Pericles

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

Ethical Considerations in Counter-Radicalisation

Katerina Hadjimatheou

Result Report

Pericles

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

Ethical Considerations in Counter-Radicalisation

KRIMINOLOGISCHES
FORSCHUNGSINSTITUT
NIEDERSACHSEN E.V.

This report identifies potential ethical risks and benefits of the counter-radicalisation tools developed by the PERICLES project and considers how these might be addressed during the development of those tools.

The ethical framework adopted in this report is that of European liberal democracy. The ethical implications of counter-radicalisation and of the tools in particular are discussed with reference to the values endorsed by that framework. The report considers both cross-cutting and overarching ethical implications and issues that arise distinctly with respect to specific tools.

At the time of writing, the PERICLES tools are still at very early stages of development and as such could take a range of different forms. With this in mind, the report explores issues as they might arise in relation to the different potential forms the tools may take. While this means the report is necessarily speculative, recommendations and proposals for the design and implementation of tools are made wherever possible.

Future reports in this work package will explore ethical issues identified here in more depth as they relate to the eventual design and potential application of the tools

Authors: Katerina Hadjimatheou

Coordinator:

Dr. Dominic Kudlacek
Criminological Research Institute of Lower Saxony
Lützerodestraße 9, 30161 Hannover, Germany
Mail: Dominic.Kudlacek@kfn.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	8/12/2017	First version of the report
V1.2		Second version of the report

Proposal for citation:

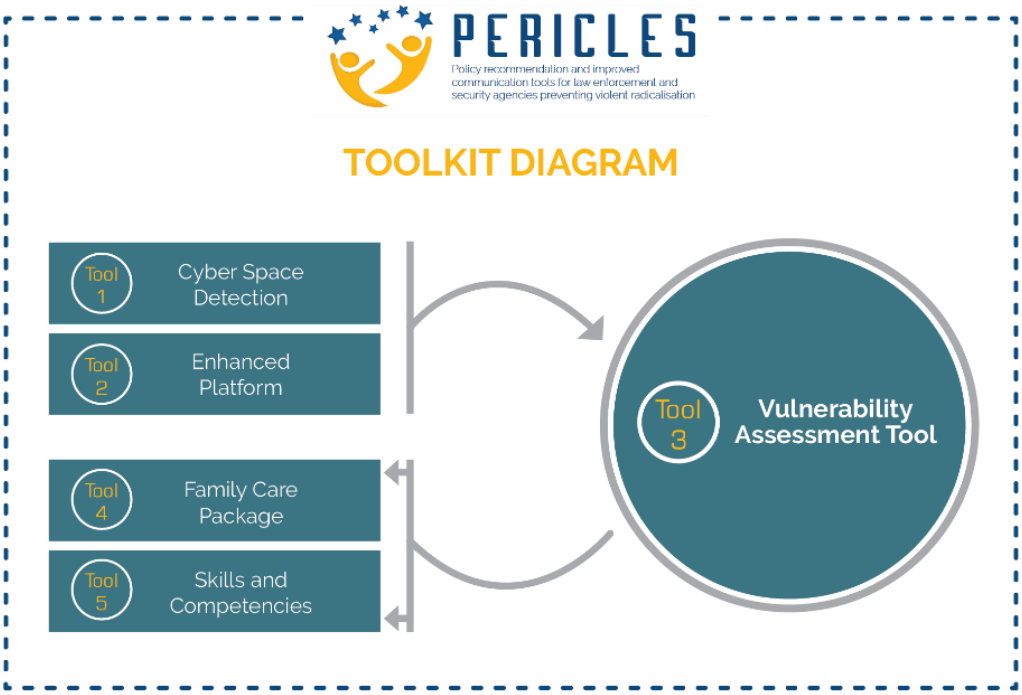
Hadjimatheou, K. (2017): Ethical Considerations in Counter-Radicalisation, PERICLES project.

TABLE OF CONTENTS

1 INTRODUCTION

This report discusses the potential ethical risks and benefits of the 5 counter-radicalisation tools developed by the PERICLES project. In brief, these tools are:

- **Cyber-space detection system.** Pericles will provide an updated cyber-space detection system based on an analysis of metadata and violent and online-radicalised communication. Social networks will be studied, with a focus on Twitter, which is one of the most popular open channels of dissemination of radical propaganda.
- **Enhanced platform of exchange.** An enhanced platform will be generated that provides end users with a more efficient interface for exchanging information and examples of best practices of strategies and tools aimed at preventing radicalisation.
- **Figure 3** – Pericles toolkit



- **Vulnerability assessment tool.** An assessment tool will be developed that combines a variety of indicators from 'at-risk' individuals and groups, to create a vulnerability assessment along with recommended actions. Indicators include behaviour, religion, family circumstances, personal factors and social networks.
- **Family care package.** Tailored material for family members of those at risk of radicalisation will be produced. Families will be provided with advice and support on how they

can detect signs of radicalisation, how to intervene during the earliest stages, and the best course of action to take.

- **Updated skills and competencies package.** An updated counter-radicalisation training course will be developed for frontline staff in order to increase their knowledge, awareness and understanding of radicalisation for better preparedness. These interactive courses will involve the latest tools and resources to cover crucial points, such as warning signs, providing advice to vulnerable individuals, and developing the skills to build resilience.

As Fig.1 illustrates, the Vulnerability Assessment Tool is key to the development of the other tools, which feed back into it in turn. In part for this reason, it receives most of the attention in this report. The Cyberspace Detection Tool is also discussed at length, because it involves monitoring and analysis of speech, which raises a set of distinct issues linked to freedom of expression. The Family Care Package raises issues linked to interference in family relationships and developmental stages of a person's life. The Enhanced Platform raise far fewer issues, because they do not involve attempts to classify or categorise people for distinct treatment according to perceived radicalisation. The Skills and Competencies tool is at too early a stage to be discussed here, but it is unlikely to raise distinct issues in itself over and above those that are already raised by the other tools.

This report begins in Section 2 by providing some background to the practice of applied ethics, the framework adopted, and the relation between ethics, law, and codes of professional ethics. This helps to contextualise and support the discussion of issues to follow. This introductory section also provides an ethical rationale for counter-radicalisation policy and practice, a brief overview of the ethical risks associated with it, and an indication of how the PERICLES tools may add ethical value to such policy and practice.

In Section 3, each tool is discussed in turn, and ethical issues are raised in relation both to its foreseen and its potential design and application. Section 5 concludes.

2 BACKGROUND AND APPROACH

This report identifies ethical issues that arise in relation to the design and application of counter-radicalisation tools.

2.1 THE ETHICAL FRAMEWORK AND BASIS FOR COUNTER-RADICALISATION

The European Union (EU) is built upon the values of liberal democracy, namely the recognition of universal human rights, equality, tolerance, and freedom. It is a founding tenet of the EU that its member states should aim to foster societies in which there is general acceptance and collective pursuit of these values, for the benefit of all citizens. Liberal political and philosophical theory defends such societies on the ground that they both treat people with dignity and respect and enable people to increase their overall wellbeing.

Violent extremists actively oppose liberal politics, not only through the use of illegal means- such as murder, violence, destruction, unrest and vandalism- but also through the promotion of hateful ideologies designed to sow discord and spread fear among people while at the same time weakening support for democratic processes. Therefore, the reasons we should care about and take part in counter-violent extremism are the same as the reasons for supporting mutual tolerance, respect, and the maintenance of liberal democracy in general. The same reasons apply to counter-radicalisation, because radicalisation is a process that remains a key driver of violent extremism, even though it does not always result in violence. Thus counter-radicalisation addresses a social problem that is linked to, but not in every case a precursor to extremist violence. As Neumann points out in a recent report, the strength of counter-radicalisation as a field of practice lies in its recognition of the distinctive role non-security measures can play in addressing the problem of extremism; specifically, he argues, ‘counter-radicalisation recognises the social roots of the problem, enables early interventions, promotes non-coercive solutions, and serves as an early warning system for emerging conflicts and grievances’ (Neumann, 2017).

As this suggests, some kind of intervention to prevent, divert, diffuse and disrupt the radicalisation process is not only permitted but arguably required, morally speaking. The fact that we still do not have many well-tested and reliable models of radicalisation processes in which to ground such interventions does not reduce the imperative to intervene; rather, it just makes it more important to put the knowledge available to careful use (Neumann, 2017: 18). The overarching aim of the PERICLES project is not only to bring such knowledge to bear on concrete counter-radicalisation policies, but also to combine it with insights into the effectiveness of such policies, and their responsiveness to user needs and in so doing to improve counter-radicalisation practice.

2.2 ETHICS, THE LAW, AND PROFESSIONAL CODES OF ETHICS

The ethical values and standards referred to in this report have their basis in liberal political theory.[1] But they are also reflected in common values, laws, and codes of ethics or of conduct of counter-radicalisation professionals. These three distinct locations of ethical guidance are outlines and distinguished here.

The ethical values of liberal democracies are encoded in their laws, including in particular human rights law. For example, the legal right to freedom of expression reflects a moral right to express and receive information, opinions, and beliefs. Legal norms thus track ethical norms. However, ethics and law remain distinct in important ways, and it is worth spelling out the distinction here in order to clarify the nature of the claims being made in this report and avoid misunderstanding.

Ethics relates to the reasons we have for deciding that something is the right or wrong thing to do. Laws specify the rules a society can legitimately coerce people into compliance with. To say that something is legally permissible is to say that it can be done without sanction. Thus legal analysis can tell us what the law requires and so what we need to do if we want to avoid being prosecuted, sued, and so on. Ethics tells us what we should or should not do if we care about doing what's right and avoiding what's wrong, even if no consequences will follow for us either way. Thus ethics and law provide different kinds of reasons for doing or not doing something.

[1] For an introduction, see the entry on Liberalism in the Stanford Encyclopaedia of Philosophy: https://plato.stanford.edu/entries/liberalism/

Ethics regulates a much broader range of actions than does the law. Many ethical norms are encoded and enforced by law, but much unethical behaviour is not regulated by the law. For example, ethics can tell us what attitudes we should adopt and which virtues and qualities we should cultivate in ourselves and others. The law is often silent on these matters. In addition, ethics tells us that setting out deliberately to insult or offend someone's religious or moral beliefs is wrong. In contrast, legal considerations arise only when the insults or offensive speech in question fall within the narrower set that are legally enforceable, say in the form of prohibitions on hate speech. Like law, ethics can describe minimal rules of acceptable conduct. But unlike law, it can also offer guidance as to what we should be encouraged to do, required to do, or praised for doing, morally speaking. It thus gives us both minimal rules and standards to aspire to. With respect to counter-radicalisation, ethics can offer advice about what kinds of things tool developers and practitioners should do if they want their actions to be both legal and examples of best practice.

Because ethics is broader than law, ethical debates have a key role to play in criticising the law, driving its reform and shaping its direction. This is particularly relevant to emerging areas of law, such as laws regulating conduct in the digital sphere. The suggestions made in this report relate for the most part to the design and implementation of counter-radicalisation tools and not to the legal rules regulating their use. However, they occasionally offer suggestions for how certain practices, such as data sharing between agencies, might be accommodated legally.

Counter-radicalisation policy is implemented by a wide range of people, including by professionals whose activities are governed by professional codes of ethics or codes of conduct. Such codes tend to include both ethical principles and standards of behaviour as well as legal requirements. They typically articulate the overarching 'mission' or purpose of the profession, and then spell out the specific set of values they are regulated by, the virtues they should cultivate, the standards they should meet, and the special responsibilities they have to those they serve or work with (Davis, 1988). When professionals implement their counter-radicalisation policy, they will be bound by their own ethical values, the laws currently in place, and the norms and requirements of their professional code. While codes of ethics will guide all professional activity, including that linked to counter-radicalisation, professional standards-setting bodies may also produce guidance specific to discreet areas of work. For example, on 28th Nov 2017 the UK's British Psychological Society published draft ethical guidelines for psychologists working with extremism, violent extremism and terrorism. These guidelines tell professionals how to apply the standards of their code in the counter-radicalisation context.

As will be discussed in what follows, the provisions of codes of ethics and specific guidelines on counter-radicalisation may in some cases appear to conflict with the requirements of counter-radicalisation policy and law, giving rise to ethical dilemmas. In addition, because counter-radicalisation policy tends to be conceived of and designed as a collective task, it requires the collaboration of agents with differing professional cultures and codes of ethics. The ethical code of one counter-radicalisation professional may in some cases conflict with that of another. For example, the duties of police to 'the public' may in some cases appear to justify action that would, if carried out by a psychologist, conflict with the latter's duty to the individual. Ethical reasoning and analysis can be deployed in order to help to resolve these conflicts.

2.3 OVERVIEW OF ETHICAL RISKS OF COUNTER-RADICALISATION

Earlier, it was suggested that counter-radicalisation is a legitimate area of activity for governments, the third sector and beyond, because of the harms radicalisation poses to the values of liberal democracy. But of course not all and any means are legitimate for use in the pursuit of counter-radicalisation.

There is a vibrant ongoing debate in both the research and political communities about what measures constitute appropriate and proportionate counter-radicalisation interventions. Disagreement arises in part because there is no straightforward way to distinguish in practice between radicalisation that leads to violence and radicalisation that does not; yet in principle it is far easier to justify measures that involve some kind of monitoring or intrusion when they are designed to prevent or disrupt the former than the latter. What is more, the distinction between legitimate political expression -which is generally recognised as a fundamental right in a liberal democracy- and radical speech worthy of monitoring or disruption, is very often difficult to define let alone discern systematically in practice. Similar issues arise with respect to freedom of thought, religion, and association.

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.

More fundamentally, there is a live debate in both the academic and wider policy and practitioner community over the need for interventions that target the stage prior to a decision to engage in or materially support violent extremism, and whether 'vulnerability to radicalisation' is an appropriate concept in which to ground such interventions. With respect to the first of these concerns, much criticism has been directed towards what is often termed ‘the preventive turn’ and the promotion of the notion of “pre-crime” in national security policy and criminal law (McCulloch and Pickering, 2009). For example, the UK’s PREVENT programme has been criticised as ‘undermining fundamental tenets of the justice system by justifying an incursion into the private life of the populace, using the possibility of crime as a reason’ (O’Donnell, 2016).

Whether or not preventive interventions are justified depends on the legitimacy of their aims, given the norms of a liberal democracy, and the proportionality of the means used to achieve those. It is certainly legitimate for a state to intervene on a range of fronts to prevent terrorism and violence, including acts preparatory to violence (Sorell, 2011). However, the extent to which this also justifies intervention with non-violent radicalised behaviour or views depends on a number of factors, chief amongst which is the extent to which a link between such behaviour and violence is demonstrable. The more tenuous that link is, the more difficult it is to justify concerted state-sponsored intervention of any sort. For instance, 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 is unlikely that counter-radicalisation intervention with respect to the latter would be justified.

Finally, while overall the findings of social science and other disciplines can make a significant contribution to the production of well-informed, proportionate and efficient counter-radicalisation policy and practice, risks arise from its potential misinterpretation or misapplication. Violent extremist and terrorist groups have diverse ideologies and diverse political histories. It would be unwise to assume that the lessons learned from counter-radicalisation in one context will apply straightforwardly in another. Even within broad movements, factions and cultures may play an important role in shaping radicalisation and, it follows, in what makes an appropriate response to it.

We explore below how some of these risks might arise in connection with the eventual development of the PERICLES tools.

2.4 THE POTENTIAL ADDED ETHICAL VALUE OF THE PERICLES COUNTER-RADICALISATION TOOLS

In this section we consider the rationale for developing these tools and how they might represent ethical progress with respect to existing approaches to counter-radicalisation. The social science and computer science that informs the development of the PERICLES tools is still evolving. Its key concepts remain contested and the evidence it has produced evidence remains relatively sparse and patchy. The PERICLES project recognises these challenges. The tools it produces do not claim to resolve them. However, they do aim to reduce these risks and improve on existing alternatives used in current practice. They do this by supporting counter-radicalisation actors to make better-informed, more transparent decisions within the scope of their existing powers and competencies. They make the results of research and personal professional insight into radicalisation available in an accessible and operational, real-time form to relevant actors, including in particular LEAs and families affected by radicalisation. As the second report released by this project- the PERICLES Gap Analysis- pointed out, existing tools are often insufficiently informed by the little evidence that is available (Kudlacek et al. 2017a). What is more, their effectiveness is insufficiently well monitored or established. PERICLES aims to produce tools that are underpinned not only by the most robust knowledge available, but also by awareness of the needs of users and relevant stakeholders. The tools produced will also be designed in ways that facilitate the monitoring and evaluation of their effectiveness, so as to produce a better knowledge base for future interventions.

The Family care package aims to support good future practice in the following ways:

a) By being based on state-of-the art scientific evidence
b) By involving active consultation of the different target group and a needs assessment of stakeholders including LEAs, families of radicalised persons, former radicals and convicted terrorists
c) By being informed by expert opinions
d) By building on best practices

The Vulnerability Assessment tool aims to support interventions[2] that are more responsive to the needs of users and that support their decisions by:

a) making use of better information and more transparent reasoning and

b) reducing ad-hoc reliance on hunches, popular myths, or misguided assumptions of different kinds while

c) being open about uncertainty and gaps in knowledge

In doing so, it aims to support fair, proportionate, respectful and effective interventions.

The Cyberspace Detection System aims to improve understanding of radicalised content online by:

a) developing a systematic, legally informed understanding of the scope and boundaries of radicalised speech

b) applying this in ways that reduce the amount of information that is detected and therefore potentially monitored by LEAs.

To the extent that the PERICLES tools achieve these aims, they contribute to ongoing ethical improvements to efforts to combat a phenomenon that is deleterious to liberal democracy.

The tools can address the issue of accountability and legitimacy by trying to be as precise and transparent as possible about what it is that is being detected or revealed, why this is necessary, and how the means adopted constitute proportionate responses. Being open and transparent about the tools and involving relevant groups in their design and evaluation should help to mitigate the risks of disproportionate or unjustified action. It should also help to reduce the risk that they are misinterpreted in ways that can lead to an erosion of trust.

[2] The project has raised questions over whether the term 'intervention' accurately reflects measures designed to help vulnerable individuals, or whether it brings damaging connotations of interference with liberties. The project will review this issue in Jan 2018 and make adjustments as necessary.

3 ETHICAL RISKS AND BENEFITS OF SPECIFIC PERICLES TOOLS

In this section we examine the ethical risks and potential benefits of the Vulnerability Assessment Tool, the Cyberspace Detection System, the Family care package and the Enhanced Platform, in turn. It is important to state clearly, however, that the tools in question are still in an early stage of development. With this in mind, we explore issues as they might arise in relation to the different potential forms the tools may take. Proposals for the design and implementation of tools and concrete recommendations are made wherever possible.

3.1 VULNERABILITY ASSESSMENT TOOL

PERICLES develops an assessment tool that combines a variety of indicators from 'at-risk' individuals and groups, to create a vulnerability assessment along with recommended actions. The tool will signal radicalisation and indicate possible measures to deal with it. Indicators include behaviour, religion, family circumstances, personal factors and social networks. Key users of the tool will be LEAs, city councils, and potentially other state institutions and agencies such as social services and schools.

Foreseen uses

Users will be able to turn to the tool to help them to follow up a concern about an individual or a group. They can input data relevant to the particular case and the tool will help support their subsequent decisions and actions with respect to that case. For example, the tool may make suggestions about the kind of information the user should seek and consider in order to make a rounded assessment of an individual or group and ensure that factors identified by social scientific research as important to the process of radicalisation are not overlooked (e.g. educational outcomes; family circumstances etc.). The tool will also indicate what type of interventions are available to assist the individual or mitigate the threat.

Potential uses

The tool may also be a living resource, populated by information inputted by a professional or set of professionals on specific individuals or groups

and combining these with models based on other data such as social network data, demographics, previous events, information on cases or theoretical relations between factors.

It is possible that the tool will not store data but will have a self-learning (AI) element that processes clusters of cases and learns from them before deleting the original data. It is also possible that the tool will both store data and have a self-learning element.

3.1.1 Ethical issues linked to the category of 'vulnerability to radicalisation

3.1.1.1 Should the targets of counter-radicalisation be treated as vulnerable, as threats, or both?

Currently, many counter-radicalisation policies and interventions conceptualise their target group as those 'vulnerable to radicalisation'. Thus this tool reflects current trends in counter-radicalisation policy and practice. Yet the appropriateness and coherence of this category has been criticised in the academic literature. Here we describe two such criticisms and consider what they imply for the design and application of the tool.

On the one hand, it has been argued that the category 'vulnerable to radicalisation' is ambivalent between vulnerability and threat, while on the other it is claimed that it is underpinned by a false dichotomy between vulnerability and threat. Those who make the latter criticism often point out that the policy discourse distinguishes between those who are on the path towards radicalisation, whom it designates as vulnerable and meriting primarily therapeutic intervention, and those who are already radicalised, whom it treats as a potential threat meriting criminal justice intervention. Thus, for example, Heath-Kelly argues that 'the language of (susceptibility to) persuasion, disadvantage and vulnerability is used to create a vulnerable potential terrorist subject separate from existing radical subjects (who prey on the weak, promote contagious ideas and take advantage of our 'open institutions')' (Heath-Kelly, 2013). This criticism implies that any attempt to reflect the coexistence of vulnerability and dangerousness in counter-radicalisation interventions would produce practices that are incoherent, because approaches to reducing these two features are mutually incompatible.

A second line of criticism suggests that there is a softening of the distinction between these categories of vulnerability and threat in counter-radicalisation policy, which posits those who risk engaging in terrorism as 'passive and vulnerable subjects who have become 'infected' or 'gripped'

by ideas, and who are thus in need of intervention and support [while] they are also and simultaneously constituted as potentially dangerous' (O'Donnell, 2016) This line of criticism implies that counter-radicalisation practices grounded in hard distinctions between the vulnerable and the threatening would not reflect the reality of radicalisation, which at least sometimes involves people becoming dangerous whilst remaining vulnerable. It raises the concern that such practices would instead neglect to address either the vulnerability of the dangerous or the dangerousness of the vulnerable and therefore fall short of what is needed for effective and fair interventions.

Are these criticisms fair, and if so, what do they imply for the design and application of a counter-radicalisation vulnerability assessment tool? We may start to address this question by observing that in reality people can be both vulnerable and dangerous. This is easiest to observe in teenagers and young adults. This age group tends to be more impressionable, emotionally unstable and immature and therefore vulnerable to manipulation than other age groups. Yet it is also more likely to engage in reckless, risky and dangerous behaviour, including violence. This suggests that we should be sceptical that the conceptual distinction between vulnerability and threat reflects any corresponding incompatibility in reality.

If people can be both vulnerable and dangerous at the same time, as indeed it seems they can, then we not only can but should design counter-radicalisation interventions in ways that recognise and respond to this fact. Preventive or disruptive measures by police need not undermine the work of other organisations offering services designed to support and offer alternatives to young people. Rather, both can be pursued at the same time without contradiction, providing of course that they are designed and coordinated thoughtfully. As this suggests, the establishment of open and continuous inter-agency dialogue and the building of partnerships between relevant agencies is vital to the development of an effective counter-radicalisation strategy.

3.1.1.2 Implications for data-sharing between agencies

It has just been claimed that police work to prevent and disrupt crime and anti-social behaviour relating to radicalisation is in principle compatible with the work of other agencies to reduce vulnerabilities to radicalisation. However, turning this in-principle compatibility into compatibility in practice depends on clear and consistent delineations of the roles and responsibilities of law enforcement agencies and other counter-radicalisation practitioners. If those roles and responsibilities become blurred, there is a real risk of incoherent and even counter-productive practice. This is at

least in part because counter-radicalisation interventions led by educators, psychologists, youth clubs and so on rely for their success on the creation of private and quasi-private associations and spaces, where views can be aired, feelings shared, and relationships formed free from surveillance or suspicion (whether perceived or real). If tools result in the sharing of information between agencies in ways that undermine the trust necessary for these interventions to work, then they may have counter-productive results, not only in the individual cases, but also more broadly, as word spreads amongst communities.

Data sharing and collaboration between criminal justice and other agencies can also challenge the ethical integrity of professionals. This may occur if they are placed under pressure to share information they were given in confidence. It may also occur if they are compelled to share information they are authorised to collect and use for one purpose with agencies who will use it for another purpose. Such crossing of professional boundaries poses potential challenges for the legitimacy of the professions. The British Society of Psychologists flag precisely this dilemma in their recent Draft Guidelines for Ethical Practice of Psychologists working with counter-terrorism, extremism, and radicalisation (BSA, 2017).

A tool should enable the sharing of expertise and insight, but data on specific individuals should not be shared beyond what is necessary to prevent imminent harm to them or others. Complete openness with targets about engagement with law enforcement agencies, as has long been practiced by the groups Hayat and Exit, which support radicals to disengage from groups, is now acknowledged as best practice, and this should be reflected in any counter-radicalisation tool. (Young et al, 2016: 22)

3.1.1.3 Does vulnerability merit emotion-based interventions by the state, and are these compatible with respect for individual decision-making and autonomy?

A separate concern arises from the operationalisation of the concept of vulnerability to radicalisation. This relates to the nature of counter-radicalisation interventions. The concept of vulnerability to radicalisation suggests manipulability and gullibility, stemming from immaturity, emotional weakness, disadvantage and dependence. Its adoption implies that those choosing a path of radicalisation are not exercising their agency to make an autonomous choice, but rather are being steered or directed by distorting environmental or emotional factors. If this is indeed what occurs in the radicalisation process, then addressing vulnerability is not primarily a matter of education, information and rational discussion. As noted in the first PERICLES report, on the state of the art, it requires an approach based on emotion regulation, at least in combination with the provision of knowledge (Kudlacek et al, 2017a). Further weight is given to this claim

by the fact that, for targets emotionally involved in certain ways of thinking, or who feel affiliated with a certain group, 'the incorporation of new knowledge could be obstructed by their emotional involvement, reducing the effectiveness of the intervention' (PERICLES D1.1: 6).

Approaches based on emotion regulation may be problematic for a number of reasons. First, emotional manipulation, fearmongering,etc are precisely the kinds of tactics extremist groups are accused of deploying to 'groom' vulnerable people into radicalisation. Liberal democratic states should distinguish themselves from these groups by reference not only to their contrasting values but also to their use of means of recruitment that respect the capacity for autonomous judgements and choices of their citizens. In a liberal democracy, people's political choices should ideally be formed autonomously, by a process of responsiveness to facts, evidence, and rational argumentation on either side of the debate; not via a process of intentional manipulation by others. This is reflected in the elements common to counter-radicalisation around the EU, including 'dialogue about ideologies, critical thinking, discussion about policy, values-based approach, voluntariness, individual guidance and support' (Kudlacek et al, 2017a).

If the individuals targeted by counter-radicalisation are deemed 'vulnerable' to radicalisation due to emotional or psychological weaknesses of some sort, the ideal approach to addressing this is to build resilience, not exploit these weaknesses to push a liberal democratic political agenda.

Of course, rational dialogue and resilience building are approaches that are also used by extremist groups seeking to attract people to their causes. For any individuals who do radicalise solely or primarily by means of such approaches, the label of vulnerability is problematic. For it suggests that anyone who is open to political persuasion is thereby vulnerable, in which case the distinction between those who are autonomous and those whose autonomy has been in some sense compromised collapses; the vast majority of citizens seem to qualify as vulnerable; and the category ceases to provide a basis for targeted protective intervention.

3.1.2 Risks of stigmatisation and knock on effects

In 2016, researchers on the EU-funded TERRA project, which examined evidence-based approaches to counter-terrorism, warned that: 'while im-

plementing well intended social initiatives, policy makers should be additionally aware that they can potentially negatively identify their target groups and unwittingly contribute to stigmatisation. Care should be exercised in all communications from the government about the population, and that ethnic and religious minorities are not singled out, negatively identified, nor labelled as suspicious, either explicitly nor implicitly.' (Young et al, 2016:10) This sound advice can usefully be directed beyond policy makers to counter-radicalisation practitioners, for reasons now discussed.

Stigmatisation can be defined as the process of marking a person or a group out as having an undesirable characteristic (Hadjimatheou, 2014: 189; Courtwright 2011; Arneson 2007). 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.

Stigmatisation of an individual as being involved in radicalisation is sometimes justified, e.g. if the individuals or groups in question are found indeed to have been engaged in violence. For example, it is right for governments and financial institutions to blacklist and therefore stigmatise a group posing as philanthropic organisation but in fact exposed as dedicated to funding violence. In such cases, the stigmatisation is both appropriate and fitting to the wrongdoing in question.

However, in some cases stigmatisation may be merely an unfortunate consequence of counter-radicalisation interventions that are not intended to be stigmatising but end up being so, sometimes inevitably. Whether a counter-radicalisation intervention is stigmatising or not and how harmful the stigmatisation is depends very much on how it is perceived, not only by those who are targets of the intervention, but also by those who identify with those targets of the intervention, and by onlookers whose opinion of those targeted might be affected. Where this kind of stigmatisation occurs, it brings certain risks that should be addressed in the design and implementation of counter-radicalisation measures. These risks arise when individuals are accurately singled out for counter-radicalisation interventions as well as when they are wrongly targeted.

If an individual has been identified incorrectly as meriting counter-radicalisation intervention, they may be stigmatised by such targeting in ways that are seriously counter-productive. Much of the reason for this has to

do with the way being stigmatised makes people feel and how these feelings in turn affect their attitudes to and relations with others in society, especially with the authorities. Individuals wrongly stigmatised as vulnerable to radicalisation may feel anger or indignation at what they perceive to be an unjust implication of wrongdoing and/or an unjust interference with their freedom to hold and express political views (Hadjimatheou, 2014). This may create knock-on social costs, by eroding their trust in the authorities or agencies in question, which in turn may fuel alienation and reduce a willingness to cooperate (EU Fundamental Rights Agency 2010: 21).

When counter-radicalisation interventions stigmatise or are perceived as stigmatising people on the basis of their membership in a certain group (e.g. religious, ethnic, political), other members of the group may also feel targeted, in virtue of their identification with the group. Thus the negative feelings and attitudes linked to stigmatisation can spread beyond the individual and increase the isolation or alienation of groups from mainstream society and the authorities.

Risks of stigmatisation-related negative consequences increase when counter-radicalisation interventions take place against the background of pre-existing stigmatisation and suspicion of specific groups in society. In such cases, even the most sensitive counter-radicalisation communication strategy on the part of the authorities may be met with cynicism instead of reassurance, at least by some of those who perceive they are being unfairly targeted. This risk is likely to be particularly high when the authority doing the surveillance has itself conducted unfair and suspicion-based surveillance or intervention in the past. For example perceived heavy handedness and excessive suspicion of British Muslims following the terrorist attacks of 2001 and 2007, such as through widespread use of stop and search by police, led many critics to argue that British Muslims were being singled out as a new ‘suspect community’ (Hickman et al, 2012; McGovern, 2010; Pantazis and Pemberton, 2009; 2011). Against this background, it is unsurprising that the government’s anti-radicalisation programme PREVENT has met with suspicion and rejection by some in the Muslim community and beyond.

For reasons such as these, counter-radicalisation interventions have been and surely will continue to be interpreted as stigmatising even when they do not involve implications of threat, or disrespect for community practices. We now consider some ways in which these risks might be addressed in the design and implementation of the vulnerability assessment tool.

3.1.2.1 Avoiding, reducing, mitigating, and justifying stigmatisation

It is important to state clearly that the stigmatisation-related risks of counter-radicalisation practices should be viewed alongside the moral risks of relevant agencies *not* engaging in counter-radicalisation. The latter may be considerably greater than the former (Sorell, 2011:5) The former can only be addressed by attempting to design counter-radicalisation practices in ways that actively involve the views of those likely to be affected, and deliver them in ways that are respectful, proportionate, and discreet.

The vulnerability assessment tool mitigates some of the risks of stigmatisation, especially perceived stigmatisation by religious groups, by operationalising knowledge and insight on ideologies of violent left-wing and right-wing extremism as well as religious extremism. In other words, the tool itself does not single out or focus exclusively upon one community; rather it addresses radicalisation as such and treats all kinds of radicalisation as equally worthy of concern and intervention. Furthermore, the tool can be useful not only in increasing awareness of genuine indicators of radicalisation, but also in debunking popular myths about radicalisation that might fuel discriminatory practices. These important aspects of the tool should be explicitly highlighted so as to avoid misinterpretations of its functionality by those it targets.

3.1.2.2 Bias and error

If the vulnerability assessment tool is developed in such a way that it can be used by law enforcement to make decisions about whom to conduct surveillance on or whom to deprive of liberty, the ethical issues become more challenging. For example, if there were an algorithm that purported to distinguish suspect people from innocents then this would need to be based on transparent, reliable, demonstrated evidence, of the kind that just is not available currently. We just do not know enough about the causes of radicalisation yet to design a tool that would not end up subjecting many non-radicalised individuals to some kind of stigmatisation or interference. Even AI tools based on apparently reliable evidence have run into a great deal of difficulty in recent years. Google's facial recognition programme, which mistakenly labelled black people in photos "gorillas" is one such notorious example.[3]

[3] See the Guardian newspaper's report on this at: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

However, this is not to say that no self-learning element to the tool would be justifiable. It depends on what the learning relates to and how good the input is.

3.2 CYBERSPACE DETECTION SYSTEM

The cyberspace detection system uses a 'pragmatic methodology' to improve the detection of radical speech- defined as tweets that contain threat or incitation or physical harm, such as the expression of contempt or hatred towards certain groups- online. This methodology involves a first stage of filtering by human supervision and then the application of computer methods to automate the process. The incorporation of the human evaluation is intended to correct some of the deficiencies of purely semantic approaches, by providing valuable insight into context and intention. This should enable more precise distinctions between what is genuinely radical content and what is, for example, humour or criticism. The analysis will focus not only on the content of the tweets but also the relationship between different tweets, the source of the tweet and certain environment variables. The tool aims to detect and make predictions about the patterns followed by radical content; it is not able to predict who will become radical.

Foreseen design and applications

The tool aims to help law enforcement agents filter content more efficiently to detect radical speech online. The tool is currently being developed with data from Twitter alone and will therefore only be applicable to that platform, at least for the duration of the project. The tool will also remain in the hands of the developers who will use it to provide data and consultancy to law enforcement partners. The tool will not be operated by law enforcement officers. Foreseen applications include:

1. Helping social media platforms like Twitter to target more effectively their efforts to find and take down illegal content/content that violates the terms of use
2. Helping LEAs filter material so as to focus better their efforts to understand determined violent or radical uses of online discourses that can materialize in concrete political actions.

3. Providing insight into other risk factors –specifically, environmental risk factors-, indicative of potential radical communication behaviours
4. Providing insight to LEAs on how to better target counter-speech

Potential design and applications

One possible development of the tool would see it being designed in a more user-friendly way so as to be operated by LEAs. Open questions include whether they would also own the data produced and if they may use it to identify and monitor individuals producing radicalised content and/or to predict how radicalised speech online is likely to appear.

Ethical purposes to which the tool could be put

There are legitimate reasons for wanting to know where radical speech is happening online, how it manifests itself, and whether and what kinds of patterns it follows, especially relative to external occurrences, such as terrorist attacks. One reason is that the internet is a key facilitator of and medium for the sharing of radicalised views as well as of certain kinds of illegal radicalised speech, such as hate speech, incitement to violence, harassment and so on. Tools that help society understand better what kinds of views are expressed when and by whom, how they are diffused, which ones are more compelling to other users and so on can help relevant stakeholders to design better policies for counter-radicalisation and beyond.

Radicalised speech online is also a legitimate topic of research. As mentioned at the start of this report, the link between violent speech and violent action is still poorly understood. Radicalised speech online represents potentially useful data, because the sheer quantity of it means it is possible to gather and analyse many messages very quickly. Properly exploited, this data could be used to better understand this link, even if what is ultimately established is that causal or even correlative links are difficult to demonstrate. Although identifying and scrutinising this material is justifiable, there are potential risks to doing so as well. These are now considered.

3.2.1 Risks of stigmatisation and offence

Subjects whose communications are categorised as ‘radical’ or ‘radicalised’ may take offense. For example, the UK’s Centre for the Analysis of Social Media has undertaken recent analysis of Islamophobic tweets in

the UK over the period of a year. Despite attempting to define Islamophobia in clear and justifiable terms, the researchers found themselves the subject of much criticism and the recipients of numerous complaints. These came primarily from individuals who objected not only to being labelled Islamophobic but also to the positing of a concept of Islamophobia, which they claimed was used to unfairly discredit legitimate (though perhaps unpopular) forms of political expression. The fact that people might take offence in this way does not suggest gathering or analysing the material is impermissible, but rather that the delineation of categories should be done thoughtfully and that overbroad categories are better avoided.

3.2.2 Ethical risks of removal of extremist material

Since 2016, the EU and individual Member States have increased the pressure on social media providers and other internet platforms to remove terrorist or extremist material from their sites. This move is based on the view that the proliferation of such content is itself a driver of radicalisation, providing the 'oxygen' extremist groups need in order to begin to normalise themselves, better recruit and grow. In response, a group of such companies announced plans to set up a system for the immediate removal of content via for an industry database of "hashes"—unique digital signatures—of extremist material banned on their platforms (Citron, 2017). The companies explained in that content will be hashed only if it involves "the most extreme and egregious terrorist images and videos … content most likely to violate all of our respective companies' content policies". Since then, reports suggest that the mechanisms used by companies for removing such content are becoming more sophisticated and effective.

There is some evidence that removal of content works. For example, a 2015 Brookings report claimed that: "the primary ISIS hashtag—the group's name in Arabic—went from routinely registering in 40,000 tweets per day or more in around the time suspensions began in September 2014, to [fewer] than 5,000 on a typical day in February. Many of those tweets consisting of hostile messages sent by parties in the Persian Gulf" (Berger and Morgan 2015, 56).

There are growing concerns that removal of offending material online could have negative consequences that might even outweigh the benefits, even when the material concerns terrorism. In particular, concerns have been raised in the policy literature that removal of content

- Removes opportunities for counter-speech, which might otherwise expose those radicalised to different views

- Isolates terrorists in ways that cut off channels for help to those vulnerable or wishing to disengage, for example by driving them onto the dark web (Cox 2015).
- Reduces public knowledge about social views
- Removes a valuable source of research and investigative journalism for better understanding of radicalisation, terrorism, and related activity
- Removes an opportunity for those radicalised to 'let off steam'
- Fuels victim mentality as those censored feel they are unjustly silenced
- Increases "the speed and intensity of radicalization for those who do manage to enter the network." (Berger and Morgan, 2015)
- Can interfere illegitimately with freedom of expression of those neither committing nor intending to commit crimes

Much could be said about the weight of each of these concerns compared to the aims and purposes served by removal. For the purposes of this report, however, it is sufficient to point out that they should be considered if the cyberspace detection tool develops in such a way as to be usable for the purposes of identifying content for removal.

3.2.3 Ethical issues relating to counter-speech and disengagement support online

The US State Department in 2012 launched a South East Asia-facing counter-radicalisation "viral peace campaign". The programme was designed "to use social media as a way of promoting community involvement and peaceful change" and "to help people craft online strategies that use a whole range of tools—including 'logic, humour, satire, [and] religious arguments'—to match the violent extremists' energy and enthusiasm" (Neumann 2013, 446). As reported by Greenberg in a 2016 analysis of counter-radicalisation online, since then Google has also developed a 'program of diversion' for those who search for certain terms, diverting them to NGO sites delivering messages supporting peace and disengagement. The Google Adwords Grants Program also gives NGOs credits with which to buy advertising space for counter-radicalisation messages that will appear at the top of certain searches identified as indicative of radicalisation (Ward-Bailey 2016). "We should get the bad stuff down, but it's also extremely important that people are able to find good information, that when people are feeling isolated, that when they go on-line, they find

a community of hope, not a community of harm," explained a Google executive (Quinn 2016).

More overt counter-speech involves government funding of organisations promoting nonviolent interpretations of ideologies and religions online, and exposing the dark realities of life as a terrorist fighter (Greenberg, 2016).

These approaches are less ethically problematic than removal of content because they do not involve interferences with freedom of expression or association, neither do they censor the open web in ways that appear to misrepresent the views of people on it, nor need they involve stigmatisation. However, there is very little evidence as yet about their effectiveness as yet.

3.2.4 Ethical risks of using the tool to identify potential terrorists

Could the cyberspace detection tool be used to identify actual violent perpetrators? It could be used to identify individuals who post radicalised content, but it is unlikely that it could be used to distinguish between those who have an intent to commit violence from those who do not. As Sageman points out '…the discourse and documents available on "jihadi websites" is rarely connected to actual violent plots' (Sageman, 2014). There is a risk that extreme radicalised speech online will be mistakenly interpreted by the justice system as evidence of intent to commit attacks. For example, Sageman expresses the concern that public ignorance of the link between speech and crime, combined with overly censorious terrorism laws, has led to some individuals being criminalised unnecessarily: 'The U.S., British, and Danish justice departments have played on juries' ignorance of terrorism and the inflammatory nature of this discourse to hire, as expert witnesses, pseudo scholars who claim that they have generated a terrorist profile from such online discourse. These self-proclaimed "experts" have helped condemn immature young people, whose only crime was boasting and bragging on the Internet, to very long prison terms' (Ibid).

This suggests that education is needed to support users of the tool, to ensure they understand the method and the results sufficiently well to make effective and proportionate use of them. In practice, this might mean either that the tool developers retain control over the tool, providing analysis and results as a service to LEAs, or that LEAs themselves run the tool, but only after adequate training.

3.3 FAMILY CARE PACKAGE

This tool will provide tailored material for family members of those at risk of radicalisation. Families will be provided with advice and support on how they can detect signs of radicalisation, how to intervene during the earliest stages, and the best course of action to take. The family care package is intended to provide evidence-based information and support to those who are often best placed to intervene to help those who are involved in radicalisation.

3.3.1 Risks of unnecessary interference

As our discussion of stigmatisation implies, social services, psychological and security interventions do not come without ethical risks. For this reason at least, unnecessary interventions should be minimised. With respect to the Family Care Package, there will always be a risk that services will be offered too early or to families that do not need them. Managing this risk while also pro-actively alerting families to the services on offer is a balancing exercise. Below, risks relating to specific kinds of potential unnecessary interference are considered.

3.3.1.1 Unnecessary interference with natural developmental stage of youth

It is not uncommon for teenagers and young people to explore extremist views at this stage in their lives. Indeed, some argue that it is ‘a natural aspect of the development of adult characteristics’ (Coppock and McGovern: 2015). Democratic theory recognises a value in young people experimenting with different views, developing critical abilities and a sense of social justice, as this helps prepare them for a lifelong role as active citizens. Yet it is difficult for families (and indeed for researchers) to know when involvement with such views goes beyond natural experimentation and becomes dangerous for the individual and/or society. It is important, therefore, that the interventions recommended seek to avoid inadvertently ‘stymying this particular area of development, seeing it as a threat’ (Ibid.).

3.3.1.2 Unnecessary interference based on misinterpretation of behaviour

The Family Care Package aims to direct families to the services that are appropriate to their own individual circumstances. But sometimes it may be difficult to determine what type of intervention is likely to be most appropriate to those circumstances. For example, trauma of the kind that refugees from war zones may have experienced may provoke behaviour that appears to indicate a propensity to violence. In this way, a child may engage in fantasy play in which they act out scenes of violence that they

have witnessed in the past. In such a case, deciding which kind of intervention to recommend involves assessing whether in the particular instance, acting out violence in play is a reliable indication of an intention to carry out such violence in reality. If such assessments result in recommendations for security interventions that turn out to be unnecessary, this may have a negative impact on the individual and the family in question. While such risks cannot be entirely dispersed, they can be reduced by means of measures such as evidenced-based approaches to risk assessment, peer review of risk assessment, monitoring of the effect of interventions, and sound evaluation and review.

3.3.2 Risks related to collaboration between families and law enforcement agencies

The tool specifies that it will help collaboration between families and law enforcement. It is vital that good communication and mutual, well-founded trust between these groups is established, but it is difficult to achieve in practice. The unequal power dynamics between families and police and the historical and political context in which collaboration takes place may have distorting effects on the relationship. With this in mind, it is important that families who use the tool do so voluntarily and can choose not to collaborate closely with police if they choose not to.

3.4 ENHANCED PLATFORM

This tool is envisaged as a source of information on counter-radicalisation and a place to exchange ideas and best practices between counter-radicalisation professionals. It will help users develop informed and evidence-based counter-radicalisation strategies. Learn from experiences. The tool is likely to be a closed online system, with access granted only to stakeholders who have been granted rights.

There are very few potential ethical risks with this tool, and none that sound training could not resolve. One potential risk is that a stakeholder might end up applying methods, best practices or lessons learned from one radicalisation context to another, without considering sufficiently the particularities of the case and distinctions between the contexts. In doing so, they may end up misinterpreting actions and stigmatising groups or individuals wrongly as radicalised. Sharing of best practice across borders should be accompanied by efforts to take these into account, so as to avoid enacting ineffective or counter-productive interventions.

4 CONCLUSION

A number of recommendations have been made in this report. These relate to the design of tools, the training of those who use them, and their application in practice. It has been suggested that tools that categorise and classify individuals for counter-radicalisation intervention risk stigmatising those individuals. The design of such tools should therefore aim to reduce the need for classification, increase the accuracy of classifications, and/or develop ways of reducing the negative impact on those affected. Machine learning and predictive analytic functions are especially risky when they have direct effects on specific individuals, and it is therefore important that they be based on clean data and sound evidence and that there is transparency and oversight in their development and application. More generally, the engagement and involvement of stakeholders, including those likely to be targeted by tools, in their design and in the design of the interventions to follow is an important means of reducing ethical risks of offense, stigmatisation, discrimination, amongst others, and of improving effectiveness. A further issue that cuts across all the tools is the need for sound interpretation of the social science and the relevance of lessons learned and best practice across radicalised groups, cultures, populations and jurisdictions. Careful training, awareness-raising, and sensitisation of users to the ethical implications of the tools are important means to reduce these risks. These issues and more will be discussed in more depth and with more reference to the specific application of the tools in the following report.

REFERENCES

Berger and Morgan (2015) “The ISIS Twitter Census,” Brookings Analysis Paper no. 20, March. At: https://www.brookings.edu/wp content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf.

British Society of Psychologists

Citron, D. (2017) ‘What to Do about the Emerging Threat of Censorship Creep on the Internet’. CATO Policy Analysis. At: https://www.cato.org/publications/policy-analysis/what-do-about-emerging-threat-censorship-creep-internet#_idTextAnchor03

Cox, J. ISIS now has a propaganda site on the dark web. 16 November 2015. Motherboard. Available from http://motherboard.vice.com

Davis, M. (1998) *Thinking Like an Engineer: Studies in the Ethics of a Profession*, Oxford University Press.

Greenberg, K. (2016) ‘Counter-Radicalization via the Internet’. *Annals of the American Society of Political and Social Science*, 668(1)

Hadjimatheou, K. (2014) ‘The Relative Moral Risks of Untargeted and Targeted Surveillance’ *Ethical Theory and Moral Practice*, 17:187–20

Heath-Kelly, C. (2013). ‘Counter-terrorism and the counter-factual: Producing the ‘radicalisation’ discourse and the UK PREVENT strategy’. *The British Journal of Politics and International Relations*, 15: 394–415.

Kudlacek, D. et al. (2017a) PERICLES Project Report ‘Prevention of radicalisation in selected European countries: comprehensive report on the state of the art in counter-radicalisation’. At: http://project-pericles.eu/wp-content/uploads/2017/10/Pericles-D1.1-Findings-Report.pdf

Kudlacek, D. et al. (2017b) PERICLES Project Report ‘Gap analysis on counterradicalisation measures’. At http://project-pericles.eu/wp-content/uploads/2017/10/Pericles-D1.2-Gap-Analysis-Report.pdf

Miller, C. and Smith, J. (2017) Anti-Islamic Content on Twitter. Demos. At: https://www.demos.co.uk/project/anti-islamic-content-on-twitter/

Neumann, P. (2017) ‘Countering Violent Extremism and Radicalisation that Lead to Terrorism: Ideas, Recommendations, and Good Practices

from the OSCE Region'. At http://icsr.info/wp-content/uploads/2017/09/Countering-Violent-Extremism-and-Radicalisation-that-Lead-to-Terrorism-2.pdf

Neumann, P. (2013) 'The trouble with radicalization'. *International Affairs*, 89(4): 873–893

Sageman, M. (2014) 'The stagnation in terrorism research'. *Terrorism and Political Violence,* 26(4): 565-580

Sarma, K. M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist*, 72(3), 278-288.

Sorell, T. (2011) 'Preventive Policing, Surveillance, and European Counter-Terrorism' *Criminal Justice Ethics*, 30(1) 1-22.

Sorell, T. (2016) 'Online grooming and preventive justice'. *Criminal Law and Philosophy* 11(4) 704-724

Thaler, R. and Sunstein, C. (2008) *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press.

Quinn B. 2 February 2016. Google to point extremist searches toward anti-radicalisation websites. The Guardian. At: https://www.theguardian.com/uk-news/2016/feb/02/google-pilot-extremist-anti-radicalisation-information

O'Donnell, A. (2016) 'Contagious ideas: vulnerability, epistemic injustice and counterterrorism in education' *Educational Philosophy and Theory*.

Ward-Bailey J. 3 February 2016. How Google plans to fight extremism through search advertising. The Christian Science Monitor. At https://www.csmonitor.com/Technology/2016/0203/How-Google-plans-to-fight-extremism-through-search-advertising

Young, H. et al. (2016) 'Evidence-based policy advice' Terra Project Final Report. At: http://terratoolkit.eu/wp-content/uploads/2016/12/TERRA-Evidence-based-Policy-Advice_English_Final-Report.pdf