AI for Defence: Readiness, Resilience and Mental Health

AUTHOR INFO REDACTED FOR REVIEW PHASE

Artificial Intelligence (AI) is a cross-cutting technology that is making a major impact on behavioural analysis in both the defence and mental health domains. Employing AI well could boost readiness and resilience of military personnel. This article explores how AI is being used today in research and practice for Mental Health in the Defence domain. We identify key challenges that exist and signpost the important trends and directions of travel that could 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[1] has looked across a ten-year period at the self-reported mental health status of serving (i.e. regulars) and ex-serving (i.e. veterans) personnel from the Iraq and Afghanistan conflicts. Veterans are defined as those serving one paid day in the Armed Forces. This work found a significant prevalence for Post-Traumatic Stress Disorder (PTSD) (6%) and alcohol misuse (10%) in addition to common mental health disorders (21%) such as depression and anxiety.

Research[2][3] has identified significant differences between veterans and the general military population in the rates of PTSD. It is estimated that 7.4% of UK veterans suffered from PTSD, whereas the rate in the general military population is 4%. It is important to note that this prevalence rate is not uniform across groups. In 2014/16 among serving regular personnel, the overall prevalence is 5% but amongst veterans, the rate is higher at 7%. Quantifying the difference in the prevalence of specific mental health disorders for soldiers compared to 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 for the better. AI, such as Natural Language Processing (NLP), has recently been applied to mental health problems such as experiments with mobile phone-based screening of college students for depression[4][5], social media-based early risk prediction of depression[6] and therapy and counselling support around suicidal ideation in lifeline conversations[7][8]. However, research papers and practice reports around using AI for non-military clinical mental health and AI for

[1] Stevelink, S. et al, 'Mental health outcomes at the end of the British involvement in the Iraq and Afghanistan conflicts: A cohort study. The British Journal of Psychiatry', 213(6), 690-697. 2018. doi:10.1192/bjp.2018.175
[2] Kok, B.C. et al, Posttraumatic stress disorder associated with combat service in Iraq or Afghanistan: reconciling prevalence differences between studies. JNMD. 2012;200(5):444–450
[3] Hines, L.A. et al, Posttraumatic stress disorder post Iraq and Afghanistan: prevalence among military subgroups. Can J Psychiatry. 2014 Sep;59(9):468-79. doi: 10.1177/07067437140590090
[4] Tlachac, M.L. et al, 'StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic', Interact. Mob. Wearable Ubiquitous Technol. 6, 2, Article 76 (July 2022), 32 pages. https://doi.org/10.1145/3534604
[5] Tsakalidis, A. et al, Combining Heterogeneous User Generated Data to Sense Well-being, COLING 2016, https://aclanthology.org/C16-1283/
[6] Wu, J. et al, 'Exploring Social Media for Early Detection of Depression in COVID-19 Patients', Web Conference 2023 (WWW '23), ACM, 3968–3977. https://doi.org/10.1145/3543507.3583867
[7] Wang, Z. et al, 'Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations', EACL 2023, pages 1436–1446, 2023, https://aclanthology.org/2023.eacl-main.105
[8] Chim, J. et al, Overview of the CLPsych 2024 Shared Task: Leveraging Large Language Models to Identify Evidence of Suicidality Risk in Online Posts, CLPsych 2024, https://aclanthology.org/2024.clpsych-1.15/

military personnel-based mental health have traditionally been published in separate venues leading to silos of excellence within each community that are rarely bridged[9]. There have been some recent attempts to bridge these communities though, such as the 2023 Workshop on AI, Defence and Mental Health[10]. This article continues this work, highlighting recent examples of excellence and key challenges from both communities and signposting some exciting directions of travel for the future that have the potential to bring both areas closer together.

AI Research and Practice for Mental Health in the Defence Domain

For readers less familiar with the modern deep learning-based AI models discussed in this section there are some excellent articles explaining some of the basics. These include an introduction to Convolutional Neural Networks (CNNs)[11], medium-sized pre-trained Transformer models such as BERT and GPT[12,13] and the most recent Transformer-based Large Language Model (LLM) architectures such as GPT-4 and Stanford's Alpaca[14]. A CNN model usually takes a 2-dimensional 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 1-dimensional 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. Pretrained 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.

Studies have shown that the suicide rate among members of the United States Armed Forces rose soon after the onset of military operations in Iraq and Afghanistan, and that in 2012 suicide was the second-leading cause of death among serving military personnel, higher than combat related causes[15]. 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)[16], Convolutional Neural Network

[9] Leightley D, Murphy D. Personalised digital technology for mental health in the armed forces: the potential, the hype and the dangers. BMJ Mil Health. 2023 Feb;169(1):81-83. doi: 10.1136/military-2022-002279
[10] Workshop on AI and Defence: Readiness, Resilience and Mental Health, 2023, DSTL AI Fest 5 and Turing AI UK-2023 fringe event, https://www.southampton.ac.uk/~sem03/AI-and-Defence-2023.html
[11] https://medium.com/thecyphy/train-cnn-model-with-pytorch-21dafb918f48, accessed 6 October 2023
[12] https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021, accessed 6 October 2023
[13] https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-2-bf2403804ada
[14] https://crfm.stanford.edu/2023/03/13/alpaca.html, accessed 6 October 2023
[15] Gregg Zoroya. Army, Navy suicides at record high. 2012. USA Today. http://www.usatoday.com/story/news/nation/2012/11/18/navy-suicides-army/1702403/, accessed 4 October 2023.
[16] Thompson, P. Bryan, C. Poulin, C. Predicting military and veteran suicide risk: Cultural aspects, CLPsych 2014, ACL, https://aclanthology.org/W14-3201

(CNN) and Transformer-based pre-trained language models with[17] and without[18] 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[19] and Transformer-based text summarization models trained on transcript data from lifeline conversations[20]. Reported accuracy for military personnel suicided risk prediction using unstructured text in Electronic Health Records and social media such as Reddit or question answer websites is as high as 89%.

For military veterans with PTSD, small scale studies[21,22] 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. Additionally, Support Vector Machines[23] and Linguistic Inquiry Word Count (LIWC) lexicon-based n-gram language models[24] have been explored for PTSD classification on small datasets with reported accuracies around 60% ± 7.2%. A study[25] focused on the UK military explored not just the role of machine learning, but which features contributed to the outcome. The authors found, when trying to predict PTSD outcome, that alcohol misuse, gender and deployment status all contributed to the predictive performance of Random Forest models.

Outside the military context there has been a lot of work exploring AI for mental health, often focussing on publicly available datasets from social media sources including Twitter and Reddit which 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 we discuss next, studies into digital interventions such as AI-driven mental health moderation[26,27] have shown toxicity in conversations can be reduced and characteristics such as civility and supportiveness increased.

Standard machine learning algorithms such as Random Forest models and Support Vector Machine models have been used for some time to classify PTSD and depression[28] and

---

[17] Rawat, B.P.S. and Yu, H. 'Parameter Efficient Transfer Learning for Suicide Attempt and Ideation Detection', LOUHI 2022, ACL, https://aclanthology.org/2022.louhi-1.13

[18] Park, S. et al, 'Suicidal Risk Detection for Military Personnel', EMNLP 2020, ACL, https://aclanthology.org/2020.emnlp-main.198

[19] Shing, H. et al, 'Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings', CLPsych 2018, ACL, https://aclanthology.org/W18-0603

[20] Wang, Z. et al, 'Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations', EACL-2023, ACL, https://aclanthology.org/2023.eacl-main.105

[21] Evans, H. et al, 'Understanding the Care Ecologies of Veterans with PTSD', CHI '20, ACM, https://doi.org/10.1145/3313831.3376170

[22] Barish, G. et al, 'A Mobile App for Patients and Those Who Care About Them: A Case Study for Veterans with PTSD + Anger', PervasiveHealth'19, ACM, https://doi.org/10.1145/3329189.3329248

[23] Byers, M. and Metsis, V. 'Text Analysis for Understanding Symptoms of Social Anxiety in Student Veterans', AAAI-2021, AAAI, https://doi.org/10.1609/aaai.v35i18.17975

[24] Coppersmith, G. Dredze, M. Harman, C. 'Quantifying Mental Health Signals in Twitter', CLPsych 2018, ACL, https://aclanthology.org/W14-3207/

[25] Leightley, D. Williamson, V. Darby, J. Fear, N.T. Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort, Journal of Mental Health, 28:1, 34-41, 2019, DOI: 10.1080/09638237.2018.1521946

[26] Wadden, D. et al, 'The Effect of Moderation on Online Mental Health Conversations', ICWSM-2021, AAAI, https://doi.org/10.1609/icwsm.v15i1.18100

[27] Sharma, A. et al, Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 5, 46–57, 2023. https://doi.org/10.1038/s42256-022-00593-2

[28] Saleem, S. et al 'Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts', COLING-2012, ACL, https://dl.acm.org/doi/abs/10.1145/3442381.3450097

identify mental health risk markers in electronic health records[29], 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. These are used for tasks such as mental health classification[30], online counselling behaviour prediction[31] and support to re-write conversational posts more empathically[32]. The strong performance of pre-trained Transformer-based models can also be enhanced[33] by adding model layers trained using n-gram features based on clinical psycholinguistic lexicons. For longitudinal mental health tasks, such as classification of mood changes and mental health risk over time, post embedding methods involving path signatures[34], Hawkes processes[35], as well as the use of Transformers, Bi-directional Long-Short Term Memory (BiLSTM)[36] and multi-task learning[3738] have been successfully used. Pre-trained Transformers can also be used in more complex models such as multi-channel CNN's[39], which utilizes the encoded distance from a post's embedded text to embedded symptom descriptions and achieves accuracies of up to 95% for classifying mental health issues including depression, anxiety, and bipolar disorders.

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[40] 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[41] explored two-step domain adaption of Transformer-based models to try and overcome these problems, first adapting a general pre-trained model to social media writing style and then to mental health in general before finally finetuning the model for the target mental health topic.

---

[29] Alvarez-Mellado, E. et al 'Assessing the Efficacy of Clinical Sentiment Analysis and Topic Extraction in Psychiatric Readmission Risk Prediction', LOUHI 2019, ACL, https://aclanthology.org/D19-6211/
[30] Ji, S. et al 'MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare', LREC-2022, ACL, https://aclanthology.org/2022.lrec-1.778/
[31] Li, A. et al 'Understanding Client Reactions in Online Mental Health Counseling', ACL-2023, ACL, https://aclanthology.org/2023.acl-long.577
[32] Sharma, A. et al 'Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach', WWW '21, ACM, https://doi.org/10.1145/3442381.3450097
[33] Zanwar, S. 'The Best of Both Worlds: Combining Engineered Features with Transformers for Improved Mental Health Prediction from Reddit Posts', SMM4H-2022, ACL, https://aclanthology.org/2022.smm4h-1.50/
[34] Tseriotou, T. et al 'Sequential Path Signature Networks for Personalised Longitudinal Language Modeling', Findings of the ACL 2023, ACL, https://aclanthology.org/2023.findings-acl.310/
[35] Hills, A. Tsakalidis, A. Liakata, M. Time-Aware Predictions of Moments of Change in Longitudinal User Posts on Social Media, ECML PKDD Workshop, AALTD 2023, https://doi.org/10.1007/978-3-031-49896-1_19
[36] Tsakalidis, A. et al 'Identifying Moments of Change from Longitudinal User Text', ACL-2022, ACL, https://aclanthology.org/2022.acl-long.318/
[37] Azim, T. et al 'Detecting Moments of Change and Suicidal Risks in Longitudinal User Texts Using Multi-task Learning', CLPsych-2022, ACL, https://aclanthology.org/2022.clpsych-1.19/
[38] Singh, G.L. at al, ConversationMoC: Encoding Conversational Dynamics using Multiplex Network for Identifying Moment of Change in Mood and Mental Health Classification, ML4CMH 2024, AAAI, https://ceur-ws.org/Vol-3649/
[39] Song, H. et al 'A Simple and Flexible Modeling for Mental Disorder Detection by Learning from Clinical Questionnaires', ACL-2023, ACL, https://aclanthology.org/2023.acl-long.681/
[40] Harrigian, K. Aguirre, C. Dredze, M. 'Do Models of Mental Health Based on Social Media Data Generalize?', Findings of EMNLP-2020, ACL, https://aclanthology.org/2020.findings-emnlp.337/
[41] Aragon, M. et al 'DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media', ACL-2023, ACL, https://aclanthology.org/2023.acl-long.853/

Although much prior work has focussed on textual data, multi-modal methods have been explored including embeddings of Flickr images[42] and BitChute video content[43] using models such as a 3D-CNN and Vision Transformer (ViT). There is also a body of research exploring Internet of Things (IoT)[44] 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, with 100,000's of users across 18 major cities, to show that circadian rhythms associated with sleep quality can be extracted from social media usage data[45] as a useful feature for mental health analysis. Other studies have explored the use of mobile phones to predict stress[46] and use of Fitbit devices to predict depression[47]. 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[48] and modern Transformer-based chatbots to uplift the sentiment of online dialogue[49]. Work has also explored passive use of digital assistants, such as using interactions with Alexa devices in the home to classify risk of mental health issues[50]. In a military context, studies[51] have explored using mobile apps, behaviour change and machine learning to successfully reduce alcohol consumption, over a two to three month period, for help-seeking military veterans and serving personnel.

Trends, Challenges and Open Questions

Methods

There is a trend towards using AI research in the mental health domain, especially around NLP-based approaches. After a few years delay, AI research is finding its way into experiments involving a military context such as military PTSD, especially with medium-sized pre-trained Transformers like BERT and GPT. Large Language Models (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 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

---

[42] Xu, Z. Pérez-Rosas, V. Mihalcea, R. 'Inferring Social Media Users' Mental Health Status from Multimodal Information', LREC-2020, ACL, https://aclanthology.org/2020.lrec-1.772/

[43] Das, M. et al 'HateMM: A Multi-Modal Dataset for Hate Video Classification', ICWSM-2023, AAAI, https://doi.org/10.1609/icwsm.v17i1.22209

[44] https://en.wikipedia.org/wiki/Internet_of_things

[45] Zhou, K. et al 'How Circadian Rhythms Extracted from Social Media Relate to Physical Activity and Sleep', ICWSM-2023, AAAI, https://doi.org/10.1609/icwsm.v17i1.22202

[46] Adler, D. et al, 'Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies', PloS one vol. 17,4 e0266516. 27 Apr. 2022, doi:10.1371/journal.pone.0266516

[47] Adler, D. et al, 'Identifying Mobile Sensing Indicators of Stress-Resilience', IMWUT-2021, ACM, https://dl.acm.org/doi/10.1145/3463528

[48] Lee, Y. Yamashita, N. Huang, Y. 'Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional', HCI-2021, ACM, https://doi.org/10.1145/3392836

[49] Saha, T. et al, 'A Shoulder to Cry on: Towards A Motivational Virtual Assistant for Assuaging Mental Agony', NAACL-2022, ACL, https://aclanthology.org/2022.naacl-main.174/

[50] Chua, H. Caines, A. Yannakoudakis, H. 'A unified framework for cross-domain and cross-task learning of mental health conditions', NLP4PI, ACL, https://aclanthology.org/2022.nlp4pi-1.1/

[51] Leightley, D. 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. 10,6 e38991. 20 Jun. 2022, doi:10.2196/38991

unhelpful mental health interventions, this AI adoption delay is probably also a result of a more cautious approach by researchers and practitioners. 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 around 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-focussed mental health context. Even for non-military studies[52] there is very little work yet around LLMs for mental health and this is a research gap that offers an exciting direction of travel 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 conformation bias will need to be addresses first as discussed later. For example, mental health LLM applications that exhibit confirmation bias around 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 around widely available devices such as wearable fitness technology, smartphones and AI chatbot apps for self-help purposes[53]. 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 making it easier for people to recognize problems and subsequently reach out to a medical professional. As evidenced from this articles authors personal experience, in the military mental health 'first responders 'often take the form of trusted friends and AI-based services could be well positioned to take on this role, although trust by users in the confidentiality of such services remains to be seen. 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. Open questions around this trend include the wider liability issues around AI, such as who is responsible when and if mental health warning signs are missed[54].

Data

With regards 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[55] 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

---

[52] Singh, L.G. et al, Extracting and Summarizing Evidence of Suicidal Ideation in Social Media Contents Using Large Language Models, CLPsych 2024, https://aclanthology.org/2024.clpsych-1.20/

[53] Darcy, A. et al, Evidence of Human-Level Bonds Established With a Digital Conversational Agent: Cross-sectional, Retrospective Observational Study JMIR Form Res 2021;5(5), doi: 10.2196/27868

[54] Nghiem, J. 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 Form Res 2023, doi: 10.2196/47380

[55] Leightley, D. 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, Volume 113, 2018, Pages 17-25, ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2018.02.012

identified[56]. 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.

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[57]. Clinical experts such as psychiatrists have been asked to label social media datasets for AI training purposes, but 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 around 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 application 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 we expect to see more of in the future.

Errors, bias and ethics

It has been well reported[58] that generative AI, especially the latest LLMs, suffer from problems such as hallucinations, loss of information, encoded bias and privacy concerns. Problems around NLP model bias have been seen for many years now and can be hard to reliably detect due to weak evaluation regimes that lack authenticity[59]. 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 manifest 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, making it hard for humans to identify these errors. Information loss, in the form of missing key points, often occurs when LLMs are asked to generate summaries. With regards 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.

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

---

[56] Chui, Z. et al, Mental health problems and admissions to hospital for accidents and injuries in the UK military: A data linkage study. PLOS ONE 18(1), 2023, https://doi.org/10.1371/journal.pone.0280938

[57] Pater, J.A. et al, Social Media is not a Health Proxy: Differences Between Social Media and Electronic Health Record Reports of Post-COVID Symptoms, Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 148, 2023, https://doi.org/10.1145/3579624

[58] Menick et al. Teaching language models to support answers with verified quotes, DeepMind, 2022, https://arxiv.org/abs/2203.11147

[59] Tsakalidis, A. at al, Can We Assess Mental Health Through Social Media and Smart Devices? Addressing Bias in Methodology and Evaluation, ECML PKDD 2018, https://doi.org/10.1007/978-3-030-10997-4_25

military which includes soldiers such as the Brigade of Gurkhas, who are recruited from Nepal[60], and commonwealth citizens recruited for roles in the military such as Fijians[61]. This demonstrates the need for approaches which 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 bias. 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[62].

There is also the unique military culture to consider, where rude is not always rude but it can simply be part of the bonding or establishing camaraderie within a military group[63]. Humour can function as a "disciplinary technology" within the British defence context, amongst others[64]. These examples reflect quite different patterns of communication than existing 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 application is that LLMs currently do not report any uncertainty around the output they generate, and so the bias, error and false negative rates are completely unknown. For a mental health application this means it is very hard to quantify the risk of potential warning signs being missed. There is active research into improving LLMs in this regard and progress will come over time, but at the current moment 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[65]. 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 Bill[66] 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. 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 can AI competition 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

[60] https://www.army.mod.uk/who-we-are/corps-regiments-and-units/brigade-of-gurkhas/
[61] https://assets.publishing.service.gov.uk/media/5e3c0a4c40f0b609169cb58f/FOI201909980-_Number_of_serving_Fijians_Redacted.pdf
[62] Fenech, G. Thomson, G. Defence against trauma: women's use of defence mechanisms following childbirth-related trauma, Journal of Reproductive and Infant Psychology, 33:3, 268-281, 2015, DOI: 10.1080/02646838.2015.1030731
[63] Caddick, N. Smith, B. Phoenix, C. Male combat veterans' narratives of PTSD, masculinity and health, Sociology of Health and Illness, 37:1, 97-111, 2015, DOI: 10.1111/1467-9566.12183
[64] Godfrey, R. Soldiering on: Exploring the role of humour as a disciplinary technology in the military. Organization, 23(2), 164-183. 2016, https://doi.org/10.1177/1350508414533164
[65] https://www.gov.uk/government/topical-events/ai-safety-summit-2023
[66] https://bills.parliament.uk/bills/3137

exemplar for this type of commercial/academic engagement is the work done by Stanford's Alpaca team[67] on Meta's Llama LLM[68].

Moving from lab experiments to real deployments

In a military context, there are a lot of 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 2022-2027[69] recognises the importance of mental health, but we argue an increased awareness of the impact of mental health on military readiness is needed at a policy level to allow funding to enact change. There are also structural issues around how useful data around mental health assessment is often held in data silos, and an open question about the need for rebalancing the trade-offs[70] between the right to privacy for soldiers verses the military's duty of care for those same soldiers.

Military spending is quite rightly scrutinised closely to ensure the taxpayer receives value for money. Competition for resource is fierce between 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 which 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 like Headspace[71]. This is helpful, but there is much more to be gained. Equally, in the authors opinion, the British military has to 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 can we best evaluate the psychological impact AI adoption is having on soldiers, both positive and negative. This might require methodologies that include external independent risk assessment, as internal military assessment is likely to be focussed on short-term warfighting readiness at the expense of longer-term mental health risks. Moreover, the ability to get soldiers to sign up for trials which may expose mental health vulnerabilities which can lead to career limiting downgrades may be difficult.

Conclusion

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

There are 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 increasing

[67] https://crfm.stanford.edu/2023/03/13/alpaca.html
[68] https://llama.meta.com/
[69] https://assets.publishing.service.gov.uk/media/62b3333dd3bf7f0af6480740/Defence_People_Health_and_Wellbeing_Strategy.pdf
[70] Adler, D.A. et al, Burnout and the Quantified Workplace: Tensions around Personal Sensing Interventions for Stress in Resident Physicians. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 430, 2022, https://doi.org/10.1145/3555531
[71] https://www.headspace.com

pervasiveness of smartphones and wearable fitness technology able to monitor the human body are trends to note. We would recommend more studies using this type of AI with military veterans as a promising direction for future research.

We also think helping military personnel 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, could 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, we recommend 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 which leads to challenges around racial and gender bias. 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. We would recommend more research is performed in this direction, 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.

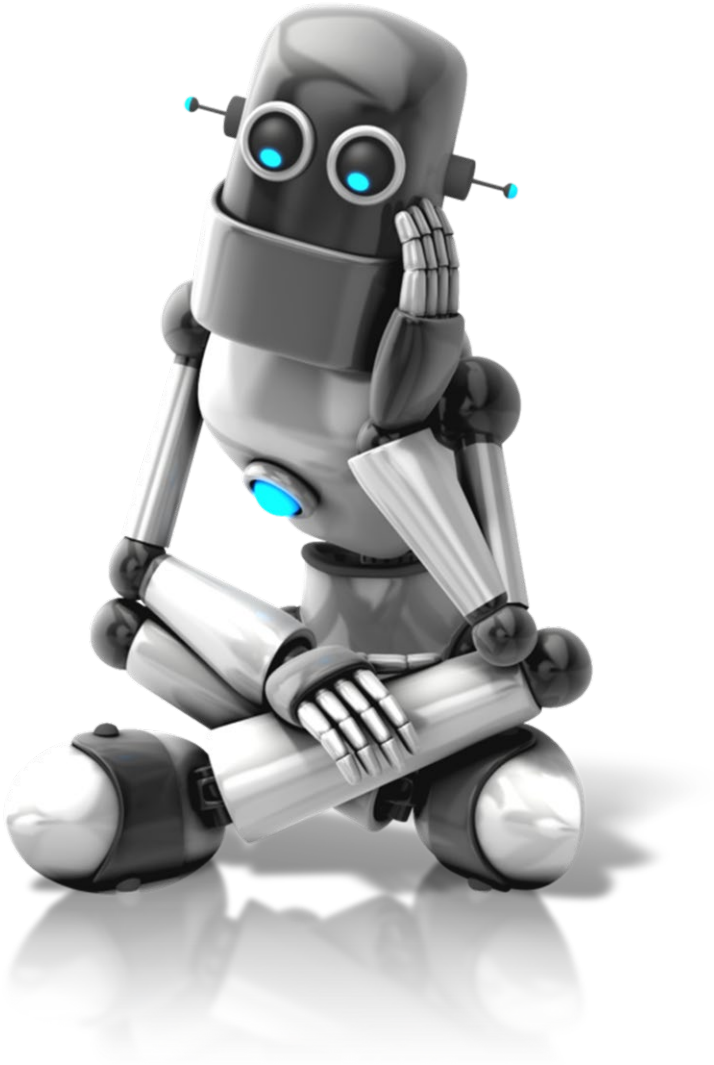
Lastly, we also observe that to date only limited work has been performed on how best to evaluate the psychological impact battlefield AI adoption will have on soldiers, both positive and negative. We recommend that external independent risk assessment and evaluation of the mental health impacts from 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.

We are at an exciting moment in time – the latest AI models could 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 this potential is realized fully.

Acknowledgements

FUNDER INFO REDACTED FOR REVIEW PHASE

Figures



Artwork for article (optional for RUSI to use – authors have full rights for use of image)