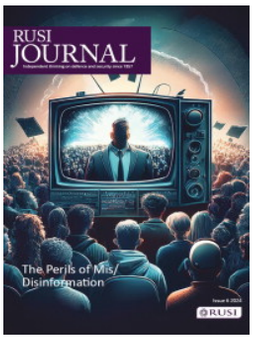

The RUSI Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rusi20

AI for Defence: Readiness, Resilience and Mental Health

Stuart E Middleton, Daniel Leightley, Patrick Hinton, Sarah Ashbridge, Daniel A Adler, Alec Banks, Maria Liakata, Brant Chee & Ana Basiri

To cite this article: Stuart E Middleton, Daniel Leightley, Patrick Hinton, Sarah Ashbridge, Daniel A Adler, Alec Banks, Maria Liakata, Brant Chee & Ana Basiri (2024) AI for Defence: Readiness, Resilience and Mental Health, The RUSI Journal, 169:6, 52-62, DOI: 10.1080/03071847.2024.2424780

To link to this article: https://doi.org/10.1080/03071847.2024.2424780

Published online: 26 Nov 2024.

Research Article
RUSI Journal

AI for Defence

Readiness, Resilience and Mental Health

Stuart E Middleton, Daniel Leightley, Patrick Hinton, Sarah Ashbridge, Daniel A Adler, Alec Banks, Maria Liakata, Brant Chee and Ana Basiri

AI is a cross-cutting technology that is having a major impact on behavioural analysis in both the defence and mental health domains. Employing AI well may boost the readiness and resilience of military personnel. Stuart Middleton and his co-authors explore how AI is being used today in research and practice for mental health in the defence domain. They identify key current challenges, and signpost the important trends that may help to build bridges between these domains for the ultimate benefit of both.

The mental health of any soldier influences their operational effectiveness and can have a significant impact on resilience and readiness. A comprehensive study looked across a 10-year period at the self-reported mental health status of British serving (that is to say, regulars) and ex-serving (veterans)[1] personnel from the Iraq and Afghanistan conflicts.[2] It found a significant prevalence of PTSD (6%) and alcohol misuse (10%) in addition to common mental health disorders (21%), such as depression and anxiety.

Research has identified significant differences in the rates of PTSD between veterans and the general military population.[3] It is estimated that 7.4% of UK veterans suffered from PTSD, whereas the rate in the general military population is 4.8%.[4] It is important to note that this prevalence rate is not uniform across groups. Quantifying the difference in the prevalence of specific mental health disorders for soldiers compared with the general population is important, as it allows AI mental health support solutions to be better targeted for the military context.

The use of digital technology, most notably AI, is transforming society. AI, such as natural language processing (NLP), has recently been applied to mental health problems such as in experiments with mobile phone-based screening of college students for depression,[5] social media-based early

1. Veterans are defined as those serving one paid day in the British armed forces.
2. Sharon A M Stevelink et al., 'Mental Health Outcomes at the End of the British Involvement in the Iraq and Afghanistan Conflicts: A Cohort Study', *British Journal of Psychiatry* (Vol. 213, No. 6, 2018), pp. 690–97.
3. See Brian C Kok et al., 'Posttraumatic Stress Disorder Associated with Combat Service in Iraq or Afghanistan: Reconciling Prevalence Differences Between Studies', *Journal of Nervous and Mental Disorder* (Vol. 200, No. 5, 2012), pp. 444–50; see also Lindsey A Hines et al., 'Posttraumatic Stress Disorder Post Iraq and Afghanistan: Prevalence among Military Subgroups', *Canadian Journal of Psychiatry* (Vol. 59, No. 9, 2014), pp. 468–79.
4. Stevelink et al., 'Mental Health Outcomes at the End of the British Involvement in the Iraq and Afghanistan Conflicts'.
5. M L Tlachac et al., 'StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students During the Coronavirus Pandemic', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (Vol. 6, No. 2, Article 76, 2022), pp. 1–32; Adam Tsakalidis et al., 'Combining Heterogeneous User Generated Data to Sense Well-being', *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (2016), pp. 3007–18.

Effective and responsible deployment of AI from the defence and mental health domains has the potential to boost the readiness and resilience of military personnel. *Generated by AI. Courtesy of Adobe Firefly*

risk prediction of depression[6], and therapy and counselling support around suicidal ideation in lifeline conversations.[7] However, research papers and practice reports on using AI for non-military clinical mental health and using AI for military personnel-based mental health have traditionally been published in separate venues, leading to silos of excellence within each community that are rarely bridged.[8] Nonetheless, there have been some recent attempts to bridge these gaps, such as the 2023 workshop on AI, defence and mental health by the Alan Turing Institute.[9] This article continues this work, highlighting recent examples of excellence and key challenges identified by both communities. It further signposts likely trends that have the potential to bring both areas closer together.

AI Research and Practice for Mental Health in the Defence Domain

Studies have shown that in 2012, suicide was the second-leading cause of death for US armed forces serving military personnel, higher than combat-

6. Jiageng Wu et al., 'Exploring Social Media for Early Detection of Depression in COVID-19 Patients', in Ying Ding et al. (eds), *WWW '23: Proceedings of the ACM Web Conference 2023* (New York, NY: Association for Computing Machinery, 2023), pp. 3968–77.
7. Zhong-Ling Wang et al., 'Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations', in Andreas Vlachos and Isabelle Augenstein (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (Dubrovnik: Association for Computational Linguistics, 2023), pp. 1436–46; see also Jenny Chim et al., 'Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts', in Andrew Yates et al. (eds), *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (St Julians: Association for Computational Linguistics, 2024), pp. 177–90.
8. Daniel Leightley and D Murphy, 'Personalised Digital Technology for Mental Health in the Armed Forces: The Potential, the Hype and the Dangers', *BMJ Mil Health* (Vol. 169, No. 1, 2023), pp. 81–83.
9. Remarks made in 'Workshop on AI and Defence: Readiness, Resilience and Mental Health', online workshop, 31 March 2023, <https://www.southampton.ac.uk/~sem03/AI-and-Defence-2023.html>, accessed 3 July 2024.

DOI: 10.1080/03071847.2024.2424780

related causes.[10] Understandably, the use of AI for suicide risk prediction has therefore been well studied. In the context of military personnel, NLP models such as decision-tree forest, support vector machine (SVM),[11] convolutional neural network (CNN) and transformer-based pre-trained language models with[12] and without[13] techniques such as prompt-tuning and adapter layers have been explored. Similar NLP methods have been used to classify suicide ideation/risk for non-military personnel, including CNN models applied to Reddit posts[14] and transformer-based text summarisation models trained on transcript data from lifeline conversations.[15] Reported accuracy for military personnel suicide risk prediction using unstructured text in electronic health records and social media such as Reddit or question-and-answer websites is as high as 92%.[16]

For military veterans with PTSD, small-scale studies have explored how AI and app-based intermediaries can be used to support individuals in a military environment to both seek out and get support managing their care.[17] Additionally, support vector machines and linguistic inquiry word count lexicon-based n-gram language models[18] have been explored for PTSD classification on small datasets with reported accuracies around 60% ± 7.2%.[19] A study focused on the UK military explored not just the role of machine learning, but which features contributed to the outcome.[20] The authors found, when trying to predict PTSD outcome using machine learning models such as random forest applied to demographic and self-reported psychometric measures of users, that the features with the highest contribution to prediction performance were alcohol misuse, gender and deployment status.

Outside the military context, there has been a lot of work exploring AI for mental health, often focusing on publicly available datasets from social media sources including Twitter and Reddit that have been annotated for mental health issues such as depression and drug abuse, as well as suicide and PTSD. In addition to classification of mental health behaviour types, which is discussed in detail next, studies into digital interventions for mental health applications, such as peer-level mental health

10. Gregg Zoroya, 'Army, Navy Suicides at Record High', *USA Today*, 18 November 2012, <http://www.usatoday.com/story/news/nation/2012/11/18/navy-suicides-army/1702403/>, accessed 4 October 2023.
11. Paul Thompson, Craig Bryan and Chris Poulin, 'Predicting Military and Veteran Suicide Risk: Cultural Aspects', in Philip Resnik, Rebecca Resnik and Margaret Mitchell (eds), *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD: Association for Computational Linguistics, 2014), pp. 1–6.
12. Bhanu Pratap Singh Rawat and Hong Yu, 'Parameter Efficient Transfer Learning for Suicide Attempt and Ideation Detection', in Alberto Lavelli et al. (eds), *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)* (Abu Dhabi: Association for Computational Linguistics, 2022), pp. 108–15.
13. Sungjoon Park et al., 'Suicidal Risk Detection for Military Personnel', in Bonnie Webber et al. (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: Association for Computational Linguistics, 2020), pp. 2523–31.
14. Han-Chin Shing et al., 'Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings', in Kate Loveys et al. (eds), *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (New Orleans, LA: Association for Computational Linguistics, 2018), pp. 25–36.
15. Wang et al., 'Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations'.
16. Park et al., 'Suicidal Risk Detection for Military Personnel'.
17. Hayley Evans et al., 'Understanding the Care Ecologies of Veterans with PTSD', in Regina Bernhaupt et al. (eds), *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI: Association for Computing Machinery, 2020), pp. 1–15; Greg Barish et al., 'A Mobile App for Patients and Those Who Care About Them: A Case Study for Veterans with PTSD + Anger', *PervasiveHealth'19: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare* (New York, NY: Association for Computing Machinery, 2019), pp. 1–10.
18. Glen Coppersmith, Mark Dredze and Craig Harman, 'Quantifying Mental Health Signals in Twitter', in Philip Resnik, Rebecca Resnik and Margaret Mitchell (eds), *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD: Association for Computational Linguistics, 2014), pp. 51–60.
19. Morgan Byers and Vangelis Metsis, 'Text Analysis for Understanding Symptoms of Social Anxiety in Student Veterans', *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 18, 2021), pp. 15958–59.
20. Daniel Leightley et al., 'Identifying Probable Post-traumatic Stress Disorder: Applying Supervised Machine Learning to Data from a UK Military Cohort', *Journal of Mental Health* (Vol. 28, No. 1, 2019), pp. 34–41.

moderation input during online mental health chatroom discussions,[21] have shown that toxicity in conversations can be reduced and characteristics such as civility and supportiveness increased.

Standard machine learning algorithms such as random forest models and SVM models have been used for some time to classify PTSD and depression,[22] and identify mental health risk markers in electronic health records,[23] with the aim to help mental health professionals identify high-risk patients more efficiently. More recently, pre-trained transformer-based models have been explored to provide word, sentence and post embeddings. Embeddings are simply vectors of numbers representing a concept such as a word or sentence, and are explained in the appendix of this article. These are used for tasks such as mental health classification,[24] online counselling behaviour prediction[25] and support to rewrite conversational posts more empathically as a means to reduce online users' tendencies to respond more toxically than they might do in a face-to-face conversation.[26] The strong performance of pre-trained transformer-based models can also be enhanced by adding bidirectional long short-term memory (BiLSTM) model layers trained using n-gram features based on clinical psycholinguistic lexicons.[27] The different types of layers available to machine learning models are explained in more detail in the appendix of this article. For longitudinal mental health tasks, such as classification of mood changes and mental health risk over time, post embedding methods involving path signatures,[28] Hawkes processes,[29] as well as the use of transformers, BiLSTM[30] and multitask learning[31] have been successfully used. For example, multi-task learning might train a model on two

21. David Wadden et al., 'The Effect of Moderation on Online Mental Health Conversations', *International AAAI Conference on Web and Social Media* (Vol. 15, 2021), pp. 751–63; Ashish Sharma et al., 'Human–AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support', *Nature Machine Intelligence* (No. 5, 2023), pp. 46–57.
22. Shirin Saleem et al., 'Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts', in Martin Kay and Christian Boitet (eds), *Proceedings of COLING 2012* (Mumbai: COLING 2012 Organizing Committee, 2012), pp. 2375–88.
23. Elena Alvarez-Mellado et al., 'Assessing the Efficacy of Clinical Sentiment Analysis and Topic Extraction in Psychiatric Readmission Risk Prediction', in Eben Holderness et al. (eds), *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* (Hong Kong: Association for Computational Linguistics, 2019), pp. 81–86.
24. Shaoxiong Ji et al., 'MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare', in Nicoletta Calzolari et al. (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association, 2022), pp. 7184–90.
25. Anqi Li et al., 'Understanding Client Reactions in Online Mental Health Counseling', in Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Toronto: Association for Computational Linguistics, 2023), pp. 10358–76.
26. Ashish Sharma et al., 'Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach', in Jure Leskovec et al. (eds), *WWW '21: Proceedings of the Web Conference 2021* (New York, NY: Association for Computing Machinery, 2021) pp. 194–205.
27. Sourabh Zanwar, 'The Best of Both Worlds: Combining Engineered Features with Transformers for Improved Mental Health Prediction from Reddit Posts', in Graciela Gonzalez-Hernandez and Davy Weissenbacher (eds), *Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task* (Gyeongju: Association for Computational Linguistics, 2022), pp. 197–202.
28. Talia Tseriotou et al., 'Sequential Path Signature Networks for Personalised Longitudinal Language Modeling', in Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki (eds), *Findings of the Association for Computational Linguistics: ACL 2023* (Toronto: Association for Computational Linguistics, 2023), pp. 5016–31.
29. Anthony Hills, Adam Tsakalidis and Maria Liakata, 'Time-Aware Predictions of Moments of Change in Longitudinal User Posts on Social Media', in Georgina Ifrim et al. (eds), *Advanced Analytics and Learning on Temporal Data*, Lecture Notes in Computer Science, Vol. 14343 (Cham: Springer, 2023), pp. 293–305.
30. Adam Tsakalidis et al., 'Identifying Moments of Change from Longitudinal User Text', in Smaranda Muresan, Preslav Nakov and Aline Villavicencio (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin: Association for Computational Linguistics, 2022), pp. 4647–60.
31. Tayyaba Azim, Loitongbam Gyanendro Singh and Stuart E Middleton, 'Detecting Moments of Change and Suicidal Risks in Longitudinal User Texts Using Multi-task Learning', in Ayah Zirikly et al. (eds), *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology* (Seattle, WA: Association for Computational Linguistics, 2022), pp. 213–18; Loitongbam Gyanendro Singh et al., 'ConversationMoC: Encoding Conversational Dynamics Using Multiplex

similar tasks at once, such as emotion detection and suicide risk classification. By training with two similar tasks, the model layer weight matrices tend to be more robust at encoding inputs than if each task was trained separately. Pre-trained transformers can also be used in more complex models such as multi-channel CNNs, which use the encoded distance from a post's embedded text to embedded symptom descriptions and achieve accuracies of up to 95% for classifying mental health issues including depression, anxiety and bipolar disorders.[32]

Not all mental health areas have large amounts of labelled data with which to train models. There is evidence that domain adaption, where AI models trained on one topic are then applied to another topic, is difficult for mental health[33] due to the presence of significant topical and temporal bias within training datasets from underlying differences in user posting behaviours from different data sources. An interesting recent work[34] explored a two-step domain adaption of transformer-based models to try to overcome these problems, first adapting a general pre-trained model to social media writing style and then to mental health in general before finally fine-tuning the model for the target mental health topic. This two-step domain adaptation approach improved model accuracy compared to simple fine-tuning alone.

Although much prior work has focused on textual data, multimodal methods have been explored, including embeddings of Flickr images[35] and BitChute video content[36] using models such as a 3D-CNN and vision transformer. There is also a body of research exploring internet of things (IoT) approaches to support mental health using mobile and wearable devices and digital assistants. Wearable smartwatch devices have been used alongside social media posts in a large-scale study, involving several hundred thousand users across 18 major cities, to show that circadian rhythms associated with sleep quality can be extracted from social media usage data[37] as a useful way to analyse mental health. Other studies have explored the use of mobile phones to predict stress[38] and use of Fitbit devices to predict depression.[39] Digital assistants have also been used for active interventions, including simple template-driven chatbots to encourage self-disclosure of mental health issues to a professional[40] and modern transformer-based chatbots to uplift the sentiment of online dialogue.[41] Researchers have also explored passive use of digital assistants, such as using interactions with Alexa devices at home to classify risk

Network for Identifying Moment of Change in Mood and Mental Health Classification', *CEUR Workshop Proceedings* (Vol. 3649, 2024), pp. 42–56.

32. Hoyun Song et al., 'A Simple and Flexible Modeling for Mental Disorder Detection by Learning from Clinical Questionnaires', in Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Toronto: Association for Computational Linguistics, 2023), pp. 12190–206.
33. Keith Harrigian, Carlos Aguirre and Mark Dredze, 'Do Models of Mental Health Based on Social Media Data Generalize?', in Trevor Cohn, Yulan He and Yang Liu (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020* (Stroudsburg, PA: Association for Computational Linguistics, 2020), pp. 3774–88.
34. Mario Aragon et al., 'DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media', in Rogers, Boyd-Graber and Okazaki (eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 15305–18.
35. Zhentao Xu, Verónica Pérez-Rosas and Rada Mihalcea, 'Inferring Social Media Users' Mental Health Status from Multimodal Information', in Nicoletta Calzolari et al. (eds), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association, 2020), pp. 6292–99.
36. Mithun Das et al., 'HateMM: A Multi-Modal Dataset for Hate Video Classification', *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 17, 2023), pp. 1014–23.
37. Ke Zhou et al., 'How Circadian Rhythms Extracted from Social Media Relate to Physical Activity and Sleep', *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 17, 2023), pp. 948–59.
38. Daniel A Adler et al., 'Machine Learning for Passive Mental Health Symptom Prediction: Generalization Across Different Longitudinal Mobile Sensing Studies', *PloS One* (Vol. 17, No. 4, 2022).
39. Daniel A Adler et al., 'Identifying Mobile Sensing Indicators of Stress-Resilience', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (Vol. 5, No. 2, 2021), pp. 1–32.
40. Yi-Chieh Lee, 'Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional', *Proceedings of the ACM on Human-Computer Interaction* (Vol. 4, No. CSCW1, 2020), pp. 1–27.
41. Tulika Saha et al., 'A Shoulder to Cry On: Towards A Motivational Virtual Assistant for Assuaging Mental Agony', in Marine Carpuat, Marie-Catherine de Marneffe and Ivan Vladimir Meza Ruiz (eds), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, WA: Association for Computational Linguistics, 2022), pp. 2436–49.

of mental health issues.[42] In a military context, studies have explored using mobile apps, behaviour change and machine learning to successfully reduce alcohol consumption, over a period of two to three months, in help-seeking military veterans and serving personnel.[43]

Trends, Challenges and Open Questions

Methods

There is a trend of using AI research in the mental health domain, especially the use of NLP-based approaches. After a delay of a few years, AI research is finding its way into experiments involving military contexts such as military PTSD, especially with medium-sized pre-trained transformers such as BERT and GPT. LLMs including GPT-4, Bard and LLaMa have not, to the knowledge of the authors, been broadly applied in this area.

Therefore, an increased awareness of the latest AI approaches might help military practitioners to speed up that adoption process in the mental health domain. Given the critical nature of mental health applications, and the potential consequences of missing or unhelpful mental health interventions, a cautious approach by researchers and practitioners is likely to have contributed to this delay. Often medium-sized AI models, such as BERT, are carefully fine-tuned and tested for robustness before any deployments with humans are conducted. This delayed adoption trend is most clear for the latest and most powerful LLMs including GPT-4, Bard and LLaMa, which have not yet, to the knowledge of the authors, been explored in a military-focused mental health context. Even for non-military studies,[44] there is currently very little work on LLMs for mental health; this is an interesting research gap worth exploring further given the growing evidence that LLM performance is significantly better than smaller pre-trained transformer models for many text classification and reasoning tasks. However, current LLM weaknesses around bias due to under-representation in the training data, highly plausible hallucinations and confirmation bias will need to be addressed first, as discussed below. For example, mental health LLM applications that exhibit confirmation bias on topics such as suicidal ideation are clearly dangerous and would need to have strong guardrails applied to ensure interventions, such as recommending seeking professional medical support, are triggered early during such conversations.

Beyond NLP, there is a growing AI trend for using IoT in mental health, especially those that draw on widely available devices such as wearable fitness technology, smartphones and AI chatbot apps for self-help purposes.[45] Avoiding the need to talk to humans about stigmatised issues such as mental health – especially in the military where there are fears of it being career limiting – can remove a barrier to getting help and make it easier for people to recognise problems and subsequently reach out to a medical professional. As evidenced from the personal experience of the authors of this article, in the military, mental health 'first responders' often take the form of trusted friends; AI-based services could be well positioned to take on this role, although it remains to be seen whether users have trust in the confidentiality of such services. It is likely that AI companies will begin to provide 24/7 self-monitoring services for things such as quality of sleep and digital biomarkers, which could then be used to provide evidence to help clinical mental health assessments. There are a number of open questions on this trend that closely mirror the wider liability issues regarding the use of AI, such as who is responsible if and when mental health warning signs are missed.[46]

42. Huikai Chua, Andrew Caines and Helen Yannakoudakis, 'A Unified Framework for Cross-domain and Cross-task Learning of Mental Health Conditions', in Laura Biester et al. (eds), *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)* (Abu Dhabi: Association for Computational Linguistics, 2022), pp. 1–14.
43. Daniel Leightley et al., 'Evaluating the Efficacy of the Drinks: Ration Mobile App to Reduce Alcohol Consumption in a Help-Seeking Military Veteran Population: Randomized Controlled Trial', *JMIR mHealth and uHealth* (Vol. 10, No. 6, 2022).
44. Loitongbam Gyanendro Singh et al., 'Extracting and Summarizing Evidence of Suicidal Ideation in Social Media Contents Using Large Language Models', in Andrew Yates et al. (eds), *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)* (St Julians: Association for Computational Linguistics, 2024), pp. 218–26.
45. Alison Darcy et al., 'Evidence of Human-Level Bonds Established with a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study', *JMIR Formative Research* (Vol. 5, No. 5, 2021).
46. Jodie Nghiem et al., 'Understanding Mental Health Clinicians' Perceptions and Concerns Regarding Using Passive Patient-Generated Health Data for Clinical Decision-Making: Qualitative Semistructured Interview Study', *JMIR Formative Research* (Vol. 7, 2023).

Data

On training data for AI, studies using electronic health records have arguably been the most reliable since there is clear evidence available of a clinical diagnosis for patients and the context of a full medical history. One such study[47] linked the electronic healthcare records of a representative sample of UK military personnel across England, Scotland and Wales with self-reported questionnaire data. This linkage enabled, for the first time, the reasons for admission into hospital from military personnel to be analysed and risk factors identified.[48] However, electronic health record data can be hard to get from hospitals and will always be orders of magnitude smaller in volume than social media content. As such, studies using such records are somewhat limited.

> Avoiding the need to talk to humans about stigmatised issues such as mental health can remove a barrier to getting help and make it easier for people to recognise problems.

On the other hand, social media is often used to train AI and offers the opportunity to observe people at much earlier stages of a mental health problem, well before they engage with a doctor or hospital. However, the quality of the mental health annotations is much lower.[49] Clinical experts such as psychiatrists have been asked to label social media datasets for AI training purposes. However, the human effort involved means these datasets are typically limited to around a thousand posts at most. At much larger scales, labels can be inferred from the social media forum titles and the existence of posts from users self-reporting a mental health problem. This type of data is large in volume but has quality issues relating to subjective bias from self-reporting.

There is now clear evidence, especially with transformer-based approaches, that AI model performance in terms of accuracy increases with the volume of training data all the way to web-scale training for LLMs. An open research question is whether low-quality, high-volume content can train AI better than high-quality, low-volume datasets for critical applications where robust, reliable and consistent models are needed. This is an active research area within AI and although web-scale LLMs are currently exciting the world in terms of performance, the use of high-quality, low-volume datasets to fine-tune and improve LLM robustness is something that is increasingly expected to be seen in the future.

Errors, Bias and Ethics

It has been well reported that generative AI, especially the latest LLMs, suffer from problems such as hallucinations, loss of information, encoded bias and privacy concerns.[50] Problems of NLP model bias have been seen for many years and can be hard to reliably detect due to weak evaluation regimes that lack authenticity.[51] Because LLMs are trained on web-scale datasets, the inherent historical and regional biases that exist within webpages and articles are 'baked in' during the training process and then manifested in the answers AI generates. Hallucinations are a well-known LLM issue currently, where LLMs introduce factual errors embedded within otherwise very plausible human-like answers. Such errors are difficult for humans to identify. Information loss, in the form of missing key points, often occurs when LLMs are asked to generate summaries. On privacy, clever prompting of LLMs can trick them into revealing people mentioned in the original training data despite attempts by LLM vendors to add guardrails to stop this.

47. Daniel Leightley et al., 'Integrating Electronic Healthcare Records of Armed Forces Personnel: Developing a Framework for Evaluating Health Outcomes in England, Scotland and Wales', *International Journal of Medical Informatics* (Vol. 113, 2018), pp. 17–25.
48. Zoe Chui et al., 'Mental Health Problems and Admissions to Hospital for Accidents and Injuries in the UK Military: A Data Linkage Study', *PLoS One* (Vol. 18, No. 1, 2023).
49. Jessica A Pater, 'Social Media is not a Health Proxy: Differences Between Social Media and Electronic Health Record Reports of Post-COVID Symptoms', *Proceedings of the ACM on Human-Computer Interaction* (Vol. 7, No. CSCW1, 2023), pp. 1–25.
50. Jacob Menick et al., 'Teaching Language Models to Support Answers with Verified Quotes', March 2022, arXiv:2203.11147.
51. Adam Tsakalidis et al., 'Can We Assess Mental Health Through Social Media and Smart Devices? Addressing Bias in Methodology and Evaluation', in Ulf Bradfield et al. (eds), *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, Vol. 11053 (Cham: Springer, 2018), pp. 407–23.

A military example of diverse user groups – which contain demographic subsets that will be under-represented in any AI model training data – can be found in the very diverse UK military which includes soldiers such as those in the Brigade of Gurkhas, who are recruited from Nepal,[52] and Commonwealth citizens recruited for roles in the military, such as Fijians.[53] This demonstrates the need for approaches that are informed by not only British cultures, but by the full scope of cultures that are recruited into the British armed forces. The lack of training examples from these groups in the data used to train LLMs could lead to AI models that exhibit racial and gender biases. There is also a risk that AI models might not detect the different ways in which different genders present their trauma, risking the potential for some to have their symptoms overlooked or misdiagnosed.[54]

There is also the unique military culture to consider, in which perceptions of rudeness can differ and being rude can simply be part of the bonding or as a way to establish camaraderie within a military group.[55] Humour can function as a 'disciplinary technology' within the British defence context, among others.[56] These examples reflect quite different patterns of communication from those that exist on the majority of social media training data used to train LLMs. The current commercial trend is to add guardrails to flawed LLMs, but it would be better to address under-representation and even explicitly encode absence during the LLM training and fine-tuning stages, making better LLMs rather than trying to guard against problems within flawed LLMs.

Perhaps one of the biggest problems for critical applications is that LLMs currently do not report any uncertainty about the output they generate, and so the bias, error and false negative rates are completely unknown. Unlike numerical models, which can often directly report uncertainty such as confidence values or error ranges associated with predicted values like a temperature measurement, a transformer-based LLM only has a probability distribution it uses to make the selection of its next output label (for a classifier) or predicted next word (for a chatbot). For a mental health application this means it is very hard to quantify the risk of potential warning signs being missed, such as an LLM generating a summary that looks helpful but actually ignores a key piece of information. There is active research into improving LLMs in this regard; progress will come over time, but currently adoption in critical applications such as mental health assessment is likely to remain cautious, especially around use cases associated with clinical interventions.

It is worth mentioning that AI regulation is being publicly discussed to safeguard against some of these issues, including the UK-led AI Safety Summit in 2023.[57] However, the prospect of any AI regulation and legal enforcement to make AI safer is likely to take years to enact. The UK's Online Safety Act is a good example of this, taking four years from its original conception as an Online Harms White Paper in 2019 to being passed in Parliament in 2023.[58] The current behaviour of most large AI vendors is to release AI models at pace, driven by fierce international competition in the AI market, and fix any problems these AI models cause afterwards. An open challenge for AI practitioners is thus how, in this competitive environment, AI competition can be put aside to bring in independent trusted expertise for a more rigorous academic evaluation of error and bias in models. A potential best practice exemplar for this type of commercial–academic engagement

52. British Army, 'Brigade of Gurkhas', <https://www.army.mod.uk/who-we-are/corps-regiments-and-units/brigade-of-gurkhas/>, accessed 3 July 2024.
53. Ministry of Defence, FOI2019/09980 [response to request for information under Freedom of Information Act], 1 October 2019, <https://assets.publishing.service.gov.uk/media/5e3c0a4c40f0b609169cb58f/FOI201909980-_Number_of_serving_Fijians_Redacted.pdf>, accessed 3 July 2024.
54. Giliane Fenech and Gill Thomson, 'Defence Against Trauma: Women's Use of Defence Mechanisms Following Childbirth-related Trauma', *Journal of Reproductive and Infant Psychology* (Vol. 33, No. 3, 2015), pp. 268–81.
55. Nick Caddick, Brett Smith and Cassandra Phoenix, 'Male Combat Veterans' Narratives of PTSD, Masculinity and Health', *Sociology of Health and Illness* (Vol. 37, No. 1, 2015), pp. 97–111.
56. Richard Godfrey, 'Soldiering on: Exploring the Role of Humour as a Disciplinary Technology in the Military', *Organization* (Vol. 23, No. 2, 2016), pp. 164–83.
57. HM Government, 'AI Safety Summit 2023', <https://www.gov.uk/government/topical-events/ai-safety-summit-2023>, accessed 3 July 2024.
58. For details of the act's passage, see UK Parliament, 'Online Safety Act 2023', last updated 31 October 2023, <https://bills.parliament.uk/bills/3137>, accessed 3 July 2024.

is the work done by Stanford University's Alpaca team[59] on Meta's Llama LLM.[60]

Moving from Lab Experiments to Real Deployments

In a military context, there are many challenges associated with scaling up AI-based mental health application pilots to real deployments that support military personnel. In the UK, the Defence Mental Health and Wellbeing Strategy 2022–27[61] recognises the importance of mental health, but this article argues that an increased awareness of the impact of mental health on military readiness is needed at a policy level so that funding can be allocated to enact change. There are also structural issues on how useful data on mental health assessment is often held in data silos, and an open question about the need for rebalancing the trade-offs[62] between the right to privacy for soldiers and the military's duty of care for those same soldiers.

Military spending is quite rightly scrutinised closely to ensure taxpayers receive value for money. Competition for resources is fierce among government departments and within the Ministry of Defence and single services. Large capital equipment programmes often steal the funding limelight leaving less 'shiny' personnel issues wanting. Intense conflict has been missing from the British military canon for nearly a decade after the drawdowns in Iraq and Afghanistan. This has arguably allowed the focus on mental health to atrophy, although self-harm and suicide figures still make headlines. There is a danger that the embrace of technology to support mental health stops at providing generic access to applications such as Headspace.[63] While helpful, there is much more to be gained. Equally, in the authors' opinion, the British military must work hard to recruit and retain doctors and mental health professionals who are the key community to champion this subject.

Outside mental health support in the military, the use of AI on the battlefield for warfighting is growing. Another open challenge is how the psychological impact that AI adoption is having on soldiers, both positive and negative, can be evaluated. This might require methodologies that include external independent risk assessment, as internal military assessment is likely to be focused on short-term warfighting readiness at the expense of longer-term mental health risks. Moreover, it may be difficult to get soldiers to sign up for trials that might expose mental health vulnerabilities and lead to career-limiting downgrades after risk assessments on the basis of operational safety.

Conclusion

Effective and responsible deployment of AI from the defence and mental health domains has the potential to boost the readiness and resilience of military personnel. However, challenges exist before this vision can become a mainstream reality.

There is an increasing number of studies exploring AI support for mental health with military personnel, but these studies do not often use the latest AI methods. The recent emergence of powerful LLMs able to deliver human-like chatbot applications and the increasing pervasiveness of smartphones and wearable fitness technology able to monitor the human body are trends to note. This article recommends more studies with military veterans using this type of AI.

The authors of this article also think that helping military personnel to self-identify mental health issues and problem behaviours such as alcohol abuse at very early stages, well before the need for a formal clinical diagnosis, may make a significant impact to military readiness and resilience. Self-identification and self-help approaches using AI have the potential to avoid some of the career-limiting stigma associated with a clinical diagnosis appearing on a military service record. In this context, although electronic health records are the most reliable training data for AI models, this article recommends more studies involving military personnel working with data at stages prior to a formal clinical diagnosis. Specifically, this could include interactive self-help chatbots, AI for self-monitoring using wearable devices and AI for social media self-analysis.

For many AI training datasets, including mental health datasets, there is a clear lack of training examples from under-represented groups, leading to challenges in countering racial and gender bias,

59. Rohan Taori et al., 'Alpaca: A Strong, Replicable Instruction-Following Model', Stanford University, 2024, <https://crfm.stanford.edu/2023/03/13/alpaca.html>, accessed 3 July 2024.
60. Meta, 'Meet Llama 3.1', <https://llama.meta.com/>, accessed 3 July 2024.
61. Ministry of Defence, 'Defence People Health and Wellbeing Strategy 2022-2027', 22 June 2022.
62. Daniel A Adler, 'Burnout and the Quantified Workplace: Tensions Around Personal Sensing Interventions for Stress in Resident Physicians', *Proceedings of the ACM on Human-Computer Interaction* (Vol. 6, No. CSCW2, 2022), pp. 1–48.
63. See Headspace, <https://www.headspace.com>, accessed 3 July 2024.

among others. Although powerful, LLMs do not currently report any uncertainty around the output they generate, and so the bias, error and false negative rates are usually unknown. This article recommends more research is carried out on this, and that new AI deployments for military personnel come with an analysis of military group-specific under-representation bias and, where appropriate, guardrails on AI outputs.

Last, the authors also observe that, to date, only limited work has been performed on how best to evaluate the psychological impact that battlefield AI adoption will have on soldiers, both positive and negative. The authors recommend that external independent risk assessment and evaluation of the mental health impacts of new AI deployments is routinely provided to avoid an undue focus on short-term warfighting readiness at the expense of longer-term mental health risks.

This is an exciting moment. The latest AI models may significantly improve military mental health support and associated military readiness and resilience – but care is also needed to train, deploy and evaluate these powerful AI models so that this potential is realised fully. ■

Stuart E Middleton is Associate Professor in the School of Electronics and Computer Science at the University of Southampton. Stuart's research is focused on the natural language processing (NLP) areas of information extraction and human-in-the-loop NLP in domains including mental health, defence and law enforcement.

Daniel Leightley is a lecturer in Digital Health Sciences in the School of Life Course and Population Sciences at King's College London. Daniel's research is focused on developing and evaluating digital therapeutics.

Patrick Hinton is a Major in the British Army's Royal Artillery. Patrick's research interests include the integration of remote and autonomous systems into land forces, as well as the personnel issues facing military forces today.

Sarah Ashbridge is Principal Analyst at the Defence Science and Technology Laboratory (Dstl) and Affiliate Expert at RUSI. Sarah's research interests include the impact of climate change upon operational effectiveness, and mortuary affairs.

Daniel A Adler is a PhD candidate in Information Science at Cornell University. Daniel's research is focused on ubiquitous computing, human-computer interaction, applied ML/AI, and digital health.

Alec Banks is Senior Principal Scientist at the Defence Science and Technology Laboratory (Dstl). Alec's research interests focus on advanced and dependable autonomy and system safety in the defence domain.

Maria Liakata is Professor in natural language processing (NLP) at Queen Mary University of London. Maria's research is focused on NLP for social and biomedical applications, analysis of multimodal and heterogeneous data including personalised longitudinal language processing, opinion mining and summarisation, rumour verification and scientific discourse analysis.

Brant Chee is Principal Scientist at the Johns Hopkins University Applied Physics Laboratory, faculty in General Internal Medicine at Johns Hopkins University School of Medicine and affiliate faculty at the Armstrong Institute for Patient Safety and Quality. Brant's research focuses on machine learning, natural language processing, bioinformatics and healthcare.

Ana Basiri is Professor in geospatial data science and UK Research and Innovation (UKRI) Future Leaders Fellow at the University of Glasgow. Ana's research interest is focused on developing solutions that consider missingness and biases as useful sources of data.

This work was supported by the Economic and Social Research Council (ES/V011278/1), Engineering and Physical Sciences Research Council (EP/Y009800/1) and National Science Foundation Graduate Research Fellowship Program (DGE-2139899).

Appendix: Note on AI Models Based on Deep Learning

For readers less familiar with the modern deep learning-based AI models, there are some excellent articles explaining some of the basics. These include an introduction to convolutional neural networks (CNNs),[64] medium-sized pre-trained transformer models such as BERT and GPT,[65] and the most recent transformer-based large language model architectures such as GPT-4 and Stanford University's Alpaca.[66] A CNN model usually takes a 2D input, such as an image or a directed graph represented as an adjacency matrix, which is encoded as a 2D matrix (tensor), and then applies a filter, or kernel function, to merge values in local regions to produce a feature map which is helpful for a target task such as image classification.

A transformer model usually takes a 1D sequence, such as a sequence of words, encoded as a 1D vector (tensor) and applies a self-attention layer to weight some parts of the sequence more highly than others in the context of a target task such as text classification. Pre-trained CNN and transformer models will first learn sets of weights for each of their layers from a pre-training task, such as randomly masked word prediction, on a very large pre-training dataset such as every article in Wikipedia. Pre-training can encode lots of useful general knowledge into the layers of the model, such as relationships between factual entities or patterns associated with common images. After pre-training, additional layers can be added to the model and all layer weights updated during a second fine-tuning stage of training with a target task and dataset, such as mental health text classification.

64. Pranjal Soni, 'CNN Model With PyTorch For Image Classification', *Medium*, 9 January 2021, <https://medium.com/thecyphy/train-cnn-model-with-pytorch-21dafb918f48>, accessed 6 October 2023.
65. Arjun Sarkar, 'All You Need to Know about "Attention" and "Transformers" — In-depth Understanding — Part 1', Towards Data Science, 15 February 2022, <https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021>, accessed 6 October 2023; Arjun Sarkar, 'All You Need to Know about "Attention" and "Transformers" — In-depth Understanding — Part 2', Towards Data Science, 13 September 2022, <https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-2-bf2403804ada>, accessed 6 October 2023.
66. Taori et al., 'Alpaca', <https://crfm.stanford.edu/2023/03/13/alpaca.html>, accessed 3 July 2024.