autonomy may fail in different ways to humans and may do so in ways that appear rudimentary to us. The overall performance of the system must be factored into the assessment. This change in mindset does not only apply within the military, but also to the wider public. Autonomous systems, particularly in the context of the military, attract scrutiny and criticism. Public reaction to high profile failures may be strongly negative and will need to be managed accordingly.

5. ACKNOWLEDGEMENT

BMT would like to thank all those who supported the production of this report and all our customers; without their support none of this would be possible.

The authors would like to especially thank Alistair Weir, Andy Kimber, Russell Bond, Mike Williams, Will Alexander, Aaron McMaster, Chloe Yarrien, Thomas Beard, Eshan Rajabally and Ian Savage for their contributions to the content.

6. REFERENCES

1. ENDSLEY, M.R. Toward a theory of situation awareness in dynamic systems. *Human Factors,* 37(1): 32-64, 1995

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

3. J3016A: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - *SAE International*. (2016, September)

7. AUTHORS' BIOGRAPHY

Alistair Frizell is the Human Factors Capability Lead at BMT. He is Chartered with the Chartered Institute of Ergonomics and Human Factors.

