JDST., vol. 4, no. 1, p1-p11, 2021
https://doi.org/10.46713/jdst.004.01

www.jdst.eu, e-ISSN 2534-9813

Research Article

Military Dataset Processing Approaches or Trauma Risk Mitigation in Machine Learning Practitioners

Mélanie Breton [a]***, Valérie Lavigne*** [a]***, Malek Djaffri*** [a]***, Maxime Dionne*** [a]

[a] *Defence Research and Development Canada,*
2459 Route de la Bravoure, Quebec City, Quebec, G3J 1X5, Canada
http://www.drdc-rddc.gc.ca/

ABSTRACT

To develop new computer vision capabilities leveraging artificial intelligence, we will increasingly need to use operationally realistic training and validation datasets. Although operational full motion video and imagery datasets present information regarding their provenance and classification level, these designations are often not indicative of the presence of potentially offensive or traumatizing content. As machine learning and data scientists increasingly need to work with operational unsanitized operational video and imagery data, they will have a higher risk of being exposed to sensitive and traumatic content. In this paper, we first raise awareness about this risk within the Defense community. Then, we propose several approaches for mitigating machine learning practitioner's exposure to offensive and traumatizing media, including dataset preprocessing procedures and viewing tool design considerations.

✉ Corresponding Author: Tel.: + Fax: + ; E-mail:

ARTICLE INFO

RECEIVED: 31 Jan 2021
REVISED: 10 Nov 2021
ACCEPTED: 30 Nov 2021
ONLINE: 12 DEC 2021

KEYWORDS

Classification, dataset labelling, machine learning, moderation, operational data exploitation, sanitization, sensitive media, trauma risk mitigation, traumatic risk mitigation, user interaction.

I. INTRODUCTION

The last ten years offered several advancements in computer vision powered by deep learning based Artificial Intelligence (AI). To leverage these new capabilities for defence and security purposes, it is often necessary to retrain deep learning models with realistic imagery that is relevant to the problem we aim to solve. Although there is an interest in few shot learning, thus with smaller dataset, state-of-the-art deep learning model training still requires large annotated datasets.

Although most of computer vision scientists are at minimal risk of seeing traumatic images, this risk increases when working with datasets that are close to operational reality. Indeed, several ML (Machine Learning) tasks such as data exploration and model validation require the scientist to visually explore and inspect their data when done on a curated dataset, the exposure risk is low. However, the risk to see offensive media increases should "in the wild" or and uncurated dataset be used.

Relevant dataset for military applications may come from operational sources, as well as unreliable sources such as public internet, especially for capabilities related to Open-Source Intelligence (OSINT) materials. As those datasets tend to be large, it can be hard to have complete awareness of their content and the risk of containing content that is unsafe for Machine Learning (ML) practitioners must be considered.

As stated by the Canadian Defence Policy, Strong, Secure, Engaged: "The mental health and workplace well-being of defence civilians is also critical to the success of the Canadian Armed Forces" [1]. Therefore, we consider this topic as critical for artificial intelligence research applied to military relevant problems. In this paper, we first review different approaches being used within other application domains such as the commercial, academic, and public sectors. We then describe the AI capability development cycles and identify the steps that where ML practitioners are more vulnerable to unsafe content. Finally, we propose processing approaches and viewing tool design considerations identified specifically for ML activities.

II. HANDLING OFFENSIVE MEDIA: OTHER APPLICATION DOMAINS

First, let's define what we consider as offensive and traumatic media. We define offensive and traumatic media as images and/or video that are perceived, by the viewer, as violating the social contract [2]. For example, these media may contain acts of violence, graphical injury, hateful signs, racist internet memes[1], or even adult content.

Several studies already addressed the impact of vicarious witnessing on practitioners in defence and security field [3], [4], in social and medical fields [5], as well as with news reporters [6], [7]. Some studies also looked at the impact on online moderators [8], [9] working with user generated content (UGC). Although scientists [10], [11] reported on the offensive data and label found in well-known Open-Source datasets such as ImageNet [12] and Tiny images [13], we did not find any study about the effects of unsettling images on scientists working in the fields of computer science, data science, and artificial intelligence. Neither did they proposed methodology on how to work with this type of unsanitized data.

Online content moderation to ensure users are not presented offending content has been a critical issue for web platforms which present user generated content. Relying on human moderators has been shown to have a high human cost [14], [15]. Therefore, online content providers have increasingly developed and used AI-based filtering capabilities for online media moderation [16] – [21]. However, for sensitive or classified operational datasets, online solutions are unsuitable.

With the rise of data journalisms and OSINT (Open-Source Intelligence) studies, practical recommendations have been developed to protect reporters [7], [22], [23]. These include several practical tips such as limiting the time spent reviewing traumatic images, self-care routines, and limiting exposure intensity. While a lot of the coping approaches are the same for all practitioners working with unsafe materials, some gaps remain for those dealing images and videos. Although best practices suggest reducing stimuli intensity by removing sound or reducing viewing resolution and colour span, dedicated software tools are rarely proposed to help machine learning practitioners with reviewing sensitive materials for their specific projects. This is a gap in the machine learning field.

III. Machine Learning Practitioners Vulnerability in Military Artificial Intelligence Capability Development

While anyone who manipulates the data is at risk of a vicarious witnessing trauma, in this section we focus on the risks that are specific to AI capabilities development activities.

Because scientists must visually inspect images and videos at any point during their AI development cycle, there is a risk of stumbling upon traumatic or offensive media. For example, a scientist might work on developing a model for the detection and identification of explosion or fires in images or videos, or work on the

[1] An internet meme is a type of media that spreads virally through the internet, and it is used for humorous purposes, or to express sarcasm. It is usually under the form of either a GIF, an image with text, or a short video clip.

development of a model for the battle damage assessment. As a single image can cause an element of surprise [7], [8], any scientists could suffer from vicarious trauma. The nature of the machine learning scientists' tasks might be closer to the task done by news reporter and online moderators. However, the scientists' tasks might involve less time spent at looking at sensitive photos in details than the reporter do.

Table 1 summarizes the type of work related with computer vision and the associated risk of encountering trauma images. The table was built using computer vision roles and tasks related to developing AI models at Defence Research and Development Canada. We qualitatively determined the risk level using the likelihood of working directly with images or videos within our past projects. Please note that this list is not exhaustive. The practitioners' types and tasks are based on our AI development cycle which consists in 5 steps: data gathering, data labelling, model development, model audit, and model deployment. The risk is elevated for the tasks involving looking at the raw data - either for cleaning the dataset, for labelling purpose, or for validating a new trained model. This is because these practitioners act at the beginning of the processing chain and thus are on the frontline when it comes to tagging risky content in the data. Thus, they typically cannot rely on filtering performed by previous practitioners and must inspect the raw data themselves.

The authors assessed the risk with three levels: low, medium and high. The authors considered the number of images to be viewed for the task. A second factor was at what point in the workflow is the data viewed, considering whether traumatic images may have been filter before that. At the beginning of the machine learning workflow, the risk is higher, while the practitioner auditing the end results will interact will filtered images.

Table 1. Type of practitioners and the risk of encountering traumatic images.

Type	**Task**	**Risk of encountering traumatic images**
Data scientist, computer vision practitioner	Gather and collect the data	Medium risk
Data scientist, computer vision practitioner	Clean the data	Elevated risk
Data labeller	Draw bounding box around objects, tag, and classify imagery	Elevated risk
Data scientist, computer vision practitioner	Visually inspect the data prior to developing new model	Elevated risk
Data scientist, computer vision practitioner	Validate the trained model by visually inspecting the results	Medium risk

Military Dataset Processing Approaches or Trauma Risk Mitigation in Machine Learning Practitioners		
Data management engineer	Might be exposed during digital archiving.	Minimal risk
Partner, client, project manager	Receive and read the report that could include disturbing images	Minimal risk

IV. Dataset Processing Approaches

The goal of automated dataset pre-processing is to identify as many unsafe images and video frames as possible to allow users to flag them and decide whether to filter them out from their working subset of the data or prevent them from being displayed.

A. Tagging Approach

First, we want to ensure that the dataset is adequately stored, organized, and pre-processed. For this purpose, we need to develop a tagging approach to correctly annotate unsafe content. The goal is not to limit the access to the material, but to limit the element of surprise which could increase the risk of trauma [7], [8].

Although we could develop a complex naming convention for unsafe content, a simple tagging approach can be effective. A tagging hierarchy is desirable to allow for easily filtering out all unsafe content at once. In practice, we will start with using the tags that are provided by the open-source models [25-27] gathered to pre-process the data (for example: not safe for work, sexy, bare belly, etc.), and we will regroup all these tags in a tree structure under a common higher level “unsafe” tag. Therefore, we can provide a warning to a scientist who needs access to a specific media tagged as unsafe.

B. Automated Detection of Unsafe Content

For a team beginning to work with unsafe images, open-source detection and classification models can be leveraged to annotate individual images and video frames [24] – [28].

Then, models that are specifically designed to detect unsafe content can be combined to other neutral models for an increased detection capability. For example, if one suspects that a dataset may content violence on children, pre-tagging all instances of children in images with or without violence will help filtering out some traumatic imagery.

Additionally, we can leverage generic image similarity detection algorithms to verify whether a given image is similar to previously tagged unsafe content. For this purpose, we can use perceptual hashing [29] which detects similar images by comparing them with a distance metric (such as the Hamming distance or the Cosine distance). Another approach is to leverage Convolutional Neural Network (CNN) feature extraction which compares two images by comparing their vector at the output of a neural network.

C. Dataset Storage

Finally, it is advisable to store the unsafe content separately from the rest of the dataset to avoid accidental exposure. Access to this data should be limited to development work where it is necessary. In that case, it will be important to provide a viewing tool that reduces the intensity of exposure to unsafe data.

V. Viewing Tool Design Considerations

Working with unsafe image and video data may be unavoidable for some AI capability development and validation activities. There is a need to adapt viewing tools to reduce the risk of trauma when working with imagery that has been flagged as potentially presenting a risk to mental health. In this section, we present design considerations that allow for exposure minimisation and intensity reduction, when exposure is necessary (some design features concepts are illustrated in Figure 1). Because machine learning practitioners will need to work on sensitive and classified datasets, it is imperative for the viewing tool to be available offline.

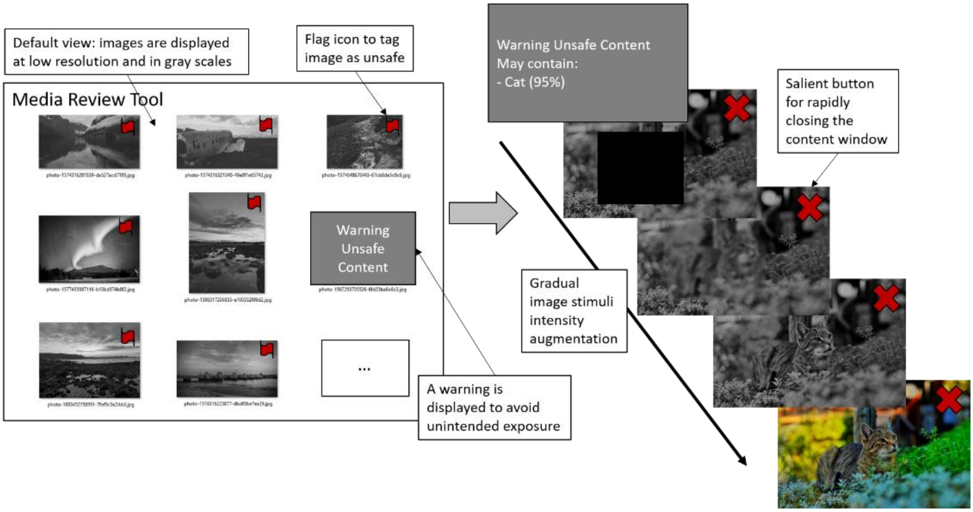

Default view: images are displayed at low resolution and in gray scales
Flag icon to tag image as unsafe
Media Review Tool
Warning Unsafe Content
...
A warning is displayed to avoid unintended exposure
Warning Unsafe Content May contain: - Cat (95%)
Salient button for rapidly closing the content window
Gradual image stimuli intensity augmentation

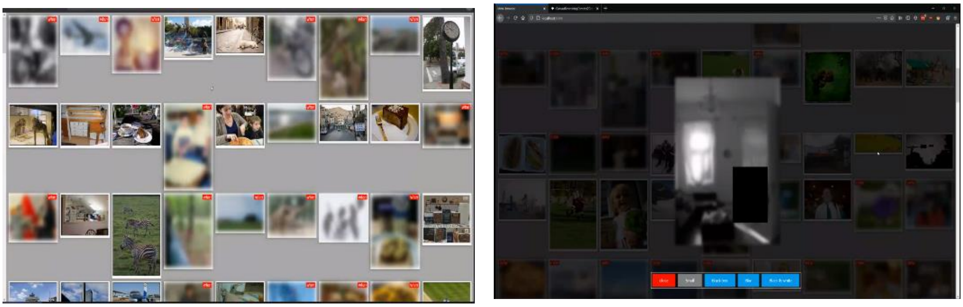
Figure 1: Illustration of reviewing tool design recommendations. A cat image is used as a placeholder for unsafe content.
(Photographies credit: https://unsplash.com/@bitcloudphotography).

Figure 2 shows our first implementation of a web interface to explore visually dataset containing images tagged as unsafe. The web interface was designed as a scrollable photo gallery using python (FastAPI, Uvicorn), HTML (Hypertext Markup Language), JavaScript, and Vue.js. The interface is designed to blur the unsafe image as well as adding a red flag at its upper-left corner. When selected for viewing, the image will have a low resolution, be desaturated, as well as blurred. If a bounding box is provided a black box will cover the unsafe area. Then as needed, the user can toggle all four options while inspecting the image.

A. Flagging Unsafe Content

It is important to provide users with the possibility to easily tag a video or an image containing unsafe content and update the metadata pertaining to that media. This annotation capability should be readily available to minimize the need to go back and review the media a second time.

1. The viewing tool should provide a flagging button to let users annotate unsafe data during dataset exploration



a) Gallery view. Unsafe images are blurred and a red flag is added at upper-left corner of the image.

b) Single image view with toggle buttons to select the image stimuli intensity (resolution, blur, color, obscuring box).

Figure 2: First version of our design concepts to review unsafe content. For this article, images were randomly tagged as unsafe.
(Images credit: random images from ImageNet dataset [12]).

B. Avoid Unexpected Exposure

As stated previously, being surprised by unexpected unsafe content can increase the risk of trauma to machine learning practitioners [7, 8]. To avoid this, the viewing tool first leverages the annotations from the preprocessing models to warn the user that the data may contain disturbing content. This supposes that the content was already reviewed by either an algorithm or by a data scientist working on that same dataset. This warning should display as much information as the model can provide regarding the nature of content detected. This first safeguard ensures that the practitioners can decide to step back and limit accidental exposition to unsafe content.

1. When data that has been tagged as potentially unsafe is loaded, a warning should be displayed to the user to avoid surprise.

C. Avoid Prolonged Exposure

A recommended best practices is limiting the time spent viewing potentially disturbing content [30, 31]. When viewing such content take regular pauses and limit the total contiguous time spent reviewing unsafe content.

1. A time limit can be set for viewing the content in a single day and enforced by a user session system. The total time could be enforced across multiple sessions.
2. To allow users to end exposure as quickly as possible in case they feel overwhelmed, the tool will also feature a salient button for rapidly closing the content window.

D. Reducing Stimuli Intensity

If the user chooses to access the content, there are several ways we can reduce its intensity to limit sensory overload [7], [22], [23].

1. If the annotation includes bounding boxes, the unsafe area should be obscured. If the user explicitly indicates that this part of the image needs to be displayed, it should be first shown partially through a blurring filter. The original unsafe content should only be displayed upon user request.
2. Unsafe images should be displayed at lower resolution by default instead of the usual full screen display.
3. Unsafe images should be converted to washed-out colors to reduce their intensity.
4. The audio track should be muted, by default, once again to limit the vicarious user experience.
5. The user must manually start a video to play it; it must not play automatically.
6. Video files should not automatically be replayed in loops.
7. There should be a feature to play video backwards. This can reduce coherence and emotional impact of unfolding events.

This feature can work both ways, i.e., it can let users edit content tags for media that was incorrectly identified as unsafe. When users interact with content that was tagged as unsafe, they should be able to comment and provide their own assessment. The original information will remain, but a human validated tag will be added to it. This provides a more granular characterization of the unsafe content. User annotations will be displayed along the warning information when other users access this content. Filtering capabilities should let the users decide whether they want to include content that was machine tagged as unsafe, but later be validated as safe by human reviewers.

This added tagging data will also be essential to improving processing models performance through models retraining and validation.

VI. Conclusion

We discussed the risks that are associated with viewing images and videos that present violent content. The mental health protection approaches that are applied in other application domains were discussed. Then, we presented our proposed data preprocessing approach and viewing tool design for the specific protection needs of machine learning practitioners who need to interact with potential unsafe datasets from military operations and unreviewed downloaded public content, such as social media data. Our next step will be to improve our proof of concept that implements the discussed approaches to protect our machine learning practitioners. Using AI will not identify all sensitive images in a dataset but will help raise awareness of sensitive content for data scientists. Although this paper focused on risks related to working with video and images, similar concerns are likely to apply to audio and textual data. Future work could focus on tool design considerations for these types of data.

REFERENCES

[1] Canada, Department of National Defence, "Strong, Secure, Engaged: Canada's Defence Policy," Ottawa, 2017. http://publications.gc.ca/site/eng/9.835971/publication.html. Accessed January 2021.

[2] A. Reid, "How are journalists at risk of vicarious trauma from UGC," Oct, 2014. https://www.journalism.co.uk/news/how-are-journalists-at-risk-of-vicarious-trauma-from-ugc-/s2/a562758/. Accessed January 2021.

[3] S. McCammon, "The warfare may be remote, but the trauma is real," *National Public Radio*, Apr, 2017. https://www.npr.org/2017/04/24/525413427/for-drone-pilots-warfare-may-be-remote-but-the-trauma-is-real. Accessed January 2021.

[4] The Psychlopaedia Team, "Witnessing trauma at work takes an emotional toll," Sept, 2016. https://psychlopaedia.org/work-and-performance/witnessing-trauma-at-work-takes-an-emotional-toll/. Accessed January 2021.

[5] J. M. Newell and G.A. MacNeil, "Professional burnout, vicarious trauma, secondary traumatic stress, and compassion fatigue," *Best Practices in Mental Health*, 6.2, pp. 57-68, Jul, 2010. https://www.ingentaconnect.com/content/follmer/bpmh/2010/00000006/00000002/art00006. Accessed January 2021.

[6] H. Ellis, "How to Prevent, Identify and Address Vicarious Trauma — While Conducting Open-Source Investigations in the Middle East," Oct, 2018. https://www.bellingcat.com/resources/how-tos/2018/10/18/prevent-identify-address-vicarious-trauma-conducting-open-source-investigations-middle-east/. Accessed January 2021.

[7] G. Rees, "Developing Your Own Standard Operating Procedure for Handling Traumatic Imagery," *Dart Centre for Journalism & Trauma*, The Journalism School at Columbia University, Dart Centre Europe, pp. 1-10, Apr, 2017. https://dartcenter.org/resources/handling-traumatic-imagery-developing-standard-operating-procedure. Accessed January 2021.

[8] —CAUTION—Graphic Description—CAUTION—
A. Chen, "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed," Oct, 2014. https://www.wired.com/2014/10/content-moderation/. Accessed January 2021.

—CAUTION—Graphic Description—CAUTION—

[9] Cambridge Consultants, “Use of artificial intelligence for online content moderation,” pp. 1-84, Jul, 2019. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf. Accessed January 2021.

[10] K. Yang, et al., "Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan, 2020.

[11] V. U. Prabhu and A. Birhane, "Large image datasets: A pyrrhic win for computer vision?", *arXiv preprint*, arXiv:2006.16923, Jun, 2020. Accessed January 2021.

[12] J. Deng, et al., "ImageNet: A large-scale hierarchical image database," *IEEE conference on computer vision and pattern recognition*, IEEE, Jun, 2009.

[13] A. Torralba, R. Fergus and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, 30.11, pp. 1958-1970, May, 2008.

[14] A. Arsht and D. Etcovitch, “The Human Cost of Online Content Moderation,” *Jolt Digest*, Harvard, Mar, 2018. https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation. Accessed January 2021.

[15] S.T. Roberts, "Commercial Content Moderation: Digital Laborers' Dirty Work," *Media Studies Publications*, 12., pp. 1-11, 2016. https://ir.lib.uwo.ca/commpub/12. Accessed January 2021.

[16] Amazon Rekognition, “Content moderation," Developer Guide. https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html. Accessed January 2021.

[17] Clarifai, “Moderation”. https://www.clarifai.com/models/moderation-image-recognition. Accessed January 2021.

[18] DeepAI, “Content Moderation API”. https://deepai.org/machine-learning-model/content-moderation. Accessed January 2021.

[19] Google Vision, “Detect explicit content (SafeSearch)”. https://cloud.google.com/vision/docs/detecting-safe-search. Accessed January 2021.

[20] P.Farley, et al., “Learn image moderation concepts,” Microsoft, April, 2020. https://docs.microsoft.com/en-us/azure/cognitive-services/content-moderator/image-moderation-api. Accessed January 2021.

[21] Picpurify, “Automatic image moderation”. https://www.picpurify.com/. Accessed January 2021.

[22] DUTCHOSINTGUY, “Vicarious trauma and OSINT – a practical guide,” *OSINT Curious Project*, Jun, 2020. https://osintcurio.us/2020/06/08/vicarious-trauma-and-osint-a-practical-guide/. Accessed January 2021.

[23] Dart Center for Journalism & Trauma, “Working with Traumatic Imagery,” Aug, 2014. https://dartcenter.org/content/working-with-traumatic-imagery. Accessed January 2021.

[24] D. Won, Z. C. Steinert-Threlkeld and J. Joo, "Protest activity detection and perceived violence estimation from social media images," *Proceedings of the 25th ACM international conference on Multimedia*, Sep, 2017. https://github.com/wondonghyeon/protest-detection-violence-estimation. Accessed January 2021.

[25] J. Mahadeokar and G. Pesavento, "Open sourcing a deep learning solution for detecting NSFW images," Aug, 2016. https://github.com/yahoo/open_nsfw. Accessed January 2021.

[26] Bedapudi6788, "NudeNet model: Neural Nets for Nudity Detection and Censoring," 2020. https://github.com/bedapudi6788/NudeNet. Accessed January 2021.

[27] Emiliantolo, "PyTorch model for NSFW detection," 2019. https://github.com/emiliantolo/pytorch_nsfw_model. Accessed January 2021.

[28] B.O. Sabat, C. Canton-Ferrer and X. Giró-i-Nieto, "Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2020*, pp. 1470-1478, 2020. https://github.com/imatge-upc/hate-speech-detection. Accessed January 2021.

[29] B. Hohannes, "ImageHash," 2020. https://github.com/JohannesBuchner/imagehash. Accessed January 2021.

[30] Holmes, E.A., James, E.L., Coode-Bate, T., and Deeprose, C., "Can playing the computer game 'Tetris' reduce the build-up of flashbacks for trauma? A proposal from cognitive science," PLOS ONE, 4.1, 2009. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004153. Accessed on April 20, 2020.

[31] Wakefield, J., "An online decency moderator's advice: Blur your eyes," BBC, 2018. https://www.bbc.com/news/technology-45664643. Accessed on April 22, 2020.