

# COUNTERING TERRORISM ONLINE WITH ARTIFICIAL INTELLIGENCE

AN OVERVIEW FOR LAW ENFORCEMENT AND  
COUNTER-TERRORISM AGENCIES IN SOUTH ASIA AND  
SOUTH-EAST ASIA



# COUNTERING TERRORISM ONLINE WITH ARTIFICIAL INTELLIGENCE

An Overview for Law Enforcement and  
Counter-Terrorism Agencies in South  
Asia and South-East Asia

*A Joint Report by UNICRI and UNCCT*



## **Disclaimer**

*The opinions, findings, conclusions and recommendations expressed herein do not necessarily reflect the views of the United Nations, the Government of Japan or any other national, regional or global entities involved. Moreover, reference to any specific tool or application in this report should not be considered an endorsement by UNOCT-UNCCT, UNICRI or by the United Nations itself.*

*The designation employed and material presented in this publication does not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area of its authorities, or concerning the delimitation of its frontiers or boundaries.*

*Contents of this publication may be quoted or reproduced, provided that the source of information is acknowledged. The authors would like to receive a copy of the document in which this publication is used or quoted.*

## **Acknowledgements**

*This report is the product of a joint research initiative on counter-terrorism in the age of artificial intelligence of the Cyber Security and New Technologies Unit of the United Nations Counter-Terrorism Centre (UNCCT) in the United Nations Office of Counter-Terrorism (UNOCT) and the United Nations Interregional Crime and Justice Research Institute (UNICRI) through its Centre for Artificial Intelligence and Robotics. The joint research initiative was funded with generous contributions from Japan.*

## **Copyright**

© United Nations Office of Counter-Terrorism (UNOCT), 2021

United Nations Office of Counter-Terrorism  
S-2716  
United Nations  
405 East 42nd Street  
New York, NY 10017  
Website: [www.un.org/counterterrorism/](http://www.un.org/counterterrorism/)

© United Nations Interregional Crime and Justice Research Institute (UNICRI), 2021

Viale Maestri del Lavoro, 10, 10127 Torino – Italy  
Website: <http://www.unicri.it/>  
E-mail: [unicri.publicinfo@un.org](mailto:unicri.publicinfo@un.org)

# FOREWORD



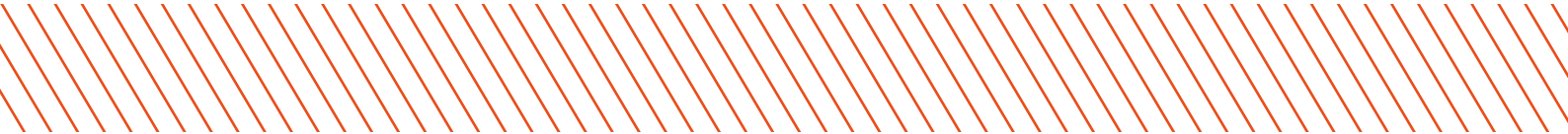
Artificial intelligence (AI) can have, and already is having, a profound impact on our society, from healthcare, agriculture and industry to financial services and education. However, as the United Nations Secretary-General António Guterres stated in his 2018 Strategy on New Technologies, “[w]hile these technologies hold great promise, they are not risk-free, and some inspire anxiety and even fear. They can be used to malicious ends or have unintended negative consequences”. AI embodies this duality perhaps more than any other emerging technology today. While it can bring improvements to many sectors, it also has the potential to obstruct the enjoyment of human rights and fundamental freedoms – in particular the rights to privacy, freedom of thought and expression, and non-discrimination. Thus, any exploration of the use of AI-enabled technologies must always go hand-in-hand with efforts to prevent potential infringement upon human rights. In this context, we have observed many international and regional organizations, national authorities and civil society organizations working on initiatives aimed at putting in place ethical guidelines regarding the use of AI, as well as the emergence of proto-legal frameworks.

This duality is most obviously prevalent online, where increased terrorist activity is a growing challenge that is becoming almost synonymous with modern terrorism. Consider that as part of 2020 Referral Action Day, Europol and 17 Member States identified and assessed for removal as many as 1,906 URLs linking to terrorist content on 180 platforms and websites in one day. Facebook has indicated that over the course of two years it removed more than 26 million pieces of content from groups such as the Islamic State of Iraq and the Levant (ISIL) and Al-Qaida. The Internet and social media are proving to be powerful tools in the hands of such groups, enabling them to communicate, spread their messages, raise funds, recruit supporters, inspire and coordinate attacks, and target vulnerable persons.

In the United Nations Global-Counter Terrorism Strategy (A/RES/60/288), Member States resolved to work with the United Nations with due regard to confidentiality, respecting human rights and in compliance with other obligations under international law, to explore ways to coordinate efforts at the international and regional levels to counter terrorism in all its forms and manifestations on the Internet and use the Internet as a tool for countering the spread of terrorism. At the same time, the Strategy recognizes that Member States may require assistance to meet these commitments.

Through the present report – a product of the partnership between the United Nations Counter-Terrorism Centre in the United Nations Office of Counter-Terrorism and the United Nations Interregional Crime and Justice Research Institute through its Centre for Artificial Intelligence and Robotics – we seek to explore how AI can be used to combat the threat of terrorism online.





Recognizing the threat of terrorism, growing rates of digitalization and burgeoning young, vulnerable and online populations in South Asia and South-East Asia, this report provides guidance to law enforcement and counter-terrorism agencies in South Asia and South-East Asia on the potential application of AI to counter terrorism online, as well as on the many human rights, technical and political challenges they will need to consider and address should they opt to do so.

Our work does not end here. Addressing the challenges identified in this report and unlocking the potential for using AI to counter terrorism will require further in-depth analysis. Our Offices stand ready to support Member States and other counter-terrorism partners to prevent and combat terrorism, in all its forms and manifestations, and to explore innovative and human rights-compliant approaches to do so.



**Vladimir Voronkov**

*Under-Secretary-General  
United Nations Office of Counter-Terrorism  
Executive Director  
United Nations Counter-Terrorism Centre*

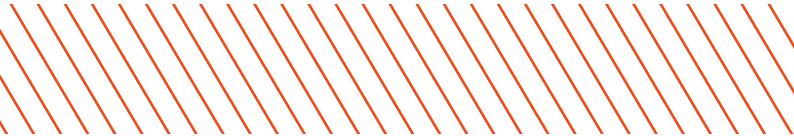


**Antonia Marie De Meo**

*Director  
United Nations Interregional Crime  
and Justice Research Institute*



# EXECUTIVE SUMMARY



The integration of digital technologies into everyday life has increased at an extraordinary pace in South Asia and South-East Asia in recent years, with the use of social media by the regions' notably young population surpassing the global average. While this trend offers a wide range of opportunities for development, the freedom of expression, political participation, and civic action, it also increases the risk of potentially vulnerable youths being exposed to terrorist online content produced by terrorist and violent extremist groups online. Additionally, given an established terrorist and violent extremist presence in South Asia and South-East Asia, law enforcement and counter-terrorism agencies in these regions are increasingly pressed to adapt to transformations in criminal and terrorist activities, as well as to how investigations into these activities are carried out.

Artificial intelligence (AI) has received considerable attention globally as a tool that can process vast quantities of data and discover patterns and correlations in the data unseen to the human eye, which can enhance effectiveness and efficiency in the analysis of complex information. As a general-purpose technology, such benefits can also be leveraged in the field of counter-terrorism. In light of this, there is growing interest amongst law enforcement and counter-terrorism agencies globally in exploring how the transformative potential of AI can be unlocked.

Considering the aforementioned trends and developments, this report serves as an introduction to the use of AI to counter terrorism online for law enforcement and counter-terrorism agencies in the regions of South Asia and South-East Asia. This report is introductory in nature as a result of the limited publicly available information on the degree of technological readiness of law enforcement and counter-terrorism agencies in these regions, which is considered likely to be indicative of limited experience with this technology. In this regard, the report provides a broad assessment of different use cases of AI, demonstrating the opportunities of the technology, as well as curbing out the challenges. The report is intended to serve as an initial mapping of AI, contextualizing possible use cases of the technology that could theoretically be deployed in the regions, whilst juxtaposing this with the key challenges that authorities must overcome to ensure the use of AI is responsible and human rights compliant. Given its introductory nature, this report is by no means intended to be an exhaustive overview of the application of AI to counter terrorism online.

The report is divided into five chapters. The first chapter provides a general introduction to the context of terrorism, Internet usage in South Asia and South-East Asia and AI. The second chapter introduces and explains some key terms, technologies, and processes relevant for this report from a technical perspective. The third chapter maps applications of AI in the context of countering the terrorist use of the Internet and social media, focusing on six identified use cases, namely: i) Predictive analytics for terrorist activities; ii) Identifying red flags of radicalization; iii) Detecting mis- and disinformation spread by terrorists for strategic purposes; iv) Automated content moderation and takedown; v) Countering terrorist and violent extremist narratives; and vi) Managing heavy data analysis demands. The fourth chapter examines the challenges that law enforcement and counter-terrorism agencies must be prepared to address in their exploration of the technology, in particular specific political and legal challenges, as well as technical issues. The report concludes with the fifth and final chapter, which provides high-level recommendations for law enforcement and counter-terrorism agencies in South Asia and South-East Asia to take onboard to support them to navigate the challenges described in terms of the use of AI to counter terrorism online.

# CONTENTS



<b>I.</b>	<b>INTRODUCTION</b>	<b>10</b>
<b>II.</b>	<b>KEY CONCEPTS, TECHNOLOGIES, AND PROCESSES</b>	<b>15</b>
i.	Artificial intelligence	15
ii.	Data	17
iii.	Machine learning	18
iv.	Deep learning	19
v.	Generative adversarial network	19
vi.	Natural language processing	20
vii.	Object recognition	20
viii.	Predictive analytics	20
ix.	Social network analysis	21
x.	Content matching technology	21
xi.	Data anonymization and pseudonymization	22
xii.	Open-source intelligence and social media intelligence	23
xiii.	Disinformation and misinformation	23
<b>III.</b>	<b>PROBING THE POTENTIAL OF AI</b>	<b>23</b>
i.	Predictive analytics for terrorist activities	24
ii.	Identify red flags of radicalization	26







- iii. Detecting mis- and disinformation spread by terrorists for strategic purposes.....27
- iv. Automated content moderation and takedown..... 29
- v. Countering terrorist and violent extremist narratives .....31
- vi. Managing heavy data analysis demands .....33

**IV. THE GREATER THE OPPORTUNITY, THE GREATER THE CHALLENGE ..... 35**

- i. Legal and political challenges..... 35
  - a. Human rights concerns.....35
  - b. Admissibility of AI-generated evidence in court..... 39
  - c. A fragmented landscape of definitions..... 40
  - d. AI governance..... 41
  - e. Public-private relations..... 42
- ii. Technical challenges in using AI tools..... 43
  - a. False positives and false negatives..... 43
  - b. Bias in the data..... 43
  - c. Explainability and the black box problem..... 44
  - d. The complexity of human-generated content ..... 45

**V. MOVING FORWARD WITH AI ..... 47**



# I. INTRODUCTION

South Asia and South-East Asia, like many parts of the world, grapple with the threat of terrorism and violent extremism. This includes both home-grown organized violent extremist groups such as Jemaah Islamiyah and internationally-oriented groups such as the Islamic State in Iraq and the Levant (ISIL, also known as Da'esh) and Al-Qaida, which includes local affiliated groups such as the Abu Sayyaf Group.

These regions have become important areas of focus for terrorist and violent extremist groups in terms of recruitment. At its height in 2015, over 30,000 foreign terrorist fighters, from more than 100 States, were believed to have joined ISIL.<sup>1</sup> Of this, more than 1,500 persons from South Asia and South-East Asia alone are believed to have travelled to ISIL-controlled territory.<sup>2</sup> By July 2017, at which point ISIL had begun to lose significant percentages of its territory, the regions of South Asia and South-East Asia saw significant numbers of foreign fighters returning to their home countries, as well as significant numbers of foreign fighters not originally from these regions relocating to South Asia and South-East Asia instead of returning to their home countries.<sup>3</sup> Radicalization poses an increasing threat in the regions. For example, the Ministry of Home Affairs and Law of Singapore recently indicated that the timeframe for recruitment has been reduced from approximately twenty-two to nine months.<sup>4</sup> The terrorist landscape in South-East Asia has also notably witnessed a growing number of women in terrorism.<sup>5</sup> This rising number of women in terrorism is believed to be linked with broader trends of increasingly self-directed terrorist attacks perpetrated by lone terrorist actors or groups.<sup>6</sup> The COVID-19 pandemic has also had an effect on the radicalization and recruitment-related phenomenon, with many Member States – including from South Asia and South-East Asia – raising concerns about radicalization in the context of the large numbers of people connected to the Internet for extended periods of time.<sup>7</sup>

The presence of terrorist and violent extremists in South Asia and South-East Asia is, however, not a new development, with these regions each having their own experience with diverse forms of national, regional and international violent extremism and having developed extensive experience in countering terrorism. The bombing of the Colombo World Trade Centre in October 1997,<sup>8</sup> the Bali bombings in October 2002,<sup>9</sup> the bombing of Super Ferry 14 in the Manila Bay in February 2004,<sup>10</sup> and the Mumbai attacks in November 2008 are all testament to the trials the regions have faced.<sup>11</sup>

Another global trend that the regions of South Asia and South-East Asia have faced in recent years is that of “digitization”. Indeed, the integration of digital technologies into everyday life in South Asia and South-East Asia is something that has been increasing at an extraordinary pace. While there is no single driver, a key factor often acknowledged is the large percentages of youths throughout the regions. Notably, more than 70.2% of the population in South Asia and

---

1 Security Council Counter-Terrorism Committee. (2021). Foreign terrorist fighters. Accessible at <https://www.un.org/securitycouncil/ctc/content/foreign-terrorist-fighters>.

2 Richard Barrett. (2017). Beyond the Caliphate: Foreign Fighters and the Threat of Returnees. The Soufan Centre, 10. Accessible at <https://thesoufancenter.org/wp-content/uploads/2017/11/Beyond-the-Caliphate-Foreign-Fighters-and-the-Threat-of-Returnees-TSC-Report-October-2017-v3.pdf>.

3 Ibid.

4 Valerie Koh. (Sept. 2017). Time taken for people to be radicalised has been shortened: Shanmugam. Today. Accessible at <https://www.todayonline.com/singapore/time-taken-people-be-radicalised-has-been-shortened-shanmugam>.

5 The Soufan Center. (Jun. 2021). Terrorism and Counterterrorism in Southeast Asia: Emerging Trends and Dynamics. The Soufan Centre Accessible at [https://thesoufancenter.org/wp-content/uploads/2021/06/TSC-Report\\_Terrorism-and-Counterterrorism-in-South-east-Asia\\_June-2021.pdf](https://thesoufancenter.org/wp-content/uploads/2021/06/TSC-Report_Terrorism-and-Counterterrorism-in-South-east-Asia_June-2021.pdf).

6 Ibid.

7 Analytical Support and Sanctions Monitoring Team submitted pursuant to resolution 2368 (2017) concerning ISIL (Da'esh), Al-Qaida and associated individuals and entities, Twenty-seventh report, S/2021/68 (3 February 2021).

8 Agence France-Presse. (Oct. 1997). 17 Die, 100 Wounded by Huge Bomb and Gunfire in Sri Lanka. The New York Times. Accessible at <https://www.nytimes.com/1997/10/15/world/17-die-100-wounded-by-huge-bomb-and-gunfire-in-sri-lanka.html>.

9 BBC News. (Oct. 2012). The 12 October 2002 Bali bombing plot. BBC News. Accessible at <https://www.bbc.com/news/world-asia-19881138>.

10 BBC News. (Oct 2004). Bomb caused Philippine ferry fire. BBC News. Accessible at <http://news.bbc.co.uk/2/hi/asia-pacific/3732356.stm>

11 Gethin Chamberlain. (Nov. 2008). Mumbai terror attacks: Nightmare in the lap of luxury. The Guardian. Accessible at <https://www.cnn.com/2013/09/18/world/asia/mumbai-terror-attacks/index.html>



South-East Asia are under 40 years old. This young population is exceptionally active online with “more than 55% using social media extensively, which is 13% more than the world’s average”.<sup>12</sup> As with younger generations across the globe, many are so-called “digital natives” – persons born or brought up during the digital age and possessing high-levels of familiarity with computers, the Internet and digital technology from an early age – and, thus, have a higher acceptance of Internet-related services and are enthusiastic users of social media. In line with this, the Association of Southeast Asian Nations (ASEAN) is, tellingly, the fastest-growing Internet market in the world, with 125,000 new users joining the Internet every day.<sup>13,14</sup> It was reported that almost 60% of the global social media users in 2020 were located in Asia.<sup>15</sup> The COVID-19 pandemic has equally played a role in expediting the process of digitization in the regions, with youths in South Asia and South-East Asia expanding their digital footprint throughout 2020 and more than 70% believing that their increased use of social media will last beyond the pandemic.<sup>16</sup> The use of social media grew the most in comparison with other online services, according to the respondents of the ASEAN Youth Survey 2020 conducted by the World Economic Forum.<sup>17</sup>



*Photo by camilo jimenez on Unsplash*

Naturally, these developments present a wide range of opportunities – if they are leveraged appropriately. The Internet and social media have demonstrated their beneficial potential in South Asia and South-East Asia on many occasions, for instance, as an accelerator of grassroots activism. Studies in South-East Asia over the last decade have demonstrated a positive correlation between social media use, political participation, and civic action in both democratic and authoritarian systems.<sup>18</sup> Unfortunately, however, the advantages of the Internet and social media that support civil society movements also make them appealing to actors with malicious intent and present a plethora of challenges for authorities in South Asia and South-East Asia, as well as globally.

- 
- 12 UNCCT-CTED. (2021). Misuse of the Internet for Terrorist Purposes in Selected Member States in South Asia and Southeast Asia and Responses to the Threat, 3.
  - 13 WEF. (2021). Digital ASEAN. Accessible at <https://www.weforum.org/projects/digital-asean>.
  - 14 WEF, in collaboration with Sea. (Jul. 2020). COVID-19 – The True Test of ASEAN Youth’s Resilience and Adaptability, Impact of Social Distancing on ASEAN Youth, 8. Accessible at [http://www3.weforum.org/docs/WEF\\_ASEAN\\_Youth\\_Survey\\_2020\\_Report.pdf](http://www3.weforum.org/docs/WEF_ASEAN_Youth_Survey_2020_Report.pdf)
  - 15 Jordan Newton, Yasmira Moner, Kyaw Nyi Nyl, Hari Prasad. (2021) Polarising Narratives and Deepening Fault Lines: Social Media, Intolerance and Extremism in Four Asian Nations. GNET, 5. Accessible at <https://gnet-research.org/wp-content/uploads/2021/03/GNET-Report-Polarising-Narratives-And-Deepening-Fault-Lines.pdf>.
  - 16 WEF. (Jul. 2020). COVID-19 – The True Test of ASEAN Youth’s Resilience and Adaptability, Impact of Social Distancing on ASEAN Youth, 9. Accessible at: [http://www3.weforum.org/docs/WEF\\_ASEAN\\_Youth\\_Survey\\_2020\\_Report.pdf](http://www3.weforum.org/docs/WEF_ASEAN_Youth_Survey_2020_Report.pdf)
  - 17 Ibid.
  - 18 Aim Signpeng & Ross Tapsell. (2020). From Grassroots Activism to disinformation; Social Media Trends in Southeast Asia. ISEAS-Yusuf Ishak Institute.

Terrorists and violent extremists around the world have adapted to the new digital paradigms of the 21<sup>st</sup> Century, learning to use information and communications technologies, and in particular online spaces and the multitude of interactive applications and social media platforms, to further their objectives – for instance, to spread hateful ideology and propaganda, recruit new members, organize financial support and operational tactics and manage supportive online communities. The use of such technologies in South Asia and South-East Asia parallels the use in other parts of the world, with social media and highly localized content tailored to local grievances and available in local languages playing a particularly important role in radicalization and recruitment. For instance, according to the former Home Affairs Minister of Malaysia, Ahmad Zahid Hamidi, social media was responsible for approximately 17 percent of ISIL recruitment in the country.<sup>19</sup> Similarly, in Singapore, of 21 nationals held for terrorism-related activities under the Internal Security Act between 2015 and 2018, 19 were radicalized by ISIL propaganda online, with the other two being radicalized by other online content encouraging participation in the Syrian conflict.<sup>20</sup> The relevance of such information and communications technologies throughout these regions in recent years can also be seen in national responses. For example, in Indonesia, the Ministry of Communications and Information Technology petitioned the encrypted messaging app Telegram to set up a specialized team of moderators familiar with Indonesian languages to specifically moderate terrorist content being spread in Indonesia.<sup>21</sup>

In addition to having to understand a widening threat landscape beyond the physical domain, law enforcement and counter-terrorism agencies are being tested and challenged to deal with extensive and complex investigations in increasingly data-heavy environments that fall outside their traditional areas of expertise. Large cases can take several years of work to search through and cross-check relevant case information, meaning that finding one key piece of information or being able to single out the most important leads for the purposes of an investigation has never been so difficult.<sup>22</sup> Law enforcement and counter-terrorism agencies across the globe therefore find themselves being pressed to keep up with the digital transformation.

Confronted by the reality that an inability to keep up may result in failing to obstruct a terrorist plot and lead to the loss of lives, there is increasing interest in exploring tools, techniques, or processes to fill operational and capacity gaps of law enforcement and counter-terrorism agencies in combatting terrorism online.<sup>23</sup> One field that is receiving considerable interest globally from both public and private sector entities confronting similar “informational overloads” is artificial intelligence (AI). As will be explained in the following chapter, AI is the field of computer science aimed at developing computer systems capable of performing tasks that would normally require human intelligence, such as visual perception, speech recognition, translation between languages, decision-making, and problem-solving. Much of the appeal of AI lies in its ability to analyze vast amounts of data – also referred to as “big data” – faster and with greater ease than a human analyst or even a team of analysts can do, and, in doing so, to discover patterns and correlations unseen to the human eye. Moreover, AI can extrapolate likely outcomes of a given scenario based on available data.<sup>24</sup>

Seen for a long time as little more than science fiction, AI is already being used throughout the public and private sector for a range of beneficial purposes. For instance, AI has played a role in helping to significantly speed up the development of messenger ribonucleic acid (mRNA)-based vaccines, such as those now being used to rein in the COVID-19 pandemic,<sup>25</sup> and is being deployed to help broker peace deals in war-torn Libya and Yemen.<sup>26</sup> The United Nations

---

19 Ibid

20 Kimberly T'ng. (Apr. 2019). Down the Rabbit Hole: ISIS on the Internet and How It Influences Singapore Youth. Accessible at [https://www.mha.gov.sg/docs/hta\\_libraries/publications/final-home-team-journal-issue-8-\(for-web\).pdf](https://www.mha.gov.sg/docs/hta_libraries/publications/final-home-team-journal-issue-8-(for-web).pdf)

21 James Hookway. (Jul. 2017). Messaging App Telegram to Boost Efforts to Remove Terror-Linked Content. Accessible at <https://www.wsj.com/articles/messaging-app-telegram-to-boost-efforts-to-remove-terror-linked-content-1500214488?mg=prod/accounts-wsj>

22 Jo Cavan & Paul Killworth. (Oct. 10, 2019). GCHQ embraces AI but not as a black box. About Intel. Accessible at <https://aboutintel.eu/gchq-embraces-ai/>.

23 INTERPOL & UNICRI. (2020). Towards Responsible AI Innovation for Law Enforcement. Accessible at: <http://unicri.it/towards-responsible-artificial-intelligence-innovation>

24 Whether those predictions are accurate or not and how that impacts the use of AI to combat the terrorist use of the Internet is analyzed in the third and fourth chapters.

25 Hannah Mayer et al. (Nov. 24, 2020). AI puts Moderna within striking distance of beating COVID-19. Digital Initiative. Accessible at <https://digital.hbs.edu/artificial-intelligence-machine-learning/ai-puts-moderna-within-striking-distance-of-beating-covid-19/>.

26 Dalvin Brown. (Apr. 23, 2021) The United Nations is turning to artificial intelligence in search for peace in war zones. The Washington Post. Accessible at <https://www.washingtonpost.com/technology/2021/04/23/ai-un-peacekeeping/>.

Secretary-General, António Guterres, has indicated that, if harnessed appropriately and anchored in the values and obligations defined by the Charter of the United Nations and the Universal Declaration of Human Rights, AI can play a role in the fulfilment of the 2030 Agenda for Sustainable Development, by contributing to end poverty, protect the planet and ensure peace and prosperity for all.<sup>27</sup>

AI can be a powerful tool in counter-terrorism, enabling law enforcement and counter-terrorism agencies to realize game-changing potential, enhancing effectiveness, augmenting existing capacities and enabling them to manage with the massive increase in data. AI can support law enforcement and counter-terrorism agencies, for example, by automating highly repetitive tasks to reduce workload; assisting analysts through predictions of future terrorist scenarios for well-defined, narrow settings; identifying suspicious financial transactions that may be indicative of the financing of terrorism; as well as monitoring Internet spaces for terrorist activity at a scale and speed beyond traditionally available human capabilities.

Excitement about the potential for societal advancement with AI is, however, tempered by growing concerns about possible adverse impacts and unintended consequences that may flow from the deployment of this technology. As a result, AI is currently the subject of extensive debate among technologists, ethicists, and policymakers worldwide. With counter-terrorism – and particularly pre-emptive forms of counter-terrorism – already at the forefront of debates about human rights protection, the development and deployment of AI raises acute human rights concerns in these contexts.<sup>28</sup>



*Photo by Shahadat Rahman on Unsplash*

Nevertheless, from a security perspective, the need for law enforcement and counter-terrorism agencies to adapt to the digital transformation certainly exists. It could even be said to be particularly pressing for South Asia and South-East Asia, given the increased rate of digitalization and the growing presence of younger, potentially vulnerable, ele-

27 António Guterres. (Sept. 2018). Secretary-General's Strategy on New Technologies. United Nations. Accessible at <https://www.un.org/en/newtechnologies/images/pdf/SGs-Strategy-on-New-Technologies.pdf>

28 Although these issues will be touch upon in this report, for further analysis see: OHCHR, UNOCT-UNCCT & UNICRI. (2021). Human Rights Aspects to the Use of Artificial Intelligence in Counter-Terrorism.

ments of local populations online, as well as established terrorist and violent extremist presence in the regions. It is, however, relevant to note that there is limited publicly available information on the degree of technological readiness of law enforcement and counter-terrorism agencies in South Asia and South-East Asia or in terms of any specific capabilities that might be available or under development.<sup>29</sup> Given the relatively nascent state of the use of AI for counter-terrorism purposes, this is likely indicative of limited knowledge or experience with this technology in these regions.

A recent survey carried out by UNICRI and INTERPOL at the Third Annual Global Meeting on AI for Law Enforcement attested to the general embryonic state of the use of AI in law enforcement at large. From 50 representatives of the law enforcement agencies around the world, 50% considered that the level of AI knowledge and expertise within their organization was “rudimentary”, 30% considered it “intermediary” and only 20% considered it to be “advanced”.<sup>30</sup>

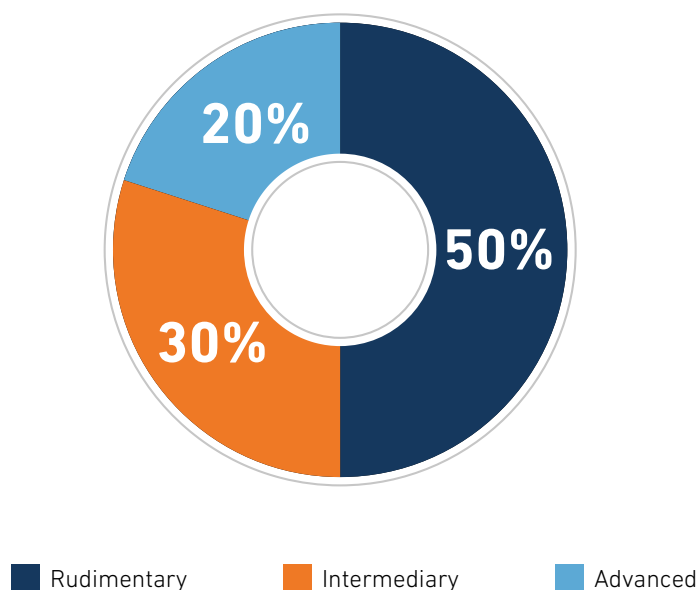


Figure 1: INTERPOL-UNICRI survey on self-perception of the level of AI knowledge and experience in law enforcement

Against this backdrop, this report aims to serve as an introduction for law enforcement and counter-terrorism agencies in South Asia and South-East Asia on the application of AI in the context of countering terrorism online. The report provides a broad assessment of different use cases, demonstrating the opportunities presented by this technology, as well as curbing out the challenges. The report is intended to serve as a mapping of AI, contextualizing possible use cases of the technology that could theoretically be deployed in the regions, whilst juxtaposing this with key challenges that authorities must overcome to ensure the use of AI is responsible and human rights compliant. Although the report is specifically intended for law enforcement and counter-terrorism agencies in South Asia and South-East Asia, its scope and structure also lend to it being a useful resource for law enforcement and counter-terrorism agencies globally.

In the preparation of this report, the United Nations Counter-Terrorism Centre (UNCCT) of the United Nations Office of Counter-Terrorism (UNOCT) and the United Nations Interregional Crime and Justice Research Institute (UNICRI) have relied predominantly on desk-based research and open-source information, such as articles, reports (including media reports), and conducted partially closed interviews with cross-sector experts from research institutes, international security organizations and non-governmental organizations.<sup>31</sup> An Expert Group Meeting on leveraging AI to combat the

29 In order to fill these knowledge gaps, UNOCT-UNCCT and UNICRI launched a regional survey of AI capabilities and technological readiness in the South Asia and South-East Asia. Findings of this exercises are expected in late 2021.

30 Insights from the Third INTERPOL-UNICRI Global Meeting on AI for Law Enforcement. (Nov. 23-27, 2020). Full results will be available in the forthcoming Third INTERPOL-UNICRI Report on AI for Law Enforcement.

31 Representatives from the following entities were interviewed: Dartmouth College’s Institute for Security, Technology, and Society; the German Federal Criminal Police Office’s Terrorism/Extremism Research Unit; the Metropolitan Police Service, United Kingdom; Moonshot; and the North Atlantic Treaty Organization (NATO).

terrorist use of the Internet and social media with a focus on South Asia and South East Asia was organized virtually by UNOCT and UNICRI on 18 March 2021 to complement the conclusions reached on the basis of the open-source information analyzed and interviews conducted, and to collect specific insights, in particular from the targeted regions.<sup>32</sup> Given the sensitive nature of security matters, several interview partners shared information anonymously and hence referencing is not always possible.

It is important to also underscore that reference to any specific tool or application in this report should not be considered an endorsement by UNOCT-UNCCT, UNICRI or by the United Nations itself. The tools and applications mentioned in this report are included solely to demonstrate the potential application of AI. UNOCT-UNCCT and UNICRI acknowledge the challenges of the development and deployment of any such AI-enabled technologies, which necessarily includes those mentioned in this report.

## II. KEY CONCEPTS, TECHNOLOGIES, AND PROCESSES

For law enforcement and counter-terrorism agencies to more accurately appreciate how AI can be used to counter terrorism online, it is essential to start by establishing a foundational understanding of the technology itself. This chapter provides a brief technical overview of the key concepts, technologies, and processes.

### i. Artificial intelligence

A fast-developing field, AI increasingly touches many areas of our lives. AI is already widely used to suggest movies and television shows on streaming platforms and to provide recommendations for online shopping. It curates and populates news feeds on social media and unlocks cell phones with facial recognition. Notwithstanding the level of integration of AI into society and the frequency of the popular use of the term, there is no universal definition of “AI”. The term is however generally understood to describe a discipline concerned with developing technological tools exercising human qualities, such as planning, learning, reasoning, and analyzing. In public debates, the term is often used interchangeably with “machine learning” and “deep learning”, which is unprecise from a technical point of view. Whereas AI describes a general domain, machine learning and deep learning are both subfields, describing specific types of AI. An AI system may thus, for instance, include sensors that capture the environment plus a machine learning algorithm to perform a certain task according to the data received. Both these separate subfields are further described in more detail below.

---

32 Participants in the Expert Group Meeting included representatives from: AWO; Infinium Robotics; INSIKT Intelligence; INTERPOL; LIR-NEasia; Omdena Inc.; SIMAVI; the Cyber Security Cooperative Research Centre; the Department of Research and Innovation, Ministry of Education, Myanmar; the Digital Rights Foundation, Pakistan; the Graduate Institute of International and Development Studies, Geneva; the Home Team Science and Technology Agency (HTX), Singapore; the Human Rights Commission of Malaysia (SUHAKAM); the Indian Police Service; the Information and Technology Division, Sri Lankan Police; the Police Policy Research Center, National Police Agency, Japan; the S. Rajaratnam School of International Studies (RSIS), Singapore; the Southeast Asia Regional Centre for Counter Terrorism (SEARCCT); the University of Tokyo, Japan; and Tisane Labs.



Photo by Firmbee.com on Unsplash

AI can also be categorized as 'narrow' or 'general'. The AI systems that exist in our world today consist of what is known as narrow AI applications, i.e., AI models that are programmed towards one specific objective such as finding the best way between A and B or matching similar images. These programs cannot be applied beyond their use case, or when a small change in the environment occurs. For example, if a chessboard had 9 lines instead of 8, the AI model would immediately be unable to play. In other words, AI algorithms are masters at picking up patterns, but they cannot yet understand and adapt to a changing world. General AI or artificial general intelligence would be able to do it but so far it only exists in science fiction. Artificial general intelligence would not be trained to a specific purpose, but its intelligence would rather be more human-like, including in analyzing, planning, communicating, and reasoning. The concept of artificial super intelligence goes even further as it refers to AI that would surpass human intelligence in all aspects. From creativity to problem-solving, super-intelligent machines would overcome our intelligence as individuals and as a society. This is the type of AI that has generated quite an amount of philosophical debate, with some experts arguing that it may even present an existential threat to humanity.

AI is often perceived as a modern development or something very futuristic, yet the concept of AI itself is in fact quite an old one, having its roots in the 1950s. Due to limited storage and processing capacities at the time, much of the early aspirations around AI never materialized, resulting in what became known as the "AI Winter" – a period when public interest and investment in AI significantly waned. New technological developments, particularly in the last decade, have allowed for cheaper solutions for mass data storage and faster processing. Combined with the "democratization" of the technology, whereby AI is becoming increasingly accessible and can be used without large investment or even with limited technical knowledge increase of experts, as well as growing access to huge volumes of data, the necessary ingredients for the dawn of a new era of AI are at last on the table.

As a concept, AI may be misleading as it seems to imply some similarity with human intelligence or human learning processes. Deep neural networks, a family of AI algorithms, are in fact inspired by the architecture of the human brain, with the construction of different layers of processing units leveraged to produce interesting results, even with unstructured or unlabeled data. However, the capabilities of AI, including deep neural networks modelled on human brains, differ immensely from human capacities. For instance, even if AI can make predictions, it cannot give meaning to the results. For example, an AI-based application to detect breast cancer can detect the illness very reliably. Through the analysis of mammography pictures, it produces results with a lower error rate than radiologists.<sup>33</sup> However, the software is not able to understand the meaning of such a diagnosis as it can only see the numbers of the pixel intensities behind that picture. Google's AI Alpha Go can similarly compute the best moves in the complex game Go, but it cannot explain why it chose certain moves and hence, give meaning to them.<sup>34</sup>

33 Karen Hao. (Jan. 3, 2020). Google's AI breast cancer screening tool is learning to generalize across countries. MIT Technology Review. Accessible at <https://www.technologyreview.com/2020/01/03/238154/googles-ai-breast-cancer-screening-tool-is-learning-to-generalize-across-countries/>.

34 Elizabeth Gibney. (2017). Self-taught AI is best yet at strategy game Go'. Nature. Accessible at <https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858>.





It is also important to note that AI models can find patterns and correlations that do not necessarily have a causal relationship but can result in the intended output. For example, an AI that has been trained to detect trains in images may, for instance, not identify the trains but rather the train tracks that also frequently appear in images used to develop the model. Results like this are based on unintended neural connections.<sup>35</sup>

While AI's ability to detect patterns and correlations can be informative and present unforeseen connections, identifying "train tracks" instead of "trains" can be dangerous in the context of counter-terrorism operations. In fact, this inability of AI to truly "understand" and only extract patterns is crucial when it comes to AI applications in the realm of counter-terrorism. Whereas an error in an algorithm designed to recommend movies to watch based on a personal history profile and the preferences of users with similar patterns can simply result in a bad movie night, errors in AI models in counter-terrorism operations can have much more dire outcomes and implications for human rights.

## ii. Data

AI makes it possible for machines to learn from experience. A machine learning model can be trained by processing large amounts of data and recognizing patterns in the data, patterns that are used to make predictions about new data points. To obtain accurate predictions, two major components are necessary: a model and data.

Data can be understood as units of information collected through observation.<sup>36</sup> For the purposes of AI, data is a set of values of variables about persons or objects, among other things, that can be qualitative (for example, colour of eyes) or quantitative (for example, age). A dataset, or database, is a collection of several units of data that are measured and reported, producing data visualizations such as graphs, tables, or images.

In order to have a good AI model, it is fundamental that both data quantity and quality are ensured. Generally speaking, simple models with large data sets are more accurate and effective than complex models with small data sets. On the other hand, it is also not useful to have a lot of data if it is inaccurate, unrepresentative, or outdated. In this sense, the quality of the outcome highly depends on the quality of the training data. This includes having data that is free of bias, as this bias can be reflected in the results.<sup>37</sup>

After constructing a dataset, data needs to be transformed in a way that the model can read it. Data can be "structured" when it is stored in a predefined format such as a spreadsheet, or "unstructured" when it consists of a conglomeration of many varied types of data that are stored in their native formats, such as in an email or a social media post. Unstructured data accounts for the overwhelming majority of all data produced. Transforming and cleaning data for use as input for a model is the most time-consuming but also the most important step in terms of guaranteeing high performance of an AI model. Given its nature, unstructured data requires more work to process, although it can also be of great value by helping entities with the right capabilities to elicit valuable insights and opportunities.

---

35 Laura Thesing, et al. (2019). arXiv. What do AI algorithms actually learn? – On false structures in deep learning. Accessible at <https://arxiv.org/abs/1906.01478>.

36 OECD. (2008). Glossary of Statistical Terms. Accessible at <https://stats.oecd.org/glossary/>

37 The challenge of bias is further discussed in chapter four below.



Photo by Chris Liverani on Unsplash

One key challenge encountered with the use of AI in the field of counter-terrorism is that the vast quantities of data needed are not always available or accessible. The reality of terrorism is that incidents are not regularly occurring events and cases of radicalization are often unique. There are several open-source databases on terrorism that can be used for the purposes of training algorithms, such as the Global Terrorism Database, which collects historical information on more than 200,000 terrorist incidents globally since 1970.<sup>38</sup> While these are very valuable sources of data, in particular in terms of making predictions about possible future types of terrorist attacks, target areas and weapons used, realizing the applications described in the present report would require more specific and near real-time datasets concerning the actions of individual terrorists or suspected terrorists, such as social media data. Naturally as most of this data is unstructured in nature, extensive work would be necessary in terms of preparing its use. One possible solution to overcome the absence of sufficient quantities of “real world” data is to use “augmented data” – i.e., artificial data produced by general adversarial networks (GANs) for the purposes of training algorithms.<sup>39</sup> Further research is, however, needed to explore the potential of augmented data in the realm of counter-terrorism.

### iii. Machine learning

Machine learning is a subfield of AI that concerns algorithms that can “learn” from data, i.e., progressively improve performance on a specific task. In contrast with other computer software, machine learning algorithms do not require explicit instructions from humans, but extract patterns and learn implicit rules from examples included in a database. The initial database is split into three groups so that the algorithm can be trained with some data examples from the training dataset, validated on the validation dataset and subsequently tested on never-before-seen examples from the test dataset.<sup>40</sup>

The types of learning examples encountered include “regression”, whereby the output is a real number; “classification”, whereby the output is a label from a finite set of options; and “ranking”, whereby the output orders objects according to their relevance, sequence labelling, autonomous behaviour etc. Depending on the type of learning example, several models can be employed including decision-trees, linear and logistic regression, neural networks, among others.

38 See: <https://www.start.umd.edu/gtd/>

39 Pierluigi Casale. (Feb. 13, 2020). How does Artificial Intelligence improve map making, Tomtom. Accessible at <https://www.tomtom.com/blog/maps/artificial-intelligence-map-making/>.

40 UNOCT-UNCCT & UNICRI. (2021). Algorithms and Terrorism: The Malicious use of Artificial Intelligence for terrorist purposes, 9.



## iv. Deep learning

Deep learning is a subfield of machine learning that deals with a smaller family of algorithms, known as deep neural networks. These are algorithms that are inspired by the human brain and that seek to learn from large amounts of data by performing a task repeatedly, each time making minor modifications to its internal features to improve the outcome. The term “deep learning” comes from several (or “deep”) layers of a neural network.<sup>41</sup> Advances in deep learning are driving progress and research in image and video processing, text analysis, and speech recognition.<sup>42</sup>

## v. Generative adversarial network

GANs consist of two neural networks that compete with one another, thereby improving both their respective performances. For example, one neural network is used to create fake faces and the other is tasked with filtering them out from a set of real ones. As the filtering system improves, so too does the face-faking one. Websites such as <https://thispersondoesnotexist.com/>, which generate a new fictional face with each refresh, are well-known examples of the use of GANs.

Besides generating realistic fake faces, GANs are also the base of the widely publicized and hotly debated phenomenon known as “deepfakes”. A portmanteau of “deep learning” and “fake media”, deepfakes are a type of synthetic media invented in 2017. They involve the use of GANs to manipulate or generate fake visual and audio content that humans or even technological solutions cannot immediately distinguish from authentic content.<sup>43</sup> Deepfakes were initially, and are still overwhelmingly, used to create pornographic content, combining the faces of female celebrities with bodies of pornographic actors.<sup>44</sup> However, deepfakes and the technology behind them have also been identified as a potentially powerful weapon in today’s disinformation wars, where people can no longer rely on what they see or hear.<sup>45</sup> This is particularly the case when considering the reach and speed of the Internet, social media and messaging applications.

Deepfakes present considerable potential for a range of malicious and criminal purposes which include: destroying the reputation and credibility of an individual; harassing or humiliating individuals online, including through the use of sexual deepfakes; perpetrating blackmail, extortion and fraud; disrupting financial markets; and stoking social unrest and political polarization.<sup>46</sup> From the perspective of terrorism, deepfakes pose threats in terms of their use in disinformation campaigns on social media to manipulate public opinion or undermine people’s confidence in potential and incumbent political representatives or state institutions.<sup>47</sup> They can also be an effective instrument for propaganda, radicalization or as a call for action.

Interestingly, AI may also be part of the solution to countering this growing AI-based societal challenge. For instance, developers, including Facebook,<sup>48</sup> Microsoft<sup>49</sup> and many more,<sup>50</sup> have developed AI models that have been trained to be able to spot AI-manipulated audio-visual content. However, there few mature publicly available tools.

---

41 Ian Goodfellow, Yoshua Bengio & Aaron Courville. (2016). Deep Learning. MIT Press. Accessible at [www.deeplearningbook.org](http://www.deeplearningbook.org).

42 Yann LeCun, Yoshua Bengio, Geoffrey Hinton. (May 27, 2015). Deep learning. Nature. 521(7553), 436–444.

43 Oscar Schwartz. (Nov. 12, 2018). You thought fake news was bad? Deep fakes are where truth goes to die. The Guardian. Accessible at <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>

44 Giorgio Patrini. (Oct. 7, 2019). Mapping the Deepfake Landscape. Accessible at <https://giorgiop.github.io/posts/2018/03/17/mapping-the-deepfake-landscape/>

45 Than Thi Nguyen, et al. (Jul. 28, 2020). arXiv. Deep Learning for Deepfakes Creation and Detection: A Survey. Accessible at <https://arxiv.org/abs/1909.11573>

46 UNOCT-UNCCT. (2021). Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes.

47 Mika Westerlund. (2019). The Emergence of Deepfake Technology: A Review, Technology Innovation Management Review, 9(11). Accessible at [https://timreview.ca/sites/default/files/article\\_PDF/TIMReview\\_November2019%20-%20D%20-%20Final.pdf](https://timreview.ca/sites/default/files/article_PDF/TIMReview_November2019%20-%20D%20-%20Final.pdf)

48 Facebook. (June 16, 2021) Reverse engineering generative models from a single deepfake image. Facebook AI. Accessible at <https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image/>

49 Tom Burt. (Sep. 1, 2020). New Steps to Combat Disinformation. Microsoft. Accessible at <https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/>

50 See, for example: <https://platform.sensity.ai/deepfake-detection>.

## vi. Natural language processing

Natural language processing (NLP) is a deep learning application that concerns the processing and analyzing of large amounts of natural human language data, to enable machines to read, understand and derive meaning from human languages. Tasks in NLP frequently involve speech recognition, natural language understanding, natural language generation and translation between languages. Significant advancements were made in NLP over the last decade, especially when combined with the previously described GANs, as recently demonstrated by an article co-written by the text generator GPT-3 in the British daily newspaper The Guardian<sup>51</sup> The type of architecture most used in NLP is what is known as a Recurrent Neural Network, whereby the neural network nodes are connected along a temporal sequence. Relying on an internal memory that processes sequences of inputs, it is possible for an algorithm to extract morphosyntax and semantics function based on a sequence of words.<sup>52</sup>



Photo by Amador Loureiro on Unsplash

## vii. Object recognition

Object recognition is a subcategory of computer vision that uses deep learning algorithms to process pictures and identify geometrical shapes and, ultimately, objects. Complex algorithms based on what are known as Convolutional Neural Networks use multiple layers of locally connected nodes to progressively extract higher-level features from raw input.<sup>53</sup> If the input is an image, the first layers of the neural network may, for instance, identify lines and curves, while the last layers may identify letters or faces. A prominent variation of object recognition is facial recognition – a biometric technology capable of identifying persons of interest in images or videos by comparing and analyzing patterns, shapes and proportions of their facial features and contours with faces within a database.

## viii. Predictive analytics

Predictive analytics seeks to anticipate likely future or unknown events by analyzing what has already happened and, from this, extrapolating likely outcomes in related contexts. It includes a variety of statistical techniques from data

51 GPT-3 (Sep. 8, 2020) A robot wrote this entire article. Are you scared yet, human?. The Guardian. Accessible at <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>; Will Douglas Heaven. (Feb. 24, 2021). Why GPT-3 is the best and worst of AI right now. MIT Technology Review. Accessible at <https://www.technologyreview.com/2021/02/24/1017797/gpt3-best-worst-ai-openai-natural-language/>.

52 UNOCT-UNCCT & UNICRI. (2021). Algorithms and Terrorism: The Malicious use of Artificial Intelligence for terrorist purposes.

53 Kunihiko Fukushima. (1980). Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics. 36(4), 193–202. Accessible at <https://link.springer.com/article/10.1007/BF00344251>.



mining and machine learning that analyze current and historical facts as the foundation of its predictions.<sup>54</sup> Predictive models capture relationships among a range of variables to allow assessment of risk associated with a particular set of conditions. Machine learning based tools can present findings using simple charts, graphs, and scores that indicate the probability of events in the future, guiding decision-making processes.

## ix. Social network analysis

Social network analysis (SNA) is an extensive approach for understanding and modelling network structures and the behaviour of actors therein.<sup>55</sup> Social networks can be physical – as in a disease transmission network – or digital – like in a social media friendship network. A fundamental step for analyzing social networks is to encode network data into low-dimensional representations, typically using a set of nodes to represent individual elements of the network and a set of links or edges between these nodes which correspond to pairwise relations. Data mining and machine learning can be used to develop network representation methods, allowing for the mapping of communities, identifying main actors or groups within a community and further applications such as classifying, linking predictions, detecting anomalies and clustering.<sup>56</sup>

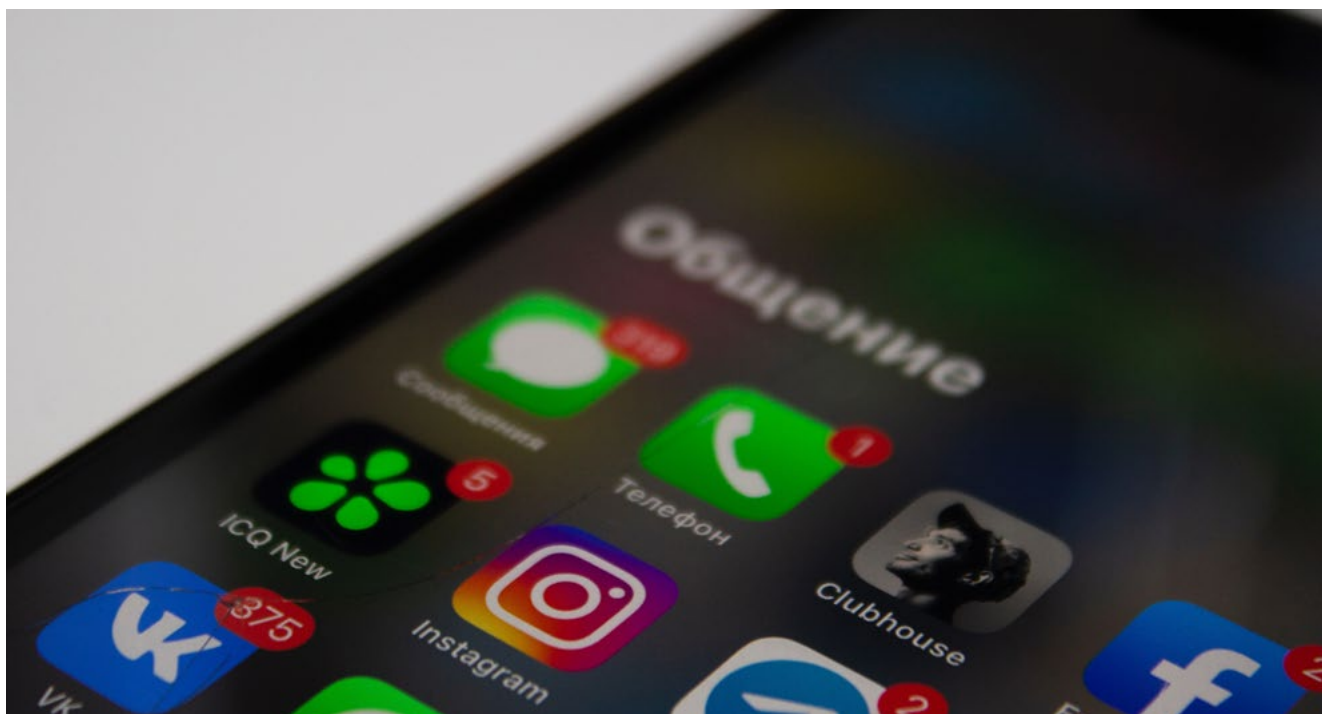


Photo by Anton Maksimov juvnsky on Unsplash

## x. Content matching technology

Image and video matching makes use of “hashing” technology to shorten an input of any length into a fixed string of text – a so-called hash – which is like a digital fingerprint. The short version of the input can then be compared to other files to find duplicates and prevent them from being shared. This technology can be used to help identify identical

54 M. Irfan Uddin, et al. (2020). Prediction of Future Terrorist Activities Using Deep Neural Networks. Complexity. Accessible at <https://doi.org/10.1155/2020/1373087>.

55 Roberto Musotto. (2020). From Evidence to proof: Social Network Analysis in Italian Criminal Law of Justice, in Special Collection on Artificial Intelligence. UNICRI. Accessible at [http://www.unicri.it/Publications/Artificial-Intelligence-AI\\_Collection](http://www.unicri.it/Publications/Artificial-Intelligence-AI_Collection).

56 Qiaoyu Tan, Ninghao Liu, Xia Hu. (2019). Deep Representation Learning for Social Network Analysis, in Front. Big Data. Accessible at <https://www.frontiersin.org/articles/10.3389/fdata.2019.00002/full>.

harmful content, such as extremist, terrorist, or pornographic content, in real-time at a high scale. While AI allows the identification of unknown or “first-generation” harmful content, hashing technology permits the detection of already-identified content so that when the same material is shared several times through several platforms it can be processed and removed at a faster rate. In this sense, this technology complements the functionalities of AI in this field and can be combined to allow for more efficient content moderation. PhotoDNA is an example of this technology that was initially developed to identify child sexual abuse material and is now more widely used on other illegal content.<sup>57</sup> In December 2016, Facebook, Twitter, Google and Microsoft announced plans to tackle extremist content such as terrorist recruitment videos and violent terrorist imagery using PhotoDNA.<sup>58</sup>

## xi. Data anonymization and pseudonymization

Given that the AI applications presented in this report entail the processing of personal data, techniques such as anonymization and pseudonymization are key for the protection of the right to privacy. Although sharing some similarities, anonymization and pseudonymization are distinct concepts. Data anonymization consists of the transformation of personal data into anonymous data so that the individuals or groups of individuals to which the data belongs are no longer identifiable in the data. Anonymization can be achieved through different methods, including through pseudonymization. Pseudonymization consists of the removal or substitution of all direct identifiers with other unique identifiers in such a way that unique individuals are still distinguishable in a data set, but their identity cannot be traced back without access to additional information. It should be noted that anonymized data is still personal data as long as the anonymization is not irreversible, as with appropriate skills or technology the data could be linked back to individuals or groups.<sup>59</sup>



*Photo by Malik Earnest on Unsplash*

---

57 See <https://www.microsoft.com/en-us/photodna>

58 Facebook. (Dec. 5, 2016) Partnering to Help Curb Spread of Online Terrorist Content. Accessible at <https://about.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/>

59 United Nations Development Group. (2017). Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda. Accessible at <https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda>.



## xii. Open-source intelligence and social media intelligence

Open-source intelligence (OSINT) is the method of gathering, analyzing, and interpreting publicly available data. OSINT sources can vary from media (print newspapers, television, etc.), Internet (online publications, blogs, and other social media websites), public government data, professional and academic publications, commercial data or what is sometimes called “grey literature” (technical reports, patents, newsletters, etc.).<sup>60</sup> OSINT tools help to navigate, analyze, and visualize material, varying from searching for specific keywords to interactions of accounts.<sup>61</sup>

Social media intelligence (SOCMINT) is a subcategory of OSINT that focuses on intelligence gathering on social media. In this case, SOCMINT tools allow organizations to analyze conversations, respond to social signals and synthesize social data points into meaningful trends and analysis.

## xiii. Disinformation and misinformation

The rise of the “fake news” phenomenon has made disinformation and misinformation household terms. The difference between dis- and misinformation lies in the awareness about its inaccuracy. Disinformation is false information that is spread intentionally to create harm. Misinformation, however, is false information disseminated unintentionally. A further related term is mal-information, which is information that is based on reality but used strategically to incite harm.<sup>62</sup> Notably, these phenomena have increased considerably during the COVID-19 pandemic.<sup>63</sup> Although none of these terms are directly related to the field of AI, AI can significantly exacerbate the effects of disinformation and misinformation or play a central role in combatting them.

# III. PROBING THE POTENTIAL OF AI

Having taken note of the threat of terrorism in South Asia and South-East Asia, examined key trends and developments related to the “digitalization” in these regions and looked at what this means in terms of the challenges with which law enforcement and counter-terrorism agencies must increasingly grapple, this chapter seeks to present how some of the AI technologies described in the previous chapter can be applied.

In this regard, a series of six use cases for the deployment of AI-enabled technology in counter-terrorism have been identified, each of which will be presented in an introductory format below to support conceptualization of the application of AI in the field of countering terrorism online. These use cases have been selected due to their prominence in discourse amongst stakeholders within the AI community in general, as well as for their specific potential to be applied in combatting terrorism online. This should not however be considered an exhaustive summary of the application of AI for counter-terrorism purposes.

The examples of specific tools provided in this chapter predominantly originate from Europe and the United States as these are regions where much of the research and development in the domain of AI and counter-terrorism is concentrated. Before proceeding further, it is important to underscore once again that reference to any specific tool or application in this report should not be considered an endorsement by UNOCT-UNCCT, UNICRI or by the United Nations itself. The tools and applications mentioned in this report are included solely to demonstrate the potential application of AI.

---

60 Jeffrey Richelson. (2016). *The US Intelligence Community*. Routledge.

61 For an introduction, see: <https://www.osintessentials.com>.

62 UNESCO. (2018). *Journalism, ‘Fake News’ and Disinformation: A Handbook for Journalism Education and Training*. Accessible at <https://en.unesco.org/fightfakenews>.

63 UNICRI. (Nov. 2020). *Stop the virus of disinformation*. Accessible at [http://unicri.it/sites/default/files/2021-01/misuse\\_sm\\_0.pdf](http://unicri.it/sites/default/files/2021-01/misuse_sm_0.pdf).

## i. Predictive analytics for terrorist activities

The application of predictive analytics for counter-terrorism could, in some ways, be described as the “Holy Grail” for security forces, enabling them to transcend a traditionally reactionary approach to terrorism and become more proactive by anticipating future terrorist activities and intervening before an attack occurs. To allow for this, an AI model would need to be fed large quantities of real-time data regarding the behaviour of a terrorist or a suspected individual. By analyzing this data, such a model could potentially, for instance, make predictions regarding the likely future activities of these individuals. Given the massive growth over the past decade in the amount of data regarding individuals’ behaviour online, especially on social media, there has been growing interest in exploring how social media data collected regarding individuals’ behaviour online can be used to predict terrorist activities.

Given the unpredictability of human behaviour and the current state of technological development, the application of algorithms to predict behaviour at an individual level is likely to remain of very limited value.<sup>64</sup> Additionally, human rights experts and civil society organizations have pointed towards several ethical concerns regarding potential entry points for discriminatory judgement and treatment. The large quantities of data concerning an individual required for the algorithm to accurately function further give rise to concerns about the possibility of unwarranted mass surveillance.

Predictive analytics can, however, still contribute to countering terrorism, but in a different manner. Rather than monitoring individuals online and forecasting their behaviour, predictive models informed by statistics from online sources that have been thoroughly anonymized or at least pseudonymized to protect user privacy could be used to identify trends or forecast the future behaviour of terrorists. This analysis based on aggregated data can be helpful to support security and intelligence agencies prioritizing scarce resources as operational support, making strategic decisions or providing warnings to the competent authorities. The examples below illustrate how predictive analytics create deep insights on the network structure of terrorist groups, predicting fragmentations and creating policies aimed at reducing attacks.



Photo by Kevin Ku on Unsplash

INSIKT Intelligence, a tech start-up active in this domain, employs different machine learning models to detect potential threats online through NLP and SNA techniques performed on open-source content acquired from social media and other sources.<sup>65</sup> The insights derived as a result of the text and network analysis are then used to identify potentially dangerous content and possible threats or prescribe patterns of relationships between individuals or organizations. Using SNA, INSIKT assesses the activity of a group of users in a network, determining nodes of influence and levels/effectiveness of information diffusion from, for example, propaganda across these networks. This open-source-

64 Alexander Babuta, Marion Oswald & Ardi Janjeva. (Apr. 2020). Artificial Intelligence and UK National Security: Policy Considerations. RUSI Occasional Papers. Accessible at [https://static.rusi.org/ai\\_national\\_security\\_final\\_web\\_version.pdf](https://static.rusi.org/ai_national_security_final_web_version.pdf).

65 See <https://www.insiktintelligence.com/>



based analysis allows law enforcement agencies to make use of big data and prioritize limited resources on potential threats. Data anonymization or pseudonymization also enhances compliance with data protection principles, such as those included in the ASEAN Framework on Personal Data Protection or the Privacy Framework of the Asia-Pacific Economic Cooperation.<sup>66</sup>

Automated modelling of terrorist networks through the use of systematically collected data about an organization can support counter-terrorism efforts by identifying priorities and the most effective strategies to “influence” the behaviour of terrorists. For example, researchers have, for instance, applied AI including SNA algorithms to predict fragmentations of terrorist groups to create deeper insights on how and when terrorist groups such as ISIL and Al-Qaida split.<sup>67</sup>

In one interesting study in 2013, behavioural forecast models were applied to undertake a comprehensive and extensive assessment of Lashkar-e-Taiba – a group associated with Al-Qaida, which was responsible for numerous attacks in Pakistan, India, and Afghanistan and reached global notoriety with the Mumbai attacks in 2008.<sup>68</sup> Based on data regarding the group’s ideology, history and other pertinent facts, the researchers applied “temporal probabilistic” rules to determine what actions would be required to thwart Lashkar-e-Taiba’s campaign and reduce the group’s lethality. The basis of the research were behavioural models, so-called Stochastic Opponent Modeling Agents (SOMA) Rule Learning Algorithm. The research resulted in the generation of policy recommendations that could be used to create an environment “around” Lashkar-e-Taiba conducive to reducing attacks.<sup>69</sup>

Another research initiative of interest concerns a machine learning predictive analytics model linking network structure to lethality. The algorithm, known as “Shaping Terrorist Online Network Efficacy” or “STONE” produces forecasts on who is likely to succeed in a certain position within a given terrorist network if an actor is forcibly removed through an intervention by security forces; how will a network restructure itself when multiple actors are removed; and finally, how to manipulate the network in order to minimize the “expected lethality” of a network.<sup>70</sup> Predicting network responses to different interventions is useful for intelligence and security agencies to determine how to most effectively deploy limited resources by directing them to a particularly promising area.

These examples show how predictive analytics can function in cases where enough data is available to inform counter-terrorism operations. Online data, in particular social media data, can theoretically be a game changer for predictive analytics, offering an entire new dimension of publicly available or open-source data on terrorist organizations, their members, as well as actions of other actors that could influence their behaviour. Nevertheless, as noted, there are real challenges with respect to the use of social media data to predict future individual behaviour of terrorists that are likely to hamper the application of predictive analytics in this regard. Law enforcement and counter-terrorism agencies seeking to explore the use of predictive analytics in this manner should recall that the results produced reflect probabilities and, necessarily, fall short of constituting actionable evidence. Notwithstanding this, while using predictive analytics to forecast terrorist activities may be technically and ethically challenging, as well as practically challenging to act upon, predictive analytics may find use for law enforcement and counter-terrorism agencies in drawing investigator’s attention to particular patterns or, as the case may be, abnormalities in patterns that may arise and merit further consideration.

---

66 ASEAN. (Nov. 2016). Telecommunications And Information Technology Ministers Meeting, Framework on Personal Data Protection. Accessible at <https://asean.org/storage/2012/05/10-ASEAN-Framework-on-PDP.pdf>. APEC. (2015). Privacy Framework. Accessible at [https://www.apec.org/Publications/2017/08/APEC-Privacy-Framework-\(2015\)](https://www.apec.org/Publications/2017/08/APEC-Privacy-Framework-(2015)).

67 See <https://www.insiktintelligence.com/>.

68 Venkatraman Siva Subrahmanian, et al. (2013). Computational analysis of terrorist groups: Lashkar-e-Taiba, Springer Science & Business Media. Springer.

69 Ibid, 156.

70 Francesca Spezzano, Venkatraman Siva Subrahmanian & Aaron Mannes. (2014). Reshaping Terrorist Networks, Communications of the ACM, 57(8), 60-69; Francesca Spezzano, Venkatraman Siva Subrahmanian & Aaron Mannes. (2013). STONE: Shaping Terrorist Organizational Network Efficiency, Proc. 2013 International Conference on Advances in Social Network Analysis and Mining (ASONAM 2013).

## ii. Identify red flags of radicalization

A further AI use case for countering terrorism online concerns the use of AI-powered technology to help identify individuals at risk of radicalization in online communities to facilitate appropriate investigation and intervention, which, as already noted, is an increasingly pertinent phenomenon online. It is also one that is impossible to detect using traditional law enforcement methods.

While radicalization is a complex social phenomenon and the path to radicalization is very personal and often political, machine learning techniques such as NLP can provide valuable support to law enforcement and counter-terrorism agencies, as well as for that matter relevant other actors within the community such as social workers. NLP can be used, for instance, to identify keywords that may indicate the state of radicalization of a social media account or the vulnerability of an individual to terrorist narratives online. It can also be helpful to recognize specific behavioural patterns of individuals, such as consuming or searching for terrorist and violent extremist content which fits radicalization indicators.

The European Union (EU)-funded Real-time Early Detection and Alert System for Online Terrorist Content (RED-Alert) Project is one example of a tool aiming to detect early stages of radicalization while seeking to cater for high privacy and security standards. RED-Alert uses NLP, SNA and complex event processing to collect, process, visualize and store online data related to terrorist groups, including early stages of radicalization based on social media content.<sup>71</sup> The tool supports the search for known keywords or subjects in content that has not been identified as relevant yet. Additionally, the tool includes an anonymization and de-anonymization process of the data that adapts to organizational processes of the law enforcement agencies, which seems promising also for other areas dealing with sensitive data. The RED-Alert project concluded in late 2020, with law enforcement agencies involved in piloting the platform indicating that it offered a significant improvement over tools they currently use.<sup>72</sup> Notwithstanding this, it is important to note that the platform has been used only in a test phase and thus its operability outside testing environments remains to be seen.



*Photo by Kai Pilger on Unsplash*

<sup>71</sup> See <https://cordis.europa.eu/project/id/740688>.

<sup>72</sup> International Institute for Counter-Terrorism. (Oct. 29,2020). Red-Alert Final conference recording. Accessible at <https://www.youtube.com/watch?v=l6ndaCudyJk>.



Moonshot, a tech start-up based in the United Kingdom that specializes in countering violent extremism, provides a further good example of the role AI can play in identifying people vulnerable to radicalization. Like the RED-Alert project, Moonshot aims to identify individuals vulnerable to radicalization online, but rather than facilitate interception and investigation by law enforcement and counter-terrorism agencies, Moonshot seeks to connect these vulnerable individuals with “positive messages”. This application is further addressed in the section below on countering terrorist narratives.

AI can also help monitor global shifts, which can indicate fruitful soil for radicalization online. For instance, the German government-funded research project “Monitoring System and Transfer Platform Radicalization” (MOTRA) is currently working on a comprehensive monitoring tool to analyze aggregated data to monitor high-profile societal developments. The objective is to detect changes in attitudes, which can potentially serve as an early indicator of criminal activity. Systematic monitoring enables faster identification and classification of new trends and serves as the basis for prognostic statements enabling the development of a security policy that is evidence-based, repressive and preventive. The design of the methodology and technology is still, however, under development and, hence, there is limited information available. Initial results are expected from MOTRA in 2023.<sup>73</sup>

As these examples demonstrate, AI-enabled technology can be beneficial in supporting analysts to identify potential vulnerabilities to radicalization online. Nevertheless, it must be emphasized that automated assessments of vulnerability to radicalization prompt very serious ethical concerns. Moreover, it must also be said that the technology is far from being in a position whereby it could replace the work of experienced security professionals. Finally, it is important to acknowledge that, even if the technology was in such an advanced state that it could be leveraged confidently and reliably, activities such as this in the preventative domain do not necessarily always provide grounds for law enforcement intervention.

### iii. Detecting mis- and disinformation spread by terrorists for strategic purposes

The phenomenon of mis- and disinformation is not new. Yet, the capability for such “fake news” content to reach wide audiences at relatively low costs via online means is unprecedented. While the fabrication or distortion of information is not necessarily illegal, it can certainly be harmful and has the potential to contribute to the spread of terrorist or violent extremist narratives into the mainstream discourse. During the COVID-19 pandemic, for instance, such terrorists or violent extremists created and amplified misleading content on a large-scale, by taking advantage of vulnerabilities in the social media ecosystem and by manipulating people through conspiracy narratives and fake news to *inter alia* undermine trust in the government and, at the same time, reinforce non-state actors’ extremist narratives and recruitment strategies.<sup>74</sup>

For mis- and disinformation to take root however, it must first be widely spread online and brought into contact with vulnerable users. While humans play a major role in spreading mis- and disinformation, so-called bots compound the scope and scale of the problem. Short for “robot”, bots are a type of software application that operates online and performs repetitive tasks. Chatbots, a type of bots, can, for instance, simulate basic conversation and, for this reason, are often used on websites to facilitate and perform rudimentary customer services. According to one study in 2017, there were as many as 23 million bots on Twitter, 140 million bots on Facebook and around 27 million bots on Instagram.<sup>75</sup> Groups such as ISIL have proven themselves proficient at employing bots on social media to automate the dissemination of their propaganda.<sup>76</sup>

---

73 Bundeskriminalamt. (2021). Projektbeschreibung. Bundeskriminalamt Accessible at [https://www.bka.de/DE/UnsereAufgaben/Forschung/ForschungsprojekteUndErgebnisse/TerrorismusExtremismus/Forschungsprojekte/MOTRA/Projektbeschreibung/projektbeschreibung\\_node.html](https://www.bka.de/DE/UnsereAufgaben/Forschung/ForschungsprojekteUndErgebnisse/TerrorismusExtremismus/Forschungsprojekte/MOTRA/Projektbeschreibung/projektbeschreibung_node.html)

74 UNICRI. (Nov. 2020). Stop the virus of disinformation. Accessible at [http://unicri.it/sites/default/files/2021-01/misuse\\_sm\\_0.pdf](http://unicri.it/sites/default/files/2021-01/misuse_sm_0.pdf).

75 Carolina Alves de Lima Salge & Nicholas Berente. (Sep. 2017). Is That Social Bot Behaving Unethically?. Communications of the ACM, 60(9), 29–31. Accessible at <https://cacm.acm.org/magazines/2017/9/220438-is-that-social-bot-behaving-unethically/fulltext?mobile=false>.

76 J. M. Berger & Jonathan Morgan. (March 2015). The ISI Twitter Census, The Brookings Institution. Accessible at [https://www.brookings.edu/wp-content/uploads/2016/06/isis\\_twitter\\_census\\_berger\\_morgan.pdf](https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf).



*Photo by Daria Nepriakhina on Unsplash*

While it is not likely to stem the flow of mis- and disinformation in its entirety, the identification of fake or bot accounts created with the intention of spreading fake news or to steer debates in certain directions presents a possible entry point for combating significant percentages of terrorist spread mis- and disinformation. Researchers investigating Twitter bots have suggested that tweets made by bots concern very narrow topics, whereas tweets by humans tend to be more diversified in terms of content. In this regard, it was hypothesized that AI tools could be used to automatically identify bots.<sup>77</sup>

Government Communications Headquarters (GCHQ), the British intelligence and security organization, recently announced that it will leverage AI-supported tools to detect and identify fake accounts that spread false news. GCHQ will also use technology to automate fact-checking through validation against trusted sources, to detect and block botnets, and to identify groups of Internet trolls, known as “troll farms”, and other sources of misinformation.<sup>78</sup>

Although not directly of relevance to law enforcement and counter-terrorism agencies, a further application of the use of AI to counter dis- and misinformation online is demonstrated by the NewsGuard initiative. Founded in 2018, NewsGuard is a journalism and technology company that assesses the credibility of news and information websites and tracks online misinformation.<sup>79</sup> The analysis of the trustworthiness is conducted by journalists and experienced editors. NewsGuard used entries from their misinformation dataset and combined with NLP and other machine learning techniques to detect false news across platforms.<sup>80</sup>

---

77 Shashank Gupta. (Dec. 16, 2017). A Quick Guide to Identify Twitterbots Using AI, Hackernoon Online. Accessible at <https://hackernoon.com/a-quick-guide-to-identify-twitterbots-using-ai-c3dc3a7b817f>.

78 GCHQ. (2021). Pioneering a New National Security - The Ethics of Artificial Intelligence. Accessible at: <https://www.gchq.gov.uk/files/GCHQAIpaper.pdf>.

79 Brian Stelter. (March 4, 2018). This start-up wants to evaluate your news sources. CNN Business.

80 NewsGuard. (2020). The Social Media Platforms Say They Can't Keep Up With the Volume of Hoaxes. Now, They Can. NewsGuard Online. Accessible at <https://www.newsguardtech.com/press/the-social-media-platforms-say-they-cant-keep-up-with-the-volume-of-hoaxes-now-they-can/>



Similarly, in Sri Lanka, researchers are identifying and annotating misinformation with a machine learning tool. According to the preliminary research findings, a machine learning model has been trained with as little as 1,600 articles, with up to 97% accuracy in classifying misinformation across one million articles. The continued multistage research will be conducted in English, Bengali and Sinhala. Through the application of a simpler machine learning model, as opposed to a neural network, the researchers in question aim to obtain traceable results without compromising accuracy significantly.<sup>81</sup> Although not designed for law enforcement or counter-terrorism purposes, the promising results of such studies may, in time, feed into practical applications for security forces to counter terrorist use of mis- and disinformation.

#### iv. Automated content moderation and takedown

Social media channels and webpages that host multi-media content fight against terrorist misuse of their services in different ways. In many ways, how they do so depends on how terrorists misuse their platforms. For example, a website that enables the sharing of large files may undertake a different approach than a messaging app that allows for encrypted communications. One well-known response by social media companies to tackle terrorist and extremist content is “deplatforming”. As defined by the Oxford Dictionary, deplatforming is “the action or practice of preventing someone holding views regarded as unacceptable or offensive from contributing to a forum or debate, especially by blocking them on a particular website”. It is often used as a response by social media companies to repeat violations of the terms of service or the community standards of their platforms. Another approach often used is a technique known as “shadow-banning”.<sup>82</sup> This refers to the content being either deleted entirely or having its visibility significantly limited, without the user that published the content being made aware.

Such tools and approaches are, naturally, not without controversy. Notably, concern has arisen with respect to the possibility of subjective, disparate, and potentially biased enforcement by different platforms resulting from the absence of internationally agreed upon definitions of what constitutes terrorism and violent extremism.<sup>83</sup> Moreover, with limited transparency regarding the tools and approaches used by social media companies and the absence of details regarding their implementation, it is also difficult to really assess their effectiveness. In fact, the effectiveness of deplatforming is quite debated,<sup>84</sup> as it may in fact increase grievances and provide victimization narratives for banned users and communities. It may also contribute to encouraging the migration of such individuals from well-regulated large platforms with considerable content moderation resources to less regulated smaller platforms and niche services, whose capabilities to address such challenges online may be more limited. The resulting fragmented landscape can, in fact, compromise the collective efforts to combat the problem in general.

It is pertinent to note that it is also not always straightforward to disrupt terrorist activities by simply deleting an account or limiting the visibility of content. For instance, the perpetrators of the Christchurch terrorist attack in 2019 live-streamed the attack – an innovation that presented new challenges for content moderators. Although Facebook successfully removed the original video 12 minutes after the recording ended, the video went viral and, in the 24 hours that followed, there were as many as 1.5 million attempts to upload copies of the video across the globe.<sup>85</sup> Christchurch unequivocally demonstrated the need of being able to not only delete accounts and limit the visibility of content, but also to be able to operate across different platforms in order to prevent the strategic distribution of terrorist content in a timely manner.<sup>86</sup> In the aftermath of Christchurch, the Global Internet Forum to Counter Terrorism (GIFCT) created a shared industry database of hashes of terrorist propaganda with the intention of supporting coordinated takedown of such content across platforms, while adhering to data privacy and retention policies.<sup>87</sup>

---

81 Conversation with Yudhanjaya Wijeratne, LIRNEasia on 16 March 2021. Official documentation of the research is still to be published.

82 Nathan Chandler. (2019). Why Do Social Media Platforms practice Shadowbanning. howstuffworks. Accessible at <https://computer.howstuffworks.com/internet/social-networking/information/shadowbanning.htm>.

83 Human Rights Watch. (Jul. 30, 2020). Joint Letter to New Executive Director. Global Internet Forum to Counter Terrorism. Accessible at <https://www.hrw.org/news/2020/07/30/joint-letter-new-executive-director-global-internet-forum-counter-terrorism>.

84 Ryan Greer. (2020). Weighing the value and risks of Deplatforming. GNET Insights. Accessible at <https://gnet-research.org/2020/05/11/weighing-the-value-and-risks-of-deplatforming/>.

85 Facebook. (Mar. 18, 2019). Update on New Zealand. Accessible at <https://about.fb.com/news/2019/03/update-on-new-zealand/>.

86 Marie Schroeter. (2021). Trugschluss Algorithmen, IP Spezial (1). Deutsche Gesellschaft für Auswärtige Politik, 54-56.

87 GIFCT. (2021). Joint Tech Innovation. Accessible at: <https://gifct.org/joint-tech-innovation/>.



*Photo by u j e s h on Unsplash*

Automated content moderation has risen to prominence as a pragmatic approach to dealing with vast amounts of user-generated content online and the speed at which certain content can go viral. Private companies leverage different forms of automated solutions to curate and moderate content online, either by removing or downgrading the content or by redirecting users to other content. For instance, Facebook relies on machine learning to prioritize which content needs to be reviewed first.<sup>88</sup> Posts that violate the company's policies are flagged either by users or by machine learning filters, which includes everything from spam to hate speech and content that "glorifies violence". Since 2020, the company elected to deal with clear-cut cases by automatically removing the post or blocking the account.<sup>89</sup> Only content that does not obviously violate the company's policy is reviewed by a human content moderator. For countering terrorism and violent extremism, Facebook relies on different tools, including AI language models to understand text that might be advocating for terrorism, which is often language and group-type specific. The platform reports that human expertise is still needed for nuanced understanding of how terrorism and violent extremism manifests around the world.<sup>90</sup> Other popular platforms like Twitter<sup>91</sup> and YouTube<sup>92</sup> also rely on AI to quickly remove comments that violate the companies' rules.

Content moderating AI models are trained to filter out specific content that fits certain criteria. However, in their current form, they suffer from inherent limitations. For instance, a machine learning model trained to find content from one terrorist organization may not work for another because of language and stylistic differences in their propaganda. As MIT's journalist Karen Hao puts it, "An algorithm that has learned to recognize Holocaust denial can't immediately spot, say, Rohingya genocide denial".<sup>93</sup> As noted previously in this report, AI must be trained on data. Without such data, it cannot filter out content. At the same time, as will be noted in the following chapter, AI faces serious challenges with respect to layers of complexity in language use, particularly irony and humour, which can significantly hamper the efficacy of automated content moderation. Current models are also often trained on major languages and are therefore less reliable for minority languages such as those mostly spoken in the South Asia and South-East Asia region. In this regard, automated content moderation can or rather should not be truly fully automated. Human oversight of the review and decision-making processes remains a necessity. Notwithstanding the limitations, automated content moderation solutions are increasingly considered indispensable in the private sector in light of the sheer amount of content published online each and every day.

---

88 Mike Schroepfer. (Nov. 13, 2019). Community Standards report. Facebook AI. Accessible at <https://ai.facebook.com/blog/community-standards-report/>.

89 James Vincent. (Nov. 13, 2020). Facebook is now using AI to sort content for quicker moderation. The Verge. Accessible at <https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation>.

90 Erin Saltman. (Aug. 27, 2020). Countering terrorism and violent extremism at Facebook: Technology, expertise and partnerships. Observer Research Foundation. Accessible at <https://www.orfonline.org/expert-speak/countering-terrorism-and-violent-extremism-at-facebook/>.

91 Twitter Public Policy. (May 15, 2019). Addressing the abuse of tech to spread terrorist and extremist content. Twitter. Accessible at [https://blog.twitter.com/en\\_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c.html](https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c.html).

92 Jigsaw. (2020). Countermeasures in practice. Accessible at <https://jigsaw.google.com/the-current/white-supremacy/countermeasures/>

93 Karen Hao. (Mar. 11, 2021). How Facebook got addicted to spreading misinformation. MIT Technology Review. Accessible at: <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>.



Naturally, given the nature of the problem at hand, national authorities have traditionally had limited command over content moderation and takedown. In response to mounting criticism that under existing forms of self-regulation social media companies are failing to tackle the spread of such content, several countries have sought to adopt national legislation aimed at compelling companies to do more. For example, this can be achieved by regulating how fast content must be taken down and by setting incentives and punishing non-compliance. Germany's Network Enforcement Act,<sup>94</sup> for instance, is one example of national legislation, which gained notoriety in 2017 when it became one of the first such instruments adopted.<sup>95</sup> Another notable example is the recently adopted EU regulation on addressing the dissemination of terrorist content online, which compels removal of terrorist content within one hour of receipt of removal notices.<sup>96</sup>

Some national authorities have also begun to adopt a more "hands-on" approach. For instance, law enforcement and counter-terrorism agencies in several countries, as well as within Europol,<sup>97</sup> have begun to try to have online terrorist and violent extremist content removed themselves by using the content flagging mechanisms in the platforms to report content as an infringement of the platforms' terms of service. In this regard, just as AI has risen to prominence in terms of how social media platforms automate content moderation, AI can also play a role in enhancing the capabilities of such Internet referral units within law enforcement and counter-terrorism agencies to sift through the vast amounts of content generated. Notably, in this regard, the United Kingdom Home Office and ASI Data Science announced in 2018 the development of new technology that leverages machine learning to analyze audio-visual content online to determine whether it could be ISIL propaganda.<sup>98</sup> According to the Home Office, tests have shown the tool can automatically detect such content with 99.995% accuracy. The methodology behind the tool has not been made public.

## v. Countering terrorist and violent extremist narratives

While removing terrorist and violent extremist content from the Internet or social media may, to a certain degree, be effective in preventing the spread of harmful narratives, it is exceptionally controversial and raises serious human rights concerns. Moreover, it is important to note that content removal does not contribute in any way to addressing the root causes of terrorism and leaves vulnerable individuals at risk. Beyond just identifying vulnerable individuals, NLP and machine learning algorithms can play an even more proactive role in countering terrorism online. AI can be used to analyze users' behaviour and direct them to specific content conducive to countering terrorist narratives or to trigger ad-targeting tactics.<sup>99</sup>

Moonshot, with its innovative "Redirect Method", which was piloted in 2016 with Google's Jigsaw, uses automated risk assessment and NLP to identify vulnerable audiences based on their online search behaviour.<sup>100</sup> For instance, if people are searching for specific content that matches pre-defined indicators, advertisements with positive, de-radicalizing content are triggered and users are redirected to ads and curated videos on a YouTube channel designed to refute ISIL propaganda. In one such video, a woman in ISIL-controlled territory secretly recorded her life, giving unique firsthand insights into the reality of life under ISIL rule.<sup>101</sup> The counter-narrative content is curated in cooperation with non-governmental organizations on the ground in order to tailor the messages adequately to local contexts. This includes non-ideological content and contact details to help-lines and one-on-one interventions.<sup>102</sup>

---

94 In German, "Netzwerkdurchsetzungsgesetz".

95 Ben Knight. (Jan. 1, 2018) Germany implements new internet hate speech crackdown. DW. Accessible at <https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590>

96 European Parliament. (Apr. 28, 2021). New rules adopted for quick and smooth removal of terrorist content online. European Parliament. Accessible at <https://www.europarl.europa.eu/news/en/press-room/20210422IPR02621/new-rules-adopted-for-quick-and-smooth-removal-of-terrorist-content-online>.

97 Europol. (2021). EU Internet Referral Unit EU IRU. Europol. Accessible at <https://www.europol.europa.eu/about-europol/eu-internet-referral-unit-eu-iru>.

98 Home Office & The Rt Hon Amber Rudd. (Feb. 13, 2018). New Technology revealed to help fight terrorist content online. GOV.UK. Accessible at <https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online>.

99 Joan Barata. (Feb. 24, 2021). Moderating Terrorist and Extremist Content. VoxPol. Accessible at <https://www.voxpol.eu/moderating-terrorist-and-extremist-content/>.

100 See <https://moonshotteam.com/redirect-method/>.

101 CNN. (Sep. 25, 2014). Raqqa woman secretly films her life under ISIS rule. CNN. Accessible at [https://www.youtube.com/watch?v=0qK-Krjw0QgI&list=PL0l4bTGBHIMdNkDF061NF6OjCL\\_Gld6Va&t=94s](https://www.youtube.com/watch?v=0qK-Krjw0QgI&list=PL0l4bTGBHIMdNkDF061NF6OjCL_Gld6Va&t=94s)

102 Moonshot CVE. (2020). Social Grievances and Violent Extremism in Indonesia. Moonshot CVE. Accessible at <https://moonshotcve.com/indonesia-social-grievances-violent-extremism/>.



*Photo by laura adai on Unsplash*

While it is difficult to measure the impact of approaches such as the Redirect Method, it is perhaps nevertheless promising that, during the pilot of the programme, targeted individuals that were redirected to such content watched more than half a million minutes of content.<sup>103</sup>

Notably, Moonshot has also conducted a variety of programmes using micro-targeted advertisement online for deradicalization purposes, producing useful insights on digital deradicalization strategies in South Asia and South-East Asia. For example, in 2020, Moonshot conducted an experiment in Indonesia connecting vulnerable individuals with psychosocial support.<sup>104</sup> The research found that users receiving psychosocial support engaged more frequently with non-ideological deradicalizing content than with ideological deradicalizing content.

The London-based Institute for Strategic Dialogue also uses AI-driven social media advertisement tools to counter ISIL narratives online and in 2015 launched a curated counter-narrative campaign tackling ISIL.<sup>105</sup> In cooperation with the Network Against Violent Extremism and supported by Facebook and Twitter, the institute targeted vulnerable people especially at risk of being targeted by ISIL propaganda.

The potential for AI to play a role in countering terrorist narratives online is clear, especially when it comes to reaching individuals and groups at risk. Nevertheless, the AI tools in question can only be but one part of the solution. AI can help to connect the dots, but truly countering terrorist narratives online requires a much more nuanced understanding of individuals' paths to radicalization that such tools can afford. Furthermore, the important role of civil society organizations and such initiatives in these processes cannot be discounted.

<sup>103</sup> Naomi LaChance. (Sep. 8, 2016). Google program to deradicalize jihadis will be used for right-wing American extremist next. The Intercept. Accessible at <https://theintercept.com/2016/09/07/google-program-to-deradicalize-jihadis-will-be-used-for-right-wing-american-extremists-next/>.

<sup>104</sup> Moonshot CVE. (Dec. 2020). Social Grievances and Violent Extremism in Indonesia. Moonshot CVE. Accessible at <https://moonshotcve.com/indonesia-social-grievances-violent-extremism/>.

<sup>105</sup> Tanya Silverman, et al. (2016). The impact of counter-narratives. Institute for Strategic Dialogue. Accessible at [https://www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives\\_ONLINE\\_1.pdf](https://www.isdglobal.org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE_1.pdf).





## vi. Managing heavy data analysis demands

Analysis of both the Berlin Christmas market attack in 2016 and the Manchester Arena bombing in 2017 attest that datasets should be used to cross-reference information and validate patterns in order to identify relevant connections, as in both instances the attackers were already in local authorities records as subjects of interest.<sup>106</sup> Notwithstanding this, given the growing volume and velocity of data collected through law enforcement processes, in particular in the context of online investigations, such analyses are often not possible.

Whether it is for the purpose of countering terrorism online or offline, AI can undoubtedly play a significant part in enhancing the capabilities of national authorities to process large quantities of data in an effective manner and, in doing so, to optimize the necessary amount of human and financial resources allocated for any specific situation. More specifically, AI can be used to extract relevant information, filter, and triage data to help prioritize the analysis of vast sets of data that may identify vital investigative leads and help save lives.

The analysis of audio-visual content is one task that requires considerable specialized human resources. With the massive expansion of smart video recording capabilities in recent years and in law enforcement's own surveillance capabilities, including through closed-circuit television (CCTV) and the use of body-worn cameras (or "bodycams") and patrol drones, there has been a dramatic increase in the quantity of video footage requiring analysis. In the context of terrorism, it is very well established that terrorist groups and individuals make extensive use of the medium of video and actively share and disseminate such content online. Moreover, when considering online counter-terrorism investigations, it is also pertinent to consider that, according to CISCO, by 2021, an estimated 82% of consumer Internet traffic will be video.<sup>107</sup> In light of this, the work of digital forensics detectives is immense and growing.

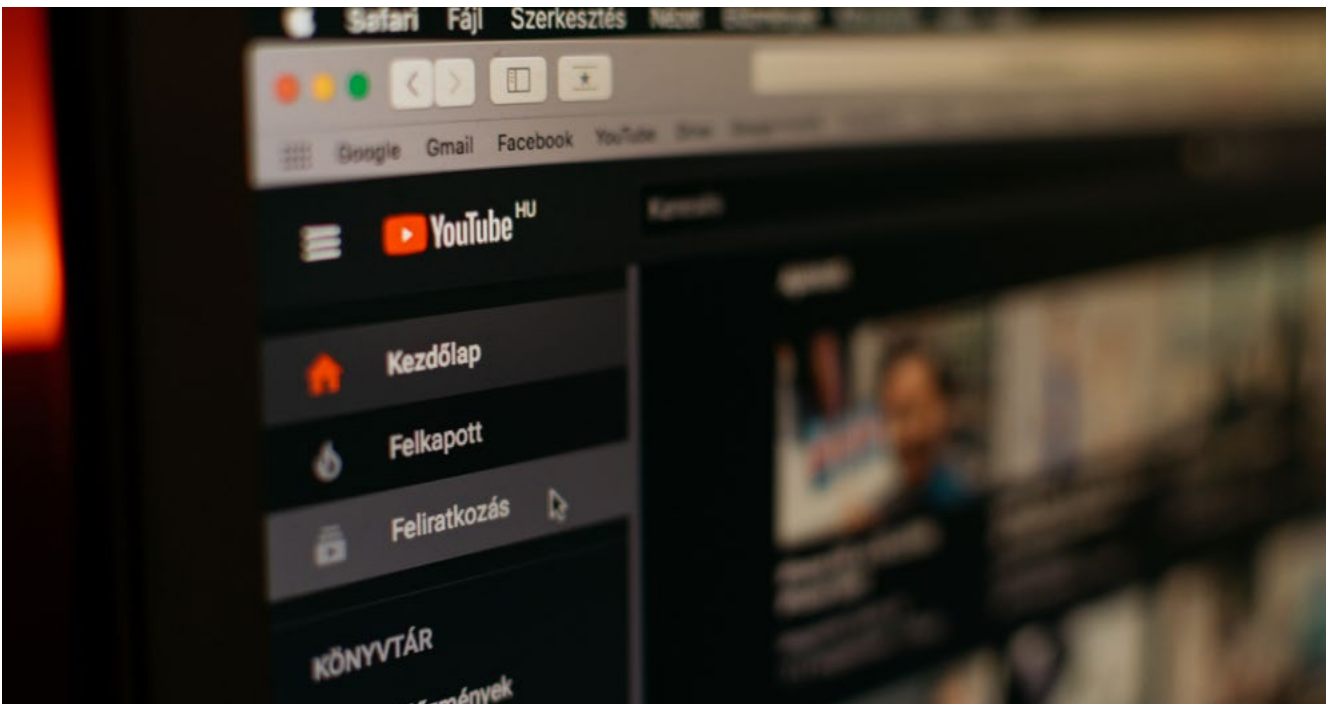


Photo by Szabo Viktor on Unsplash

106 Regarding Manchester see: David Anderson. (Dec. 2017). Attacks in London-Manchester March-June 2017. Accessible at [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/664682/Attacks\\_in\\_London\\_and\\_Manchester\\_Open\\_Report.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/664682/Attacks_in_London_and_Manchester_Open_Report.pdf). Regarding Berlin, see: Bruno Jost. (Oct. 2017). Abschlussbericht des Sonderbeauftragten des Senats für die Aufklärung des Handelns der Berliner Behörden im Fall AMRI, Der Sonderbeauftragte des Senats von Berlin.

107 Cisco. (2020). Cisco Annual Internet Report (2018-2023). Accessible at <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf>.

Facial recognition, a biometric application of AI-powered object recognition, is one technology that offers considerable promise in terms of being able to analyze video footage to identify persons of interest. The adoption of facial recognition technology has increased dramatically over the past few years, fueled by the rapid improvement of machine learning. Facial recognition software requires high-quality photos for a reference database to identify targeted individuals. Several law enforcement agencies around the world have already begun experimenting with the use of the technology. In 2017 and 2018, for instance, the German Federal Police piloted facial recognition-enabled CCTV in one of its busiest train stations in Berlin. The final report of the pilot confirmed the promise of using the technology to identify relevant subjects such as terrorists or violent offenders in crowded public places.<sup>108</sup> INTERPOL also operates a facial recognition system that contains facial images received from more than 179 countries. Face images in notices and diffusions requested by member countries are searched and stored in the facial recognition system in line with INTERPOL's Rules on the Processing of Data.<sup>109</sup> Tools such as facial recognition can also be particularly useful for forensic investigators analyzing video evidence collected online. At the same time however, the rising use of facial recognition technology by authorities across the globe has contributed to considerable negative feedback from human rights groups and civil society organizations that have flagged concerns around the potential of the technology to, inter alia, reflect and reinforce pre-existing biases in law enforcement.<sup>110,111</sup>

---

108 Bundespolizeipräsidium. (2018). Abschlussbericht Teilprojekt 1 Biometrische Gesichtserkennung. Accessible at [https://www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011\\_abschlussbericht\\_gesichtserkennung\\_down.pdf?\\_blob=publication-File](https://www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011_abschlussbericht_gesichtserkennung_down.pdf?_blob=publication-File).

109 INTERPOL. (2021). Facial Recognition. INTERPOL. Accessible at: <https://www.interpol.int/en/How-we-work/Forensics/Facial-Recognition>

110 Amnesty International. (Jan. 26, 2021). Ban dangerous facial recognition technology that amplifies racist policing. Amnesty International. Accessible at <https://www.amnesty.org/en/latest/news/2021/01/ban-dangerous-facial-recognition-technology-that-amplifies-racist-policing/>.

111 Access Now, et al. (Jun. 7, 2021). Open letter calling for a global ban on biometric recognition technologies that enable mass and discriminatory surveillance. Access Now. Accessible at <https://www.accessnow.org/cms/assets/uploads/2021/06/BanBS-Statement-English.pdf>.

## IV. THE GREATER THE OPPORTUNITY, THE GREATER THE CHALLENGE

As demonstrated through the six AI use cases described in the preceding chapter, the potential application of AI for countering the threat of terrorism online is evidently large and thus merits close consideration by law enforcement and counter-terrorism agencies. Notwithstanding this, the application of AI should not be seen as a “quick and easy” solution by any means, as has been alluded to at several stages throughout this report. Although certainly an attractive technology, there are wide-ranging legal, political, and technical challenges that all law enforcement and counter-terrorism agencies must be aware of in advance as well as during the deployment of AI-based technologies. While the actions required to address many of these challenges may fall more with the policymakers in Member States or with the technology provider, it is essential that law enforcement and counter-terrorism agencies are aware of the full spectrum of challenges that AI can present and, as appropriate, to put in place necessary oversight mechanisms.

As with previous chapters, what follows is not intended to be an exhaustive overview of the challenges that may be encountered with AI. Rather, it is intended to be an introductory overview of several key cross-cutting areas of concern. It is important to note that each of the use cases described in the previous chapter will also bring their own unique challenges. Accordingly, law enforcement and counter-terrorism agencies seeking to explore any of these challenges will need to further conduct a thorough assessment on a case-by-case basis to better understand specific challenges presented by each individual use.

### i. Legal and political challenges

#### a. Human rights concerns

Chief among the concerns around the use of AI-enabled technologies in counter-terrorism is the very real and serious potential for the deployment of these technologies to hamper human rights and fundamental freedoms. The United Nations General Assembly has long asserted that States are obliged to respect and fulfil human rights and to protect individuals against abuses by non-State actors in the counter-terrorism context.<sup>112</sup> Moreover, the Human Rights Council has reaffirmed in its resolution on the promotion, protection and enjoyment of human rights on the Internet that human rights apply online as much as they do offline.<sup>113</sup> Law enforcement and counter-terrorism agencies aiming to explore the development of AI capabilities, therefore, need to ensure that they do so in a human rights compliant manner. In addition to refraining from using AI in an ostensibly unlawful manner, such as deploying AI-enabled surveillance systems beyond what is necessary and proportionate for a legitimate aim, authorities need to consider other less overt risks to human rights associated with the use of AI, such as of the potential for machine learning algorithms using biased data to compound bias through automated processes and therefore produce discriminatory outputs.

From the wide spectrum of human rights that can be impacted by the wrongful use of AI-enabled technologies, the right to privacy, freedom of thought and expression and non-discrimination are often identified as the most affected rights.<sup>114</sup> This derives from characteristics inherent to AI. As already noted, the development of AI-enabled technologies requires extensive amounts of data to train the models. Some of this data is likely to be personal data. The United Nations General Assembly Resolution 68/167 on the right to privacy in the digital age describes the “unlawful or arbitrary collection of personal data” as a highly intrusive act that could violate “the rights to privacy and to freedom of

---

112 United Nations General Assembly. (Mar. 11. 2008). Resolution adopted by the General Assembly on 18 December 2007, 62/159. Protection of human rights and fundamental freedoms while countering terrorism. A/RES/62/159. Accessible at <https://undocs.org/en/A/RES/62/159>; United Nations Secretary-General. (Aug. 28, 2008). The protection of human rights and fundamental freedoms while countering terrorism: Report of the Secretary-General. A/63/337. Accessible at <https://www.securitycouncilreport.org/un-documents/document/terrorism-a-63-337.php>.

113 Human Rights Council. (June 29, 2012). Resolution 20/...: The promotion, protection and enjoyment of human rights on the Internet. A/HRC/20/L.13. Accessible at <https://undocs.org/A/HRC/20/L.13>.

114 See, for example: Kathleen McKendrick. (Aug. 2019). Artificial Intelligence Prediction and Counterterrorism, Chatham House. Accessible at <https://www.chathamhouse.org/sites/default/files/2019-08-07-AICounterterrorism.pdf>; Boaz Ganor. (2019). Artificial or Human: A New Era of Counterterrorism Intelligence. Studies in Conflict & Terrorism.

expression and may contradict the tenets of a democratic society”.<sup>115</sup> Moreover, the data fed into the algorithms is often contaminated with human biases and the deployment of these models may result in the amplification of those biases and therefore impact the right to non-discrimination.

Notably, however, not all AI-enabled technologies pose a similar risk to human rights. As observed in chapters two and three, AI is a broad field, and the different types of applications impact human rights in different degrees.

The rights to privacy, freedom of thought and expression and non-discrimination are all provided for under the Universal Declaration of Human Rights, as well as under various international and regional treaties, including the widely-ratified International Covenant on Civil and Political Rights and the ASEAN Human Rights Declaration.<sup>116</sup> Given the relevance of these rights in the online space, this sub-section focuses on key concerns related to the impact of using AI-enabled technologies to counter terrorism online.<sup>117</sup> However, law enforcement and counter-terrorism agencies should bear in mind that other human rights may be impacted by the use of AI tools to counter terrorism online. For instance, the right to the presumption of innocence and fair trial rights are also potentially highly impacted through the use of AI-generated evidence, which, given the opacity of AI, may be difficult for defendants to challenge in court.<sup>118</sup>

Another general observation to bear in mind is that most human rights are not absolute and can be limited if certain requirements are met. In general, limitations to human rights provided in the International Covenant on Civil and Political Rights – including privacy, freedom of expression and non-discrimination – are allowed when such limitations are legally established in the law and are necessary and proportionate to achieve a legitimate aim.<sup>119</sup>

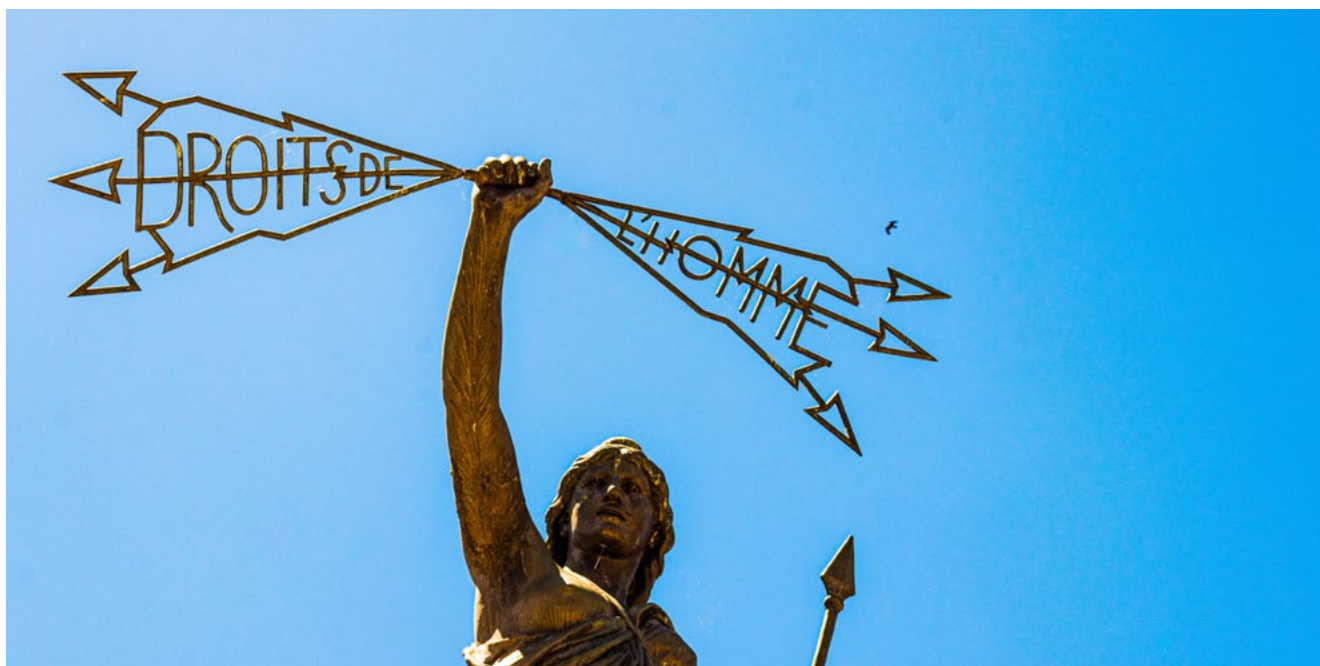


Photo by DDP on Unsplash

115 United Nations General Assembly. (Jan. 21, 2014). Resolution adopted by the General Assembly on 18 December 2013. 68/167. The right to privacy in the digital age. A/RES/68/167. Accessible at <https://undocs.org/A/RES/68/167>.

116 ASEAN. (2012). ASEAN Declaration of Human Rights. Accessible at: <https://asean.org/asean-human-rights-declaration/>.

117 For a more comprehensive overview of the human rights considerations to the use of AI in counter-terrorism, please see: OHCHR, UN-OCT-UNCCT & UNICRI. (2021). Human Rights Aspects to the Use of Artificial Intelligence in Counter-Terrorism.

118 Aleš Završnik. (2020). Criminal justice, artificial intelligence systems, and human rights. ERA Forum (2020) 20, 567–583. Accessible at <https://doi.org/10.1007/s12027-020-00602-0>.

119 American Association for the International Commission of Jurists. (1984). Siracusa Principles on the Limitation and Derogation of Provisions in the International Covenant on Civil and Political Rights Annex. E/CN.4/1984/4 (1984). Accessible at <https://www.refworld.org/docid/4672bc122.html>.



As described in the report of the former Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue, the right to privacy protects each individual's "private sphere", an "area of autonomous development, interaction and liberty" where they are safeguarded "from state intervention and from excessive unsolicited intervention by other uninvited individuals". The right to privacy also entails "the ability of individuals to determine who holds information about them and how is that information used".<sup>120</sup> To give substance to the right to privacy in a growingly digital world, data privacy frameworks have been approved at regional and national levels across the globe. Many countries, including in South Asia and South-East Asia, have enacted personal data protection laws.<sup>121</sup> As highlighted in the Report of the United Nations High Commissioner for Human Rights on the right to privacy in the digital age, while these laws and frameworks vary in content, most of them follow a set of common principles, including that personal data processing should be "fair, lawful and transparent", as well as limited to what is "necessary and proportionate to a legitimate aim". Whenever possible, techniques such as data anonymization and pseudonymization should be implemented.<sup>122</sup>

AI tools involving the mass gathering of untargeted personal data inherently risk intruding upon the privacy of individuals. As noted in the UNOCT and INTERPOL Handbook on Online Counter-Terrorist Investigations "Massive, non-targeted and indiscriminate collection of data is unlikely to meet the requirements for necessity and proportionality. Similarly, the retention of data for longer than is necessary can breach international human rights or data protection laws." For this reason, "Laws establishing automatic collection, storage and use of personal data should include details of the purpose for which the data is collected, the way in which it is used, who has access to the data, the purpose(s) for which it may be used and the length of time that the data may be stored."<sup>123</sup> On the other hand, even when States define conditions under which these intrusions are legally justified, they must take into account that the information collected can be exploited and misused by ill-intended actors.

Although the risks of AI to privacy are frequently stressed, it is sometimes asserted that AI-enabled tools can increase the privacy of individuals. The argument suggests that automated collection and assessment of information makes people less vulnerable to human review of personal data and leads to more effectively targeted operations, focusing on data related to suspicious individuals.<sup>124</sup> On the other side of the debate, counter-contentions assert that "algorithmic profiling" can be considered more intrusive than human profiling and is therefore riskier for human rights.<sup>125</sup> It is also argued that generally, the collection of mass data online in order to gather intelligence has the potential to substantially stress the relationship between governments and citizens as it damages the essence of a trustworthy regime. State-led intrusion through the supervision of online platforms transforms states into "digital authoritarian states".<sup>126</sup>

Nevertheless, the type of technology used and, therefore, the required collection and storage of potentially private data have different implications on the right to privacy. Automated tools that support law enforcement in analyzing lawfully collected data and finding patterns or matching faces from a suspect with open-source or otherwise lawfully collected data from social media accounts must be differentiated from technologies that seek to monitor social media and other online sources in real-time to identify relevant leads for counter-terrorism operations. While the first set of tools may comply with the requirements for limiting individuals' privacy, the latter are too intrusive of the privacy of mostly innocent individuals to be allowed under human rights law.

The rights to freedom of thought and expression include both an "internal" freedom to unconditionally hold and change thoughts, conscience, religion, beliefs or opinions and an "external" freedom to manifest those thoughts, conscience, religion, beliefs, or opinions.

---

120 United Nations General Assembly. (Apr. 17, 2013). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue. A/HRC/23/40. Accessible at <https://undocs.org/A/HRC/23/40>

121 Ameya Paratkar. Data Privacy Laws in South East Asia. Accessible at <https://blog.mithi.com/data-privacy-laws-in-south-east-asia/>; Ikigai Law. (Dec. 2019). New era of data protection regulation in South Asia. Accessible at <https://www.ikigailaw.com/new-era-of-data-protection-regulation-in-south-asia/>

122 Human Rights Council. (Aug. 3, 2018). The right to privacy in the digital age: Report of the United Nations High Commissioner for Human Rights. A/HRC/39/29. Accessible at <https://undocs.org/A/HRC/39/29>.

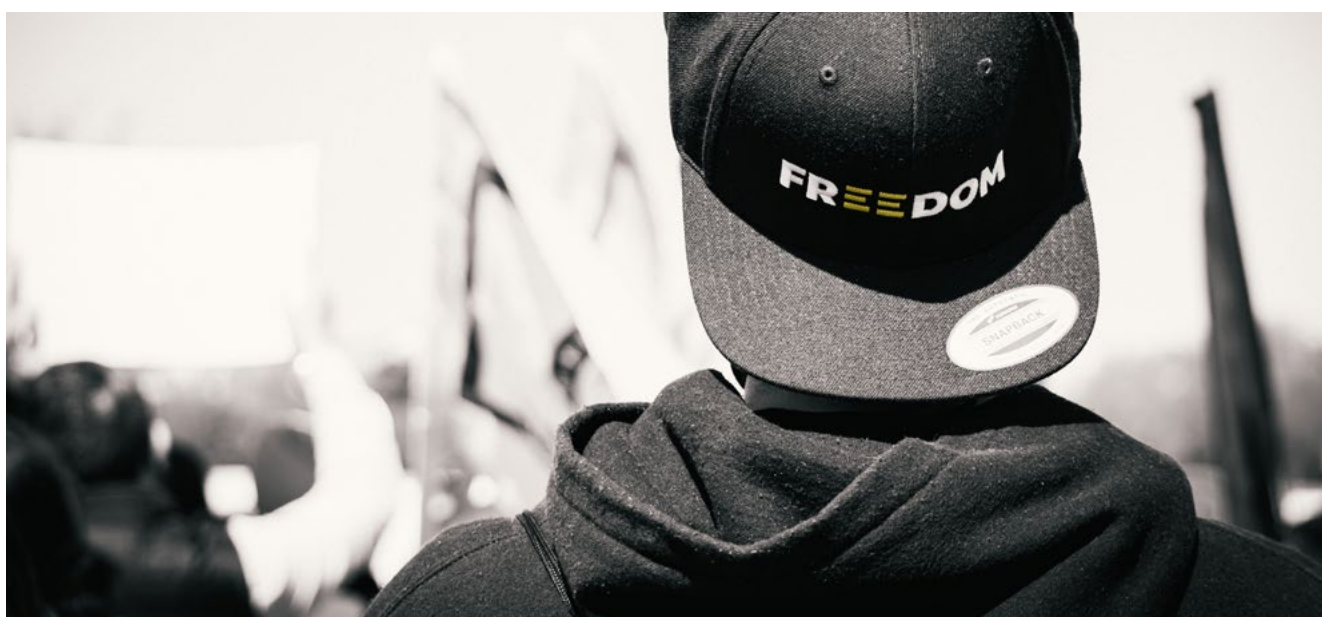
123 INTERPOL & UNOCT. (2021). Using the Internet and Social Media For Counter Terrorism Investigation, 2<sup>nd</sup> Edition, 13.

124 Alexander Babuta, Marion Oswald & Ardi Janjeva. (Apr. 2020). Artificial Intelligence and UK National Security: Policy Considerations. RUSI Occasional Papers. Accessible at [https://static.rusi.org/ai\\_national\\_security\\_final\\_web\\_version.pdf](https://static.rusi.org/ai_national_security_final_web_version.pdf).

125 Ibid.

126 Boaz Ganor. (2019). Artificial or Human: A New Era of Counterterrorism Intelligence, Studies in Conflict & Terrorism, 5.

States can limit the “external” freedom to express beliefs or put them into action if the respective requirements are met. Therefore, under certain circumstances, content moderation and takedown, including through AI solutions, may be permitted. Tech Against Terrorism, an initiative launched by the United Nations Counter-Terrorism Committee Executive Directorate (CTED), has called upon policymakers to provide clear regulation on what is illegal content and how to proceed with it, avoiding the take-down of content protected under freedom of expression. In this regard, it is worth noting that extremists and terrorists increasingly opt to communicate through smaller platforms, making use of a diverse environment of online services from cloud storage to encrypted messenger services to escape the hurdles of tighter measures and take-downs of malicious content in bigger online platforms. As noted above, smaller companies often lack the financial and human resources to prevent their platforms from being exploited by malicious actors. Clear regulation on what is considered terrorist and other malicious content and how to deal with it, combined with capacity-building activities for smaller platforms, can support the companies’ efforts to combat this risk of terrorist exploitation. Nonetheless, the enduring lack of a universally accepted definition of terrorism explained in the subsequent sub-section remains an obstacle for States, as well as smaller and bigger companies alike, to properly navigate the tension between the need to ensure content moderation and the right to freedom of expression.



*Photo by Gayatri Malhotra on Unsplash*

The right to freedom of thought, in its “internal” dimension, is an absolute right, which cannot be limited or interfered with by States in any circumstance.<sup>127</sup> This is especially relevant to note when considering the potential of certain technologies, including AI-enabled technologies, to influence individuals’ thoughts.<sup>128</sup> Deploying automated applications to counter the terrorist use of the Internet by monitoring user-generated content online can infringe the freedom of thought as it can trigger actions based on the individuals’ behaviour online, which represent thoughts. AI-enabled tools interpret daily interactions online and judge upon them as “outwards signs of inner thoughts to indicate who is a security risk to allow law enforcement or security services to take preventative action to stop a thought translating into action”.<sup>129</sup> Similarly, law enforcement and counter-terrorism agencies aiming to use AI-enabled technologies to direct users at risk of radicalization to counter-narrative content need to consider the possibility that these technologies could result in unlawful interference with the right to freedom of thought through the manipulation of the way that the targeted users think.<sup>130</sup>

127 Human Rights Committee. (Sep. 27, 1993). General comment No. 22 (48) (art. 18). CCPR/C/21/Rev.1/Add.4 Accessible at <https://undocs.org/CCPR/C/21/Rev.1/Add.4>.

128 Susie Allegre. (2017). Rethinking Freedom of Thought for the 21<sup>st</sup> century, *European Human Rights Law Review*, 3, 224. Accessible at [https://susieallegre.com/wp-content/uploads/2020/04/Alegre%20from%202017\\_EHRLR\\_Issue\\_3\\_Print\\_final\\_0806%5B6745%5D.pdf](https://susieallegre.com/wp-content/uploads/2020/04/Alegre%20from%202017_EHRLR_Issue_3_Print_final_0806%5B6745%5D.pdf).

129 Ibid, 229.

130 Ibid, 227-228.



The right to non-discrimination guarantees that no person shall be treated less favourably because they hold certain protected characteristics such as race, gender, ethnic origin, or religion.<sup>131</sup> Differential treatment can be justified if it is for a legitimate aim, which must be reasoned and achieved through proportionate means. Counter-terrorism measures online targeting individuals based on their protected characteristics could therefore violate the right to non-discrimination. This holds especially true for AI tools monitoring individual behaviour online, screening for indicators that could act as proxies for the protected characteristics. In fact, traces of protected characteristics are often hidden. Moreover, the enduring problem of bias in AI, further explained below, often impacts the right to non-discrimination for instance by disproportionately targeting individuals belonging to certain disadvantaged groups. Examples from other areas aside from the sensitive online intelligence gathering have demonstrated repeatedly how AI-enabled systems can discriminate against vulnerable groups. There is no reason to believe that systems for law enforcement agencies will not suffer from the same weaknesses. Employing AI in public bodies is often justified with increasing efficiency, yet this cannot justify unequal treatment.

## b. Admissibility of AI-generated evidence in court

Although in principle there are several AI uses cases that can be of relevance for law enforcement and counter-terrorism agencies to counter online terrorism online, the practical utility of any such use case is directly linked with the ability for it to contribute to prosecution before a court. If actions taken on the basis of patterns and correlations identified by AI systems are not deemed evidentiary in nature, the benefit of using AI tools is necessarily limited from the perspective of the end-users.

In light of the technical challenges that will be presented in subsequent sections, in particular the opacity and the risk of bias with AI, evidence derived from the use of AI may fall short in terms of demonstrating reliability and authenticity – aspects traditionally considered important for the admissibility of evidence. As with all forms of digital evidence, AI-generated evidence is susceptible to being modified either intentionally or unintentionally. Moreover, the human rights concerns highlighted in the preceding section may affect the proportionality of the AI-enabled processes used to collect evidence. In other words, the advantages of using AI tools for the interests of security and justice and the probative value of the AI-generated evidence may not justify the risks to human rights posed by the deployment of AI systems by law enforcement and counter-terrorism agencies.<sup>132</sup>

While the implications of this vary depending on each country's legal system and framework, with a lack of legal certainty and a growing body of legal discourse developing around the topic,<sup>133</sup> the use of AI tools to gather evidence is likely to face procedural challenges in courts throughout the world – at least for the foreseeable future. Recognizing this, law enforcement agencies keen on exploring the application of AI have called for additional guidance on the admissibility of AI-derived evidence in court that assesses the impact and results of the specific use of AI tools while ensuring the respect for human rights and rule of law.<sup>134</sup>

Notably, in the well-known 2016 COMPAS case in the United States, concerning a predictive analytics model used by judges and parole officers to assess a criminal defendant's likelihood to re-offend, the Court permitted the use of the AI-generated risk assessment with certain limitations, namely that concerns regarding the accuracy of the risk assessment must be made clear and risk assessment scores could not be used to determine the threshold question of whether to incarcerate a person or the severity of the sentence.<sup>135</sup>

---

131 United Nations and the Rule of Law, Equality and Non-discrimination. Accessible at <https://www.un.org/ruleoflaw/thematic-areas/human-rights/equality-and-non-discrimination/>.

132 INTERPOL. (2019). Global Guidelines for Digital Forensics Laboratories.

133 Daniel Seng & Stephen Mason. (2021). Artificial Intelligence and Evidence. Singapore Academy of Law Journal. Accessible at <https://journalonline.academyPublishing.org.sg/Journals/Singapore-Academy-of-Law-Journal-Special-Issue/Current-Issue/ctl/eFirstSALPD-FJournalView/mid/503/ArticleId/1602/Citation/JournalsOnlinePDF>

134 INTERPOL & UNICRI. (2019). Artificial Intelligence and Robotics for Law Enforcement. Accessible at: <http://unicri.it/artificial-intelligence-and-robotics-law-enforcement>

135 James X. Dempsey. (Aug. 2020). Artificial Intelligence: An Introduction to the Legal, Policy and Ethical Issues. The Berkeley Center for Law & Technology. Accessible at [https://www.law.berkeley.edu/wp-content/uploads/2020/08/Artificial-Intelligence-An-Introduction-to-the-Legal-Policy-and-Ethical-Issues\\_JXD.pdf](https://www.law.berkeley.edu/wp-content/uploads/2020/08/Artificial-Intelligence-An-Introduction-to-the-Legal-Policy-and-Ethical-Issues_JXD.pdf)

As has already been noted, however, even in the event that AI-based evidence does not meet legal thresholds required for use in court, AI tools may nevertheless be of use for law enforcement and counter-terrorism agencies combatting terrorism online by drawing the investigators' attention to particular patterns or abnormalities in patterns that may result in the identification of other forms of admissible evidence.

### c. A fragmented landscape of definitions

It is well established that there is no universally accepted definition of terrorism, and this is a reality that is unlikely to change in the near future. There is equally a lack of clarity or understanding on the precise processes of radicalization.<sup>136</sup> These realities have several serious implications in terms of using AI to counter terrorism online, hindering the use of the technology.<sup>137</sup>

A lack of common understanding of what defines terrorism necessarily results in a lack of clarity regarding what constitutes terrorist content online. Practitioners seeking to moderate content must therefore cater for differing interpretations of what defines terrorism and, of course, the politics that comes with it. Tech Against Terrorism recently highlighted the urgency of addressing the challenge of the absence of a definition of terrorism, recommending policymakers to contribute to counter-terrorism efforts online by providing more definitional clarity "via improved designation"<sup>138</sup>



Photo by Joshua Hoehne on Unsplash

A further challenge is that a multitude of definitions can also lead to a fragmented collection of data on terrorism and violent extremism online. Without the fundamental knowledge, comprehensive methods based on empirical data are certain to lack accuracy, consistency, and comparability. Moreover, it is important to note that radicalization or the path to terrorism is highly individual and, hence, extremely difficult to cast in a mold that can be translated into patterns for automated models to process.

Finally, at this juncture, it is perhaps important to also recall an ever-present concern, namely that authorities may seek to use national terrorism legislation to further their own political agenda by curbing the freedom of speech of human rights activists, journalists, or persons critical of the government, and this can be reflected in how social media

136 Jeffrey Monaghan & Adam Molnar. (2016). Radicalisation theories, policing practices, and "the future of terrorism?"; Critical Studies on Terrorism, Vol. 9(3), pp.393–413.

137 Kathleen Kendrick. (Aug. 2019). Artificial Intelligence Prediction and Counterterrorism. Chatham House, 31–32. Accessible at <https://www.chathamhouse.org/sites/default/files/2019-08-07-AICounterterrorism.pdf>.

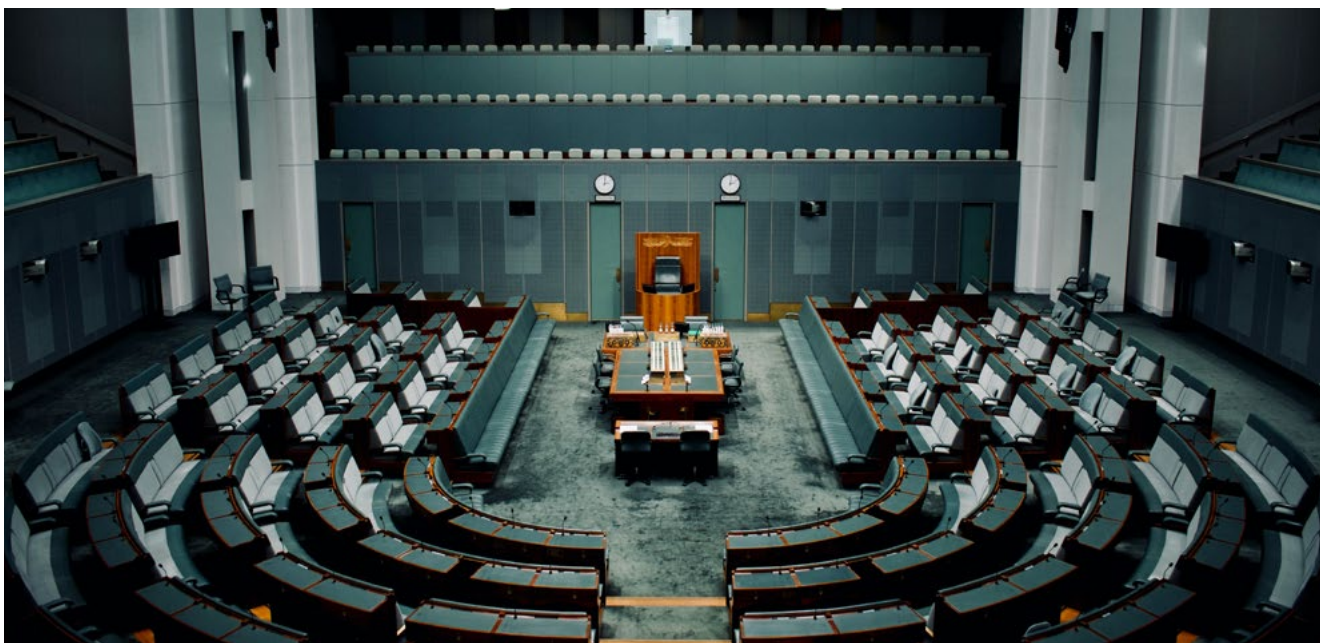
138 Tech against Terrorism. (2021). Position Paper Content Personalisation and the online dissemination of terrorist and violent extremist content. Accessible at <https://www.techagainstterrorism.org/wp-content/uploads/2021/02/TAT-Position-Paper-content-personalisation-and-online-dissemination-of-terrorist-content1.pdf>.



is regulated. These concerns could be considered acute in some parts of South Asia and South-East Asia that have had contentious histories with respect to human rights. Developments such as this are likely to have a chilling effect on civil society activism. Speaking about the challenges that situations such as this can present, Brian Fishman, who leads the effort against terrorist and hate organizations at Facebook, noted that for social media platforms to avoid indirectly acting as an extended arm of a government through their content moderation practices, they would require internationally agreed lists of designations of certain organizations or movements as terrorists.<sup>139</sup>

#### d. AI governance

A question that goes far beyond the counter-terrorism use cases presented and discussed in this report is how the use of AI in general should be regulated in order to ensure fair, just, accountable and transparent AI systems. Many debates in particular have arisen around ethical and fair AI, aiming to erase bias or systematic distortion.<sup>140</sup> The robustness of such systems – their resilience and security – is a further aspect of growing importance in policy discourses, especially in light of the increasing integration of AI into daily operations within both the public and private sector and the consequent potential of attacks being directed at the operation or functionality of AI.<sup>141</sup> What each of these concepts entails from a regulatory point of view is, however, not universally defined at this time. Nevertheless, the governance of AI would necessarily impact the standards of counter-terrorism applications.



*Photo by Aditya Joshi on Unsplash*

It is perhaps pertinent to note that in April 2021, the European Commission presented a proposal for a Regulation laying down harmonized rules on AI,<sup>142</sup> which, if adopted by the European Parliament and the Council of the EU, will directly become hard law in the EU Member States – and the first first-ever supra-national legal framework on AI. The current draft roughly establishes that the rules apply to both providers and users of AI systems, even if they are located outside of the EU, as long as the systems or their outputs are used in the EU. In this regard, although the draft AI act

139 Brian Fishman. (2019). Crossroads: Counter-Terrorism and the Internet. Texas National Security Review, Vol 2 (2), 88.

140 Stefan Feuerriegel, Mateusz Dolata & Gerhard Schwabe. (2020). Fair AI, Business Information Systems Engineering, 62, 379–384.

141 UNOCT-UNCCT & UNICRI. (Jun. 2021). Algorithms and Terrorism: The Malicious Use of AI for Terrorist Purposes.

142 European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, COM/2021/206 final. Accessible at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>.

is a proposal of the EU, it is of merit internationally for all actors exploring the use of AI.<sup>143</sup> An important takeaway from the substance of the draft is its human rights-centric approach to classifying AI systems in categories according to their “unacceptable”, “high” or “low risk” to human rights.<sup>144</sup> This categorization is behind the draft’s general prohibition of the use of live facial recognition in public spaces by law enforcement.<sup>145</sup> Notwithstanding this general prohibition, “the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack” it is one of the exceptions to the prohibition of the use of live facial recognition technology in public spaces in law enforcement.<sup>146</sup>

#### e. Public-private relations

A further potential challenge for law enforcement and counter-terrorism agencies seeking to explore the development of AI capabilities concerns the role of the private sector in countering terrorist use of the Internet and social media – although it is pertinent to note that this is perhaps an issue that extends beyond the specific domain of AI.

With the ever-increasing relevance of online communication, including in terms of counter-terrorism, the weight of the private sector is increasing immensely due to extent of both their resources and competencies vis-à-vis public actors. As a result, the public sector often turns to private entities to complement and support their efforts to counter terrorism online. This is the case not only because any terrorist content is hosted on platforms or servers operated by these private entities but also because private sector entities possess skills and capacities to intervene online that law enforcement and counter-terrorism agencies in many ways simply lack. The private sector furthermore serves as a gatekeeper, holding access to data that is needed to train and feed into AI algorithms developed by law enforcement and counter-terrorism agencies. The private sector also tends to have a higher capacity than the public sector to design and develop innovative technological tools needed to combat terrorism online and, thus, law enforcement and counter-terrorism agencies may need to partner with and/or procure tools from the private sector.

Relations between the public and private sector in the context of countering terrorism online are not straightforward due in part to their differing nature and objectives. For example, content that might prompt action by law enforcement has in the past been taken down and deleted by private companies, thereby destroying important evidence for investigators.<sup>147</sup> On one particular occasion, social media companies even deleted evidence demonstrating war crimes that could have been used by law enforcement agencies.<sup>148</sup> While it can be argued that it is preferable that malicious content be taken down, such content is currently not archived in a way accessible for investigations, unless authorities present a subpoena, court order or warrant to the companies responsible for the platforms. It can thus be unclear for law enforcement agencies whether they may be missing relevant evidence as they are not aware if such evidence may exist or, if they do, for how long the data is stored before it is deleted – something that differs from company to company.

A further issue in this context is the technical expertise required to develop, develop and maintain AI applications, which is undeniably more concentrated in the private sector and, for that matter, particularly concentrated in the hands of the large tech firms that have traditionally excelled at luring the students and faculty from universities.

In this regard, law enforcement and counter-terrorism agencies will need to contend and build improved relations with the private sector if they are to fully understand the use of the Internet and social media by terrorist groups and individuals or to access to AI tools or develop or operate their own AI tools in-house. It should however be recognized that agencies in some countries in South Asia and South-East Asia may possess limited bargaining power vis-à-vis big technology companies.

---

143 Exceptions exist under Article 2, which include AI systems developed or used exclusively for military purposes, authorities from third countries and international organizations.

144 Arts. 5-6.

145 Art. 5.

146 Ibid.

147 Stuart Macdonald, Sara G. Correia, & Amy-Louise Watkin. (2019). Regulating terrorist content on social media: Automation and the rule of law. *International Journal of Law in Context*, 15(2), 190. Accessible at <https://doi.org/10.1017/S1744552319000119>.

148 Belkies Wille. (Sep. 10, 2020). ‘Video unavailable’ Social Media Platforms Remove Evidence of War Crimes. Human Rights Watch News. Accessible at <https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes>.



## ii. Technical challenges in using AI tools

### a. False positives and false negatives

When optimizing an algorithms' accuracy, AI developers can change the threshold that sets the classification label into positive or negative. By adjusting this threshold, it is possible to control two types of errors: false positives and false negatives. False positives can be understood as instances where a positive result is incorrectly given, and false negatives can be understood as instances where a negative result is incorrectly given. As some margin of error in the output of algorithms is inevitable and it is not possible to simultaneously reduce both false positives and false negatives, there is a need to choose which correction should be prioritized. This choice is not straightforward, as both false negatives and false positives can have significant implications. In a predictive model that tries to identify terrorist actors (considered as the positive label), reducing false negatives implies increasing false positives, which translates into the acceptance of wrongly identifying someone as a terrorist. This minimizes the risk of potentially dangerous individuals passing through the algorithm unspotted but can indiscriminately burden civilians. On the other hand, optimizing for false positives increases false negatives. This approach prioritizes avoiding the incorrect identification of someone as a terrorist but has also more tolerance to relevant subjects escaping undetected. Accordingly, the precise classification of the threshold can have a significant impact on the results of the AI model deployed.



Photo by Brendan Church on Unsplash

### b. Bias in the data

Fairness in AI is a declared aim across wide parts of the tech industry, policy, academia, and civil society and one of the key principles touted in terms of the responsible use of AI. It implies that algorithmic decisions do not create a discriminatory or unjust impact on the end-users – as was the case with the aforementioned COMPAS recidivism algorithm.<sup>149</sup> An investigation by an independent newsroom – ProPublica – demonstrated that the algorithm used by judges and parole officers in the United States to assess a criminal defendant's likelihood to re-offend was biased against certain racial groups. The analysis of the tool found that African American defendants were more likely to be incorrectly judged to be at a higher risk of recidivism than Caucasian defendants, i.e. there were considerably more false positives among the African American community, while Caucasian defendants were more likely than African American defendants to be incorrectly flagged as low risk, meaning that the false-negative rate was recurrently higher in the Caucasian community creating a differential treatment for the two groups. Another high-profile example of algorithmic discrimination, this time relating to gender-based bias, came to light in 2018 when Amazon's AI automatic hiring tool was shuttered after it was seen to result in bias against women. Based on the male dominance within the tech company, Amazon's system taught itself that male candidates were preferable.<sup>150</sup>

149 Julia Angwin, et al. (2016). Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. ProPublica. Accessible at <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

150 Jeffrey Dastin. (Oct. 11, 2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Accessible at <https://www.reuters.com/article/us-amazon-com-jobs-automation-%20insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-%20idUSKCN1MK08G>.

Although a lofty aspiration, ensuring fairness in AI can be an exceptionally challenging matter. Undeniably, the technology itself is a neutral statistical and mathematical process, but AI can amplify existing biases of societies when trained with biased datasets, resulting in incidents of automated solutions discriminating against individuals, groups, and communities on prohibited grounds.<sup>151</sup>

In addition to the risk of discrimination stemming from training data that reflects inherent biases in society, it has also been suggested that a homogenous technology environment leads to blind spots when developing software, also inducing discriminatory effects. This has been behind the call for a more diverse representation in tech companies.<sup>152</sup>

Finally, it is also important to note that, whereas bias in datasets can be unintended, data can also be consciously manipulated.<sup>153</sup> Manipulated data can have the same impact as biased data being used to train a model in that it can reproduce similar harmful effects. A dataset that has been tampered with is hard to detect from the outside, especially for non-technical experts. In the realm of security, and especially counter-terrorism applications, this is a weak point that must be mitigated through strict regulation on the use and access to the system, as well as for regular monitoring of the system and its performance.

### c. Explainability and the black box problem

A further principle often underscored as essential for the responsible use of AI is the notion of explainability. Closely associated with the requirement of transparency, explainability focuses on ensuring that AI is not a so-called “black box” and that algorithmic decisions can be understood by end-users in non-technical terms.

Deep learning systems are often described as black boxes as they combine and re-combine attributes in many arbitrary ways. Once input is provided, the internal behavior that leads the system to the output may not be clear. In other words, humans cannot understand how deep learning systems produce results. This makes it difficult to trace and verify results, which can raise problems in terms of human rights, particularly the right to a fair trial and due process for an accused individual or individuals, as well as hamper accountability, which in turn can affect the right of an effective remedy of those who have had their rights unjustly restricted or violated through the deployment of AI systems. The notion of explainability, in this regard, requires that end-users are able to interpret the information extracted from the black box and understand what elements used in the machine learning model were responsible for each specific outcome.

The lack of explainability can only be mitigated with a profound technical understanding of the neural network and with the support of explainability tools for local and global decisions<sup>154</sup>, such as SHAP<sup>155</sup> or LIME.<sup>156</sup> Explainability tools can help explain and present in understandable terms the features in the data that were most important for the model and the effect of each feature on any particular output.<sup>157</sup>

### d. The complexity of human-generated content

As already stated, a prerequisite to the development of reliable AI tools is to have access to accurate and detailed labelled datasets. While many of the challenges presented in this report are global in nature, South Asia and South-East Asia presents particular additional considerations, given the regions’ great diversity in terms of languages and dia-

---

151 Tom Simonite. (Jul. 22, 2019). The best Algorithms Struggle to Recognize Black Faces Equally. Wired. Accessible at <https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/>.

152 Marie Hicks. (Oct. 12, 2018). Why techs gender problem is nothing new. The Guardian. Accessible at <https://www.theguardian.com/technology/2018/oct/11/tech-gender-problem-amazon-facebook-bias-women>.

153 UNOCT-UNCCT & UNICRI. (2021). Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes.

154 Andrew Burt. (Dec. 13, 2019). The AI Transparency Paradox. Harvard Business Review. Accessible at <https://hbr.org/2019/12/the-ai-transparency-paradox>.

155 See <https://shap.readthedocs.io/en/latest/index.html>

156 Marco Tulio. (Aug 9 2016). arXiv, “Why should I trust you?”: Explaining the Predictions of Any classifier. Accessible at <https://arxiv.org/abs/1602.04938>.

157 Harmanpreet Kaur, et al. (2020). Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. CHI 2020 Paper. Accessible at [http://www-personal.umich.edu/~harmank/Papers/CHI2020\\_Interpretability.pdf](http://www-personal.umich.edu/~harmank/Papers/CHI2020_Interpretability.pdf).



lects. By way of example, consider that, according to one estimate, there are currently approximately 2,300 languages spoken across Asia as a whole.<sup>158</sup> Although a great number of advancements have been made in NLP, the vast majority of efforts in this domain are focused on a handful of key languages: English, Chinese, Urdu, Farsi, Arabic, French, and Spanish – and, of these, English is heavily favoured. There have been few results with respect to other so-called “low resource” languages.



*Photo by Ryoji Iwata on Unsplash*

Furthermore, the languages spoken in South Asia and South-East Asia – like all languages in the world – are dynamic, constantly evolving and adapting to social developments. This natural dynamic also poses major challenges for AI since AI systems cannot adapt to changing contexts on their own, but rather require extensive amounts of data to inform adaptation. In that sense, data sets must be constantly monitored and updated as necessary. For many of the less prominent languages or dialects in the regions, the likely insufficiency of required training data can present a challenge.

An example of the practical challenges AI might encounter in this context was recently seen when algorithms designed to identify and block malicious content on Facebook were repeatedly deceived by a far-right extremist group in the United States that used an encrypted language to communicate online and openly share bomb-making guidance. Unfamiliar with the code used, the algorithms failed to detect the content.<sup>159</sup>

Context can add a further layer of complexity as it determines the finer nuances of human communication that result in irony and humour. Irony- or sarcasm-detection is a deceptively complex task, even for humans, as it varies from

---

158 Andy Kirkpatrick & Anthony J. Liddicoat. (April 2019). The Routledge International Handbook of Language Education Policy in Asia. Routledge. Accessible at <https://www.routledge.com/The-Routledge-International-Handbook-of-Language-Education-Policy-in-Asia/Kirkpatrick-Liddicoat/p/book/9781138955608>.

159 Tess Owen. (2020). The Boogaloo Bois are all over Facebook. Vice. Accessible at <https://www.vice.com/en/article/7kpm4x/the-boogaloo-bois-are-all-over-facebook>.

person to person and is highly dependent on culture,<sup>160</sup> as well as many other aspects such as facial expression and tone.<sup>161</sup> Within the NLP field of sentiment analysis, sarcasm is a specific area of study, as it cannot be classified as a negative or positive sentiment. Recent studies show that sarcasm alone can account for as much as a 50% drop in accuracy when automatically detecting sentiment.<sup>162</sup> Approaches to address this issue include adding layers of information to better capture the speaker-environment relationship such as the analysis of pragmatic features like emojis, hashtags and mentions;<sup>163</sup> harnessing context incongruity when certain features show that the context is not in agreement with the content of the text;<sup>164</sup> or utilizing user embeddings that encode stylometry and personality of the users.<sup>165</sup>

In May 2021, Twitter launched a new hate-speech warning feature that asks people to review replies with potentially harmful or offensive language. Early tests have shown that the algorithms “struggled to capture the nuance in many conversations and often didn’t differentiate between potentially offensive language, sarcasm, and friendly banter”.<sup>166</sup> Improvements in the prompting system included adding user embeddings such as the relationship of the user and the replier and the origin of the user.<sup>167</sup> An example from Sri Lanka also demonstrated how Facebook’s algorithms were not able to assess multi-layered cultural context. User-generated posts on social media prior to the Easter bombings in Colombo in April 2019 slipped through the monitoring systems because algorithms were not capable of identifying hate speech in the posts due to a complex cultural environment.<sup>168</sup> Although the technology is always evolving, this technical limitation can become a significant stumbling block for law enforcement and counter-terrorism agencies seeking to use AI to identify relevant information online.

- 
- 160 Aditya Joshi. (Aug. 11, 2016). How do Cultural Differences Impact the Quality of Sarcasm Annotation? A Case study of Indian Annotators and American Text. Proceedings of the 10<sup>th</sup> SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities, 95-99. Accessible at <https://www.aclweb.org/anthology/W16-2111/>.
- 161 Kevin Zawacki. (Jan. 22, 2015). Why can’t robots understand sarcasm?. The Atlantic. Accessible at <https://www.theatlantic.com/technology/archive/2015/01/why-cant-robots-understand-sarcasm/384714/>.
- 162 Martin Sykora, Suzanne Elysa & Thomas W. Jackson. (2020). A qualitative analysis of sarcasm, irony and related #hashtags on Twitter. Big Data & Society, July-December, 1-15. Accessible at <https://journals.sagepub.com/doi/pdf/10.1177/2053951720972735>.
- 163 Ane Berasategi. (2020). Sarcasm detection with NLP. towards data science. Accessible at <https://towardsdatascience.com/sarcasm-detection-with-nlp-cbff1723f69a>.
- 164 Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya. (2015). Harnessing Context Incongruity for Sarcasm Detection. Proceedings of the 53<sup>rd</sup> Annual Meeting of the Association for Computational Linguistics and 7<sup>th</sup> International Joint Conference on Natural Language Processing (Short Papers), 757-762. Accessible at <https://www.aclweb.org/anthology/P15-2124.pdf>.
- 165 Devamanyu Hazarika, et al. (2018). arXiv, CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. Cornell University. Accessible at <https://arxiv.org/abs/1805.06413>.
- 166 Anita Butler & Alberto Parrella. (May 5, 2021). Tweeting with consideration. Twitter Blog. Accessible at [https://blog.twitter.com/en\\_us/topics/product/2021/tweeting-with-consideration.html](https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration.html).
- 167 Ibid.
- 168 Yudhanjaya Wijeratne. (Apr. 25, 2019). The Social Media Block isn’t helping Sri Lanka, Slate. Accessible at: <https://slate.com/technology/2019/04/sri-lanka-social-media-block-disinformation.html> and Yudhanjaya Wijeratne. (May 7, 2019). Big Tech is as Monolingual as Americans. Foreign Policy. Accessible at <https://foreignpolicy.com/2019/05/07/big-tech-is-as-monolingual-as-americans/>.



## V. MOVING FORWARD WITH AI

The use of AI holds considerable potential to contribute to countering terrorism online. AI can be particularly helpful in supporting law enforcement in dealing with the large amounts of collected data by classifying new entries, extracting patterns, flagging significant information, developments or relationships, and visualizing results. It can aid in finding patterns and relations that may otherwise stay unrecognized and can, thus, greatly benefit law enforcement and counter-terrorism agencies to turn the tide in the fight against terrorism online. However, there are numerous political, legal, and technical challenges that exist, and it is paramount for law enforcement and counter-terrorism agencies in South Asia and South-East Asia to acknowledge these challenges. Considerable prudence in this regard is required of any agency exploring or considering moving forward with the application of AI.



*Photo by Sven Mieke on Unsplash*

In view of this and recognizing that continued technological progress in the field of AI is to be expected, the following recommendations are provided for law enforcement and counter-terrorism agencies in South Asia and South-East Asia. The first set of these recommendations is region-specific, while the remaining five are universal in character and thus may be of equal relevance for law enforcement or counter-terrorism agencies outside of South Asia and South-East Asia seeking to explore the use of AI.

These recommendations were compiled and categorized based on feedback collected from participants at the UN-OCT-UNICRI Expert Group Meeting. The order of recommendations should not be interpreted as indicating any particular priority.



### **Build a regional approach:**

- Recognizing that information on current AI capabilities in South Asia and South-East Asia is limited or not publicly available, further field research should be conducted to understand current capabilities and inform the shaping of tailored practical guidance on the use of AI to counter terrorism online in the regions.
- Efforts by national authorities should be invested into cataloguing existing knowledge in this field at the national level and throughout the regions and in identifying national experts in their countries or within the regions.
- Efforts by national authorities should be invested into developing databases of lawfully collected region-specific data to aid in the development of specific tools for South Asia and South-East Asia and such databases should be made available to researchers.
- A forum for the exchange of ideas, experiences, and challenges should be established at a regional level to further developing a regional corpus of knowledge and expertise.
- Dialogue and cooperation with private sector entities, in particular social media platforms, should be fostered to aid in facilitating access to relevant training data from South Asia and South-East Asia.
- The development of applications should cater for the multicultural and multilinguistic reality of South Asia and South-East Asia, which distinguishes these regions from others in many respects.



### **Acknowledge the limitations of AI:**

- Law enforcement and counter-terrorism agencies should acknowledge that:
  - ▶ AI is not perfect and may never be. There still exist many technical challenges and, while AI is getting more effective and accurate, it may take a considerable amount of time to reach levels required by both the end-users and by civil society – if this can be achieved at all.
  - ▶ There is no “one size fits all” solution. AI tools must be tailored to specific contexts and requirements.
  - ▶ AI presents results as probabilistic calculations and not infallible predictions.
  - ▶ AI cannot unilaterally eliminate terrorism and violent extremism online. Investments into AI should be flanked by efforts to prevent radicalization and violent extremism at its roots.



### **Respect human rights and protect civil society:**

- The potential human rights impact of the use of AI to counter terrorism online should be acknowledged. Human rights impact assessments should be carried out in advance of deploying any tools and the legality, proportionality and necessity of their use should be assessed on a regular and ongoing basis throughout the entire lifecycle of the AI tool.
- National laws and policies should be put in place, defining, and governing the use of AI tools to counter terrorism online to, inter alia, address the risk of the misuse of such tools, for instance, to persecute legitimate social movements, civil society organizations and journalists, as well as to minimize the risk of function creep.
- A multi-stakeholder approach with regular feedback loops should be adopted with respect to the use of AI to counter terrorism online to ensure that any possible blind spots in the design, development and application of automated tools are identified.







### **Establish technical safeguards:**

- Decisions based on AI-enabled technology should be explainable. Human rights compliance and explainability tools should be used to help overcome black box concerns and ensure accountability for any actions or decisions based on AI systems.
- Datasets should be as free of bias and as transparent as possible for the purposes of external evaluation.
- The performance of algorithms should be evaluated, and datasets should be reviewed to ensure they are accurate, representative, and regularly updated. Sensitive datasets should be evaluated by trusted independent third parties, in accordance with international and national laws and policies.



### **Ensure human oversight and accountability:**

- Thorough consideration should be given to how AI-enabled technology can be used to augment human analysts as opposed to replacing them or their functions.
- Automated decision-making should not be encouraged. Analysis and conclusions reached on the basis of AI systems should always be reviewed by a trained human, who should make the final determination.
- All personnel working with AI-based systems should understand the system and training should be provided to personnel on technical and legal aspects of the use of the tools, as well as possible associated human rights risks.
- Oversight mechanisms should be established and be equipped with sufficient technical expertise to be able to understand and appropriately assess the use of AI. These oversight mechanisms should regularly assess each instance of the use of AI systems, the legality, proportionality and necessity of their use, and the quality of the outcomes they produce. Institutional checks and balances should be guided by transparent policies.



### **Develop knowledge and build capacities:**

- Law enforcement and counter-terrorism agencies should develop their own in-house knowledge and technical capacities insofar as possible to enable them to exert full control over their system and make adjustments as needed.
- Law enforcement and counter-terrorism agencies should engage closely with the AI industry and research community to build their knowledge and understanding.
- The sharing of experiences and challenges between law enforcement and counter-terrorism agencies using AI for countering terrorism online should be encouraged to foster a community of practice founded on responsible use of the technology.
- The legal requirements surrounding individual use cases of AI to counter terrorism online should be analyzed, especially from the perspective of the impact of AI on human rights.

