Pericles

Policy recommendation and improved communication tools for law enforcement and security agencies preventing violent radicalization

# Ethical review of PERICLES deliverables

Dr. John Guelke

Result Report

Coordinator:

**Prof. Dr. Dominic Kudlacek**
University of Applied Sciences Bremerhaven
An der Karlstadt 8, 27568 Bremerhaven, Germany
Mail: dominic.kudlacek@rub.de

This project has received funding from the *European Union's Horizon 2020 research and innovation programme under* grant agreement No 740773

Document Evolution:

| **Version** | **Date** | **Note of Modification** |
|---|---|---|
| V1.1 | 10.04.2020 | First version of the report |

# TABLE OF CONTENTS

# 1 INTRODUCTION

The following comprises an ethical review of the most recent eight deliverables. Five out of these eight deliverables are deliverables describing the current state of play of the tools outputted by the Pericles project. There are also three deliverables describing modelling of implementation of the tools by Thales, validation workshops with end users, and a policy recommendations document.

## 1.1 ETHICAL FRAMEWORK

The ethical framework referred to in this review is that developed in Pericles deliverables D6.1 'Ethical Considerations in Counter Radicalisation' by Kat Hadjimatheou, and D6.2 'Ethical Case Study' by John Guelke. This framework is drawn from mainstream liberal political theory, but 'are also reflected in common values, laws, and codes of ethics or of conduct of counter-radicalisation professionals'.[1]

This framework emphasises the importance of individuals being able to live autonomous lives, with a strong presumption that these should not be interfered with unless there are weighty and compelling reasons to do so. Respecting autonomous lives means respecting rights to associate, speak, travel and enquire freely. Police surveillance of individuals arguably interferes with this autonomy. Where there is specific evidence to suggest that an individual is involved in plotting violence, such an interference can be justified relatively straightforwardly. However, much counter terrorism practice involves much more contentious scenarios, where any evidence in involvement in violence is much more partial or speculative. Counter radicalisation can be treated as a species of counter terrorism,[2] and the same issue applies: namely that the more tenuous the link to actual violent behaviour, the weaker the justification the cause of counter radicalisation can claim. And the more intrusive measures are, the more that they interfere with autonomous lives, the stronger the justification needed. Indeed, as Hadjimatheou writes: "if it were possible to distinguish reliably between the violent and non-violent amongst those who engage in otherwise legal radical behaviour, such as radicalised speech online, it

[1] Hadjimatheou, 2017, 10
[2] Hadjimatheou, 2017, 9

is unlikely that counter-radicalisation intervention with respect to the latter would be justified".[3]

Counter radicalisation as an activity is particularly liable to inflicting stigma. Hadjimatheou explains the concept of stigma thus:

> "Stigmatisation can be defined as the process of marking a person or a group out as having an undesirable characteristic. In the context of counter-terrorism and counter-radicalisation, the undesirable characteristic in question is likely to be a proclivity to indiscriminate violence, or affinity with political views which much of society deem repugnant. Counter-radicalisation interventions may lead to individuals being stigmatised as being vulnerable to manipulation by nefarious persuasive others or as being a threat to society, or both."[4]

Measures that are stigmatising often effect far beyond individuals against whom there is any particular evidence of involvement in violence. Furthermore, this risk of stigmatising effect is likely to fall upon innocent people in specific communities, raising the prospects of outcomes that are discriminatory as well as unjust:

> "When radical views overlap with or track culture, religion, or community membership, counter-radicalisation measures can risk stigmatising people beyond those they are aimed at. These issues arise most strongly in relation to counter-radicalisation measures that involve interference with liberties, such as police surveillance and monitoring. But they also arise in relation to practices of other agencies, such as schools and social services."[5]

Erroneously casting suspicion on the innocent is a problem common to counter terrorism measures widely, and historically. Adam Roberts in his (1989) writes of counter-terrorism that "[i]ts main problems arise from the fact that it involves trying to combat clandestine fighters, who may cause the most appalling carnage, but who hide among the rest of the population and are very difficult to track down. This creates a situation where there is often a strong public desire for retribution, but the proper target for such retribution is not available."[6] On the one hand this provides a pressure towards identification of suspects on weak evidence, and a pressure towards intrusions into individual privacy, as that evidence which is to be

[3] Hadjimatheou, 2017, 13
[4] Hadjimatheou, 2017, 21
[5] Hadjimatheou, 2017, 12
[6] Adam Roberts, 1989, 60

found – plots with others, weapons or explosive materials – are likely to be hidden in spaces normatively treated as private, and closed to public scrutiny.[7]

## 1.2 D3.2 Report on Stakeholders Disengagement Concepts Implementation Modelling

Deliverable D3.2 concerns an approach to modelling of implementation of PERICLES work.[8] This work is aimed at researchers and policy makers. While such modelling can be more or less accurate or useful, any responsibility for use made of such modelling could be considered to lie with those individual researcher/policy makers. As such, this deliverable raises no ethical issues.

## 1.3 D4.1 Modelling and Classifying Radical Content on Twitter (MODERAD)

Deliverable D4.1 outlines the cyberdetection tool, MODERAD – Modelling and Classifying Radical Content on Twitter. This deliverable begins by outlining the distinctive approach of the technology:

> "The main methodological novelty of this tool, compared to other semantic or syntactic detectors, is the detection of radical contents by combinating lexico-pragmatical analyses with certain (cyber-) environmental factors related to the publication of these radical contents. Therefore, it is not an automatic search methodology but rather supervised automation; its main use is to give information to police departments or justice systems of certain radical content, and the quantitative or qualitative features."[9]

The ModeRad is a developing system for modelling the likelihood that a particular tweet contains extremist content. An algorithm has been trained to distinguish radical and non radical contents:

> "When this tweet processing is finished, ModeRad will show a list of users sorted from highest to lowest according to the percentage of radical tweets published by each one. In this window there

[7] On privacy and counter terrorism see Guelke and Sorell, 2010 and on privacy more widely 2017.

[8] "In this report, we first present the methodology used for concept modelling. Next we present some ontology views example understandable by human but also usable by software agents for future uses." (Lortal, Malfreyt and Faure, 2019, 6)

[9] Dolado et al, 2019, abstract

> are two buttons, one that allows you to export the tweets collected and the list of users obtained, and another that generates a graphics report in .pdf format."[10]

In addition to this algorithm predicting which subsequent content will be radical and which not, the system enables further filtering by additional criteria: keywords, hashtags, URLs, language, date and geolocation.[11]

This system does raise ethical risks. Specifically those of stigma, error and discrimination.

The tool categorises tweets as suspected extreme content, inevitably risking stigma. Categorising media in this way is a legitimate task for both law enforcement and social media platforms themselves, both of whom will have distinct obligations to protect users from the worst examples of hate speech. Police also may be able to identify extreme accounts with individuals responsible for real world violence. Although the mere fact of categorisation is stigmatising, these assessments are not published, but are rather a filtering tool for police or companies to make further assessments themselves.

Categorisations generated by algorithms will inevitable make some erroneous categorisations, mistakenly ascribing extremist intentions to posts where this intention was not present.[12] Similarly, searching by keywords, hashtags, URLs, language, date and geolocation although a legitimate way of narrowing down the search for genuinely extreme content, can also be a source of error. People can use seemingly incriminating keywords, hashtags or URLs for innocuous reasons. And the mistaken ascriptions of content as extreme may easily track particular communities, leading to mistaken ascriptions of content as extreme being more likely in Twitter activity among certain groups.

## 1.4 D4.2 Enhanced Platform

D4.2 presents the Enhanced Platform, a public facing site explaining and presenting the Pericles Tools. It also controls access to the other tools.

The presentation of the tools for a public audience raises no ethical issues. The information is accurate and clear, and distinguishes between the different levels of development of what has been outputted to date:

[10] Dolado et al, 2019, 27
[11] Dolado et al, 2019, 13
[12] On this see also account of D5.1 below

A list of recommended tools and/or resources based on the combination of the objective and approach selected. This list clearly differentiates different tool categories and whether they are:

a) A PERICLES project tool;

b) An available tool;

c) An existing tool that is available, however, via request from the source organisation; and

d) A PERICLES research based desired tool or resource, which is a tool that does not yet exist, but the research carried out throughout the project identified it as a desirable asset.[13]

The only significant ethical risk relating to the Enhanced Platform relates to its role in gaining access to the other tools. However, none of this risk is technological. Access to the sensitive tools – MODERAD and MAVAST – is a matter of emailing a relevant contact.[14]

## 1.5 D4.3 Multi-Agency Vulnerability Assessment Support Tool (MAVAST)

D4.3 describes and presents the Multi-agency Vulnerability Assessment Support Tool, MAVAST.

This tool facilitates collaborative assessments of individuals who may be vulnerable to radicalisation themselves, or already themselves violent based on an extremist ideology. The tool elicits judgments from individuals across different agencies – police, social workers and teachers, for example – to enable them to share information about a person of interest. Specifically, it elicits judgments on observable characteristics that are indicators of a person's propensity to involvement in violent extremism:

> "In specific combinations these observable indicators are considered to be indicative for crossing the threshold between tolerated (radical) behaviours and illegal extremist (terrorist) acts. Observable indicators are generally combinations of behaviours dealing with amongst other things: changes in habits, appearance, group affiliations, capabilities, disclosures, transgressions, pathology, and traumatic past." [15]

[13] Sullivan et al, 2019, 8
[14] Sullivan et al, 2019, 19, 20
[15] Van Hemert et al, 2019, 6

The tool is aimed at actors who might be in a position to prevent a slide from vulnerability to extreme ideology and involvement in violence, rather than 'deradicalising' an individual already so involved.[16] The kinds of interventions envisioned are in the category of 'safeguarding and/ or cross-sectoral mitigating actions'.[17] Interventions relevant to someone who is already a violent extremist, is treated separately as a matter for the enforcement of the law. The categorisations, and possible actions – tagging – possible on the MAVAST system, are thus limited:

> Once the individual has been identified for further scrutiny by the appropriate authorities an appraisal process takes place to assess whether or not further actions are required, and if so, what type of actions. There are basically four potential outcomes of this appraisal process:
>
> 1. The individual is *not vulnerable*
>
> 2. The individual is *possibly vulnerable*
>
> 3. The individual *is vulnerable*
>
> 4. The individual *is violent based on extreme ideology*
>
> …
>
> Congruent with the four potential outcomes of the appraisal process, four type of actions can be advised:
>
> 1. Not vulnerable: Identification tag should be removed
>
> 2. Possibly vulnerable: More information should be collected
>
> 3. Vulnerable: Mitigation actions (interventions) should be put in place
>
> 4. Violent: Legal action is advised[18]

This tool raises a number of ethical issues. Specifically to do with stigma, error, discrimination and privacy. The ethical issues involved are acknowledged in the deliverable right from the outset.[19] For example in relation to privacy the authors write:

[16] Van Hemert et al, 2019, 9
[17] Van Hemert et al, 2019, 6
[18] Van Hemert et al, 2019, 10
[19] For discussion see particularly van Hemert et al, 2019, 11, 35, 36.

> "The EU actively supports and celebrates the Universal Declaration of Human Rights; within this context faulty stigmatization of individuals as potentially extreme violent is challenging, to say the least."[20]

Like with the MODERAD, the MAVAST deals in ascriptions of extremism. However, unlike the MODERAD MAVAST ascribes extremism, or vulnerability to extremism, to individuals rather than mere content. Arguably this is even more stigmatising, as it is a more personal judgment. However, the categorisation is not arbitrary, and it is not public. Teachers, social workers and police are all agencies to whom personal assessments of vulnerability on this point are highly relevant.

The precise relation between the observable characteristics identified and violent extremism is likely to remain an ongoing matter of research. There is no law like relation between any set of observable characteristic and involvement in violence – all good law enforcement, teaching and social work professionals know this – but this does mean that relying on observable characteristics will inevitably sometimes result in mistaken assessments. Again, good professionals know this, and treat such assessments tentatively, but it does count as a risk inherent to technologies of this kind. Furthermore, mistaken judgments of individuals as adhering to an extreme ideology may easily track particular communities, leading to mistaken ascriptions of individuals as extreme being more likely among certain groups. Again, this risk is one that good professionals should already be alive to.

Finally, this kind of technology inevitably raises privacy issues. Assessment are made of individuals across a range of varyingly intimate social relations. Inevitably sharing these for investigative purposes represents a cost to their privacy, as does having their behaviour assessed for evidence of vulnerability or susceptibility to violent behaviour. Furthermore, the sharing of these assessments digitally is arguably more intrusive than the mere fact of scrutiny and assessment, as it creates the (manageable) risk of such assessments being shared beyond those who are meant to see them, highlighting the importance of information security as this technology is implemented.[21]

[20] Van Hemert et al, 2019, 6
[21] On this see also the account of D5.1 below

## 1.6 D4.4 The Family Information Portal

D4.4 presents the Family Information Portal:

> "The Family Information Portal provides free, accessible and comprehensive information for families who are affected by radicalisation; whether they simply have some concerns or whether radicalisation is confirmed as the cause of the problems within the family. It deals with an entire spectrum of scenarios, from a family who is looking for basic information about whether a member might be radicalising, all the way through to a situation in which a family member is missing or killed in a foreign conflict zone.
>
> The Family Information Portal is intended for families affected by Islamist, Separatist, extreme Right-Wing and extreme Left-Wing terrorism.
>
> Alongside the information for families, we have also developed Guidelines for Law Enforcement Agencies who may be confronted with families who have been affected by radicalisation."[22]

As a set of fixed pages with information for families to make use of or not at their own discretion, the Family Information Portal raises no technology based ethical issues. However, information like this can raise non technology based ethical issues, primarily because families making use of this information may well be vulnerable themselves. If it presents inaccurate information or makes bad recommendations, particularly affecting other people. As such, the nuance and caution shown are appropriate – for example making clear that the indicators listed are a reason 'to be alert', rather than to conclude someone is involved in violence.[23]

The advice offered for what to do if one thinks a family member is radicalising is similarly measured. A wide range of appropriate measures are discussed:

- Try to find out what keeps your child busy

- Know your child's friend and gain information about your child's friend

- Keep lines of communication open, listen to your child and talk to them about their interests

[22] Rooze, Chakari and Young, 2020, abstract
[23] Rooze, Chakari and Young, 2020, 14

- Encourage them to participate in positive activities with local groups you trust

- Allow and encourage discussion on different topics

- Avoid confrontational discussions

- Be aware of your child's online activity and update your own knowledge

- Provide alternatives, for example: it is not necessary to travel to conflict areas in or-der to help poor and needy people. It is also possible to help them by joining a local NGO

- Contact the school

- Help your children to be critically aware of what they see on TV and internet

- Involve organisations which have experience and expertise on this field

- Ask for help from authorities.[24]

Each measure may be appropriate or effective in context. Arguably the advice to be aware of a child's online activity could be overly intrusive, but a child's claim to keep their online activity private from their parents could only ever be partial, and if a parent has reason to worry about their child being influenced by violent extremism this is more than a sufficient reason to bring this under greater scrutiny. In any case this is something that will be down to a parent's discretion in any case.

## 1.7 D4.5 Skills and Competencies Training (SCT)

D4.5 outlines the Skills and Competencies Training tool:

> "The Skills and Competencies Training (SCT) is a training curriculum aimed at supporting prevention and intervention efforts that deal with processes of potentially violent radicalisation and with incidents or indications of politically or religiously motivated violence. The focus of the training is on selective or secondary prevention of violence with emphasis on the role and function of ideology (e.g. to motivate, to justify, to normalise, to establish identity). Phenomena as well as measures are addressed; thus, the training takes both declarative knowledge ("what") and procedural knowledge ("how to") into account. It emphasizes a multi-

[24] Rooze, Chakari and Young, 2020, 14-15

> agency approach and stresses the relevance of an individual's social context."[25]

The SCT raises no technology based ethical issues. It also raises no other ethical issues, as it is addressed to a critical and literate audience not solely dependent on it as a source of information.

## 1.8 D5.1 End User Workshop Report

D5.1 reports on the End User Workshops taking place throughout the project.[26] The workshops themselves raise no ethical issues, and neither does their publication, given the anonymised matter in which all feedback is reported. The workshops did highlight a number of useful ethical issues with the tools.

First, in relation to MODERAD, the inherent difficulty of distinguishing the intent behind the uses of certain keywords was highlighted – it is difficult enough at times for a human to detect ironic or satirical repetitions of keywords or hashtags:

> "Ethical issues were touched on too, even in the development phase as tweets from unaware authors are used to inform software development. Developers and users have to keep in mind the problem of false positives. A specific example is that software tends to a literal reading and cannot easily recognise irony or satire, and discern it from serious statements. When a significant number of false positives are found, there is a risk of eating up valuable analysis time."[27]

In relation to the MAVAST, the importance of data security was highlighted, albeit as a problem for implementation:

> "Another very substantial barrier to implementation discussed in this regard is in data protection and the personal rights of the individuals discussed. Two limitations were seen: Privacy regulation makes it impossible in some cases to discuss a certain individual, and the use of classified or privileged information is not always possible (e.g. in case of medical information). Suggestions were made to integrate different access levels to the tool,

[25] Görgen, Garbert, and Wagner, 2019, 6
[26] Wagner et al, 2020
[27] Wagner et al, 2020, 8

> to run the tool only on local computers, and store data only on secure and specially approved databases."[28]

Furthermore, it may be desirable with the MAVAST specifically to enable the monitoring of relevant observable factors over time:

> "Participants saw additional potential of the tool if it could be used to document and supervise changes over time. This also connects to the assessment that often information collection in unclear cases is a continuous process."[29]

However, it should be noted that monitoring a person over a longer period of time represents a greater intrusion into their privacy, requiring weightier reasons for continuing the monitoring.

Finally end users noted possible difficulties, particularly relevant to MAVAST, with cooperation across jurisdictions with very different cultures and practices:

> "Some challenges regarding the applicability of tools with a focus on multi-professional and especially inter-agency cooperation became clear. Differences in possibilities for implementation of the tools between countries also were apparent. One example spotlighted in the workshops are differences between the Dutch police with their tradition of community-oriented and multi-agency work and the Greek police where this is yet quite uncommon."[30]

## 1.9 Final Policy Recommendations Document

An additional document has been produced identifying policy recommendations following from Pericles research throughout the project. Again, the production of policy recommendations does not raise ethical issues as such, but certain recommendations underline ethical issues identified particularly in relation to the MAVAST tool: namely the moral importance of more empirically grounded work on the processes by which individuals come to assume a violence justifying, extreme ideology, and ongoing research on the reliability of tools used in counter radicalisation:

> One key issue apparent within the field of radicalisation research is the lack of proper scientific evaluations being performed on psychometric scales and assessments tools. To a similar extent,

[28] Wagner et al, 2020, 10
[29] Wagner et al, 2020, 10
[30] Wagner et al, 2020, 14

> a large portion of the knowledge concerning path-ways to radicalisation, models and risk factors associated with radicalisation are based more on plausibility than empirical evidence. [31]

And:

> It is recommended that member states finance more projects and initiatives that evaluate the effectiveness of specific counter-radicalisation tools, especially identification tools, in order to improve its uptake by relevant practitioners.[32]

A tool like MAVAST depends for its accuracy on the observable characteristics genuinely being indicative of susceptibility to violent extremist ideology. Progress in counter radicalisation depends on being responsive to progress in social science research on this topic.

[31] Policy Recommendations Document, 2020, 4
[32] Policy Recommendations Document, 2020, 7

# 2 REFERENCES

Dolado, Alejandro Rabasa, Miriam Esteve Campello, Fernando Miro Llinares, Francisco J. Castro Toledo, Asier Moneva Pardo. 2019. ModeRad 'Modelling and Classifying Radical Content on Twitter' Pericles Deliverable D4.1

Görgen, Thomas, Matthias Garbert, Daniel Wagner. 2019. 'Skills and Competencies Training. A modular and adaptable curriculum for the further education of professionals dealing with violence prevention connected to extremist ideologie' Pericles Deliverable D4.5

Guelke, John. 2019. 'Ethical Case Study' Pericles Deliverable D6.2

Hadjimatheou, Katerina. 2017. 'Ethical Considerations in Counter-Radicalisation' Pericles Deliverable D6.1

van Hemert, Dianne, Tony van Vliet, Bob van der Vecht, Jacomien de Jong, Rinze Bruining, Ward Venrooij, Bas Keijser, Helma van den Berg and Mirjam Huis in 't Veld. 2019. 'MAVAST Multi Agency Vulnerability Assessment Support Tool' Pericles Deliverable D4.3

Lortal, Gaelle, Gilles Malfreyt and David Faure. 2019. 'Report on Stakeholders Disengagement Concepts Implementation Modeling' Pericles Deliverable D3.2

O'Sullivan, Ciarán, Sheryl Lynch, Akshay Chiddarwar and Stephen M. Purcell. 2019. 'Enhanced Platform' Pericles Deliverable D4.2

Roberts, Adam. 1989. 'Ethics, Terrorism and Counter-Terrorism' in Terrorism and Political Violence. Vol. 1. no. 1

Rooze, Magda, Ali Chakari and Holly Young. 2020. 'The Family Information Portal

With Guidelines for LEAs Dealing with Families Affected by Radicalisation' Pericles Deliverable D4.4

Wagner, Daniel, Thomas Görgen, Bas Keijser, Mirjam Huis in 't Veld, Dianne van Hemert. 2020. 'End User Workshop Report' Pericles Deliverable D5.1