Assessment Submission Cover Sheet

This Assessment Cover Sheet **must** be included on all Assessment submissions.

|  |  |
| --- | --- |
| Assignment Title | Assignment B – Portfolio Assessment |
| Module | Data Mining |
| Student Name  (same as Student Card) | Ciaran Finnegan |
| Student Number |  |
| Programme |  |
| Part-Time/Full-Time |  |
| Year of Study  (First Year, Second Year, etc) |  |

Late Submissions: Assessment submitted after the deadline will have a late penalty applied.

**Academic Integrity for assessment in TU Dublin Programmes**

Each student is responsible for knowing and abiding by TU Dublin Academic Regulations and Policies. Any student in breach of these regulation/policies will be subject to action in accordance with the University’s procedures for breaches of assessment regulations. Please refer to the General Assessment Regulations at

<https://tudublin.libguides.com/c.php?g=674049&p=4794713>

<https://www.tudublinsu.ie/advice/exams/breachesofregulations/>

All students are expected to complete their courses/programmes in compliance with University regulations. No student shall engage in any activity that involves attempting to receive a grade by means other than honest effort, for example:

1. No student shall complete, in part or in total, any examination or assessment for another person.
2. No student shall knowingly allow any examination or assessment to be completed, in part or in total, for themselves by another person.
3. No student shall plagiarise or copy the work of another and submit it as their own work.
4. No student shall falsify any data. Falsification is the invention of data, its alteration, its copying from any other source, or otherwise obtaining it by unfair means, or inventing quotations and/or references.
5. No student shall use aids or devices excluded by the lecturer in undertaking course work or assessments/ examinations.
6. No student shall knowingly procure, provide, or accept any materials that contain questions or answers to any examination or assessment to be given at a subsequent time.
7. No student shall provide their assignments, in part or in total, to any other student in current or future classes of this module/ programme unless authorised to do so by the lecturer.
8. No student shall submit substantially the same material in more than one module/programme without prior authorization.
9. No student shall alter graded assignments or examinations and then resubmit them for regrading, unless specifically authorised to do so by the lecturer.
10. All programming code and documentation, unless correctly referenced, submitted for assessment or existing in the student’s computer accounts must be the students’ original work or material specifically authorized by the lecturer.
11. Collaborating with other students to develop, complete or correct course work is limited to activities explicitly authorized by the lecturer.
12. For all group assignments, each member of the group is responsible for the academic integrity of the entire submission. Consequently, all group members must satisfy themselves that all elements of their submission adhere to the academic integrity statement points above.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University’s Regulations and Policies covering Academic Integrity (see General Assessment Regulations)*.*

Coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. If you have queries about what constitutes plagiarism, please speak to your lecturer.

|  |  |
| --- | --- |
| Student Signature |  |
| Date |  |

IMPORTANT:

* Complete the required number of tasks as defined in Assessment Handout
* The sections listed below are an example of the section headings for each task. You can use alternative headings
* Tasks 1-3: Sub-Sections 1-7 should be no longer than 8 pages (minimum 6 pages), including diagrams, images, screen captures, tables, etc. Careful selection of these is needed.
  + Code does not count to this total. Code should be added to the relevant section.
* Detailed discussion is expected. Marks are awarded based on depth of information given.
* Marks are awarded based on complexity of problem and depth of work.

# TASK 4 – *Ethics and the user of Data Science/ML/AI*

## Task 4-1 : Stop The Killer Robots – Autonomous Drone Warfare

1. **Overview of problem**

This title may sound like a bad ‘B-Movie’ but the *Campaign to Stop Killer Robots* (https://www.stopkillerrobots.org/) is an umbrella framework for 180+ organisations that is concerned about the growing potential threat of autonomous weapons systems.

In this section of the assignment, I am considering the ethical and legal issues raised by groups like *Stop The Killer Robots* (STKR) that are associated with this extreme end of the spectrum when it comes to autonomy in technology.

What can, and should, be done to ensure that the artificial intelligence development and processes underpinning ‘killer robots’ is accountable and free from abuse?

The question becomes less and less academic each day. We already have present day examples of quasi-autonomous weapons in the field, such as Israel’s *Harpy* anti-radar ‘fire-and-forget’ drone. These, and ongoing military AI development across the globe, raise an ongoing moral dilemma around such technology**[1]**.

1. **Ethical and Legal Challenges**

Paul Scharre, a former US-Army Ranger, and a director at the New American Security ‘think-tank’, wrote in 2019 in his book: *Army of None: Autonomous Weapons and the Future of War* that the Pentagon needed to shift its thinking on artificial intelligence**[2]**.

Scharre distilled his concerns down to two kinds of legal and ethical questions;

1. Machine permissibility. What is the system allowed to do on its own?
2. Machine accountability? Who takes initial (and ultimate) responsibility for what the system does on its own?

In February 2020, as a response to these types of concerns, the US DoD rolled out a list of five AI ethical principles to govern its work in this area**[3]**:

1. Responsibility and good judgement applied by military personnel in the use of AI capabilities.
2. Equitable. Bias is minimized.
3. Traceable. Reasons for AI decisions can be understood.
4. Reliable. Systems tested, secured, and robust.
5. Governable. The ability exists to easily disengage in the case of unintended behaviour.

However, the concern from the *Campaign to Stop Killer Robots* is that these effectively remain guidelines and could ultimately be used by human actors to avoid taking legal (or moral) responsibility for the actions of autonomous weapons systems**[4]**.

The campaign highlights that a new proposed international treaty to prohibit and restrict ‘killer robots’ has been endorsed by dozens of countries**[5]**. Despite this, the major powers remain resistant to new treaties, preferring to look at existing legislation and regulation **[6][7]**. This fuels scepticism in many that the desire to be first in the ‘AI Arms Race’ will lead to compromises in ethical standards.

1. **Challenges for Data Scientist**

The US DoD have declared that they want to integrate ethics into all aspects of their AI test and evaluation processes**[8]**, and thus have outlined policies for their data engineers.

In the need to be equitable and traceable, the AI test harnesses must be able to identify algorithmic bias. It must be clear what data elements are contributing to a systems decision. If a system is literally going to be targeting an individual, or group of individuals, it must be clear what criteria the machine learning model is using to make that decision.

Even for the DoD a major challenge is that AI testing is still heavily dependent on manual assessment. There is widespread engagement with the private sector and academia, but this is still seen as an area of concern. There is a lack of sophisticated toolkits to test AI-driven systems, in the view of the US DoD.

The disengagement mechanism appears to be more of a general engineering challenge in terms of capability, rather than one unique to data engineers. ‘Pulling the plug’ quickly and effectively, if needed, in the event of a suspicious decision requires a generally well-built system architecture.

Other US government departments are impressed with the AI techniques being deployed by the DoD and seek to emulate them in their own ethical artificial intelligence strategies**[9]**.

1. **Reflections**

Many voices in the *Campaign to Stop The Killer Robots* advocate an outright ban on AI technologies being used to create autonomous weapons systems. Professor Noel Sharkey has passionately argued that computers should never be in the business of killing people**[4]**.

However, Is an outright ban even remotely practical? Many in the military today believe that such a ban is impossible **[10]**.

Crucially, there are other voices in the STKR organisation, with both academic and military backgrounds, that push for governments to, at the very least, adopt the US DoD ethical principles and then enshrine this process in multi-lateral treaty agreements. Critically, it should be paramount that ‘permanent significant human control’ always remains in place **[11]**.

To me, it seems that, just like with Nuclear and Chemical weapons before them, AI-based weapons need to be comprehensively covered under dedicated international arms treaties. Such weapons will proliferate and become a great deal harder to count and verify, but pressure needs to be brought to bear on the major powers to recognise the genuine concerns of humanity in the face of ‘killer robots’.

1. **References**

[1] Winter, C., 2017. *'Killer robots': autonomous weapons pose moral dilemma | DW | 14.11.2017*. [online] DW.COM. Available at: <https://p.dw.com/p/2nT6O> [Accessed 11 December 2021].

[2] Scharre, P., 2019. *Army of None: Autonomous Weapons and the Future of War*. 1st ed. New York: W. W. Norton & Company.

[3] Barnett, J., 2020. *Pentagon adopts ethical principles for artificial intelligence*. [online] FedScoop. Available at: <https://www.fedscoop.com/dod-ai-ethics-principles/> [Accessed 11 December 2021].

[4] Wareham, M., 2020. *Robots Aren't Better Soldiers than Humans*. [online] Hrw.org. Available at: <https://www.hrw.org/node/376854/printable/print> [Accessed 11 December 2021].

[5] Wareham, M., 2020. *Killer Robots: Growing Support for a Ban*. [online] Human Rights Watch. Available at: <https://www.hrw.org/news/2020/08/10/killer-robots-growing-support-ban> [Accessed 11 December 2021].

[6] Klane, M., 2018. “U.S., Russia Impede Steps to Ban ‘Killer Robots.’”. *Arms Control Today*, [online] 48(8), pp.31-33. Available at: <https://www.jstor.org/stable/90025262> [Accessed 11 December 2021].

[7] Bowcott, O., 2015. *UK opposes international ban on developing 'killer robots'*. [online] The Guardian. Available at: <https://www.theguardian.com/politics/2015/apr/13/uk-opposes-international-ban-on-developing-killer-robots> [Accessed 11 December 2021].

[8] Barnett, J., 2020. *How the DOD is developing its AI ethics guidance*. [online] FedScoop. Available at: <https://www.fedscoop.com/jaic-alka-patel-ai-ethics/> [Accessed 11 December 2021].

[9] Nyczepir, D., 2020. *HHS AI chief sees promise in emulating the DOD*. [online] FedScoop. Available at: <https://www.fedscoop.com/hhs-ai-office-jaic/> [Accessed 11 December 2021].

[10] Nast, C., 2020. *There's No Turning Back on AI in the Military*. [online] Wired. Available at: <https://www.wired.com/story/opinion-theres-no-turning-back-on-ai-in-the-military/> [Accessed 11 December 2021].

[11] Micha, L. and Farias, P., 2021. *The evolution of disruptive technologies and lethal autonomous weapons systems: considerations from the military field*. [online] Stopkillerrobots.org. Available at: <https://www.stopkillerrobots.org/wp-content/uploads/2021/09/The-evolution-of-disruptive-technologies-and.pdf> [Accessed 10 December 2021].

## Task 4-1 : YouTube in 2021: Trying to tame the recommendation engines that radicalised millions (by accident).

1. **Overview of problem**

Guillaume Chaslot completed is doctorate in Machine Learning in 2010 and was then offered his dream job working at Google.

He began, along with other data scientists, working on the artificial intelligence algorithms that drove the YouTube recommender sidebar, and was excited by the possibilities**[1]**.

By 2013, Guillaume had been fired from YouTube. He had advocated against the changes in the recommender engine that were inadvertently pushing mis-leading and hateful content at people, all with the intention of keep viewers hooked on YouTube.

The role of YouTube, and other social media platforms, in generating a radicalised subculture of viewers is well documented. The problem that YouTube grapples with in 2021 is how address the criticisms of policies in 2014-2018 and to be a platform that promotes diversity and truthfulness. Can it handle the test of weeding out the undesirable content, and has the challenge of the Covid workplace made it more difficult to meet this ethical objective?

1. **Ethical and Legal Challenges**

The fourth episode of the 2021 New York Times ‘Rabbit Hole’ podcast contains an interview with Susan Wojcicki, the CEO of YouTube. It focuses primarily on decisions driven by her nearly a decade ago to change the way the YouTube recommendation engine worked, and partial acknowledgement that this led to unexpected (and presumably undesirable) alt-right radicalisation of significant numbers of viewers**[2]**.

The accusation laid against YouTube, both by external observers and former staff, is that around 2014 the company deliberately chose to refine its recommender algorithm with the express intention of increasing *watch time*.

Now, a sophisticated neural network model would recommend, and actively promote in the side bar, a greater range and diversity of new videos based on past viewing history. Most critically, YouTube would continue to filter out content based on the obvious, and established, criteria for banning videos - violence, nudity, and profanity - but would do little to assess content beyond those measures. Thus, an increasing array of alt-right videos were being pushed out to viewers who has started searching for videos on topics such as history and self-help**[1]**.

The alt-right videos, with their sensationalist titles, tended to generate more views and led certain viewers down a ‘rabbit hole’ of conspiracy theories and hate speech.

YouTube defended it actions at the time by declaring that it did deliberately did not take a partisan side in politics, and that it respected free speech.

However, the ethical issue continued to be that YouTube was providing a platform to hate speech and misinformation. More worryingly, the recommender engine was too often creating ‘filter bubbles’**[1]**. The model was clever enough not to fixate on cat videos, if that was the initial user search topic. However, if the user looked at an Alex Jones diatribe the YouTube model would never try and counterbalance with something from Jon Stewart.

YouTube was feeding a serious imbalance in news media, and it took until late 2019 before it began to effectively acknowledge responsibility, and it became evident and visible that YouTube was getting serious about its ethical media stance.

Of course, as welcome as these changes are, it is worth remembering that free speech is desirable, and content should be available with opposing views. Is there s danger that YouTube might inadvertently suppress content just because it is unpopular?

1. **Challenges for Data Scientist**

Probably the most obvious challenge is volume. The YouTube Community Guideline site states that 100s of hours of content are uploaded every minute**[3]**.

Compounding this metric is the fact that YouTube is still heavily reliant on human moderators. YouTube sent its workforce, including the 10,000+ moderators, home at the start of the pandemic and ramped up the scope and operations of the automated AI moderation system. However, these systems reported a significantly higher volume of ‘false positives’, taking down videos that were actually not in breach of guidelines and for which half of the decisions were subsequently reversed**[4]**. The dependency on human moderation became even more apparent when YouTube made the decision to re-introduce greater levels of human involvement in late 2020**[5]**.

This highlights that the machines are not quite ready yet to replace the human element in content moderation, and that the assessment of videos in bulk remains a complex task in need of further solutions.

That said, the company states that, as of April 2021, 94% of content breaking its rules is caught by its AI systems, and most of those videos are removed before they have 10 views**[6]**. Clearly the situation is improving but difficulties remain.

1. **Reflections**

Susan Wojcicki’s interview in 2020 with the New York Times**[2]** seemed to imply that, at the time, she doubted that changes to YouTube could do much harm. Politics seemed a very niche element of YouTube, with very low viewership.

However, the AI models that were built and deployed in 2012-2015 were designed to capture and keep attention as a primary objective. They worked extremely well and are seem by many as a contributor to the increased polarization and coarseness of global political debate in 2021.

I have two reflections;

The first is that work is still needed to enhance the AI models to prevent misinformation and hate speech being uploaded to platforms like YouTube. In December 2021, Meta announced innovations it was working on in the field of Few-Shot Learner (FSL), an AI technology to allow more rapid action on capturing harmful content**[7]**. Presumably, these are the types of approaches that YouTube will also embrace.

The second is related to the fact that since 2020 YouTube has been effectively saying that it will follow the establishment line on topics such as vaccines and hate crime. This is a positive move, but the ‘establishment’ is not always right. A recent example is the self-censorship of YouTube in Russia**[8]**. Is there not an argument that we should encourage YouTube, and other such platforms, to allow the occasional ‘edgy’ content.

1. **References**

[1] Roose, K. and Mills, A., 2021. *One: Wonderland (Published 2020)*. [online] Nytimes.com. Available at: <https://www.nytimes.com/2020/04/16/podcasts/rabbit-hole-internet-youtube-virus.html?> [Accessed 12 December 2021].

[2] Roose, K. and Miils, A., 2021. *Four: Headquarters (Published 2020)*. [online] Nytimes.com. Available at: <https://www.nytimes.com/2020/05/07/podcasts/rabbit-hole-youtube-susan-wojcicki-virus.html?showTranscript=1> [Accessed 12 December 2021].

[3] YouTube Community Guidelines and policies - How YouTube Works. 2021. *YouTube Community Guidelines and policies - How YouTube Works*. [online] Available at: <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#detecting-violations> [Accessed 12 December 2021].

[4] Lapowsky, I., 2020. *After sending content moderators home, YouTube doubled its video removals*. [online] Protocol — The people, power and politics of tech. Available at: <https://www.protocol.com/youtube-content-moderation-covid-19#toggle-gdpr> [Accessed 12 December 2021].

[5] Kraus, R., 2020. *YouTube puts human content moderators back to work*. [online] Mashable. Available at: <https://mashable.com/article/youtube-human-content-moderation> [Accessed 12 December 2021].

[6] Heilweil, R., 2021. *YouTube says it’s better at removing videos that violate its rules, but those rules are in flux*. [online] Vox. Available at: <https://www.vox.com/recode/2021/4/6/22368809/youtube-violative-view-rate-content-moderation-guidelines-spam-hate-speech> [Accessed 12 December 2021].

[7] Meta. 2021. *Meta's New AI System to Help Tackle Harmful Content | Meta*. [online] Available at: <https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/> [Accessed 12 December 2021].

[8] Bergin, M. and Chang, E., 2021. *YouTube CEO Says Google Sees Free Speech as Core Value in Russia*. [online] Bloomberg.com. Available at: <https://www.bloomberg.com/news/articles/2021-09-26/youtube-ceo-says-google-sees-free-speech-as-core-value-in-russia> [Accessed 12 December 2021].