Quantified minds: Predicting human functional state for human-machine teaming

Kate Ewing & Clare Borras

Defence Science and Technology Laboratory (Dstl), Porton Down, Salisbury, UK

ABSTRACT

A new dawn of intelligent machines has re-energised the concept of human-machine teaming (HMT) whereby humans, and autonomous systems, collaborate towards a shared operational goal. Across Defence, Human Factors specialists will be challenged to integrate human-autonomy teams into already complex systems for which knowing the functional state of human teammates will be critical to system optimisation. Presently, innovation in machine learning and data collection methods is making human cognition more available to operational settings than ever before. This paper overviews the state of the art in techniques for estimating human functional state from the perspective of designing complex military systems involving artificially intelligent (AI) agents. Considerations are provided for designers seeking to quantify variables such as mental workload, situation awareness (SA) or the level of demand upon particular communication modes, whether for system operation or design and evaluation. Finally, some examples of methods used in HMT research are presented along with a speculative look at future influences upon the specification of human functional state for use with autonomy in Defence.

KEYWORDS

Human-machine teaming, Autonomy, Psychophysiology, Defence, Complex Systems

Introduction: Autonomy in the future operating environment

Intelligent machines offer new and powerful capabilities to UK Defence. Unmanned ground vehicles (UGVs), unmanned aircraft, intelligent assistants and interfaces are being developed for integration into complex systems across military domains. Human-machine teaming (HMT) is a concept in which humans and autonomous machines collaborate within a system to achieve a shared operational goal (e.g. MOD, 2018). Designers working with human-autonomy teams (HATs) face the significant technical and Human Factors (HF) tasks of developing and integrating the AI agents and optimising team performance. Considerable attention has been paid to the development of explainable AI so that the behaviour of AI agents may be understood, however it is also necessary to understand the determinants of performance and behaviour in military personnel. This paper presents an overview of approaches and considerations for measuring human functional state for the design and operation of HMT Defence systems.

The allocation of function between humans and autonomy will be set during the design stage and might be fixed or, increasingly in future, change dynamically depending on the situation. The ability to characterise and quantify the functional state of military personnel, for example their level of mental workload, vigilance, or capacity for a particular interaction mode (e.g., speech, touch), will be indispensable to developers and critical to dynamic functional allocation aimed at optimising performance (e.g., intelligent selection of interaction mode and level of assistance to provide). Being more dynamic and 'human' than computers, intelligent systems need innovative approaches to human machine interface (HMI) design. Methodology grounded in that of human computer interaction (HCI) and overlapping with the field of human robot interaction (HRI; Huang, 2015) will need to accommodate the unique characteristics of human-autonomy teams, i.e.:

1. Teamwork. Complex systems will increasingly see humans collaborating with autonomous agents on shared goals. So, it is important to understand the impact of team-working on human functional state and cognition, how these will be measured and how human data will be used within the system. The number and type of HMT members, their relationship (e.g., supervisory, collaborative) and operational proximity (e.g. for teleoperation) will need to be taken into account.
2. Interaction. Dynamic human-autonomy teams (HATs) will require timely and accurate communications that don't surprise, confuse or add workload. Knowing how cognitive capacity affects and is affected by tighter human-machine interaction will be important. Also, novel interactive technologies (e.g., cameras for gesture and eye-tracking) will need to integrate with novel human monitoring techniques potentially leading to convergence into one technology e.g., eye gaze as a control input and source of cognitive information.
3. Operationalisation in future. The need for ecological, life-like testing of HMT design concepts often exceeds what can be achieved with state-of-the-art autonomous systems leading to use of lower cost methods such as the 'Wizard of Oz' in which the real world and autonomy can be simulated using virtual reality (VR) (Cooke et al., 2020). Designers must work with unknowns such as new possibilities offered by technique advancements and obstacles arising from the process of integration with the HAT.

Which psychological variables are needed for use in military HMT?

As for HCI, HMT will require bio-cybernetic systems to be able to optimise human performance (e.g., by keeping mental workload within acceptable levels by providing assistance at the interface or adjusting the level of automation (adaptive aiding/automation; Ewing et al., 2016; Scerbo et al., 2003). Besides mental workload, other determinants of performance such as vigilance, fatigue, stress, attention focus and situation awareness (SA) will be of interest. Some HATs will require personnel to supervise unmanned machines (e.g., ground vehicles (UGVs), drones) at a physical distance from a battlespace in which tasks are autonomy-executed, and instead be required to manage and make decisions, for which they will need good SA and the ability to monitor the autonomy's performance. Roy et al. (2020) propose that it will be necessary to monitor mental fatigue (due to long periods of focus), mind wandering and attentional disengagement (that can undermine SA), and in-attentional sensory impairments that could cause omissions (such as missing alarms). Detection of mental underload (e.g., during low demand monitoring or navigation tasks) may be important as underload can quickly become overload when an operator is insufficiently aroused, such as for an emergent threat that demands prompt action in an uncertain, possibly lethal situation (Young & Stanton, 2002). There is much research into human trust in autonomy, i.e., low trust can cause an operator to check up on an AI machine which increases their workload. Errors and unpredictability from an autonomous system can confuse, particularly if an operator is already overloaded, harming trust and system performance. Intelligent systems could also need to know a human's interaction mode capacity (e.g., visual, auditory or speech) to select an available mode and avoid overloading any one (Heard et al., 2019). Likewise, dynamic allocation of function will need timely measurement of demand upon different cognitive and behavioural resources. To summarise, HMT in complex systems will potentially require a holistic, multimodal and multidimensional input of human data to supply the information needs of intelligent adaptation.

Technical and practical challenges

Good practise principles and method criteria must inform measurement of human functional state in HMT and should expect to draw upon those established within HCI and Human Factors, i.e.:

- Human-computer interaction should be natural, multimodal, seamless, modelled on cognitive science, context aware, efficient, consistent and well timed (Huang, 2015).
- Workload measurement should be sensitive, diagnostic and unintrusive (O'Donnell & Eggemeier, 1986); also, sensitive to transient variations, repeatable with low variance and selectively sensitive to workload over other variables (Cain, 2007).
- Psychophysiology for field-based application should demonstrate ecological validity via testing in life-like simulation and real-world operational settings (Fairclough, 2017).
- System design should be human-centred (e.g., Stanton et. al., 2021).

Technical opportunities and challenges posed by human-autonomy systems and the future operating environment (FOE) could require modifications to existing criteria or entirely new criteria to emerge. For instance, human data capture technologies may need to integrate with novel interaction devices, extended reality (XR) environments and innovative wearable HMIs adopted in military settings. Designers will need to determine the adaptive logic of the military system, to ensure its response to human data inputs are of an appropriate type and magnitude, and are well timed. However, possible restrictions upon data storage, bandwidth and processing power for personnel monitoring interfaces coupled with possible increases in the quantity, noisiness and dynamism of operational data, could overload computation and compromise the timeliness and quality of the human inputs. Environmental noise (e.g. sound, movement and electromagnetic) alongside physical challenges such as dirt and cramped space will continue to plague technology in cockpits, under-water and land vehicles. Therefore, development of hardware and deep learning methods for robust classification of human functional state alongside more processing power (e.g., via quantum computing) will be sought. Sophistication at the level of deep learning algorithms may potentially afford greater simplicity at the HMI as predictive capabilities become less data-hungry and more efficient. Refinement and integration of personnel monitoring techniques into seamless, intuitive and efficient interfaces, for example by streamlining collection of human data types required for different purposes (e.g., health monitoring, workload monitoring and interaction) may become achievable from one or two, maximally rugged and minimally obtrusive, interfaces. Thus, the current expansive phase of technological exploration and innovation will need to be followed by a consolidating phase in which methods are down-selected and refined for each specific HMT setting.

Methods for predicting human functional state: 1. Physiology and task behaviour

Traditionally, psychological states impacting human performance, such as mental workload may have been measured via task performance, behaviour, subjective questionnaires or psychophysiology - or a combination of these. However, not all are suited to dynamic operational settings (i.e. questionnaires) and will have varying levels of suitability for the HMT context. The following paragraphs overview the state of the art in methods for capturing human functional state (a full review is beyond the scope of this paper).

Physiology offers a window into cognition via cognitive influences on the sympathetic nervous system (SNS). Several non-invasive psychophysiological methods can detect SNS arousal due to task demands, e.g., electrodermal activity (EDA), blood pressure, facial electromyography (fEMG) and sweat hormone profiles, however many are too susceptible to physiological confounds (e.g., from arousal due to physical exertion) for use in an uncontrolled operational environment. Variability in heart rate (HRV), however, is being used for HMT concept testing and is often selected for applied research owing to its ability to discriminate mental workload from sources of physiological arousal whilst being measurable from unobtrusive wearable ECG sensors (e.g., Martin et al., 2019). In addition to chest straps, HR can be collected via more remote means, i.e., pulse oximetry sensors, cameras/webcams (that detect skin pulse via photoplethysmography (PPG) - tiny chromatic changes in skin due to blood flow), thermal cameras (also detect skin pulse), Doppler radars, and capacitive electrodes (that detect the ECG waveform up to 40cm from the heart; Bousefsaf et al., 2014; Hinde et al., 2021).

Cameras and microphones that are remote or mounted on wearables (e.g., Google Glass) can also capture informative data-streams via eye-tracking, facial expressions and voice: Eye blinks, gaze direction and pupillometry (pupil dilation) can offer insight into a range of cognitive features e.g. fatigue, SA, learning, strategy, attention and workload (Pignoni & Komandur, 2019). Over the last decade improvements in the sensitivity of everyday cameras and the application of sophisticated deep learning algorithms are making face and eye techniques more feasible at lower cost. For example, Shishov (2019) videoed facial expressions and eye gaze with an ordinary webcam during a lab based cognitive task then trained a combination of recurrent neural network (RNN) and LSTM algorithms on the data to predict mental workload. Recently UCL researchers discriminated workload from eye-blinks recorded with an ordinary camera by extracting time-frequency blink data then applying a 2D LSTM algorithm, arguing for the superiority of 2-dimensional deep learning over traditional time-series methods for capturing complex SNS behaviour (Cho, 2021). Voice recorded with a standard microphone can also be used with machine learning to quantify stress due to cognitive load, e.g., in pilots (Hagmüller et al., 2006).

However, despite their unobtrusiveness, audio-visual recordings require a person to remain within range or must be mounted upon clothing or other wearable, and their separation from the body makes them especially vulnerable to environmental interference e.g. from light, sound, temperature or movement, making them potentially unfeasible in dynamic HMT contexts. Further development and application of deep learning is needed to clarify if remote data capture can yield sufficiently accurate, reliable estimation.

Neurotechnologies have become portable and wearable and thus available to operational settings. EEG (electro-encephalography) has superior temporal resolution that can offer timely inputs of cognitive information for HMT. Error potentials (ErrPs) are EEG features that occur when the brain perceives a mismatch from expectation and have been used for many purposes, e.g.:

- evaluation of trust in autonomy (Akash et al., 2018);
- detection and correction of robot errors and misunderstanding of human gestures (Kim et al., 2017; Krol & Zander, 2017);
- prediction of pilot auditory error for adaptive cockpits (Dehais et al., 2019);
- detection of severity and type of system errors perceived by the human (Wirth et al., 2019).

Other features of the EEG have been used by an intelligent system to affirm a human's understanding, such that information can be re-presented when the EEG indicates a lack of perception (Kirchner et al., 2013). In addition, EEG and functional near-infrared spectroscopy (fNIRS), a recent method for detecting changes in cortical blood oxygenation, can be used to quantify a broad range of performance-relevant measures of cognition and affect e.g., fatigue, alertness, mental effort, SA and modality-specific processing such as visual or auditory (e.g., Ewing et al., 2016; Mund et al., 2020; Roy et al., 2020; Solovey et al., 2012).

Neuroergonomics arguably provides a gold standard for capturing cognition, however as noted in Fairclough and Lotte (2020) few neurotechnologies have progressed beyond lab demonstrator systems to the real world. There are significant obstacles to operational deployment including signal variability (between participants, tasks and sessions) and noise - which can afflict the signal from multiple sources in operational settings (e.g., due to environment, movement, human physiology and psychology). In addition, head mounted sensor devices may be obtrusive, and timely to set up and calibrate. To overcome these challenges, machine learning methods and sensor hardware require further innovation combined with ecological HMT concept testing. Alternatively, the ability to capture brain information makes neurotechnologies attractive for validating other methods and evaluating system design in more benign settings.

In addition to bio-cybernetic approaches are those that estimate human functional state directly from task-related interactions with the system. For example, pilots' flight control inputs, drivers' steering behaviour and desk operatives' keyboard and mouse use can all be used to classify mental workload (e.g., Martin et al., 2019; Sanchez et al., 2018). Task behaviours are potentially attractive for HMT due to offering a ready-integrated, machine-based data source that is both close to task performance and partially removed from physiological and environmental noise. However, task behaviour only offers a general measure of workload impact and is less suited for specific cognitive inferences (other than where they relate to the measured behaviour). Also, it is necessary to identify and validate suitable, informative system metrics which is potentially difficult for a dynamic HMT system still at the design stage.

Methods for predicting human functional state: 2. Modelling and examples from HMT research

This section provides examples of approaches to estimation of human functional state within HMT research. Modelling, used alone or within a multi-method approach has been used within several studies. HMT will require shared mental models between human and machine to allow AI Agents to 1) identify the human mental state and 2) show a human-like understanding so they can be understood by humans (Huang, 2015). Models of mental workload have been used to stand-in for psychophysiological methods that will be integrated later on (Heard et al., 2019), and systems may use models of individual human operators to predict their behaviour (Nikolaidis et al., 2015). Models of humans' and agents' tasks, task interrelationships and goals can provide context for an intelligent system needing to diagnose operator actions or goals (Mund et al., 2020), or to provide a basis for concept evaluation purposes (Lashley et al., 2019).

One approach is to combine task models with psychophysiological data. An in-flight simulator trial by Mund and colleagues involved an adaptive assistant using workload models and psychophysiology to determine the level and mode of assistance to provide to a pilot. For every task auditory, visual, spatial, verbal demands and visual, verbal and manual interactions were modelled along with eight vectors of estimated load (based on Multiple Resource Theory; Wickens, 2002), which the assistant could use to estimate pilot load. Based on their work, the authors proposed that psychophysiological measurement was preferable where task loads could not easily be modelled, e.g. for image scanning tasks, and that psychological influences from a complex environment could be captured by recording multiple psychophysiological data streams that can be merged into a single robust measure of workload that can be used to supplement and 'tune' the modelled load parameters (Mund et al., 2020). A combined approach has also been used by NASA, i.e., algorithmic classification of physical and cognitive workload and task modelled estimations of speech, visual and auditory load were used as inputs to an AI agent tasked with normalising human workload by manipulating the functional allocation and interaction mode. The NASA team used modelling to establish a 'ground truth' in their research but propose that physiological measures should ultimately be used for all types of workload (Heard et al., 2019).

The US Air Force is also using psychophysiology within HMT. For example, eye-tracking was included with subjective measures of mental workload, SA, stress, performance and trust in the Autonomous Flight Testbed (AFT) to help understand to what extent F-35 pilots supervising multiple unmanned F-16s could fly their aircraft, supervise the F-16s and complete their mission (Holec et al., 2020). Head movements and gaze have been used to measure trust in autonomy by the DARPA Air Combat Evolution (ACE) program for pilots performing dog-fighting battle management while supervising an AI programmed to execute combat manoeuvres (DARPA, 2021). Also, integrative technology that allows multimodal human monitoring to be integrated into a complex system for HMT evaluation has been developed by the University of Iowa Operator Performance Laboratory (OPL) (Martin et al., 2019).

Future considerations and conclusion

Human Factors specialists working with complex systems in Defence and across industry will need to anticipate future trends and influences impacting how human functional state is determined for HMT purposes. Could aspects of the FOE such as new technology and hybridized warfare affect human monitoring needs? How will HMT systems and the interactions between humans and AI agents evolve as they become increasingly familiar, e.g., will humans learn to adapt their behaviour based on system feedback? Will the principle of human-centred design need to be accompanied by or fused with 'agent-centred design' to design for human-like AI agents? Also, how will public concerns about some technologies, such as invasive methods of human monitoring that involve implanted sensors (not used within UK Defence or discussed in this article), or the use of powerful AI algorithms to access (private) mental states, and more general concerns about AI agents, impact legislation around the use of human data and autonomy? Research and development in this exciting field is likely to pose more questions as well as answers.

This paper has presented an overview of methods and considerations for the quantification of human functional state in complex HMT systems. It concludes -

1. Accurate classification will potentially require a multi-method, multi-modal and multi-dimensional approach with supplementation from environmental and task data.
2. An ecological focus on specific HMT scenarios will be necessary to down-select, integrate and refine methods used.
3. Techniques should be selected according to how their strengths and limitations enable them to meet the needs of different stages of the design lifecycle.
4. Criteria and principles for human functional state estimation and HCI design will apply but may require extension for application to HMT contexts.
5. Deep learning, hardware and methodological innovation will enhance capability in the determination of human functional state in future.

Acknowledgement

This paper was produced under the MOD Future Workforce and Human Performance program.

References

Akash, K., Hu, W. L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(4), 1-20.

Bousefsaf, F., Maaoui, C., & Pruski, A. (2014, November). Remote assessment of physiological parameters by non-contact technologies to quantify and detect mental stress states. In *2014 International Conference on Control, Decision and Information Technologies (CoDIT)* (pp. 719-723). IEEE.

Cain, B. (2007). *A review of the mental workload literature.* North Atlantic Treaty Organisation (NATO) Research and Technology Organisation (RTO) TR-HFM-121-Part II, Defence Research and Development Canada; Toronto, Canada

Cho, Y. (2021, May). Rethinking eye-blink: Assessing task difficulty through physiological representation of spontaneous blinking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-12). CHI.

Cooke, N., Demir, M., & Huang, L. (2020, July). A framework for human-autonomy team research. In *International Conference on Human-Computer Interaction* (pp. 134-146). Springer.

Defense Advanced Research Projects Agency (DARPA). (2021*). Collaborative air combat autonomy program makes strides (2021)* U.S. Department of Defense https://www.darpa.mil/news-events/2021-03-18a

Dehais, F., Rida, I., Roy, R. N., Iversen, J., Mullen, T., & Callan, D. (2019, October). A pBCI to predict attentional error before it happens in real flight conditions. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (pp. 4155-4160). IEEE.

Ewing, K. C., Fairclough, S. H., & Gilleade, K. (2016). Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop. *Frontiers in Human Neuroscience, 10*, 223.

Fairclough, S. H. (2017). Physiological computing and intelligent adaptation. In M. Jeon (Ed.). *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 539-556). Academic Press.

Fairclough, S. H., & Lotte, F. (2020). Grand challenges in neurotechnology and system neuroergonomics. *Frontiers in Neuroergonomics, 1*, 2.

Hagmüller, M., Rank, E., & Kubin, G. (2006). Evaluation of the human voice for indications of workload-induced stress in the aviation environment. *EEC Note*, *18*(06).

Heard, J., Heald, R., Harriott, C. E., & Adams, J. A. (2019). A diagnostic human workload assessment algorithm for collaborative and supervisory human--robot teams. *ACM Transactions on Human-Robot Interaction (THRI), 8*(2), 1-30.

Hinde, K., White, G., & Armstrong, N. (2021). Wearable devices suitable for monitoring twenty four hour heart rate variability in military populations. *Sensors*, *21*(4), 1061.

Holec, R., Hockensmith, M., Broll, J., Wittich, C., Donadio, B., de Visser, E., & Tossell, C. (2020). *Autonomous Flight Testbed (AFT): Designing a flight simulation system to explore future Human-Machine Teaming concepts* (No. 2891). EasyChair. www.easychair.org

Huang, W. (2015). *When HCI meets HRI: the intersection and distinction.* Virginia Polytechnic Institute and State University, VA, USA.

Kim, S. K., Kirchner, E. A., Stefes, A., & Kirchner, F. (2017). Intrinsic interactive reinforcement learning using error-related potentials for real world human-robot interaction. *Scientific Reports, 7(*1), 1-16.

Krol, L. R., & Zander, T. O. (2017). Passive BCI-based neuroadaptive systems. In *Proceedings of the 7th Graz Brain-Computer Interface Conference 2017* (pp. 248-253). Verlag der TU Graz.

Lashley, H., Thorpe, A., Tylor, R., & Grabham, A. (2019). *Measuring effectiveness of Human-autonomy teaming.* North Atlantic Treaty Organisation (NATO) Science and Technology Organisation (STO) NATO-STO-MP-HFM-300. DSTL, Fareham, UK.

Martin, P., Calhoun, P., Schnell, T., & Thompson, C. (2019, June). Objective measures of pilot workload. In *63RD SETP SYMPOSIUM PROCEEDINGS (Sept. 2019).*

Ministry of Defence. (2018, May). *Human-machine teaming*. (JCN 1/18). https://www.gov.uk/government/publications/human-machine-teaming-jcn-118

Mund, D., Pavlidis, E., Masters, M., & Schulte, A. (2020, February). A conceptual augmentation of a pilot assistant system with physiological measures. In *International Conference on Intelligent Human Systems Integration* (pp. 959-965). Springer.

Nikolaidis, S., Ramakrishnan, R., Gu, K., & Shah, J. (2015, March). Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 189-196). IEEE.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K.R Boff, L. Kaufman, J.P. Thomas (Eds.), *Handbook of Perception and Human Performance: Vol. 2. Cognitive Processes and Performance.* John Wiley and Sons.

Pignoni, G., & Komandur, S. (2019, July). Development of a quantitative evaluation tool of cognitive workload in field studies through eye tracking. In *International Conference on Human-Computer Interaction* (pp. 106-122). Springer.

Roy, R. N., Drougard, N., Gateau, T., Dehais, F., & Chanel, C. P. (2020). How can physiological computing benefit human-robot interaction?. *Robotics, 9*(4), 100.

Sanchez, W., Martinez, A., Hernandez, Y., Estrada, H., & Gonzalez-Mendoza, M. (2018). A predictive model for stress recognition in desk jobs. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.

Scerbo, M. W., Freeman, F. G. & Mikulka, P. J. (2003). A brain-based system for adaptive automation. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 200-219.

Shishov, B. (2017). *Mental workload estimation on facial video using LSTM network*. Gubkin Russian State University, Moscow.

Solovey, E., Schermerhorn, P., Scheutz, M., Sassaroli, A., Fantini, S., & Jacob, R. (2012, May). Brainput: enhancing interactive systems with streaming FNIRS brain input. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 2193-2202). SIGCHI.

Stanton, N., Revell, K. M., & Langdon, P. (Eds.). (2021). *Designing interaction and interfaces for automated vehicles: User-centred ecological design and testing.* CRC Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wirth, C., Dockree, P. M., Harty, S., Lacey, E., & Arvaneh, M. (2019). Towards error categorisation in BCI: single-trial EEG classification between different errors. *Journal of Neural Engineering, 17*(1), 016008.

Young, M. S., & Stanton, N. A. (2002). Attention and automation: new perspectives on mental underload and performance. *Theoretical Issues in Ergonomics Science, 3*(2), 178-194.